БАКАЛАВРИАТ



В. И. Соловьев

АНАЛИЗ ДАННЫХ В ЭКОНОМИКЕ

Теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel

Учебник

BOOK.ru
ЭЛЕКТРОННО-БИБЛИОТЕЧНАЯ СИСТЕМА
КНОРУС • МОСКВА • 2019

УДК 51 (075.8)
ББК 22.17я73
С51

Рецензенты:
Т.М. Гатауллин, д-р экон. наук, канд. физ.-мат. наук, проф. (ГУУ),
Д.В. Сошников, канд. физ.-мат. наук, доц. (*Microsoft*)

Соловьев, Владимир Игоревич

С51 Анализ данных в экономике. Теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel : учебник / В. И. Соловьев. — Москва : КНОРУС, 2019. — 497 с. — (Бакалавриат).

ISBN 978-5-406-XXXXX-X
DOI 10.15216 / 978-5-406-XXXXX-X

Учебник подготовлен в соответствии с требованиями Федерального государственного образовательного стандарта высшего образования (ФГОС 3++) по направлениям подготовки экономики и управления.

Содержит интегрированое изложение вероятностно-статистического фундамента анализа данных, его практической реализации в Microsoft Excel, а также примеров и задач, направленных на применение инструментария описательной и предсказательной аналитики в реальных ситуациях принятия решений в экономике, финансах, операционном менеджменте, маркетинге, управлении рисками, логистике и т. д.

Отличается от учебников по теории вероятностей и математической статистике подробным рассмотрением прикладных задач анализа данных и ориентацией на практическое решение реальных задач с использованием Microsoft Excel, а от других учебников по анализу данных — обстоятельным изложением теоретико-вероятностных и статистических основ анализа данных. Кроме того, излагаются вопросы предварительной обработки и визуализации данных, редко рассматриваемые в учебной литературе.

Для студентов бакалавриата, обучающихся по направлениям подготовки экономики и управления, а также практических специалистов, желающих применять методы анализа данных в решении прикладных задач.

УДК 519.21 (075.8)
ББК 22.17я7

Соловьев Владимир Игоревич

Анализ данных в экономике.
Теория вероятностей, прикладная статистика,
обработка и визуализация данных
в Microsoft Excel

Сертификат соответствия № РОСС RU.АГ51.Н03820 от 08.09.2015.

Изд. № _____. Формат 70×100/16. Подписано в печать __.__.2018.
Гарнитура «Журнальная». Печать офсетная. Усл. печ. л. 31,25.
Уч.-изд. л. 34,35. Тираж 1000 экз.

ООО «Издательство «КноРус».
117218, г. Москва, ул. Кедрова, д. 14, корп. 2.
Тел.: 8-495-741-46-28.
E-mail: office@knorus.ru http://www.knorus.ru

Отпечатано в ПАО «Т8 Издательские Технологии».
109316, г. Москва, Волгоградский проспект, д. 42, корп. 5.
Тел.: 8-495-221-89-80.

ISBN 978-5-406-XXXXX-X

© В.И. Соловьев, 2018
© ООО «Издательство «КноРус», 2018

Оглавление

Предисловие **6**
Методические замечания для преподавателей 7
Введение **11**
Глава 1. Данные в экономике, их визуализация и предварительная обработка **16**
§ 1.1. Данные в экономике **16**
1.1.1. Объекты, признаки и таблицы 16
1.1.2. Гистограммы 19
1.1.3. Условное форматирование 26
1.1.4. Графики и диаграммы рассеяния 30
§ 1.2. Инструменты описательной статистики в Microsoft Excel **32**
1.2.1. Измерение центра распределения 32
1.2.2. Измерение разброса данных 36
1.2.3. Описательная статистика в надстройке «Анализ данных» 38
1.2.4. Диаграммы размаха 43
§ 1.3. Визуализация качественных признаков в Microsoft Excel **49**
1.3.1. Сводные таблицы и сводные диаграммы 49
1.3.2. Таблицы сопряженности и парадокс Симпсона 56
1.3.3. Иерархия признаков 59
§ 1.4. Предварительная обработка данных **61**
1.4.1. Выбросы 61
1.4.2. Пропущенные значения 64
1.4.3. Повторяющиеся строки 68
1.4.4. Синтетические признаки 70
Глава 2. Случайные события **76**
§ 2.1. Основы комбинаторики **76**
2.1.1. Правила суммы и произведения 76
2.1.2. Комбинации без повторений 77
2.1.3. Комбинации с повторениями 80
§ 2.2. Определение вероятности **81**
2.2.1. Исчисление событий 81
2.2.2. Классическая вероятностная схема 86
2.2.3. Схема геометрических вероятностей 91
2.2.4. Статистическая вероятность 95
2.2.5. Аксиоматическое построение теории вероятностей 100
§ 2.3. Условные вероятности **107**
2.3.1. Условная вероятность. Независимость событий 107
2.3.2. Формула полной вероятности 112
2.3.3. Формула Байеса 117
§ 2.4. Последовательности испытаний **121**
2.4.1. Биномиальная схема. Формула Бернулли 121
2.4.2. Формула Пуассона 126
Глава 3. Случайные величины **129**
§ 3.1. Определение случайной величины **129**
3.1.1. Понятие случайной величины 129
3.1.2. Функция распределения случайной величины 132
§ 3.2. Дискретные случайные величины и их важнейшие числовые характеристики **134**
3.2.1. Способы задания дискретных случайных величин 134
3.2.2. Математическое ожидание, дисперсия и среднее квадратическое отклонение дискретной случайной величины 140

§ 3.3. Дискретные случайные величины, часто встречающиеся в экономической практике ... 148
3.3.1. Биномиальный закон распределения ... 148
3.3.2. Биномиальная модель ценообразования финансовых инструментов ... 150
3.3.3. Геометрический закон распределения ... 155
3.3.4. Закон распределения Пуассона ... 158
3.3.5. Простейший поток событий ... 161
3.3.6. Гипергеометрический закон распределения ... 163
§ 3.4. Абсолютно непрерывные случайные величины и их важнейшие числовые характеристики ... 165
3.4.1. Способы задания абсолютно непрерывных случайных величин ... 165
3.4.2. Математическое ожидание, дисперсия и среднее квадратическое отклонение абсолютно непрерывной случайной величины ... 167
§ 3.5. Абсолютно непрерывные случайные величины, часто встречающиеся в экономической практике ... 170
3.5.1. Равномерный закон распределения ... 170
3.5.2. Показательный закон распределения ... 172
3.5.3. Нормальный закон распределения ... 174
3.5.4. Логарифмически нормальный закон распределения и ценообразование финансовых инструментов ... 180
3.5.5. Законы распределения, важные в математической статистике ... 183
3.5.6. Смеси распределений ... 185
§ 3.6. Моменты и критические границы случайной величины ... 188
3.6.1. Моменты случайной величины ... 188
3.6.2. Асимметрия и эксцесс случайной величины ... 189
3.6.3. Квантили и процентные точки случайной величины ... 190
3.6.4. Сумма под риском (VaR) ... 192
3.6.5. Медиана и мода случайной величины ... 194
§ 3.7. Меры связи случайных величин ... 195
3.7.1. Случайные векторы и условные законы распределения ... 195
3.7.2. Ковариация и коэффициент корреляции ... 199
3.7.3. Портфель финансовых инструментов ... 205
§ 3.8. Функции случайных величин ... 210
3.8.1. Функции одной случайной величины ... 210
3.8.2. Функции нескольких случайных величин ... 213
Глава 4. Предельные теоремы теории вероятностей ... 217
§ 4.1. Закон больших чисел ... 217
4.1.1. Неравенство Чебышёва ... 217
4.1.2. Теорема Чебышёва и теорема Бернулли ... 219
§ 4.2. Центральная предельная теорема ... 221
4.2.1. Теорема Леви ... 221
4.2.2. Интегральная теорема Муавра — Лапласа ... 228
4.2.3. Математические основы теории страхования ... 231
4.2.4. Моделирование случайных величин ... 234
Глава 5. Оценивание параметров ... 244
§ 5.1. Выборочный метод ... 244
5.1.1. Генеральная совокупность и выборка ... 244
5.1.2. Допустимый объем выборки для обеспечения ее репрезентативности ... 251
5.1.3. Оценка функции распределения и плотности распределения ... 257
§ 5.2. Точечные оценки параметров ... 264
5.2.1. Свойства точечных оценок ... 264
5.2.2. Методы построения точечных оценок ... 272

§ 5.3. Интервальные оценки параметров ... 275
5.3.1. Двусторонние доверительные интервалы для математического ожидания ... 275
5.3.2. Двусторонний доверительный интервал для вероятности ... 283
5.3.3. Двусторонний доверительный интервал для дисперсии ... 285
5.3.4. Поправка на конечный объем генеральной совокупности ... 288
5.3.5. Двусторонний доверительный интервал для прогноза ... 291
5.3.6. Односторонние доверительные интервалы ... 292
Глава 6. Проверка статистических гипотез ... 296
§ 6.1. Статистические гипотезы ... 296
6.1.1. Проверка гипотез и сопутствующие риски ... 296
6.1.2. Проверка гипотез с помощью доверительных интервалов ... 306
6.1.3. Наблюдаемый уровень значимости (*p*-значение) ... 308
§ 6.2. Проверка параметрических гипотез ... 309
6.2.1. Проверка гипотезы о числовом значении математического ожидания нормального закона распределения ... 309
6.2.2. Проверка гипотезы о числовом значении вероятности ... 317
6.2.3. Проверка гипотезы о числовом значении дисперсии нормального закона распределения ... 319
6.2.4. Проверка гипотезы о равенстве дисперсий двух генеральных совокупностей ... 320
6.2.5. Проверка гипотезы о равенстве математических ожиданий двух генеральных совокупностей ... 322
6.2.6. Проверка гипотезы о равенстве генеральных долей ... 327
§ 6.3. Проверка непараметрических гипотез ... 328
6.3.1. Критерии согласия ... 328
6.3.2. Критерии независимости качественных признаков ... 332
6.3.3. Критерии сдвига ... 337
§ 6.4. Дисперсионный анализ ... 339
6.4.1. Однофакторный дисперсионный анализ ... 339
6.4.2. Двухфакторный дисперсионный анализ ... 348
Глава 7. Основы машинного обучения ... 355
§ 7.1. Модели линейной регрессии ... 355
7.1.1. Парная линейная регрессия ... 355
7.1.2. Множественная линейная регрессия ... 373
§ 7.2. Модели классификации с обучением ... 385
7.2.1. Логистическая регрессия ... 385
7.2.2. Деревья решений ... 398
§ 7.3. Модели кластерного анализа ... 407
7.3.1. Иерархические агломеративные кластер-процедуры ... 407
7.3.2. Метод *K*-средних ... 413
§ 7.4. Принципы машинного обучения ... 415
7.4.1. Парадигма машинного обучения ... 415
7.4.2. Предварительная обработка данных в *Microsoft Azure ML Studio* ... 422
7.4.3. Разработка моделей машинного обучения и оценка их качества ... 443
Заключение ... 480
Список использованной литературы ... 481
Приложения ... 483
Приложение 1. Материалы для контроля знаний ... 483
Приложение 2. Законы распределения случайных величин ... 489
Приложение 3. Интервальные оценки ... 492
Приложение 4. Проверка гипотез ... 494
Об авторе ... 497

Предисловие

Мир меняется стремительно, существенно изменилось и продолжает трансформироваться общество, экономика, производство, торговля, финансы, сфера услуг. При этом большинство изменений в значительной степени основаны на технологиях обработки и анализа данных. Сегодня анализ данных позволяет не только оптимизировать операционную деятельность и анализировать риски, но и создавать новые ценностные предложения, открывать новые рынки.

В мире, где конкурируют не товары и услуги, а модели бизнеса, где успех зависит, прежде всего, от способности понять потребности клиентов и адаптироваться к ним, преподавание университетской математики будущим экономистам и менеджерам должно в первую очередь быть направлено на формирование у студентов устойчивых компетенций в области анализа данных и обоснования принимаемых решений.

Понимая это, мы в Финансовом университете существенно изменили преподавание математических дисциплин, поставив основной целью обеспечение проактивной интеграции подготовки студентов в области математических методов, ИТ-инструментария и реальных практических задач из области профессиональной деятельности будущих выпускников.

На первом курсе, не сокращая фундаментальных основ математического анализа и линейной алгебры, мы скорректировали последовательность изложения разделов дисциплины «Математика» так, чтобы необходимый экономистам математический аппарат вводился параллельно изучению разделов микро- и макроэкономики, в которых этот аппарат применяется. Кроме того, в учебные планы добавлена дисциплина «Компьютерный практикум», в которой студенты решают практические задачи по темам, изучаемым в курсе математики, с помощью средств пакета Microsoft Excel и открытой среды R.[1]

Данный учебник служит методической поддержкой дисциплины «Анализ данных», которую студенты Финансового университета изучают на втором курсе. *Цель преподавания данной дисциплины — формирование у студентов способности собирать, анализировать и обрабатывать данные в целях создания новой ценности для бизнеса.*

Далее, на третьем курсе, студентам преподается дисциплина «Прикладные задачи машинного обучения»[2] (в которой продолжается изучение методов анализа и обработки данных применительно к созданию систем предсказательной аналитики) и дисциплина «Финтех: Инфраструктура и модели бизнеса»[3] (посвященная разработке моделей бизнеса в финансовом секторе, основанных на применении технологий обработки и анализа данных).

Учебник «Анализ данных в экономике» подготовлен в соответствии с требованиями Федерального государственного образовательного стандарта высшего образования (ФГОС 3++) по направлениям подготовки экономики и управления.

При этом **книга ориентирована не только на студентов, но и на специалистов-практиков,** желающих применять анализ данных в реальных задачах из области финансов, операционного менеджмента, маркетинга, управления рисками и т. д.

---

[1] См. *Зададаев С.А.* Математика на языке *R*: Учебник. — М.: Прометей, 2018.

[2] См. *Соловьев В.И.* Прикладные задачи машинного обучения в экономике и финансах: Практический курс с применением технологий *Microsoft Azure ML* и соревнованиями на платформе *Kaggle*: Учебник. — М.: КНОРУС, 2018.

[3] См. *Соловьев В.И.* Финтех: Инфраструктура и модели бизнеса: Учебник. — М.: КНОРУС, 2018.

Книга получилась достаточно объемная, но не из-за того, что она перегружена теорией, а потому, что в ней подробно описывается практика решения реальных задач с конкретными примерами анализа реальных данных, подробными пояснениями и большим количеством рисунков. Файлы с примерами, разобранными в учебнике, можно скачать по адресу https://is.gd/stat_book.

Автор выражает благодарность своим коллегам по департаменту анализа данных, принятия решений и финансовых технологий Финуниверситета, высказавшим замечания к рукописи учебника. Особенная признательность канд. физ.-мат. наук, доценту В.И. Глебову, который внимательно прочитал рукопись и исправил значительное количество имевшихся в ней недочетов, фактически взяв на себя функции научного редактора книги.

Автор также благодарен за доброжелательную конструктивную критику рецензентам — д-ру экон. наук, канд. физ.-мат. наук, профессору кафедры математических методов в экономике и управлении, заместителю директора центра цифровой экономики Государственного университета управления Т.М. Гатауллину и канд. физ.-мат. наук, доценту, ведущему технологическому евангелисту корпорации *Microsoft* Д.В. Сошникову. Естественно, ответственность за все оставшиеся ошибки и неточности остается за автором.

Методические замечания для преподавателей

В результате изучения данного учебника студент должен:

знать:

- основные понятия теории вероятностей, в том числе исчисление случайных событий и случайных величин, а также предельные теоремы теории вероятностей;
- основные понятия математической статистики, включая статистическую теорию выборочного метода, точечного и интервального оценивания параметров, проверки параметрических и непараметрических гипотез;

уметь:

- собирать данные из различных источников;
- визуализировать данные о количественных и качественных признаках с помощью современных технологий;
- проводить предварительную обработку данных, в том числе обработку выбросов, пропущенных значений и повторяющихся строк, генерацию синтетических признаков;
- на основе содержательной постановки практических задач выбирать вероятностные модели, наилучшим образом подходящие для исследования соответствующих случайных признаков;
- моделировать с помощью метода статистических испытаний случайные признаки, описывающие экономические системы;
- вычислять точечные и интервальные оценки параметров случайных признаков с помощью современных технологий;
- проверять статистические гипотезы, формулируя их в зависимости от последствий ошибок и проводя вычисления с использованием современных технологий;

владеть навыками:

- применения методов анализа данных к решению комплексных задач из области профессиональной деятельности;

иметь представление:

- о практических задачах, в которых использования методов теории вероятностей, прикладной статистики и обработки данных позволяет создавать новую ценность, основанную на описательной и предсказательной аналитике;
- о задачах регрессии, классификации и кластерного анализа и их применении в предсказательной аналитике;

быть компетентным:

- в формулировании задач сбора, анализа и обработки данных на основании бизнес-требований;
- в использовании современных информационных технологий сбора, анализа и обработки данных с целью описания существующего состояния организаций и рынков, а также прогнозирования будущих состояний и выработки оптимальных решений, основанных на анализе данных;
- в практической интерпретации и визуализации результатов анализа данных.

Данный учебник существенно отличается от учебников по теории вероятностей, математической статистике и анализу данных.

Во-первых, курс содержит несколько сквозных примеров, основанных на реальных данных. На этих примерах постепенно происходит изучение различных вероятностно-статистических методов и моделей, а также техник обработки, визуализации и анализа данных, и в конце формируется навык проведения комплексного анализа реальных прикладных экономических задач с использованием Microsoft Excel. Этот навык в дальнейшем может оказать студентам существенную пользу и при выполнении курсовых и дипломных работ, и в реальной практической деятельности.

По итогам изучения дисциплины студент должен выполнить индивидуальное расчетно-графическое задание, приведенное в **прил. 1.** Также в **прил. 1** приведены примеры вариантов экзаменационных заданий (в Финансовом университете студенты сдают два экзамена — по первой и второй частям курса).

Во-вторых, последовательность изложения разделов не вполне традиционна по сравнению с большинством изданных ранее учебников.

В российских учебниках теории вероятностей и математической статистики обычно изучаются вопросы теории, подкрепленные разбором упрощенных примеров и задач, предполагающих вычисления с помощью ручки и бумаги.

При этом существуют отдельные пособия, в которых обсуждаются вопросы использования Microsoft Excel, R, Python, SAS, SPSS, Statistica и других вычислительных средств для решения задач, связанных с анализом данных, но в таких пособиях основное внимание уделяется компьютерной реализации изучаемых методов без погружения в содержательную суть статистических методов..

Есть и учебники под названием «Анализ данных», но в таких учебниках под анализом данных, как правило, понимается только предсказательная аналитика, и вопросы описательной аналитики, а тем более теоретико-вероятностных и математико-статистических основ анализа данных, в таких учебниках не обсуждаются.

Наконец, предварительная обработка данных систематически не обсуждается практически ни в одном учебнике.

Данный учебник задуман как интеграция вероятностно-статистического фундамента анализа данных, его практической реализации в Microsoft Excel, а также примеров и задач, направленных на применение инструментария описательной и предсказательной аналитики в реальных ситуациях принятия решений в экономике, финансах, операционном менеджменте, маркетинге, управлении рисками, логистике и т. д.

Первая глава начинается с изучения понятий объектов и признаков. Демонстрируются примеры сбора реальных наборов данных, в том числе опубликованных в открытом доступе. На этих примерах изучаются методы предварительной обработки и визуализации данных. При этом сразу же начинают вводиться интуитивные представления о случайных величинах и их распределениях, о роли закона больших чисел, центральной предельной теоремы и нормального закона распределения в практическом применении вероятностно-статистических методов анализа данных, а также об описательной статистике и ее реализации в Microsoft Excel.

Вторая глава, посвященная случайным событиям, достаточно традиционна. В ней изучаются основы комбинаторики, классический, геометрический и статистический подходы к определению вероятности, аксиоматическое построение теории вероятностей, понятие условной вероятности, а также модели последовательностей испытаний. В отличие от многих учебников, для объединения и пересечения событий используются соответственно символы «$\cup$» (а не «+») и «$\cap$» (а не «$\cdot$»). Приводится достаточно большое число примеров простейших и более сложных практических задач, иллюстрируется технологическая реализация их решения в Microsoft Excel. В частности, подробно разбирается задача о задача о разборчивой невесте, задача о разорении, задача о разделе ставки, парадокс Монти Холла и др.

Третья глава посвящена изучению случайных величин. Достаточно серьезное внимание уделяется определению понятия случайной величины и ее функции распределения. При этом, в отличие от большинства российских учебников, где под функцией распределения понимается функция $F_X(x) = \mathbf{P}\{X < x\}$, в данном учебнике функция распределения случайной величины X определяется как $F_X(x) = \mathbf{P}\{X \leqslant x\}$. Данные два способа определения случайной величины полностью эквивалентны, а использование функций распределения $F_X(x) = \mathbf{P}\{X \leqslant x\}$ связано с тем, что в большинстве программных продуктов (в том числе в пакете Microsoft Excel, статистических библиотеках открытых сред R и Python, коммерческих продуктах SAS и SPSS)

вычисление значений функций распределения проводится по формуле $\mathbf{P}\{X \leqslant x\}$. Еще одна традиция, которую мы меняем, связана с использованием обозначений математического ожидания и дисперсии случайной величины. В российских учебниках математическое ожидание случайной величины X чаще всего обозначают «$\mathbf{M}X$», реже «$\mathbf{E}X$», а дисперсию — «$\mathbf{D}X$». Однако ни в мировой научной литературе, ни в статистических пакетах такие обозначения никогда не используются: математическое ожидание и дисперсия случайной величины X обозначаются соответственно «$\mathbf{E}(X)$» и «$\mathbf{Var}(X)$».[1] Именно эти обозначения приняты в данном учебнике. В третьей главе разобрано много практических примеров случайных величин, возникающих в прикладных задачах экономики, финансов и менеджмента. В частности, подробно описывается биномиальная модель ценообразования финансовых инструментов, использование для моделирования цен финансовых инструментов случайных величин, распределенных по логарифмически нормальному закону, вычисление суммы под риском (**VaR**). Также вводится понятие смеси распределений и приводятся практические примеры задач, в которых удобно моделировать доходы с помощью смешанных случайных величин. Случайные векторы обсуждаются достаточно компактно, но, вместе с тем, в соответствующем параграфе вводятся не только понятия ковариации и коэффициента корреляции, но и понятия условного закона распределения, условного математического ожидания и условной дисперсии. Применение случайных векторов иллюстрируется задачей об оптимальном портфеле финансовых инструментов. В конце главы изучаются функции случайных величин и, в частности, суммы случайных величин, демонстрируется суммирование нескольких независимых одинаково распределенных случайных величин. По всем изучаемым темам приводятся сведения о реализации соответствующих вычислений в Microsoft Excel и приводятся практические примеры расчетов.

Четвертая глава посвящена предельным теоремам теории вероятностей. Доказывается закон больших чисел в форме теорем Чебышёва и Бернулли, обсуждается статистическая устойчивость выборочных средних и относительных частот. Далее приводится без доказательства центральная предельная теорема в форме теоремы Леви и практические примеры исследования средних значений случайных признаков, встречающихся при исследовании экономических объектов. Также изучается интегральная теорема Муавра — Лапласа как следствие из теоремы Леви, практически исследуется сходимость биномиальных случайных величин к нормальному закону, приводятся примеры применения теоремы Муавра — Лапласа в экономических и управленческих задачах. Локальная теорема Муавра — Лапласа в учебнике не рассматривается, так как на практике вычисление в Microsoft Excel вероятностей появления конкретного числа успехов в последовательности испытаний Бернулли с помощью функции БИНОМ.РАСП дает точные результаты и при малом, и при большом числе наблюдений, и надобность в приближенном вычислении таких вероятностей отсутствует. В качестве одной из иллюстраций применения центральной предельной теоремы достаточно подробно рассматривается математическая теория страхования жизни. Завершается четвертая глава изложением теории и практики применения метода статистических испытаний для моделирования случайных признаков в экономике. Приводятся практические примеры моделирования систем массового обслуживания, в том числе пример имитационного моделирования выручки торгового предприятия, представляющей собой сумму случайного числа нескольких случайных величин. Реализация процедур моделирования с использованием вычислительных средств доступна студентам и дает представление о возможности исследования достаточно сложных случайных признаков с помощью метода статистических испытаний. Естественно, и имитационное моделирование, и другие вычисления проводятся средствами Microsoft Excel.

В пятой главе изучается теория и практика статистического оценивания параметров. Вначале подробно обсуждается выборочный метод, способы организации выборок для обеспечения репрезентативности, обсуждается сходимость выборочных средних к нормальному закону. Затем изучаются выборочные аналоги функции распределения и плотности распределения, точечные оценки параметров, их свойства и методы построения. После этого излагается теория и практика интервального оценивания параметров. Помимо традиционно изучаемых двусторонних доверительных интервалов для математического ожида-

[1] На английском языке слово *dispersion* означает «вариация, разброс значений», а слово «дисперсия» переводится на английский язык как *variance*.

ния, дисперсии и вероятности, приводятся доверительные интервалы для вероятности редких событий, доверительные интервалы для прогноза (следующего элемента выборки), односторонние доверительные интервалы, в том числе, в применении к оценке суммы под риском (**VaR**). Приводится достаточно много примеров практических расчетов в Microsoft Excel.

В шестой главе, посвященной проверке гипотез, помимо традиционно изучаемой проверки параметрических гипотез о числовом значении математического ожидания, дисперсии и вероятности, а также о равенстве математических ожиданий, дисперсий и генеральных долей, изучается также проверка непараметрических гипотез с помощью критериев согласия, критериев независимости качественных признаков и критериев сдвига, а также однофакторный и двухфакторный дисперсионный анализ. Кроме того, большое внимание уделено рискам, сопутствующим проверке гипотез, понятиям уровня значимости, мощности и *p*-значения (наблюдаемого уровня значимости). Все это иллюстрируется практикой проверки гипотез в реальных задачах маркетинга, операционного менеджмента, управления рисками и т. д. При этом, естественно, вычисления проводятся с помощью средств Microsoft Excel.

Седьмая глава знакомит читателя с основами машинного обучения. В этой главе изучаются модели парной и множественной линейной регрессии, логистической регрессии и деревьев решений, кластерного анализа. Затем обсуждается парадигма машинного обучения, метрики оценки качества моделей классификации и регрессии и приводятся два практических примера, в которых подробно описывается процесс разработки системы кредитного скоринга и системы прогнозирования продаж в облачном сервисе Microsoft Azure ML Studio. При этом данная глава носит характер вводной и имеет целью заинтересовать читателя в более подробном изучении прикладных задач машинного обучения.

Опыт проведения занятий по данному учебнику показал, что за счет переориентации на активное использование Microsoft Excel и R для автоматизации рутинных вычислений удается не только дать больше практических иллюстраций фундаментальным вероятностно-статистическим основам анализа данных, но и гораздо больше внимания уделить вопросам постановки задач на основании потребностей бизнеса, сбора, предварительной обработки и визуализации данных, практической интерпретации и использования результатов.

При этом весь материал учебника укладывается в годовой курс: первый семестр отводится на изучение инструментов визуализации и предварительной обработки данных (глава 1), а также теории вероятностей (главы 2—4), второй семестр целиком посвящен изучению описательной аналитики (главы 5—6) и введению в предсказательную аналитику (глава 7).

Еще одно отличие учебника состоит в том, что он фактически совмещен с практикумом и содержит более 450 задач для самостоятельного решения. Основная часть этих задач требует анализа данных в конкретных ситуациях принятия решений из области финансов, маркетинга, операционного менеджмента и т. д. В процессе решения таких задач студент не только закрепляет и углубляет теоретические знания, но и учится применять эти знания при постановке и решении реальных экономических задач.

В данном учебнике экономические, финансовые, управленческие приложения методов анализа данных выходят на первый план, серьезный акцент делается не только на методы решения задач, но и на постановку задач анализа данных с точки зрения бизнеса, построение вероятностных и статистических моделей, проведение расчетов в Microsoft Excel, оценку качества полученных результатов, их визуализацию и экономическую интерпретацию.

Отличительной особенностью учебника является и то, что помимо основного материала, изложенного обычным шрифтом, приводятся дополнительные материалы повышенной сложности для тех, кто хочет изучить соответствующий раздел более подробно. Такие дополнительные материалы выделены петитом. При этом есть и дополнительные теоретические материалы, и дополнительные задачи, значительная часть из которых предлагалась в различные годы на Всероссийских олимпиадах по теории вероятностей и математической статистике для студентов экономических вузов.[1]

Учебник состоит из глав, параграфов и пунктов. Пункты, рисунки, таблицы и формулы нумеруются трехступенчато: номер главы, номер параграфа, номер пункта, рисунка, таблицы или формулы. Конец доказательства или решения обозначается знаком «❑». В конце каждого пункта приводятся задачи для решения на практических занятиях и самостоятельной работы, которые имеют сквозную нумерацию по всему учебнику.

[1] См. *Репин О.А., Суханова Е.И., Ширяева Л.К.* Задачи всероссийских студенческих олимпиад по теории вероятностей и математической статистике. — СПб.: Лань, 2011.

Введение

В результате информационно-технологической революции мир уже достаточно сильно изменился и продолжает меняться. Сегодня почти у каждого человека есть по нескольку гаджетов, с которыми он не расстается, постоянно ищет и получает информацию в интернете, делится ею в социальных сетях, работает в корпоративных информационных системах, совершает покупки в интернет-магазинах, получает государственные услуги и т. д.

В 2017 г. интернетом пользовались 3,7 млрд из 7,5 млрд жителей Земли, в России — 103,0 из 146,8 млн.[1] Люди не просто «живут» в интернете, они оставляют множество «цифровых следов»:

- историю посещений интернет-ресурсов;
- историю действий в социальных сетях;
- историю переписки в электронной почте и в мессенджерах;
- историю голосовых и видеозвонков;
- историю платежей;
- историю перемещений;
- историю покупок;
- данные об интересах, намерениях и т. д.

Многие банки, инвестиционные, страховые, телекоммуникационные и торговые компании, медицинские учреждения, государственные органы и другие организации уже сегодня собирают все «следы» клиента в единую картину, чтобы лучше понимать, что клиента интересует в данный конкретный момент, и предлагать те продукты и услуги, которые клиенту действительно нужны.

В последние несколько лет происходит настоящий бум интернета вещей (*IoT, Internet of Things*), когда к интернету подключаются датчики освещенности и светильники, датчики температуры, радиаторы отопления и кондиционеры, датчики движения и скорости автомобилей, камеры видеонаблюдения и видеофиксации, станки, машины, бытовая техника и другие устройства. Ожидается, что к 2020 г. к интернету будет подключено 75 млрд устройств, от каждого из которых идет непрерывный поток данных.

Уже появилось много примеров организаций, которые с помощью инструментов анализа данных, поступающих с датчиков машин, станков, двигателей, сократили в десятки раз количество аварийных ситуаций: вместо того, чтобы ждать, когда система выйдет из строя, и высылать аварийную бригаду для ремонта, эти организации предсказывают заранее, где может случиться сбой, и предотвращают его.

И конечно, все современные организации собирают данные в информационных системах управления производством и запасами (*ERP*), в системах управления взаимоотношениями с клиентами (*CRM*), в системах электронного документооборота, управления знаниями и т. д.

В результате появляется очень много *данных*, т. е. сведений, зафиксированных в определенных (как правило, но не обязательно, компьютерных)

[1] Источник — *Internet World Stats*: *Usage and Population Statistics / Miniwatts Marketing Group*: http://internetworldstats.com.

форматах и пригодных для дальнейшего использования. Каждые два года объем данных, которым обладает человечество, увеличивается в десять раз. Чтобы данные приносили пользу, из них нужно извлекать информацию[1] и знания,[2] а для этого их нужно анализировать. Анализ данных открывает целый новый мир закономерностей нашего мироздания. И это, как мы увидим, не просто абстрактные закономерности.

Основные методы анализа данных основываются на теории вероятностей и прикладной статистике. Многие считают теорию вероятностей наукой, трудной для понимания. В общем, это действительно так. Некоторые считают эту науку абстрактной и далекой от практики. А вот это — заблуждение.

На самом деле мы сталкиваемся с вероятностными и статистическими закономерностями весьма часто и уже к окончанию школы имеем весьма солидный опыт осмысления различных ситуаций именно с вероятностных позиций. Ведь отвечаем же мы на вопрос приятеля: «Каковы твои шансы поступить в университет?» что-то вроде: «Думаю, процентов восемьдесят будет». И при этом прекрасно друг друга понимаем (хотя объяснить, что именно имел в виду отвечающий, и что именно понял спрашивающий, можем не всегда).

Мы можем ответить не только на довольно простые вопросы (типа «Сколько раз в среднем выпадет герб, если десять раз подбросить монету?», «Что более вероятно при бросании игральной кости — выпадение четной цифры или выпадение шестерки?»), так и на довольно сложные (например, «Какие ценные бумаги надежнее — государственные или корпоративные?»). Значит, к моменту поступления в университет мы уже многое знаем о вероятностных и статистических закономерностях, но пока на интуитивном уровне. Изучение теории закрепит, прояснит и весьма расширит эти знания, добавив много принципиально нового.

Большинство закономерностей, на которые мы обращали внимание до сих пор, являлись *детерминированными*, т. е. были построены по принципу «Если сложился комплекс условий S, то обязательно произойдет событие A». Например, «Если воду при нормальных условиях нагреть до 100 °С, то она обязательно закипит» или «Если на Земле подбросить вверх камень, то он обязательно упадет». Не может быть так, чтобы пять человек подбросили камни, у четверых они упали вниз, а у одного завис в воздухе.

Но бывают и случайные события. Например, пять человек подбросили по монете — и что, у всех обязательно должен выпасть герб? Да ничего подобного. Событие (например, выпадение герба) потому и называется *случайным*, что оно может произойти, а может и не произойти в результате некоторого опыта (в данном случае — в результате подбрасывания монеты).

Еще пример. Остановим на улице случайного человека и, извинившись, спросим, есть ли у него накопительный вклад в Сбербанке. Вряд ли можно точно спрогнозировать его ответ. А уж тем более невозможно точно сказать, какая сумма у него на этом депозите размещена.

---

[1] *Информация* — это результат преобразования и анализа данных при решении конкретных задач. Например, в базах данных хранятся различные данные, а по определенному запросу система управления базой данных выдает требуемую информацию.

[2] *Знания* — это обработанная, зафиксированная и проверенная практикой информация, которая использовалась и может многократно использоваться для принятия решений.

Итак, невозможно достоверно предсказать, произойдет ли случайное событие. В одном, единичном опыте — да, предсказать невозможно. Но если посмотреть соответствующую статистику, то окажется, что в начале 2017 г. суммарный объем вкладов физических лиц в Сбербанке России составлял 11 372,2 млрд руб.[1] при численности населения Российской Федерации, равной 146,8 млн человек,[2] а в начале 2018 г. — 12 069,5 млрд руб. при численности населения, равной 146,9 млн человек. Это означает, что в начале 2017 г. в среднем на одного жителя России приходился вклад в размере

$$\frac{11\,372{,}2 \cdot 10^9}{146{,}8 \cdot 10^6} = 77\,467{,}30 \text{ руб.},$$

а в начале 2018 г. — в размере

$$\frac{12\,069{,}5 \cdot 10^9}{146{,}9 \cdot 10^6} = 82\,161{,}30 \text{ руб.},$$

т. е. на 6% больше (учитывая, что инфляция в 2017 г. была зафиксирована на уровне 2,51%,[3] можно констатировать, что люди стали больше сберегать, т. е. жить стало лучше).

Но ведь и для банка в целом не имеет особого смысла информация о размере вклада конкретного человека. Для принятия решений банку нужна статистическая информация — обобщенная и усредненная.

Итак, огромный массив случайных событий — взаимодействий жителей России со Сбербанком — имеет, оказывается, четкую закономерность: в среднем каждый житель России в 2018 г. имел примерно 82,2 тыс. руб. на депозите в Сбербанке. Такие закономерности называются *статистическими*.

Теория вероятностей изучает свойства массовых случайных событий, которые могут многократно повторяться при воспроизведении определенного комплекса условий. На теорию вероятностей опирается *математическая и прикладная статистика*, которая изучает методы сбора, систематизации и обработки результатов наблюдений массовых случайных событий с целью выявления статистических закономерностей.

Анализ данных — это процесс обнаружения в имеющихся данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Из определения понятно, что в основном анализ данных концентрируется на практическом применении статистических методов не только для того, чтобы делать практические выводы описательного характера (описательная или дескриптивная аналитика), но и для того, чтобы предсказывать будущее поведение изучаемых объектов (предсказательная или предиктивная аналитика) и давать рекомендации по принятию решений (предписывающая или прескриптивная аналитика).

Важнейшие особенности анализа данных по сравнению с традиционной статистикой таковы:

---

[1] Источник — Росстат: http://www.gks.ru/bgd/regl/b18_11/IssWWW.exe/Stg/d02/23-10.doc.
[2] Источник — Росстат: http://www.gks.ru/bgd/regl/b18_11/IssWWW.exe/Stg/d01/05-01.doc.
[3] Источник — Росстат: http://www.gks.ru/free_doc/new_site/prices/potr/I_ipc.xlsx.

- данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы; поэтому понимание данных в конкретных приложениях требует значительных интеллектуальных усилий;
- сами алгоритмы анализа данных могут обладать «элементами интеллекта», в частности, способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений; разработка таких алгоритмов также требует значительных интеллектуальных усилий;
- процессы переработки «сырых» данных в информацию, а информации в знания уже не могут быть выполнены по старинке вручную и требуют нетривиальной автоматизации.

Статистика и анализ данных — не абстрактные, а чисто практические науки. При этом значительная часть методов анализа данных основывается на двух «чудесах». Предположим, что есть какой-то случайный признак[1] X, например, ущерб по договору обязательного страхования автогражданской ответственности, и есть 10 000 владельцев автомобилей, застраховавших свою ответственность. Пусть X_i — ущерб по i-му договору. Какой будет ущерб по каждому из договоров в течение года, в момент заключения договоров точно предсказать невозможно. Но оказывается, что средний ущерб на один договор можно предсказать более точно. Если величины $X_1, X_2, \ldots, X_n$ независимы и в определенном смысле одинаковы, то среднее значение при увеличении n стремится к совершенно определенному числу. Поэтому если разделить страхователей на однородные группы, чтобы в каждую группу вошли страхователи одного пола и возраста, ездящие на примерно одинаковых автомобилях и имеющие одинаковое количество страховых случаев в прошлом, то мы как раз и получаем ситуацию, когда $X_1, X_2, \ldots, X_n$ независимы и одинаково распределены[2]. Поэтому средний ущерб

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

для этих 10 000 человек будет примерно равен среднему ущербу по 8000 договорам, заключенным людьми из той же группы в прошлом году.(с поправкой на инфляцию). Это «чудо» называется законом больших чисел. Именно на этом «чуде» основана вся страховая деятельность.

Можно рассмотреть совершенно другой признак — размер чека покупателя определенного магазина. Кто завтра придет в магазин, что купит и какую сумму оставит, сегодня сказать нельзя. Но благодаря закону больших чисел можно утверждать, что поскольку в магазине независимо друг от друга совершают покупки много покупателей (конечно, мы говорим о магазине, в который ежедневно ходят достаточно много людей), примерно одинаковых по материальному положению и социальному статусу, то средний чек завтра будет примерно такой же, как средний чек сегодня, примерно такой же, как средний чек вчера, и т. д.

---

[1] Под *признаком* понимается определенная характеристика объекта.

[2] Понятие распределения является одним из основных в анализе данных. Точное определение этого понятия будет дано в главе 3.

Второе «чудо» называется центральной предельной теоремой. Эта теорема утверждает, что сумма большого числа независимых одинаково распределенных величин

$$X_1 + X_2 + \cdots + X_n,$$

распределена по так называемому нормальному закону. Мы не знаем, закон распределения стоимости покупок, совершенных каждым конкретным покупателем, но если покупателей много, они совершают покупки примерно одинаково и независимо друг от друга, то суммарная выручка будет иметь нормальный закон распределения. Неизвестно, как устроен ущерб по конкретному договору страхования гражданской ответственности автовладельца, но если страхователей много, они ездят примерно одинаково и независимо друг от друга, то суммарный ущерб будет распределен приблизительно по нормальному закону.

Основные понятия теории вероятностей были заложены в XV—XVII вв. Л. Пачоли, Дж. Кардано, Б. Паскалем и П. Ферма. Эти понятия зародились в результате попыток математически описать азартные игры и выработать рекомендации для игроков. В дальнейшем теория вероятностей и математическая статистика развивались в работах Я. Бернулли, Т. Байеса, А. Муавра, П. Лапласа, К. Гаусса, С. Пуассона, В. Я. Буняковского, П. Л. Чебышёва, А. М. Ляпунова, А. А. Маркова, Е. Е. Слуцкого, С. Н. Бернштейна, Р. Мизеса, В. Госсета (писавшего под псевдонимом *Стьюдент*), Р. Фишера, К. Пирсона, А. Вальда, Ю. Неймана, А.Н. Колмогорова, А. Я. Хинчина, В. И. Романовского, Б. В. Гнеденко, Ю. В. Линника, Ю. В. Прохорова, А. Н. Ширяева и др.

Особенно отметим академика А. Н. Колмогорова, впервые сформулировавшего в 1933 г. законченную систему аксиом теории вероятностей, из которой вытекают другие положения этой науки.

Сегодня аппарат анализа данных уже активно используется в банках, инвестиционных и страховых компаниях, на производственных и торговых предприятиях, в телекоммуникациях, медицине, сельском хозяйстве, государственных органах и т. д.

Описательная, предсказательная и предписывающая аналитика помогает существенно повысить эффективность решения задач кредитного скоринга, борьбы с мошенничеством, разработки торговых стратегий на финансовых рынках, создания рекомендаций и персонализированных предложений для клиентов, привлечения новых и борьбы с оттоком существующих потребителей, мониторинга и превентивного ремонта оборудования, медицинской диагностики и выбора схем лечения и др.

Многие такие задачи рассматриваются в этой книге в качестве примеров применения изучаемых методов.

Глава 1. Данные в экономике, их визуализация и предварительная обработка

§ 1.1. Данные в экономике

1.1.1. Объекты, признаки и таблицы

Данные — это совокупность сведений, зафиксированных на определенных носителях в форме, пригодной для постоянного хранения, передачи и обработки.

Во многих практических задачах мы имеем дело со структурированными данными, которые можно представить в форме таблицы, каждая строка которой соответствует определенному объекту, а каждый столбец — определенному признаку.

Объектами могут выступать люди, животные, города, страны, акции, кредиты, товары и т. д. Например, можно составить таблицу, в которой строки будут соответствовать студентам определенной группы, и для каждого из студентов будут зафиксированы четыре признака: возраст, пол, рост и вес.

Можно для тех же студентов составить другую таблицу, в которой признаками будут выступать оценки, полученные в предыдущем семестре.

В третьей таблице можно для каждого города России зафиксировать такие признаки, как численность населения, уровень среднедушевого дохода и суммарный объем кредитов, выданных жителям этого города некоторым банком.

В четвертой таблице можно зафиксировать для заемщиков некоторого банка их пол, возраст, уровень образования и размер выданного кредита.

Признаки бывают *количественными*, т. е. выражаемыми в числовой форме, и *качественными* — не имеющими числового выражения. Например, рост студента и количество отличных оценок, полученных им в прошлом семестре, — это количественные признаки, а пол заемщика и его уровень образования с качественные признаки.

В свою очередь, количественные признаки могут быть *дискретными* (принимающие значения из некоторого конечного или счетного множества, как правило, из множества целых чисел) и *непрерывными* (которые могут принимать любое вещественное значение из определенного интервала или на всей числовой прямой). Так, рост студента является непрерывным признаком, а количество «пятерок», полученных в предыдущем семестре, — дискретным.

Качественные признаки бывают номинальными и порядковыми. *Номинальные* признаки — это такие, которые можно отличать друг от друга, но нельзя сравнивать. *Порядковые* признаки сравнивать можно. Например, пол заемщика — это номинальный признак, а уровень образования — порядковый.

Не все данные являются структурированными. Так, например, поисковый запрос в Яндексе по словосочетанию «анализ данных» выдает 70 млн результатов, среди которых текстовые статьи, новости, изображения, видео, места на карте, товары в магазинах — эти данные невозможно представить в виде таблицы.

Далее в этой книге будут рассматриваться структурированные данные.

Пример 1.1.1. В табл. 1.1.1 представлены данные о кредитах, выданных некоторым банком (файл **«Кредиты.xlsx»**). В этой таблице объектами выступают за-

ключенные банком кредитные договора, и для каждого договора зафиксированы определенные признаки:

- номер договора;
- статус кредита (возвращен или нет);
- размер кредита (в рублях);
- срок кредита (краткосрочный или долгосрочный);
- кредитный рейтинг (по данным бюро кредитных историй);
- годовой доход (в рублях);
- стаж на текущем месте работы;
- недвижимость (в собственности, в ипотеке или в аренде);
- цель кредита;
- размер ежемесячного платежа по кредиту (в рублях);
- срок кредитной истории (в годах);
- срок с последнего нарушения условий кредитного договора (в месяцах);
- число кредитных карт, имеющихся у заемщика;
- число нарушений условий кредитных договоров.

При этом номер договора, кредитный рейтинг, срок с последнего нарушения условий кредитного договора, число имеющихся у заемщика кредитных карт и число допущенных заемщиком нарушений условий кредитных договоров — это дискретные признаки; размер выданного кредита, размер годового дохода, размер ежемесячного платежа по кредиту и срок кредитной истории — непрерывные признаки.

Цель кредита представляет собой номинальный признак — нельзя сказать, что кредит на покупку автомобиля больше или лучше, чем кредит на ремонт жилья. Для порядковых признаков сравнение возможно: с точки зрения банка возвращенный кредит лучше, чем невозвращенный; статус обладания недвижимостью также можно упорядочить: наличие недвижимости в собственности лучше, чем в ипотеке, а недвижимость в ипотеке лучше, чем арендованная. Но сказать, что ипотека в какое-то определенное число раз хуже, чем собственность, нельзя.

В данном случае срок кредита — это порядковый признак, а не числовой, поскольку он может принимать два значения: краткосрочный и долгосрочный, и очевидно, что срок краткосрочного кредита меньше, чем долгосрочного. Стаж на текущем месте работы — это также порядковый признак, так как помимо значений «1 год», «2 года», «3 года», ..., «9 лет» есть еще «< 1 года» и «10+ лет»: непонятно, во сколько раз «10+ лет» больше, чем «2 года» или чем «< 1 года». ❑

Задачи

1. Для каждого студента группы в таблице записаны: оценка по математике в первом семестре; размер обуви; любимый фильм; расстояние от места проживания до места учебы. Какие из этих признаков являются: *а*) количественными; *б*) качественными; *в*) дискретными; *г*) непрерывными; *д*) номинальными; *е*) порядковыми?

2. Приведите два — три примера наборов данных, структурированных в табличной форме.

Таблица 1.1.1

Данные о выданных кредитах

Номер договора	Статус кредита	Размер кредита	Срок кредита	Кредитный рейтинг	Годовой доход	Стаж на текущем месте	Недвижимость	Цель	Ежемесячный платеж	Срок кредитной истории	Срок с последнего нарушения	Число кредитных карт	Число нарушений кредитных договоров
1	погашен	445 412,00 ₽	краткосрочный	709	1 167 493,00 ₽	8 лет	в ипотеке	ремонт жилья	5 214,74 ₽	17,2		1	1
2	погашен	262 328,00 ₽	краткосрочный			10+ лет	в ипотеке	консолидация кредитов	33 295,98 ₽	21,1	8	3	1
3	погашен		краткосрочный	741	2 231 892,00 ₽	8 лет	в собственности	консолидация кредитов	29 200,53 ₽	14,9	29	2	1
4	погашен	347 666,00 ₽	долгосрочный	721	806 949,00 ₽	3 года	в собственности	консолидация кредитов	8 741,90 ₽	12		9	0
5	погашен	176 220,00 ₽	краткосрочный			5 лет	в аренде	консолидация кредитов	20 639,70 ₽	6,1		1	0
6	не погашен	206 602,00 ₽	краткосрочный	729	896 857,00 ₽	10+ лет	в ипотеке	консолидация кредитов	16 367,74 ₽	17,3		6	0
7	погашен	217 646,00 ₽	краткосрочный	730	1 184 194,00 ₽	< 1 года	в ипотеке	консолидация кредитов	10 855,08 ₽	19,6	10	1	1
8	не погашен	648 714,00 ₽	долгосрочный			< 1 года	в ипотеке	приобретение жилья	14 806,13 ₽	8,2	8	1	1
9	погашен	548 746,00 ₽	краткосрочный	678	2 559 110,00 ₽	2 года	в аренде	консолидация кредитов	18 660,28 ₽	22,6	33	4	3
10	погашен	215 952,00 ₽	краткосрочный	739	1 454 735,00 ₽	< 1 года	в аренде	консолидация кредитов	39 277,75 ₽	13,9		2	0
11	погашен		краткосрочный	728	714 628,00 ₽	3 года	в аренде	консолидация кредитов	11 851,06 ₽	16	76	16	0
12	погашен	541 970,00 ₽	краткосрочный			10+ лет	в ипотеке	ремонт жилья	23 568,55 ₽	23,2		2	0
13	погашен		краткосрочный	740	776 188,00 ₽	< 1 года	в собственности	консолидация кредитов	11 578,22 ₽	8,5	25	6	2
14	погашен		краткосрочный	743	1 560 907,00 ₽	4 года	в аренде	консолидация кредитов	17 560,37 ₽	13,3		1	1
15	погашен	234 124,00 ₽	краткосрочный	727	693 234,00 ₽	10+ лет	в аренде	консолидация кредитов	14 211,24 ₽	24,7	46	1	1
16	погашен	449 020,00 ₽	долгосрочный			9 лет	в собственности	консолидация кредитов	18 904,81 ₽	19,4		8	0
17	не погашен	653 004,00 ₽	долгосрочный			7 лет	в ипотеке	консолидация кредитов	14 537,09 ₽	20,5		9	0
18	погашен	666 204,00 ₽	долгосрочный	723	1 821 967,00 ₽	10+ лет	в ипотеке	консолидация кредитов	17 612,24 ₽	22	34	1	1
19	погашен	66 396,00 ₽	краткосрочный			10+ лет	в аренде	консолидация кредитов	9 898,81 ₽	27,1		2	1
20	погашен	390 390,00 ₽	краткосрочный	747	1 791 738,00 ₽	8 лет	в ипотеке	ремонт жилья	2 478,55 ₽	22,7		6	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

1.1.2. Гистограммы

Мы уже обсуждали, что в значительной степени статистические методы основаны на том, что многие случайные признаки, с которыми мы встречаемся, распределены по нормальному закону. Чтобы узнать, имеет ли некоторый признак нормальный закон распределения, как правило, начинают с построения так называемых гистограмм.

Гистограмма — это фигура, состоящая из прямоугольников, в основании которых лежат интервалы группирования признака, а высота каждого прямоугольника соответствует количеству значений, попавших в соответствующий интервал. Если некоторый признак распределен по нормальному закону, то его гистограмма имеет вид симметричного колокола, как на рис. 1.1.1.

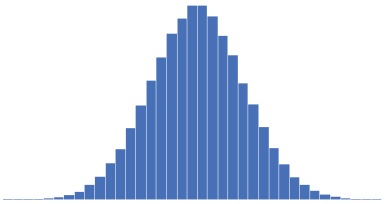


Рис. 1.1.1. *Гистограмма среднего чека в сети продовольственных магазинов*

Пример 1.1.2. Построим гистограмму роста студентов группы. Для этого создадим простую веб-форму для опроса, например, на платформе Microsoft Office Forms: https://forms.office.com. В форме будет единственный вопрос: «Ваш рост» (рис. 1.1.2, *а*). Результаты опроса студентов некоторой группы представлены в файле «Рост студентов.xlsx» (рис. 1.1.2, *б*).

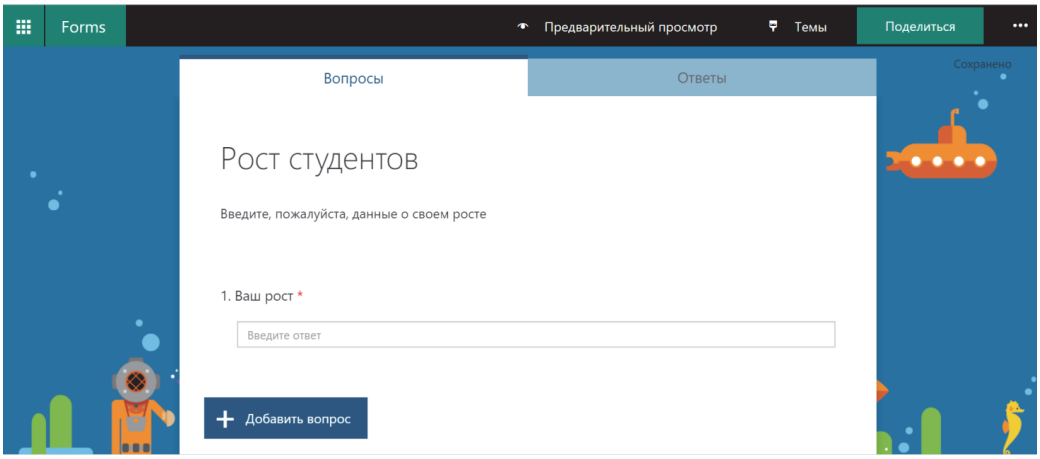
Чтобы построить гистограмму роста студентов, выделим соответствующий столбец E, на вкладке «Вставка» нажмем кнопку «Вставить статистическую диаграмму», далее выберем тип статистической диаграммы — «Гистограмма». После того, как гистограмма будет построена, нажмем на нее правой кнопкой мыши и в появившемся всплывающем меню выберем «Добавить подписи данных» — она примет вид как на рис. 1.1.2, *в*). По гистограмме мы видим, что из 30 студентов группы пятеро имеют рост от 167,0 до 172,5 см, у 11 человек рост составляет от 172,5 до 178,0 см, еще 11 человек имеют рост в границах от 178,0 до 183,5 см, оставшиеся три студента имеют рост от 183,5 до 189,0 см.

Видно, что гистограмма похожа на симметричный колокол нормального закона распределения, однако кажется, что выглядит она слишком просто: в ней всего четыре интервала, длина которых, равная 5,5 см, была выбрана автоматически. Попробуем сделать больше интервалов. Для этого нажмем правой кнопкой на ось абсцисс, выберем «Формат оси» и укажем длину интервала, равную 3 см (рис. 1.1.2, *г*). Гистограмма роста студентов стала больше похожа на колокол.

Если попробовать выбрать еще более короткие интервалы (на гистограмме, представленной на рис. 1.1.2, *д*, длина интервала равна 1,5 см), то мы увидим, что сходство гистограммы с колоколом ослабилось: во многих интервалах слишком мало значений, а в три из интервалов не попал рост ни одного студента. ❑

Замечание. Раньше, когда данных было мало, для выбора длины интервала группирования руководствовались различными правилами: — правилом Стерджеса, правилом Скотта,

правилом Фридмана — Диакониса. Оказалось, что при больших объемах данных они работают хуже, чем интуитивный подбор длины интервала или количества интервалов. На практике лучше всего попробовать несколько вариантов и выбрать из них тот, при котором гистограмма выглядит наиболее гладкой.



а)

	A	B	C	D	E
1	Время начала	Время выполнения	Почта	Имя	Ваш рост
2	08.09.2016 15:37:38	08.09.2016 15:37:58	anonymous		182,00
3	08.09.2016 15:38:02	08.09.2016 15:38:05	anonymous		175,00
4	08.09.2016 15:38:14	08.09.2016 15:38:18	anonymous		186,00
5	08.09.2016 15:38:20	08.09.2016 15:38:24	anonymous		175,00
6	08.09.2016 15:38:26	08.09.2016 15:38:36	anonymous		188,00
7	08.09.2016 15:38:38	08.09.2016 15:38:41	anonymous		177,00
8	08.09.2016 15:38:51	08.09.2016 15:38:54	anonymous		176,00
9	08.09.2016 15:38:56	08.09.2016 15:38:58	anonymous		178,00
10	08.09.2016 15:39:00	08.09.2016 15:39:03	anonymous		183,00
11	08.09.2016 15:39:11	08.09.2016 15:39:14	anonymous		185,00
12	08.09.2016 15:39:15	08.09.2016 15:39:18	anonymous		167,00
13	08.09.2016 15:39:20	08.09.2016 15:39:23	anonymous		180,00
14	08.09.2016 15:39:30	08.09.2016 15:39:33	anonymous		182,00
15	08.09.2016 15:39:34	08.09.2016 15:39:37	anonymous		179,00
16	08.09.2016 15:39:40	08.09.2016 15:39:42	anonymous		176,00
17	08.09.2016 15:39:44	08.09.2016 15:39:53	anonymous		179,00
18	08.09.2016 15:39:55	08.09.2016 15:39:57	anonymous		172,00
19	08.09.2016 15:39:59	08.09.2016 15:40:01	anonymous		177,00
20	08.09.2016 15:40:09	08.09.2016 15:40:12	anonymous		183,00
21	08.09.2016 15:40:14	08.09.2016 15:40:16	anonymous		168,00
22	08.09.2016 15:40:18	08.09.2016 15:40:20	anonymous		180,00
23	08.09.2016 15:40:29	08.09.2016 15:40:32	anonymous		179,00
24	08.09.2016 15:40:33	08.09.2016 15:40:36	anonymous		172,00
25	08.09.2016 15:40:38	08.09.2016 15:40:40	anonymous		177,00
26	08.09.2016 15:40:51	08.09.2016 15:40:54	anonymous		175,00
27	08.09.2016 15:40:56	08.09.2016 15:40:59	anonymous		173,00
28	08.09.2016 15:41:00	08.09.2016 15:41:04	anonymous		179,00
29	08.09.2016 15:41:17	08.09.2016 15:41:20	anonymous		176,00
30	08.09.2016 15:41:22	08.09.2016 15:41:26	anonymous		179,50
31	08.09.2016 15:41:28	08.09.2016 15:41:37	anonymous		172,00

б)



в)



г)



д)

Рис. 1.1.2. *Опрос студентов об их росте: веб-форма (а), результаты опроса (б), гистограмма, построенная автоматически (в), гистограмма с длиной интервала, равной 4 см (г), гистограмма с длиной интервала, равной 1, 5 см (д)*

Пример 1.1.3. Рассмотрим динамику поведения котировок акций российских компаний. Данные о котировках акций на Московской бирже можно получить у официальных распространителей биржевой информации: компаний *Bloomberg*, *Thomson Reuters*, Росбизнесконсалтинг и др. Часть из них распространяют информацию по платной подписке, другая часть предоставляет эту информацию на своих сайтах бесплатно. Например, на сайте информационного агентства «МФД-ИнфоЦентр» данные о котировках можно получить на странице https://mfd.ru/export/ (рис. 1.1.3).

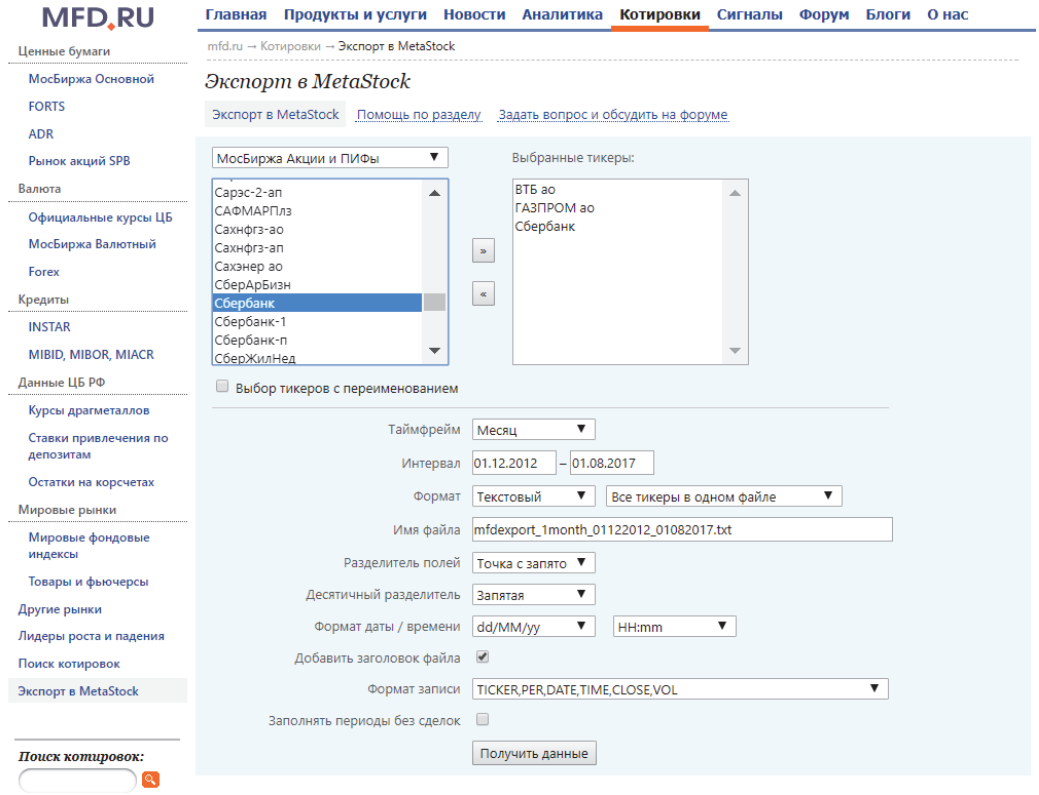


Рис. 1.1.3. *Запрос на получение котировок акций на сайте информационного агентства «МФД-ИнфоЦентр»*

Далее нужно выбрать **торговую площадку** (например, **«Мосбиржа Акции и ПИФы»** и инструменты (например, **«ВТБ ао»**, **«ГАЗПРОМ ао»**, **«Сбербанк»**). Можно получать данные в различных временных масштабах (таймфреймах) — месячные, дневные, часовые, минутные, тиковые данные и т. д. Выберем **месячный таймфрейм**, **интервал** с 01.12.2012 по 01.08.2017, **текстовый формат** файла. В качестве **разделителей полей** выберем точку с запятой, **даты** попросим представить в формате **«dd/MM/yy»**, а в качестве **формата файла** укажем: **«TICKER, PER, DATE, TIME, CLOSE, VOL»**, чтобы получить для каждого месяца из диапазона с декабря 2012 г. по август 2017 г. значения шести признаков:

- TICKER — название акции;
- PER — период (в нашем случае во всех строках это будет **«M»** — месяц);
- DATE — дата;
- TIME — время;

- CLOSE — цена закрытия (т. е. цена, по которой была совершена последняя сделка в данном периоде);
- VOL — объем торговли.

В результате мы получим текстовый файл (файл **«Акции.txt»**), первые строки которого выглядят так:

```
<TICKER>; <PER>; <DATE>; <TIME>; <CLOSE>; <VOL>
ВТБ ао; M; 01/12/12; 00:00; 0,05359; 423765060000
ВТБ ао; M; 01/01/13; 00:00; 0,05582; 620659410000
ВТБ ао; M; 01/02/13; 00:00; 0,05588; 956736250000
ВТБ ао; M; 01/03/13; 00:00; 0,0497; 684089840000
ВТБ ао; M; 01/04/13; 00:00; 0,04929; 1203587690000
……
```

Если мы откроем этот файл в пакете Microsoft Excel, то автоматически запустится мастер преобразования текстовых файлов в электронные таблицы, в котором нужно указать, что данные представлены в формате **с разделителями**, **в первой строке содержатся заголовки** признаков (рис. 1.1.4, *а*), **символом — разделителем является точка с запятой** (рис. 1.1.4, *б*), а **столбцы представлены в общем формате** (рис. 1.1.4, *в*).

В результате откроется таблица, как на рис. 1.1.5, *а*.

Удалим столбцы <PER> и <TIME>, которые не содержат полезной для нас информации: признак <PER> во всех строках содержит значение **«M»** — месяц, признак <TIME> во всех строках соответствует полуночи.

Далее вырежем данные об акциях Газпрома и вставим их справа от данных об акциях ВТБ за соответствующие даты, затем вырежем данные об акциях Сбербанка и вставим их справа от данных об акциях Газпрома.

Укажем **форматы ячеек**: для цен акций ВТБ — число с пятью знаками после запятой, для цен акций Газпрома и Сбербанка — число с двумя знаками после запятой, для объемов всех трех акций — число без знаков после запятой, но с разделителем групп десятичных разрядов.

Переименуем столбцы и поменяем их местами так, чтобы вначале шли данные о ценах, потом об объемах: см. рис. 1.1.5, *б*.

Теперь отформатируем наши данные как таблицу. Для этого нажмем на кнопку **«Форматировать как таблицу»**, расположенную на вкладке **«Главная»**. Главное, что дает такое форматирование, — не внешний вид, а возможность работы с формулами для столбцов.

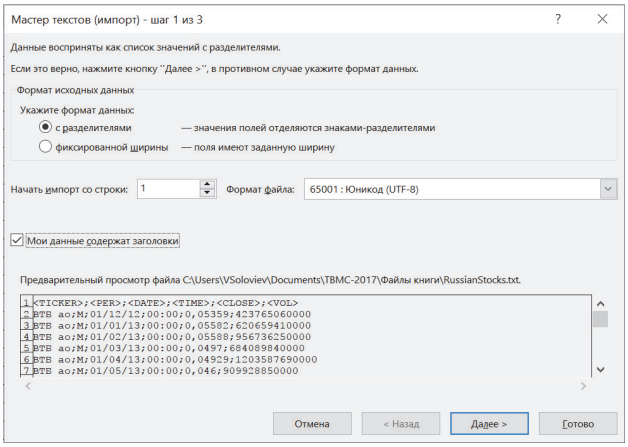
Рассчитаем для каждой из трех акций месячную доходность

$$r_t = \frac{s_t - s_{t-1}}{s_{t-1}},$$

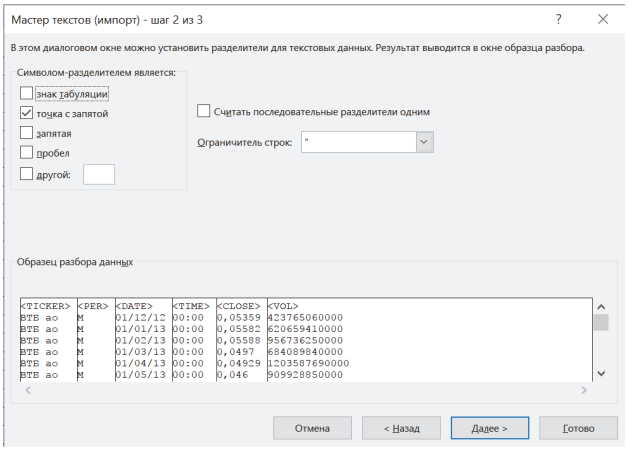
которая показывает, на сколько процентов[1] цена в данном месяце s_t больше, чем цена в прошлом месяце s_{t-1}.

Для этого вставим в таблицу три новых столбца: **«ВТБ – доходность»** (справа от столбца **«ВТБ – объем»**), **«Газпром – доходность»** (справа от столбца **«Газпром– объем»**) и **«Сбербанк – доходность»** (справа от столбца **«Сбербанк – объем»**).

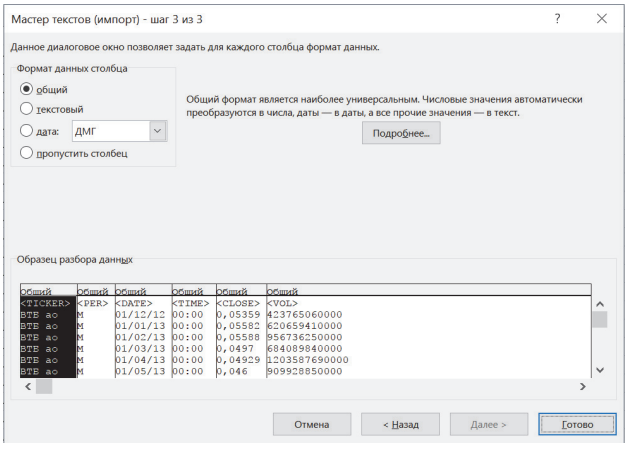
[1] *Процент* — это обозначение сотой доли, т. е. 1% = 0,1 = 1/100, 3% = 0,03 = 3/100 и т.д.



а)



б)



в)

Рис. 1.1.4. *Преобразование текстового файла в таблицу Microsoft Excel: выбор формата файла* (*а*)*, символа — разделителя* (*б*) *и формата столбцов* (*в*)

	A	B	C	D	E	F
1	<TICKER>	<PER>	<DATE>	<TIME>	<CLOSE>	<VOL>
2	ВТБ ао	M	01.12.2012	0:00	0,05359	4,23765E+11
3	ВТБ ао	M	01.01.2013	0:00	0,05582	6,20659E+11
4	ВТБ ао	M	01.02.2013	0:00	0,05588	9,56736E+11
5	ВТБ ао	M	01.03.2013	0:00	0,0497	6,8409E+11
6	…	…	…	…	…	…

а)

	A	B	C	D	E	F	G
1	Дата	ВТБ – цена	Газпром – цена	Сбербанк – цена	ВТБ – объем	Газпром – объем	Сбербанк – объем
2	01.12.2012	0,05359	143,70	92,94	423 765 060 000	638 968 460	1 191 987 680
3	01.01.2013	0,05582	142,09	109,59	620 659 410 000	495 283 520	1 723 351 580
4	01.02.2013	0,05588	137,40	104,57	956 736 250 000	733 770 920	1 612 212 000
5	01.03.2013	0,04970	134,08	98,86	684 089 840 000	953 973 070	1 691 490 080
…	…	…	…	…	…	…	…

б)

	A	B	C	D	E	F	G	H	…
1	Дата	ВТБ – цена	Газпром – цена	Сбербанк – цена	ВТБ – объем	Газпром – объем	Сбербанк – объем	ВТБ – доходность	…
2	01.12.2012	0,05359	143,70	92,94	423 765 060 000	638 968 460	1 191 987 680	#ЗНАЧ!	…
3	01.01.2013	0,05582	142,09	109,59	620 659 410 000	495 283 520	1 723 351 580	0,0416	…
4	01.02.2013	0,05588	137,40	104,57	956 736 250 000	733 770 920	1 612 212 000	0,0011	…
5	01.03.2013	0,04970	134,08	98,86	684 089 840 000	953 973 070	1 691 490 080	–0,1106	…
…	…	…	…	…	…	…	…	…	…

в)

Рис. 1.1.5. *Преобразование данных о котировках акций: исходные данные* (*а*), *преобразованные данные* (*б*), *данные, отформатированные в виде таблицы с добавленными столбцами доходностей* (*в*)

Введем в ячейку H3, соответствующую доходности акций ВТБ в январе 2013 г., формулу

=([@[ВТБ – цена]]–B2)/B2

и заметим, что эта формула автоматически распространилась на весь столбец, соответствующий данному признаку. Правда, в строке, соответствующей декабрю 2012 г., вместо числа написано «#ЗНАЧ!», а в остальных строках стоят нули.

Дело в том, что формат в новом столбце автоматически устанавливается таким же, как в столбце слева. В данном случае слева у нас стоит объем торгов, который отформатирован, как число с нулем знаков после запятой, поэтому и доходность отформатировалась как целое число. Выделим столбец «ВТБ – доходность» и изменим формат ячеек на число с четырьмя знаками после запятой.

Теперь нули превратились в положительные и отрицательные дробные числа (например, в январе 2013 г. доходность по акциям ВТБ составила 0,0416, т. е. цена акций ВТБ на конец января 2013 г. на 4,16% больше, чем цена на конец декабря 2012 г.).

Аббревиатура «#ЗНАЧ!» означает, что значение доходности за декабрь 2012 г. неизвестно. Это действительно так, поскольку для вычисления доходности в данном месяце необходимо знать значение цены в этом месяце и в предыдущем, а цена за ноябрь 2012 г. в нашем файле отсутствует. Таким образом, мы вычислили доходности за все месяцы, кроме декабря 2012 г. Аналогичным способом рассчитаем доходности по акциям Газпрома и Сбербанка (рис. 1.1.5, *в*). Удалим значения «#ЗНАЧ!», оставив соответствующие ячейки пустыми.

Добавим в таблицу еще три столбца: **«ВТБ – логарифм объема»**, **«Газпром – логарифм объема»** и **«Сбербанк – логарифм объема»**. В этих столбцах с помощью функции LN рассчитаем натуральные логарифмы объемов торгов соответствующих акций.

Теперь построим гистограммы для всех характеристик акций. Для единообразия каждую из гистограмм построим с восемью интервалами группирования (для этого после того, как гистограмма будет построена, выберем **«Формат оси»** в меню, вызываемом нажатием правой кнопки на оси абсцисс, и укажем 8 в качестве **количества интервалов**).

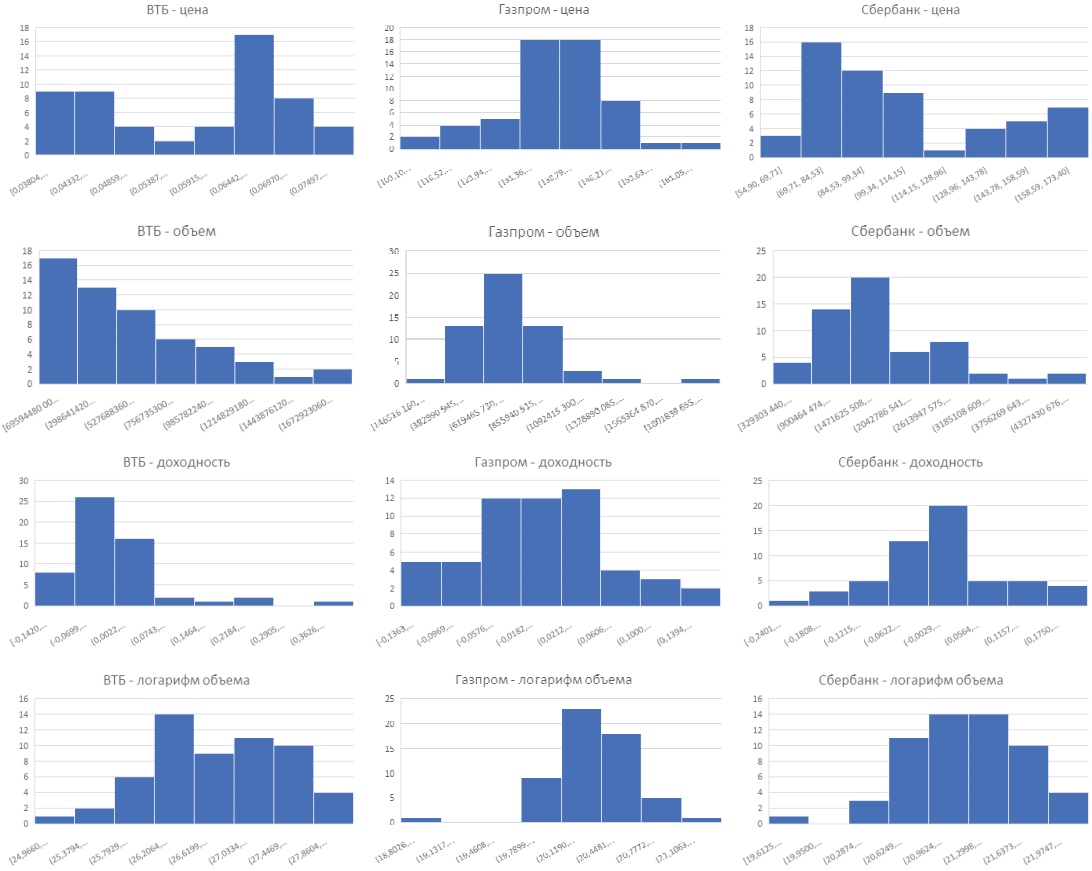


Рис. 1.1.6. *Гистограммы цен, объемов торгов, доходностей и логарифмов объемов торгов для акций ВТБ, Газпрома и Сбербанка*

Анализ результатов (рис. 1.1.6) позволяет сделать следующие заключения:

- гистограмма цены акций Газпрома похожа на симметричный колокол, в отличие от гистограмм цен акций ВТБ и Сбербанка, которые вообще ни на какой колокол не похожи;
- гистограммы объемов торгов по всем трем акциям (в особенности по акциям ВТБ) с к о ш е н ы в п р а в о, т. е. правая часть «колокола» является более пологой, чем левая часть;
- гистограммы доходностей акций Газпрома и Сбербанка напоминают кривую нормального закона, а гистограмма доходности акций ВТБ скошена вправо;

- гистограммы логарифмов объемов торгов по всем трем акциям (в особенности по акциям ВТБ) скошены влево, т. е. более пологой является левая часть «колокола».

Результаты сохранены на рабочем листе «Гистограммы» в файле «Акции.xlsx». ❑

Задачи

3. Известно, какое количество потребительских кредитов было выдано в каждый из 20 последних рабочих дней: 60; 20; 70; 70; 30; 20; 50; 50; 40; 60; 30; 40; 30; 50; 50; 60; 50; 60; 40; 40. Постройте гистограмму количества выданных потребительских кредитов и опишите ее.

4. С помощью веб-формы соберите данные о росте студентов своей группы, постройте гистограмму роста студентов и опишите ее: является ли она симметричной или скошенной, похожа ли она на колокол нормального распределения.

5. Соберите недельные данные о ценах закрытия и объемах торгов по трем акциям, обращающимся на Московской бирже, с 01.01.2010 по сегодняшний день. Рассчитайте для каждой акции недельные доходности и натуральные логарифмы объемов торгов. Постройте гистограммы всех показателей и опишите их: являются ли они симметричными или скошенными, похожи ли они на колокол нормального распределения.

1.1.3. Условное форматирование

Когда мы имеем дело с каким-либо набором данных, перед глазами у нас таблица, в которой много ячеек с числовыми и, возможно, текстовыми значениями. Как правило, посмотрев на такую таблицу невозможно определить никаких закономерностей в изучаемых данных.

На самом деле, как правило, много полезной информации «лежит на поверхности», и такую информацию можно получить с помощью визуализации данных, поскольку визуальная информация воспринимается человеком гораздо лучше, чем сухие числа

Визуализация исходных данных полезна для формирования ожиданий от анализа данных, для определения степени пригодности данных к анализу, для выдвижения предварительных гипотез о закономерностях, которым подчиняются данные, для выбора процедур обработки и анализа данных.

Наглядность визуализированных данных позволяет легко находить в них ошибочные значения, определять шаблоны и тенденции.

Одним из важных инструментов визуализации данных является условное форматирование, с помощью которого можно легко выделять необходимые ячейки или диапазоны, подчеркивать необычные значения и визуализировать данные с помощью гистограмм, цветовых шкал, наборов значков, а также любых сколь-угодно особенных или сложных правил, задаваемых пользователем.

Условное форматирование изменяет вид диапазона ячеек на основе определенных условий — *правил форматирования*. При истинности заданного усло-

вия диапазон ячеек форматируется по определенному шаблону. При этом таких правил может быть несколько.

С помощью условного форматирования можно выбрать несколько максимальных или минимальных значений, можно выделить значения, большие или меньшие определенного порога, можно отобрать уникальные, или наоборот, повторяющиеся значения и др.

Наиболее часто применяются три типа условного форматирования:

- *цветовые шкалы*, которые автоматически определяют максимальное и минимальное значение в диапазоне и закрашивают каждую ячейку цветом, соответствующим значению, изображая что-то вроде тепловой карты; например, максимальное значение может быть закрашено ярко зеленым цветом, минимальное — ярко красным, а цвет остальных ячеек тем ближе к зеленому, чем ближе значение данной ячейки к максимальному значению в диапазоне, и тем ближе к красному, чем ближе значение данной ячейки к минимуму в диапазоне;
- *наборы значков*, которые используются для классификации данных по трем — пяти категориям, разделенным пороговыми значениями так, что каждый значок соответствует определенному диапазону значений; например, в наборе значков «» зеленая стрелка, направленная вверх, может соответствовать высоким значениям, желтая стрелка, направленная в сторону, — средним значениям, а красная стрелка, направленная вниз, — низким значениям;
- *гистограммы*, которые позволяют в каждую числовую ячейку добавить прямоугольник, длина которого определяется отношением значения в данной ячейке к максимальномузначению в выделенном диапазоне.

Замечание. К сожалению, существует терминологическая путаница: гистограммы в условном форматировании и статистические гистограммы, которые рассматривались в предыдущем пункте, это принципиально разные инструменты визуализации данных. Более правильно было бы назвать гистограммы в условном форматировании как-нибудь по-другому, но использование одного и того же названия для разных инструментов уже вошло в практику. Примем это как данность.

Рассмотрим применение условного форматирования на примере.

Пример 1.1.4. Воспользуемся инструментами условного форматирования для визуализации данных о котировках акций из примера 1.1.3.

Выделим столбец B, в котором содержатся цены закрытия торгов по акциям ВТБ, на вкладке «Главная» нажмем кнопку «Условное форматирование», далее выберем в качестве типа условного форматирования «Гистограммы» и, например, «Градиентная заливка». Аналогичным образом поступим со столбцами C и D, в которых содержатся цены акций Газпрома и Сбербанка.

Замечание. Важно, что нельзя сделать это за один шаг, выделив сразу все три столбца — в таком случае длина прямоугольников будет рассчитываться по отношению к самой большой цене среди всех трех акций!

Теперь выделим столбец E, содержащий данные об объемах торгов акциями ВТБ, нажмем кнопку «Условное форматирование» (на вкладке «Главная»), далее выберем в качестве типа условного форматирования «Цветовые шкалы» и выберем шкалу от красного до зеленого. Аналогично поступим со столбцами F и G, в которых содержатся объемы торгов акциями Газпрома и Сбербанка.

Данные о доходностях визуализируем с помощью стрелок:

- зеленая стрелка вверх будет соответствовать доходностям, большим или равным 0,1;

- желтая стрелка по диагонали вверх — доходностям, большим или равным 0,05, но меньшим 0,1;
- желтая горизонтальная стрелка — доходностям, большим или равным −0,05, но меньшим 0,05;
- желтая стрелка по диагонали вниз — доходностям, большим или равным −0,1, но меньшим −0,05;
- красная стрелка вниз будет соответствовать доходностям, меньшим −0,1.

Поскольку в данном виде условного форматирования используются абсолютные пороги доходности (−0,1; −0,05; 0,05; 0,1), а не относительные значения из диапазона, можно выделить сразу все три столбца с доходностями (H, I, J).

Теперь нажмем на кнопку «Условное форматирование», далее выберем вариант «Создать правило». В появившемся окне (рис. 1.1.7) выберем в качестве типа правила «Форматировать все ячейки на основании их значений», в качестве стиля формата — «Наборы значков», в качестве стиля значков — «Направления». По умолчанию стрелки разделяют весь диапазон значений на пять равных по численности частей: 20% самых больших значений обозначаются зеленой стрелкой вверх, 20% самых маленьких значений — красной стрелкой вниз и т. д. Мы хотим использовать другие правила, введем их, как показано на рис. 1.1.7 и нажмем «OK».

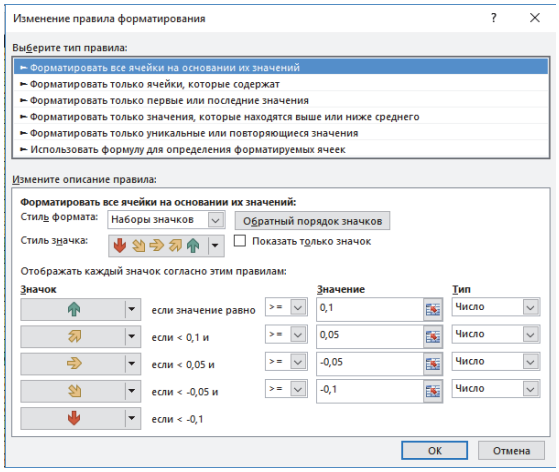


Рис. 1.1.7. *Правило условного форматирования*

Результат условного форматирования представлен на рис. 1.1.8.

Условное форматирование с помощью гистограмм позволяет увидеть тенденции в данных о ценах акций. Например, цена акций ВТБ показывала тенденцию к снижению от 0,05359 руб. за акцию в декабре 2012 г. до 0,03865 руб. в апреле 2014 г., далее в мае 2014 г. произошел подъем до 0,04790 руб., но цена не удержалась на этом уровне и продолжила снижаться до 0,03804 в сентябре 2014 г. В этом месяце был зафиксирован исторический минимум цены акций ВТБ за рассматриваемый период — прямоугольник в данной ячейке самый короткий. В октябре 2014 г. цена акций ВТБ незначительно увеличилась — до 0,03993 руб., в ноябре 2014 произошло увеличение до 0,04670 руб., в декабре 2014 г. цена акций ВТБ резко увеличилась до 0,06700 руб. и колебалась около этого значения до мая 2015 г., когда произошел очередной резкий скачок вверх — до 0,08025 руб. (в этом месяце зафиксирована самая большая цена за весь рассматриваемый период — прямоугольник в данной ячейке самый длинный). После этого началась тенденция снижения цены акций ВТБ, кото-

рая продолжалась до конца рассматриваемого периода. В августе 2017 г. акции ВТБ стоили 0,06199 руб. за шт. — больше, чем в декабре 2012 г., значительно больше, чем в точке исторического минимума, но существенно меньше, чем в точке исторического максимума.

	A	B	C	D	E	F	G	H	I	J
1	Дата	ВТБ - цена	Газпром - цена	Сбербанк - цена	ВТБ - объем	Газпром - объем	Сбербанк - объем	ВТБ - доходность	Газпром - доходность	Сбербанк - доходность
2	01.12.2012	0,05359	143,70	92,94	423 765 060 000	638 968 460	1 191 987 680			
3	01.01.2013	0,05582	142,09	109,59	620 659 410 000	495 283 520	1 723 351 580	0,0416	-0,0112	0,1791
4	01.02.2013	0,05588	137,40	104,57	956 736 250 000	733 770 920	1 612 212 000	0,0011	-0,0330	-0,0458
5	01.03.2013	0,04970	134,08	98,86	684 089 840 000	953 973 070	1 691 490 080	-0,1106	-0,0242	-0,0546
6	01.04.2013	0,04929	124,15	99,11	1 203 587 690 000	986 321 760	1 576 313 810	-0,0082	-0,0741	0,0025
7	01.05.2013	0,04600	123,40	99,05	909 928 850 000	813 775 000	1 514 417 130	-0,0667	-0,0060	-0,0006
8	01.06.2013	0,04701	109,10	93,68	773 944 750 000	769 293 390	1 603 984 540	0,0220	-0,1159	-0,0542
9	01.07.2013	0,04661	128,61	95,23	689 989 310 000	879 462 790	1 721 040 080	-0,0085	0,1788	0,0165
10	01.08.2013	0,04450	131,90	88,23	424 102 620 000	635 494 510	1 414 145 480	-0,0453	0,0256	-0,0735
11	01.09.2013	0,04267	144,15	97,86	770 174 060 000	1 035 314 530	1 926 407 650	-0,0411	0,0929	0,1091
12	01.10.2013	0,04435	150,40	102,74	1 095 103 330 000	1 120 379 070	1 935 002 670	0,0394	0,0434	0,0499
13	01.11.2013	0,04628	143,10	103,07	1 117 892 430 000	1 035 969 280	1 659 044 880	0,0435	-0,0485	0,0032
14	01.12.2013	0,04966	138,75	101,17	883 378 010 000	1 050 788 060	1 563 416 090	0,0730	-0,0304	-0,0184
15	01.01.2014	0,04544	145,16	94,70	562 228 330 000	1 056 510 690	1 427 259 190	-0,0850	0,0462	-0,0640
16	01.02.2014	0,04213	139,20	91,16	531 533 830 000	1 110 186 930	1 510 458 530	-0,0728	-0,0411	-0,0374
17	01.03.2014	0,03960	135,50	83,80	1 718 001 780 000	2 038 314 440	4 898 591 710	-0,0601	-0,0266	-0,0807
18	01.04.2014	0,03865	128,77	72,50	1 581 562 680 000	1 377 698 400	4 013 046 200	-0,0240	-0,0497	-0,1348
19	01.05.2014	0,04790	141,70	84,50	1 337 467 810 000	1 168 270 410	3 001 439 250	0,2393	0,1004	0,1655
20	01.06.2014	0,04110	148,96	84,50	1 219 984 610 000	885 913 470	2 008 494 660	-0,1420	0,0512	0,0000
21	01.07.2014	0,03980	132,00	73,60	1 064 257 460 000	1 004 959 980	2 551 370 010	-0,0316	-0,1139	-0,1290
22	01.08.2014	0,03840	131,95	73,21	699 479 420 000	851 294 800	3 076 887 590	-0,0352	-0,0004	-0,0053
23	01.09.2014	0,03804	137,90	75,52	534 436 210 000	797 896 850	2 891 411 920	-0,0094	0,0451	0,0316
24	01.10.2014	0,03993	141,50	76,23	552 675 470 000	857 933 800	2 905 609 940	0,0497	0,0261	0,0094
25	01.11.2014	0,04670	142,86	72,25	1 215 936 210 000	691 989 260	2 030 051 460	0,1695	0,0096	-0,0522
26	01.12.2014	0,06700	130,31	54,90	1 901 970 000 000	983 856 510	4 337 561 310	0,4347	-0,0878	-0,2401
27	01.01.2015	0,06894	143,82	61,50	472 099 150 000	652 102 830	2 691 982 770	0,0290	0,1037	0,1202
28	01.02.2015	0,06800	152,95	75,91	435 810 560 000	919 520 350	3 583 789 870	-0,0136	0,0635	0,2343
29	01.03.2015	0,06000	138,90	62,88	380 858 460 000	718 069 290	2 785 656 310	-0,1176	-0,0919	-0,1717
30	01.04.2015	0,06550	153,50	76,90	670 781 500 000	821 523 460	3 217 030 850	0,0917	0,1051	0,2230
31	01.05.2015	0,08025	139,00	73,50	1 045 976 670 000	472 091 480	1 830 904 250	0,2252	-0,0945	-0,0442
32	01.06.2015	0,07900	145,85	72,35	757 627 400 000	533 051 950	1 885 405 260	-0,0156	0,0493	-0,0156
33	01.07.2015	0,07200	142,50	72,30	355 294 170 000	543 868 920	2 690 621 070	-0,0886	-0,0230	-0,0007
34	01.08.2015	0,06900	148,19	74,50	297 785 390 000	654 231 910	2 488 111 940	-0,0417	0,0399	0,0304
35	01.09.2015	0,06770	134,55	75,30	288 953 870 000	646 257 900	2 008 042 110	-0,0188	-0,0920	0,0107
36	01.10.2015	0,07235	135,75	90,53	406 369 030 000	727 388 150	2 849 625 200	0,0687	0,0089	0,2023
37	01.11.2015	0,07100	138,00	102,90	293 559 920 000	785 173 850	2 286 927 960	-0,0187	0,0166	0,1366
38	01.12.2015	0,07970	136,09	101,26	349 150 430 000	632 658 380	1 880 909 280	0,1225	-0,0138	-0,0159
39	01.01.2016	0,07370	136,60	96,50	393 514 230 000	614 480 400	2 060 145 470	-0,0753	0,0037	-0,0470
40	01.02.2016	0,07340	141,40	107,00	284 381 410 000	642 613 120	2 184 006 710	-0,0041	0,0351	0,1088
41	01.03.2016	0,07668	147,75	109,90	303 739 850 000	731 620 780	1 959 737 430	0,0447	0,0449	0,0271
42	01.04.2016	0,07005	168,47	123,55	351 481 170 000	941 606 540	2 125 196 160	-0,0865	0,1402	0,1242
43	01.05.2016	0,06840	145,50	132,56	274 461 660 000	662 677 050	1 387 771 330	-0,0236	-0,1363	0,0729
44	01.06.2016	0,06800	139,51	133,00	282 574 220 000	538 549 300	1 550 840 130	-0,0058	-0,0412	0,0033
45	01.07.2016	0,06741	137,30	139,15	152 728 630 000	505 531 930	1 224 653 180	-0,0087	-0,0158	0,0462
46	01.08.2016	0,06849	134,95	143,50	225 458 160 000	472 860 350	1 150 874 110	0,0160	-0,0171	0,0313
47	01.09.2016	0,07210	134,90	145,34	360 228 510 000	567 536 440	1 118 608 200	0,0527	-0,0004	0,0128
48	01.10.2016	0,06775	138,84	147,40	159 481 010 000	409 341 480	777 345 030	-0,0603	0,0292	0,0142
49	01.11.2016	0,06940	148,80	158,70	234 598 360 000	735 320 710	1 113 951 960	0,0244	0,0717	0,0767
50	01.12.2016	0,07400	154,55	173,25	553 056 090 000	680 484 920	1 204 467 020	0,0663	0,0386	0,0917
51	01.01.2017	0,06897	149,80	172,20	300 093 660 000	508 472 070	989 614 480	-0,0680	-0,0307	-0,0061
52	01.02.2017	0,06608	134,00	156,00	136 916 720 000	618 092 220	817 013 500	-0,0419	-0,1055	-0,0941
53	01.03.2017	0,06625	127,90	159,80	266 410 390 000	754 410 290	980 688 220	0,0026	-0,0455	0,0244
54	01.04.2017	0,06671	136,75	165,20	207 910 220 000	750 182 400	965 518 550	0,0069	0,0692	0,0338
55	01.05.2017	0,06615	120,28	155,93	168 505 230 000	683 155 600	825 457 660	-0,0084	-0,1204	-0,0561
56	01.06.2017	0,06400	118,49	145,59	252 186 300 000	682 977 210	1 249 106 940	-0,0325	-0,0149	-0,0663
57	01.07.2017	0,05969	116,10	164,53	179 099 730 000	572 372 360	1 056 892 340	-0,0673	-0,0202	0,1301
58	01.08.2017	0,06199	118,68	173,40	69 594 480 000	146 516 160	329 303 440	0,0385	0,0222	0,0539

Рис. 1.1.8. *Условное форматирование данных о котировках акций*

Цветовые шкалы, с помощью которых мы отформатировали данные об объемах торгов, позволяют не только увидеть наибольшие и наименьшие значения, но и понять, как часто в данных происходят пиковые скачки. Так, по акциям ВТБ наибольший объем торгов (1 901 970 000 000 шт.) был зафиксирован в декабре 2014 г., наименьший объем (69 594 480 000 шт.) — в августе 2017 г., при этом до июня 2015 г. объемы торгов были существенно выше (цвет соответствующих ячеек скорее зеленый, чем красный, то есть объемы торгов в этот период были ближе к максимуму, чем к минимуму), чем с июля 2015 г. (с июля 2015 г. цвет ячеек стал ближе к красному, чем к зеленому).

Форматирование с помощью стрелок позволяет увидеть резко выделяющиеся значения доходностей. Так по акциям ВТБ сразу видно, что самые большие по модулю отрицательные значения доходности были зафиксированы в марте 2013 г. (−0,1106), в июне 2014 г. (−0,1420) и в марте 2015 г. (−

0,1176). Самые большие положительные значения доходности в мае 2014 г. (0,2393), в ноябре 2014 г. (0,1695), в декабре 2014 г. (0,4347), в мае 2015 г. (0,2252), в декабре 2015 г. (0,1225). В остальные месяцы доходность колебалась от −0,1 до 0,1.

Результаты сохранены на рабочем листе «Данные» в файле «Акции.xlsx». ❑

Каждый тип условного форматирования может быть полезен, какой из них использовать в каждом конкретном случае — зависит от исследователя.

Задачи

6. Проведите условное форматирование данных о потребительских кредитах из задачи 3, выбрав подходящий тип форматирования.

7. Проведите условное форматирование данных о росте студентов из задачи 4, выбрав подходящий тип форматирования.

8. Проведите условное форматирование данных об акциях из задачи 5 (не обязательно использовать те же типы форматирования, что в примере 1.1.4).

1.1.4. Графики и диаграммы рассеяния

Классический способ визуализации данных — построение *графиков*, изображающих зависимости между признаками в виде линий на плоскости. В анализе данных чаще всего строят графики зависимости признаков от времени (естественно, это возможно, только когда в наборе данных присутствует признак, связанный со шкалой времени — дата, порядковый номер даты и т. п.).

В общем же случае, когда анализируется возможная взаимосвязь двух признаков, как правило, функциональная зависимость между этими признаками отсутствует, но могут наличествовать (или нет) какие-либо более сложные статистические связи. Поэтому в общем случае для того, чтобы выдвинуть гипотезу о наличии зависимости между признаками, строят *диаграммы рассеяния*, изображающие значения двух признаков в виде точек на плоскости. Каждой строке набора данных на диаграмме рассеяния соответствует точка, координата x которой соответствует значению одного из признаков в данной строке, а координата y — значению второго признака.

Пример 1.1.5. Построим графики зависимости цен закрытия и объемов торгов от времени по данным о котировках акций ВТБ, Газпрома и Сбербанка из примера 1.1.2.

Заметим, что удобно построить три линии на одном рисунке, однако поскольку цены акций ВТБ имеют порядок нескольких копеек, а цены акций ВТБ и Газпрома — порядок сотни рублей, то если построить на одном рисунки графики зависимости от времени вспех трех акций, то график цены акций ВТБ сольется с осью абсцисс. Чтобы этого не произошло, добавим к таблице исходных данных еще один столбец, назовем его «1000*(ВТБ – цена)» и определим значения с помощью формулы «=[@[ВТБ – цена]]*1000». Очевидно, признак «1000*(ВТБ – цена)» увеличивается и уменьшается в те же даты и в то же количество раз, что и признак «ВТБ – цена».

Теперь, чтобы построить на одном рисунке три графика, выделим (держа нажатой клавишу «Ctrl») столбцы «Дата», «Газпром – цена», «Сбербанк – цена» и «1000*(ВТБ – цена)», затем на вкладке «Вставить» нажмем кнопку «Вставить график или диаграмму с областями» и выберем тип диаграммы — «Линия». В результате получим график, как на рис. 1.1.9.

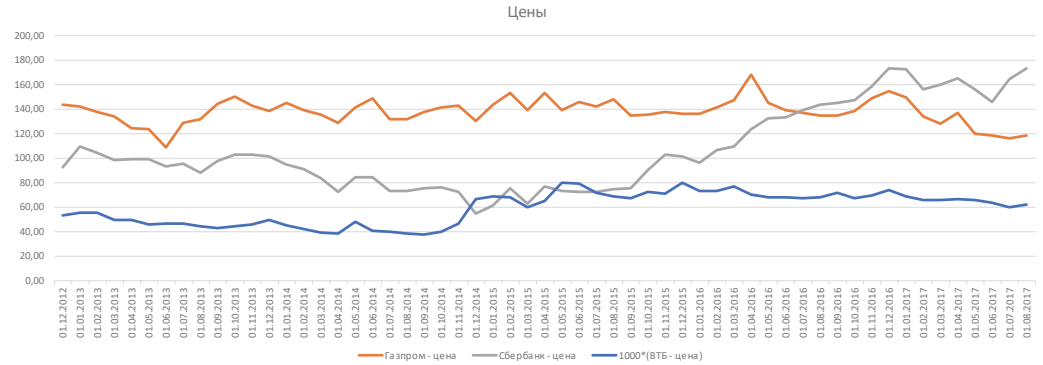


Рис. 1.1.9. *Графики зависимости цен акций от времени*

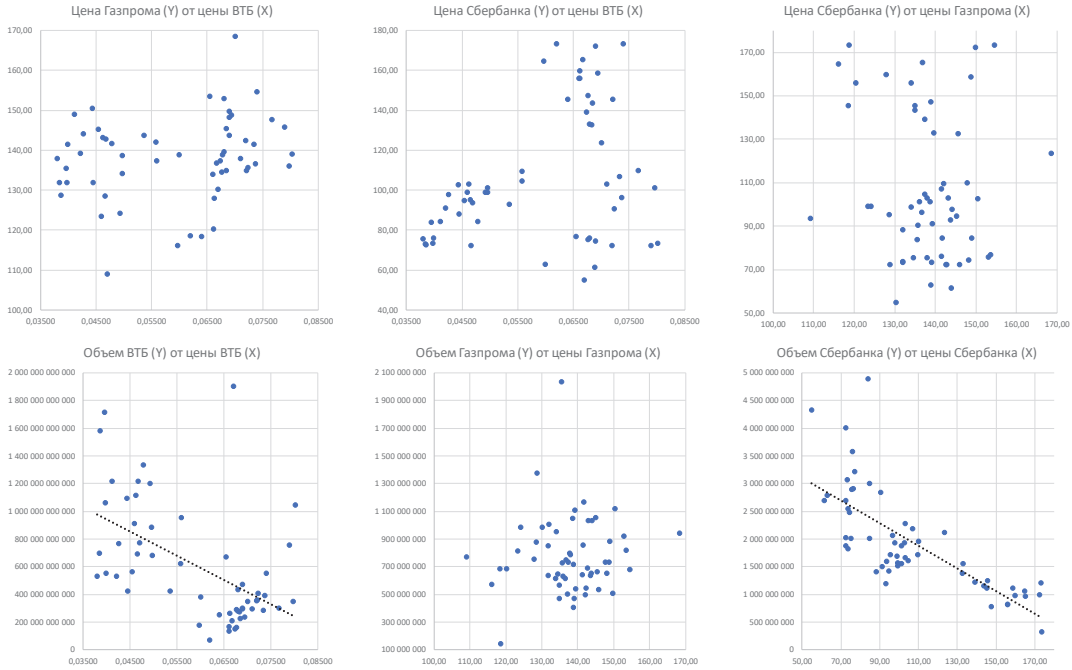


Рис. 1.1.10. *Диаграммы рассеяния*

Построим диаграмму рассеяния для анализа возможной взаимосвязи между ценами акций ВТБ и Газпрома. Выделим (держа нажатой клавишу «Ctrl») столбцы «ВТБ – цена» и «Газпром – цена», затем на вкладке «Вставить» нажмем кнопку «Вставить точечную (X, Y) или пузырьковую диаграмму» и выберем тип жиаграммы — «Точечная». Результат представлен на рис. 1.1.10.

Аналогичным образом построим диаграммы рассеяния для анализа возможнох зависимостей цен акций ВТБ и Сбербанка; цен акций Газпрома и Сбербанка; объемов торгов по акциям от цен (рис. 1.1.10; результаты также сохранены в файле «Акции.xlsx» на рабочих листах «Графики» и «Диаграммы рассеяния»).

Анализ графика цен акций Газпрома не позволяет выдвинуть никаких предположений о том, как цены акций Газпрома зависят от времени. Но для цен акций ВТБ и Сбербанка можно выделить два периода: до 1 декабря 2014 г. и после

1 декабря 2014 г. Акции ВТБ в первом периоде колебались около цены 0,05 руб. за шт., а в декабре 2014 г. произошел скачок и во втором периоде акции ВТБ стали стоить около 0,07 руб. за шт. Что касается акций Газпрома, то в первом периоде их цена демонстрировала тенденцию с снижению, а во втором — к увеличению.

Анализ диаграмм рассеяния дает возможность предположить, что объемы торгов акциями ВТБ и Сбербанка зависят от цен соответствующих акций: чем больше цена, тем меньше объем (на рис. 1.1.10. на соответствующие диаграммы рассеяния добавлены пунктирные линии, указывающие направления зависимостей[1]). Что касается взаимосвязи цен акций Сбербанка и ВТБ, то по диаграмме рассеяния видно, что при небольшой цене акций ВТБ (до 0,06 руб. за шт.) этих двух акций рост цены одной акции сопровождается ростом цены другой акции. Если же акции ВТБ стоят дороже, чем 0,06 руб. за шт., то акции Сбербанка могут как расти, так и уменьшаться в цене при росте цены акций ВТБ.

Между остальными парами признаков зависимости «не просматриваются». ❑

Задачи

9. По данным о выданных кредитах из задачи 3 постройте график зависомости количества выбранных кредитов от дня. Есть ли основания считать, что наблюдается тенденция к увеличению или уменьшению количества кредитов?

10. По данным об акциях из задачи 5 постройте графики зависимости цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов от времени. Прослеживаются ли какие-нибудь зависимости данных признаков от времени?

11. По данным об акциях из задачи 5 постройте диаграммы рассеяния для каждой пары признаков. Между какими признаками можно предположить наличие зависимости?

§ 1.2. Инструменты описательной статистики в Microsoft Excel

1.2.1. Измерение центра распределения

Описательная статистика дает ответы на вопросы о том, что общего есть в анализируемых данных и какие в них есть различия.

Предположим, что нам известны данные о некотором признаке X на n объектах:

$$x_1, x_2, \ldots, x_n,$$

и мы хотим определить наиболее типичное, наиболее ожидаемое, значение этого признака. Для ответа на этот вопрос часто используются три характеристики:

- *среднее*

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n};$$

- *мода* x_{mod}, т. е. наиболее часто встречающееся значение; мода может быть не единственна (если несколько значений встречаются одинаковое количество раз, а другие значения встречаются меньшее число раз) и может не существовать (если все элементы встречаются одинаковое количество раз);

[1] Добавить линию тренда на диаграмму рассеяния можно, выделив точки рассеяния, вызвав правой кнопкой всплывающее меню и выбрав пункт «Добавить линию тренда».

- *медиана* x_{med}, т. е. серединный элемент данных; если в наборе данных о данном признаке X нечетное число элементов, то в качестве медианы выбирается такой элемент, что половина всех оставшихся элементов не больше него, а другая половина — не меньше, если же мы имеем дело с четным числом элементов, то в качестве медианы выбирается среднее из таких двух элементов, что половина оставшихся чисел не больше них, а другая половина — не меньше.

Пример 1.2.1. В течение 14 дней записывалось, сколько кредитов на покупку автомобилей было выдано в некотором банке в каждый из этих дней: 5; 6; 6; 7; 11; 7; 5; 2; 8; 7; 4; 8; 9; 10. Упорядочим эти числа в порядке неубывания: 2; 4; 5; 5; 6; 6; 7; 7; 7; 8; 8; 9;10; 11. Вычислим среднее:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \\ = \frac{2+4+5+5+6+6+7+7+7+8+8+9+10+11}{14} = \frac{95}{14} \approx 6{,}785714.$$

Мода — это наиболее часто встречающийся элемент набора данных. В нашем примере три раза встречается число 7, остальные элементы встречаются реже. Поэтому выборочная мода $x_{mod} = 7$.

Медиана в данном примере также равна $x_{med} = 7$, поскольку слева от двух выделенных жирным семерок находится шесть чисел, и справа от этих двух семерок также находится шесть чисел:

2; 4; 5; 5; 6; 6; **7; 7;** 7; 8; 8; 9; 10; 11.

Отметив на гистограмме числа выданных кредитов среднее, моду и медиану (рис. 1.2.1),[1] замечаем, что распределение симметрично, мода и медиана совпадают, среднее незначительно отличается. Все эти величины можно использовать в качестве характеристики центра, вокруг которого распределены значения признака: с среднем каждый день выдавалось около 7 кредитов, в самый удачный день было выдано 7 кредитов, в половине дней было выдано не меньше семи кредитов, в половине дней было выдано не больше 7 кредитов. ❑

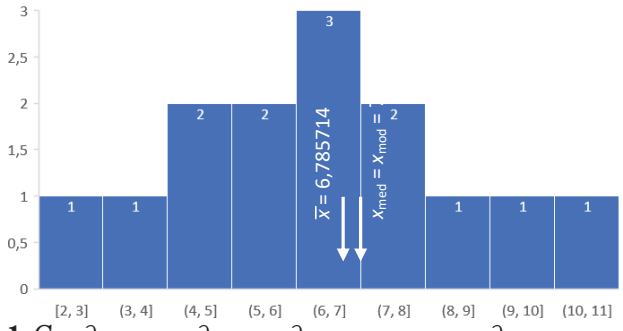


Рис. 1.2.1. *Среднее, мода и медиана числа выданных кредитов*

[1] Чтобы на диаграмме нарисовать стрелку, нужно на вкладке **«Вставка»** нажать кнопку **«Иллюстрации»**, далее выбрать **«Фигуры»** в качестве типа иллюстраций и вставить стрелку.

Замечание. Для качественных признаков вычисление среднего и медианы невозможно, поэтому среднюю тенденцию качественных признаков, как правило, характеризуют модой — наиболее часто встречающимся значением.

В Microsoft Excel для вычисления среднего и медианы можно использовать функции СРЗНАЧ и МЕДИАНА. А для расчета моды есть целых две функции. МОДА.ОДН возвращает моду, если она единственна, и первое из модальных значений, если их несколько. МОДА.НСК возвращает в виде вертикального массива все моды, если их несколько (эту функцию нужно использовать как формулу массива, т. е. сначала выделить массив подходящего размера, а после ввода формулы нажать не просто <Enter>, а <Ctrl> + <Shift> + <Enter>).

Для признаков, которые распределены симметрично относительно вертикальной оси, среднее, медиана и мода близки друг к другу, приблизительно совпадают. В случае же, когда распределение случайной величины не является симметричным, мода и медиана являются достаточно важными характеристиками. Проиллюстрируем это двумя примерами.

Пример 1.2.2. Пусть X — годовой доход (в ден. ед.) работника одной крупной отрасли, а Y — годовой доход (в ден. ед.) работника другой крупной отрасли, причем средние этих признаков совпадают и равны $\bar{x}=\bar{y}=150\,000$ ден. ед., а медианы различаются: $x_{\text{med}} = 20\,000$ ден. ед.; $y_{\text{med}} = 80\,000$ ден. ед. Несмотря на то, что средний доход в обеих отраслях равен, работники второй отрасли «в среднем» богаче, чем работники первой отрасли (поскольку половина работников второй отрасли имеет доход, больший 80 000 ден. ед., а более богатая половина работников первой отрасли имеет доход, больший 20 000 ден. ед.). ❑

Пример 1.2.3. Предположим, что производитель одежды собирается выпустить новую женскую юбку, для чего проводит опрос женщин относительно идеальной, с их точки зрения, длины юбки. Пусть X — длина, названная случайно отобранной женщиной. Если выпустить юбку длины $\bar{x}$, то, скорее всего, ни одна из женщин не будет полностью удовлетворена, поскольку почти наверняка $\bar{x}$ не совпадет ни с одним из названных женщинами значений. Если выпустить юбку длины x_{med}, то окажется, что ровно половина женщин предпочитает юбки более короткие, чем эта, а другая половина — более длинные. А вот юбка длиной x_{mod} удовлетворит максимальное количество женщин. ❑

Всегда нужно задумываться о содержательном смысле тех характеристик, которые мы вычисляем. Следующие два примера иллюстрируют ситуации, когда среднее, мода и медиана не несут практической пользы и нужно использовать другие показатели.

Пример 1.2.4. Предположим, что мы выбираем, где разместить наши сбережения. Можно открыть вклад в некотором банке, который в течение последних четырех лет стабильно начислял в конце года 5% от суммы, которая была в начале года. Альтернативным вариантом является размещение средств в инвестиционном фонде, который год назад и три года назад продемонстрировал 70%-ную доходность, а два года назад и четыре года назад показал 50%-ные убытки: табл. 1.2.1. Если принимать решение на основании средней доходности, то можно решить, что инвестиционный фонд в среднем позволяет заработать больше. Но это не так — см. табл. 1.2.2, из которой видно, что если бы мы четыре года назад вложили в банк 100 руб., то сегодня у нас было бы 121,55 руб., а если бы мы отнесли четыре года назад 100 руб. в инвестиционный фонд, то сегодня у нас было бы 72,25 руб. — меньше, чем мы вложили!

Таблица 1.2.1

Доходность активов

Банковский вклад	*Инвестиционный фонд*
0,05	−0,5
0,05	0,7
0,05	−0,5
0,05	0,7
$\overline{x}=\frac{0,05+0,05+0,05+0,05}{4}=0,05$	$\overline{y}=\frac{-0,5+0,7-0,5+0,7}{4}=0,10$
$x_{\text{med}}=\frac{0,05+0,05}{2}=0,05$	$y_{\text{med}}=\frac{-0,5+0,7}{2}=0,1$
x_{mod} не существует	y_{mod} не существует

Таблица 1.2.2

Динамика стоимости активов

Банковский вклад	*Инвестиционный фонд*
100,00 + 0,05·100,00 = 105,00	100,00 − 0,5 · 100,00 = 50,00
105,00 + 0,05·105,00 = 110,25	50,00 + 0,7 · 50,00 = 85,00
110,25 + 0,05·110,25 = 115,76	85,00 − 0,5 · 85,00 = 42,50
115,76 + 0,05·115,76 = 121,55	42,50 + 0,7 · 42,50 = 72,25

Все дело в том, что сумма на счете увеличивается ежегодно не на 0,05 руб., а в 1,05 раза. Точно также активы в инвестиционном фонде в двух годах увеличиваются в 1 + 0,7 = 1,7 раза, а в двух годах увеличиваются в 1 − 0,5 = 0,5 раза. Поэтому для того, чтобы правильно принять решение об инвестициях, лучше всего перейти от доходностей — темпов прироста активов — к темпам роста (табл. 1.2.3) и вычислить среднее геометрическое темпа роста для каждого из вариантов ($\textit{среднее геометрическое } n$ чисел $x_1, x_2, \ldots, x_n$ — это корень n-й степени из произведения данных чисел: $\left(x_1 x_2 \cdots x_n\right)^{1/n}$).

Таблица 1.2.3

Темпы роста активов

Банковский вклад	*Инвестиционный фонд*
1 + 0,05 = 1,05	1 − 0,5 = 0,5
1 + 0,05 = 1,05	1 + 0,7 = 1,7
1 + 0,05 = 1,05	1 − 0,5 = 0,5
1 + 0,05 = 1,05	1 + 0,7 = 1,7
Среднее геометрическое темпа роста $(1,05\cdot1,05\cdot1,05\cdot1,05)^{1/4}=1,05$	Среднее геометрическое темпа роста $(0,5\cdot1,7\cdot0,5\cdot1,7)^{1/4}=(0,5\cdot1,7)^{1/2}=\sqrt{0,85}=0,9219$

Среднее геометрическое показывает, что сумма, размещенная на вкладе, каждый год в среднем растет в 1,05 раза, а сумма, размещенная в инвестиционном фонде, каждый год растет в среднем в 0,92195 раза. Теперь легко видеть, что 100 руб. в банке через четыре года превращаются в $100 \cdot 1,05^4 = 121,55$ руб., а в инвестиционном фонде — в $100 \cdot 0,92195^4 = 72,25$ руб. ❑

Пример 1.2.5. На горячей линии банка работают два сотрудника — Петя и Маша. Петя на разговор с клиентом тратит минуту, а Маша — шесть минут. Вычислим, сколько в среднем длится обслуживание клиента на горячей линии.

Казалось бы, все просто: (1 + 6)/2 = 3,5 мин. Но это неправильный ответ! если бы в среднем один клиент обслуживался 3,5 минуты, то в час два сотрудника обслужили бы всего $2 \cdot 60/3{,}5 \approx 34$ клиентов. На самом же деле, только Петя обслуживает 60 клиентов в час, а еще 10 клиентов в час обслуживает Маша.

Усреднять нужно не время обслуживания клиента, а скорость обслуживания. Получается, что вместе Петя и Маша обслуживают 70 клиентов в час, а в среднем каждый из них обслуживает в час 35 клиентов. Отсюда заключаем, что в среднем каждый клиент обслуживается за 60/35 = 1,71 мин.

Таким образом, мы вычислили *среднее гармоническое* — величину

$$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}},$$

обратную к среднему от чисел, обратных к исходным. ❑

В Microsoft Excel среднее геометрическое и среднее гармоническое можно вычислить с помощью функций СРГЕОМ и СРГАРМ. Но на самом деле эти функции используются гораздо реже, чем СРЗНАЧ.

Задачи

12. Вычислите среднее, моду и медиану по данным о температуре в Москве в первые 15 дней сентября 2016 г.: 14; 16; 17; 14; 13; 12; 15; 15; 12; 13; 14; 17; 11; 10; 10. Постройте гистограмму температуры и нанесите на нее среднее, моду и медиану.

13. Вычислите среднее, моду и медиану по данным о потребительских кредитах из задачи 3.

14. Вычислите среднее, моду и медиану по данным о росте студентов, полученным в задаче 4. Нанесите среднее, моду и медиану на гистограмму роста студентов.

1.2.2. Измерение разброса данных

Чтобы делать выводы на основании данных, среднюю тенденцию знать недостаточно, нужно еще понимать, насколько далеко от средней тенденции может отклониться исследуемый признак.

Наиболее часто используется две меры отклонения от средней тенденции:

- *размах* (или *интервал*) $R_X = x_{(\max)} - x_{(\min)}$ — разность между наибольшим и наименьшим элементами;
- *стандартное отклонение*

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

Стандартное отклонение является очень важной характеристикой разброса в том случае, когда признак имеет нормальный закон распределения:

- 68,26% всех значений такого признака отклоняются от среднего не больше чем на одно стандартное отклонение;

- 95,45% всех значений такого признака отклоняются от среднего не больше чем на два стандартных отклонения;
- 99,73% всех значений такого признака отклоняются от среднего не больше чем на три стандартных отклонения.

Замечание. В случае, если распределение признака отличается от нормального, такие заключения могут быть о ш и б о ч н ы м и. Соответствующие примеры мы позже обязательно рассмотрим. Также впоследствии мы разберемся в смысле формулы для стандартного отклонения.

Пример 1.2.6. Пусть в результате опыта некоторый признак X принял три значения: -1; 0; 1, а другой признак Y принял значения -2; 0; 2.

Средние у этих признаков совпадают: $\overline{x} = (-1+0+1)/3 = 0$, $\overline{y} = (-2+0+2)/3 = 0$.

Мода не существует ни у X (так как каждое из значений -1; 0; 1 встречается по одному разу), ни у Y (каждое из значений -2; 0; 2 встречается по одному разу).

Медианы у X и Y совпадают: $x_{\text{med}} = y_{\text{med}} = 0$.

Поскольку наименьший элемент $x_{(\min)} = -1$, а наибольший элемент $x_{(\max)} = 1$, размах $R_X = x_{(\max)} - x_{(\min)} = 1 - (-1) = 2$, а так как $y_{(\min)} = -2$, а $y_{(\max)} = 2$, размах $R_Y = y_{(\max)} - y_{(\min)} = 2 - (-2) = 4$.

Вычислим стандартные отклонения признаков X и Y:

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{(-1-0)^2 + (0-0)^2 + (1-0)^2}{3-1}} =$$

$$= \sqrt{\frac{(-1)^2 + 0^2 + 1^2}{2}} = \sqrt{\frac{1+0+1}{2}} = \sqrt{\frac{2}{2}} = \sqrt{1} = 1;$$

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}} = \sqrt{\frac{(-2-0)^2 + (0-0)^2 + (2-0)^2}{3-1}} =$$

$$= \sqrt{\frac{(-2)^2 + 0^2 + 2^2}{2}} = \sqrt{\frac{4+0+4}{2}} = \sqrt{\frac{8}{2}} = \sqrt{4} = 2.$$

Итак, средние и медианы у признаков X и Y совпадают, моды у обоих признаков не существуют, а размах и стандартное отклонение у признака X меньше, чем у призника Y. ❑

Пример 1.2.7. В условиях примера 1.2.1 вычислим размах количества кредитов, выдаваемых в день: $R = x_{(\max)} - x_{(\min)} = 11 - 2 = 9$ (в самый удачный день было выдано одиннадцать кредитов, в самый неудачный день — два кредита).

«Ручное» вычисление стандартного отклонения по формуле

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \left(\frac{(2-6{,}78...)^2 + (4-6{,}78...)^2 + 2\cdot(5-6{,}78...)^2 + 2\cdot(6-6{,}78...)^2}{14-1} + \right.$$

$$\left. + \frac{3\cdot(7-6{,}78...)^2 + 2\cdot(8-6{,}78...)^2 + (9-6{,}78...)^2 + (10-6{,}78...)^2 + (11-6{,}78...)^2}{14-1} \right)^{1/2} =$$

$$= 2{,}3916062,$$

конечно, возможно, но лучше воспользоваться в Microsoft Excel функцией СТАНД.ОТКЛОН.В.

Таким образом, количество выдаваемых в день кредитов варьируется от двух до десяти, размах равен 8, и если предположить, что количество выдаваемых в день кредитов распределено по нормальному закону, то с надежностью 68% можно гарантировать, что количество кредитов будет в границах $6{,}786 \pm 2{,}392$ (или от 4,39 до 9,18), с надежностью 95,45% — в границах $6{,}786 \pm 2 \cdot 2{,}392$ (или от 2,00 до 11,57), с надежностью 99,73% — в границах $6{,}786 \pm 3 \cdot 2{,}392$ (или от −0,39 до 13,96).

Это означает, что если банк работает 365 дней в году, то 348 дней в году выдается от 2 до 12 кредитов ($365 \cdot 0{,}9545 = 348{,}393$), а дни, в которые выдается 14 кредитов или больше, наступают реже, чем раз в два года [$365 \cdot (1 - 0{,}9973)/2 = 0{,}492$][1]. ❑

Задачи

15. Вычислите размах и стандартное отклонение по данным о температуре в Москве из задачи 12.

16. Вычислите размах и стандартное отклонение по данным о потребительских кредитах из задачи 3.

17. Вычислите размах и стандартное отклонение по данным о росте студентов, полученным в задаче 4.

1.2.3. Описательная статистика в надстройке «Анализ данных»

В пакете Microsoft Excel есть надстройка **«Пакет анализа»**, которая позволяет автоматизировать многие из задач статистического анализа данных. Если в правой части вкладки **«Данные»** есть кнопка **«Анализ данных»**, то это означает, что данная надстройка в Microsoft Excel установлена. Если же кнопка **«Анализ данных»** отсутствует, то нужно выбрать в окне управления параметрами Excel (меню **«Файл»** — **«Параметры»**) пункт **«Надстройки»** (рис. 1.2.2, *а*), далее в нижней строке **«Управление»** выбрать **«Надстройки Excel»**, нажать кнопку **«Перейти...»**, в появившемся окне (рис. 1.2.2, *б*) отметить флажок **«Пакет анализа»** и нажать **«ОК»** — после этого кнопка **«Анализ данных»** должна появиться на вкладке **«Данные»**.

Первым инструментом надстройки **«Пакет анализа»**, с которым мы познакомимся, будет программа **«Описательная статистика»**. Воспользуемся ею для вычисления основных статистических характеристик уже известных нам наборов данных.

Пример 1.2.8. Вычислим статистические характеристики данных о количестве выданных кредитов из примеров 1.2.1, 1.2.7. Введем данные из примера 1.2.1 на рабочий лист Microsoft Excel в виде столбца (рис. 1.2.3, *а*) и запустим программу **«Описательная статистика»**.

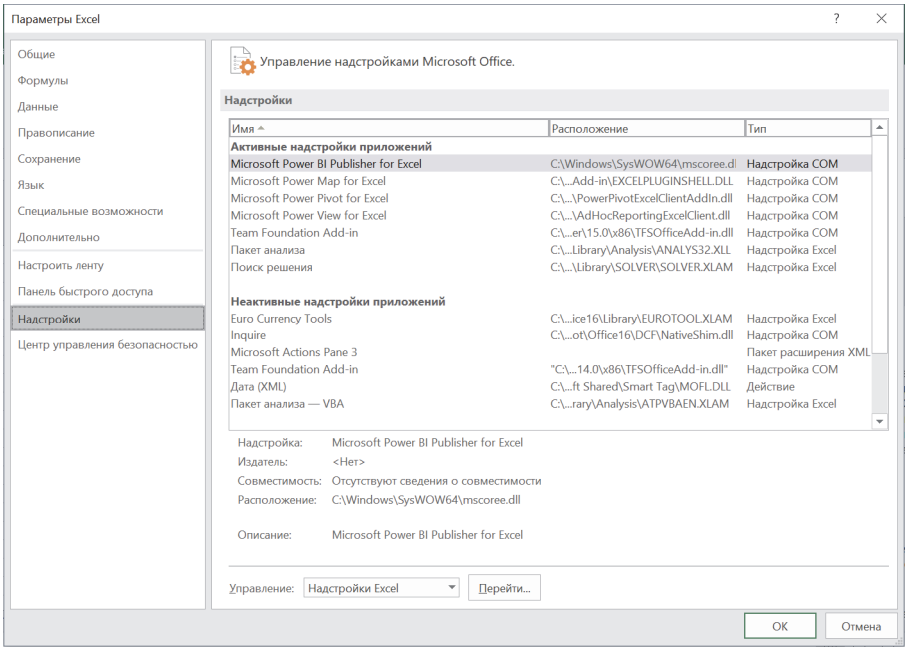
Для этого нажмем на кнопку **«Анализ данных»** (напомним, что она расположена в правой части вкладки **«Данные»**), в меню выбора инструментов анализа данных выберем соответствующий инструмент **«Описательная статистика»** (рис. 1.2.3, *б*) и нажмем **«ОК»**.

[1] Вероятность выхода за границы интервала $(-0{,}39;\ 13{,}96)$ равна $1 - 0{,}9973$, при этом с вероятностью $(1 - 0{,}9973)/2$ происходит выход за левую границу, и с вероятностью $(1 - 0{,}9973)/2$ — за правую границу данного интервала.

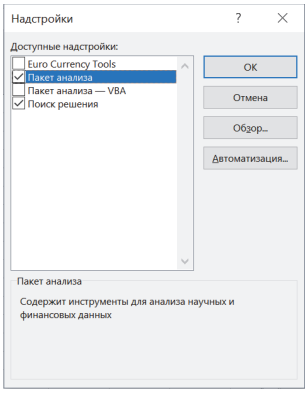
В окне ввода параметров программы **«Описательная статистика»** (рис. 1.2.3, *в*) укажем **входной интервал** (ссылку на ячейки **A1:A15**, содержащие данные о количестве выданных кредитов).

Отметим, что данные **сгруппированы по столбцам**, а поскольку первая строка входного интервала содержит заголовок, отметим флажок **«Метки в первой строке»**). Установим флажок для генерации **итоговой статистики** — набора основных выборочных характеристик.

Укажем, что результаты работы программы необходимо вывести **на новый рабочий лист**.



а)

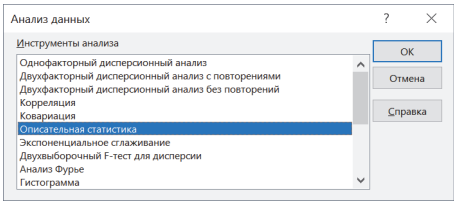


б)

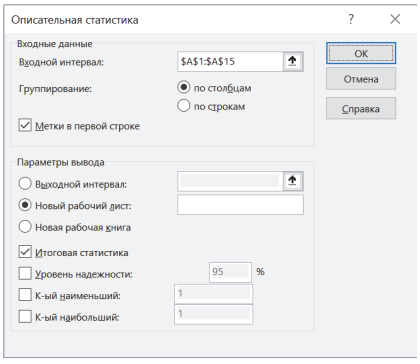
Рис. 1.2.2. *Управление надстройками Microsoft Excel: окно управления параметрами Excel* (*а*), *выбор надстроек* (*б*)

	A
1	Количество выданных кредитов
2	5
3	6
4	6
5	7
6	11
7	7
8	5
9	2
10	8
11	7
12	4
13	8
14	9
15	10

а)



б)



в)

	A	B
1	*Количество выданных кредитов*	
2		
3	Среднее	6,785714
4	Стандартная ошибка	0,639184
5	Медиана	7
6	Мода	7
7	Стандартное отклонение	2,391606
8	Дисперсия выборки	5,71978
9	Эксцесс	0,110799
10	Асимметричность	–0,13817
11	Интервал	9
12	Минимум	2
13	Максимум	11
14	Сумма	95
15	Счет	14

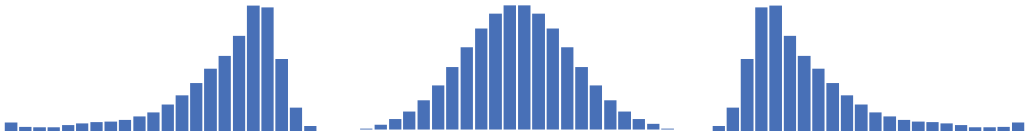
г)

Рис. 1.2.3. *Вычисление статистических характеристик данных о выданных кредитах: исходные данные (а); выбор инструмента анализа данных (б); ввод параметров (в); результаты расчетов (г)*

Результаты вычислений, представленные на рис. 1.2.3, *г*, совпали с итогами расчетов, проведенных в примерах 1.2.1, 1.2.7: **среднее** оказалось равным 6,785714; **медиана** и **мода** равны 7; **стандартное отклонение** равно 2,391606. В поле **«Интервал»** указан размах данных (9), поля **«Минимум»** и **«Максимум»** содержат наименьший и наибольший элементы (соответственно, 2 и 11), поле **«Сумма»** — сумму всех элементов (95), а поле **«Счет»** — количество элементов (14).

Обратим внимание еще на один показатель, рассчитанный программой. В поле **«Асимметричность»** находится *коэффициент асимметрии*, который показывает, насколько распределение набора данных является скошенным:

- у распределений, скошенных влево, коэффициент асимметрии отрицателен (пример такого распределения представлен на рис. 1.2.4, *а*);
- у идеально симметричных распределений коэффициент асимметрии равен нулю (рис. 1.2.4, *б*);
- у распределений, скошенных вправо, коэффициент асимметрии положителен (рис. 1.2.4, *в*).



а) *б*) *в*)

Рис. 1.2.4. *Гистограммы признаков с различными коэффициентами асимметрии: отрицательным* (*а*)*; нулевым* (*б*)*; положительным* (*в*)

Замечание. Обычно вывод о существенности скоса делают в том случае, когда коэффициент асимметрии по модулю больше единицы, в противном случае распределение считают примерно симметричным. На самом деле, для того, чтобы обосновать вывод о наличии или отсутствии скоса, требуется более сложный анализ, включающий в себя проверку соответствующей статистической гипотезы, но на практике такой анализ проводят очень редко.

В нашем примере коэффициент асимметрии равен −0,13817, что говорит в пользу предположения о симметричности распределения количества выданных кредитов, которое можно сделать на основе визуального анализа гистограммы числа выданных кредитов, построенной на рис. 1.2.1. ❑

Пример 1.2.9. Вычислим с помощью инструмента **«Описательная статистика»** характеристики данных о росте студентов из примера 1.1.2 — результаты представлены на рис. 1.2.5 (они содержатся на рабочем листе **«Описательная статистика»** в файле **«Рост студентов.xlsx»**),

	A	B
1	*Ваш рост*	
2		
3	Среднее	177,6833
4	Стандартная ошибка	0,898429
5	Медиана	177,5
6	Мода	179
7	Стандартное отклонение	4,920897
8	Дисперсия выборки	24,21523
9	Эксцесс	0,059917
10	Асимметричность	–0,08915
11	Интервал	21
12	Минимум	167
13	Максимум	188
14	Сумма	5330,5
15	Счет	30

Рис. 1.2.5. *Статистические характеристики данных о росте студентов*

Из этих результатов видно, что рост тех 30 студентов, которые приняли участие в опросе, оказался в границах от 167 до 188 см (с размахом, равным 21 см); средний рост студентов, принявших участие в опросе, составляет 177,6833 см, и поскольку стандартное отклонение роста составило 4,920897 см, можно утверждать, что если в группе есть студенты, которые не принимали участие в опросе, то с надежностью 99,73% их рост больше $177{,}68 - 3 \cdot 4{,}92 = 162{,}9$ см, но меньше $177{,}68 + 3 \cdot 4{,}92 = 192{,}4$ см (и с надежностью 95,45% рост студентов будет находиться в границах от $177{,}68 - 2 \cdot 4{,}92 = 167{,}8$ см до $177{,}68 + 2 \cdot 4{,}92 = 187{,}5$ см).

Коэффициент асимметрии, равный −0,08915, подтверждает предположение о симметричности распределения роста студентов, сделанное в примере 1.1.2 при анализе гистограммы, представленной на рис. 1.1.2, *д*. ❑

Пример 1.2.10. Проанализируем теперь данные о котировках акций российских компаний из примера 1.1.3.

Чтобы вычислить значения статистических характеристик не одного, а сразу нескольких показателей, в окне ввода параметров инструмента **«Описательная статистика»** (рис. 1.2.3, *в*) можно указать сразу все столбцы (в нашем случае $B$1:$M$58).

Результаты вычисления статистических характеристик акций компактно представлены в табл. 1.2.4.

Таблица 1.2.4

Статистические характиристики данных о котировках акций

	Среднее	*Медиана*	*Мода*	*Стандартное отклонение*	*Коэффициент асимметрии*	*Размах*	*Минимум*	*Максимум*
Цены								
ВТБ	0,0593	0,0655	0,0680	0,0130	−0,25	0,0422	0,0380	0,0803
Газпром	138,0054	138,7500	#Н/Д	10,4476	−0,23	59,3700	109,1000	168,4700
Сбербанк	106,4088	98,8600	84,5	33,5254	0,65	118,5000	54,9000	173,4000
Объемы								
ВТБ	603 220 561 403,51	435 810 560 000	#Н/Д	424 585 354 374,88	**1,20**	1 832 375 520 000	69 594 480 000	1 901 970 000 000
Газпром	777 164 205,26	731 620 780	#Н/Д	277 538 408,88	**1,71**	1 891 798 280	146 516 160	2 038 314 440
Сбербанк	1 948 336 739,82	1 723 351 580	#Н/Д	917 670 320,35	**1,08**	4 569 288 270	329 303 440	4 898 591 710
Доходности								
ВТБ	0,0065	−0,0086	#Н/Д	0,0945	**2,15**	0,5767	−0,1420	0,4347
Газпром	−0,0012	−0,0032	#Н/Д	0,0678	0,16	0,3152	−0,1363	0,1788
Сбербанк	0,0155	0,0064	#Н/Д	0,0940	0,14	0,4744	−0,2401	0,2343
Логарифмы объемов								
ВТБ	26,8862	26,8005	#Н/Д	0,7169	−0,14	3,3080	24,9660	28,2739
Газпром	20,4106	20,4108	#Н/Д	0,3658	−1,10	2,6327	18,8026	21,4354
Сбербанк	21,2818	21,2675	#Н/Д	0,4849	−0,50	2,6997	19,6125	22,3122

Заметим, что мода у большинства признаков не определена. Это связано с тем, что значения цен, доходностей, объемов очень редко повторяются.

При визуальном анализе гистограмм цен в примере 1.1.3 мы заметили: «гистограмма цены акций Газпрома похожа на симметричный колокол, в отличие от гистограмм цен акций ВТБ и Сбербанка, которые вообще ни на какой на колокол не похожи». Однако рассчитанные значения коэффициентов асимметрии цен всех трех финансовых инструментов по модулю меньше единицы, т. е., возможно, что скошенность распределений цен всех трех акций незначительна.

Что касается объемов торгов, то значения коэффициентов асимметрии (1,20; 1,71; 1,08) подтверждают предположение о том, что распределения объемов торгов по всем трем акциям скошены вправо, при этом наиболее сильный скос у объемов торгов акциями ВТБ.

В примере 1.1.3 мы заметили, что гистограмма доходности акций ВТБ скошена вправо, а гистограммы доходностей акций Газпрома и Сбербанка сим-

метричны. Значения коэффициентов асимметрии подтверждают это: 2,15 у ВТБ, 0,16 и 0,14 соответственно у Газпрома и Сбербанка.

Средние и медианные значения цен акций и объемов торгов сами по себе мало что говорят об их сравнительных инвестиционных характеристиках, ведь акции номинированы в разных масштабах. А вот сравление медианных доходностей может быть полезно. В нашем примере видно, что в половине месяцев акции ВТБ и Газпрома демонстрировали отрицательную доходность, в отличие от акций Сбербанка.

Результаты сохранены на рабочем листе **«Описательная статистика»** в файле **«Акции.xlsx»**. ❑

Задачи

18. С помощью программы **«Описательная статистика»** определите среднее, медиану, моду, размах, стандартное отклонение и коэффициент асимметрии по данным о температуре в Москве из задачи 12.

19. С помощью программы **«Описательная статистика»** определите среднее, медиану, моду, размах, стандартное отклонение и коэффициент асимметрии по данным о потребительских кредитах из задачи 3.

20. С помощью программы **«Описательная статистика»** определите среднее, медиану, моду, размах, стандартное отклонение и коэффициент асимметрии по данным о росте студентов, полученным в задаче 4.

21. Найдите основные статистические показатели для данных об акциях из задачи 5.

1.2.4. Диаграммы размаха

Как мы знаем, медиана x_{med} делит организованный в порялке неубывания ряд значений признака X на две половины, так что все значения из левой половины не больше x_{med}, а все значения из правой половины — не меньше x_{med}. По аналогии с медианой определяются *квартили*[1] — числа, делящие организованный в порялке неубывания ряд значений признака X на четыре равные по численности части:

- 25% всех значений признака X не больше *первой* (или *нижней*) *квартили* $x_{0,25}$, а оставшиеся 75% значений — не меньше $x_{0,25}$;
- 50% всех значений признака X не больше *второй* (или *средней*) *квартили* $x_{0,50}$, а оставшиеся 50% значений — не меньше $x_{0,50}$ (очевидно, вторая квартиль — это то же самое, что медиана);
- 75% всех значений признака X не больше *третьей* (или *верхней*) *квартили* $x_{0,75}$, а оставшиеся 25% значений — не меньше $x_{0,75}$.

Чтобы найти квартили признака X, нужно, расположив значения этого признака в порядке неубывания, найти медиану — это будет средняя квартиль $x_{0,50}$. Затем нужно найти медиану для части значений данного признака, не больших медианы, — это будет нижняя квартиль $x_{0,25}$. Аналогично нужно определить медиану для части значений данного признака, не меньших медианы, — это будет верхняя квартиль $x_{0,25}$.

Разность

$$IQR = x_{0,75} - x_{0,25}$$

[1] Слово *квартиль* (*quartile*) пришло из французского языка и имеет женский род.

называется *межквартильным размахом*[1] и служит характеристикой разброса значений изучаемого признака (как и стандартное отклонение s_X).

Любые значения, не попадающие в отрезок

$$[x_{0,25} - 1{,}5IQR;\ x_{0,75} + 1{,}5IQR],$$

называются *выбросами* (выбросы обязательно подлежат отдельному изучению).

Для визуализации распределения данных по квартилям используется диаграмма размаха, которую еще называют диаграммой «Ящик с усами»[2]. Эта диаграмма (рис. 1.2.6) строится следующим образом.

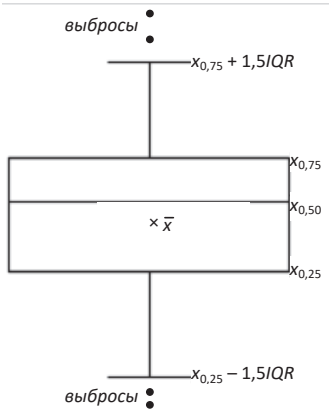


Рис. 1.2.6. *Диаграмма размаха*

«Ящик» на этой диаграмме — это прямоугольник, нижней и верхней границами которого служат соответственно нижняя квартиль $x_{0,25}$ и верхняя квартиль $x_{0,75}$. Внутри «ящика» на уровне медианы $x_{\text{med}} = x_{0,50}$ проводят горизонтальную линию, разделяющую его на две части, а среднее отмечают крестиком. Ширина прямоугольника выбирается из эстетических соображений. Из «ящика» выходят «усы», определяющие границы отрезка

$$\left[\max\left\{x_{(\min)};\ x_{0,25} - 1{,}5IQR\right\};\ \min\left\{x_{0,75} + 1{,}5IQR;\ x_{(\max)}\right\}\right].$$

Если имеются выбросы, т. е. значения, не попадающие в этот отрезок, то каждый выброс отмечается на диаграмме отдельной точкой.

Диаграмма размаха очень проста для понимания, поэтому часто используется для визуализации данных при их первичном анализе. На диаграмме видно, каковы у данного признака типичные значения: половина значений находится в «ящике», и практически все значения (за исклюбчением выбросов) находятся в границах «усов». Если медиана и среднее совпадают и находятся посередине «ящика», то это говорит о симметричности распределения признака, если же медиана смещена к какому-либо концу «ящика» или если у верхнего и нижнего «усов» различается длина, то распределение признака скошено.

Пример 1.2.11. В примере 1.2.1 мы рассматривали признак X — количество кредитов на покупку автомобилей, выданных банком за каждый из 14 дней:

$$5;\ 6;\ 6;\ 7;\ 11;\ 7;\ 5;\ 2;\ 8;\ 7;\ 4;\ 8;\ 9;\ 10.$$

Упорядочив эти числа в порядке неубывания:

[1] *IQR* — сокращение от английского *interquartile range* (межквартильный размах).

[2] От английского названия *box and whisker*.

$$2;\ 4;\ 5;\ 5;\ 6;\ 6;\ \mathbf{7;\ 7;}\ 7;\ 8;\ 8;\ 9;\ 10;\ 11.$$

мы нашли медиану $x_{\text{med}} = 7$ (слева от двух выделенных жирным семерок находится шесть чисел, и справа от этих двух семерок также находится шесть чисел).

Таким образом, средняя квартиль $x_{0,50} = 7$ нам уже известна.

Чтобы определить нижнюю квартиль $x_{0,25}$, найдем медиану левой половины упорядоченных значений признака X:

$$2;\ 4;\ 5;\ \mathbf{5;}\ 6;\ 6;\ 7.$$

Видим, что $x_{0,25} = 5$.

Теперь найдем медиану правой половины упорядоченных значений признака X:

$$7;\ 7;\ 8;\ \mathbf{8;}\ 9;\ 10;\ 11.$$

Получаем $x_{0,75} = 8$.

Межквартильный размах

$$IQR = x_{0,75} - x_{0,25} = 8 - 5 = 3.$$

Поскольку

$$x_{0,25} - 1{,}5IQR = 5 - 1{,}5 \cdot 3 = 0{,}5;\ \max\left\{x_{(\min)};\ x_{0,25} - 1{,}5IQR\right\} = \max\left\{2;\ 0{,}5\right\} = 2;$$

$$x_{0,75} + 1{,}5IQR = 8 + 1{,}5 \cdot 3 = 12{,}5;\ \min\left\{x_{0,75} + 1{,}5IQR;\ x_{(\max)}\right\} = \min\left\{12{,}5;\ 11\right\} = 11,$$

выбросы в данном примере отсутствуют.

Теперь мы можем построить диаграмму размаха (рис. 1.2.7). ❑

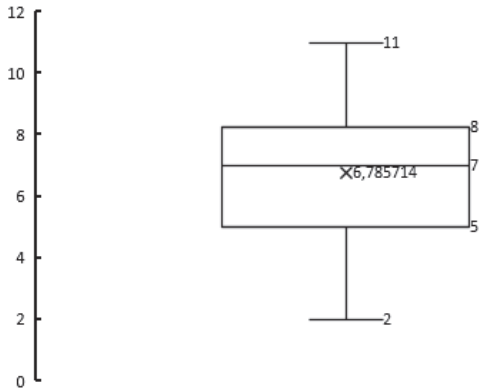


Рис. 1.2.7. *Диаграмма размаха количества кредитов в примере 1.2.11*

Замечание. При определении квартилей возникает вопрос, что делать с медианой: включать ее и в левую, и в правую половины упорядоченных значений изучаемого признака, либо не включать ни в какую. В первом случае говорят об *инклюзивной медиане*, во втором случае — об *эксклюзивной*. При построении диаграммы размаха в Microsoft Excel по умолчанию применяется алгоритм расчета квартилей с эксклюзивной медианой, но если на диаграмме вызвать правой кнопкой мыши всплывающее меню и выбрать в нем пункт **«Формат ряда данных…»** (или дважды щелкнуть мышью по диаграммме), то в появившейся справа области формата ряда данных (рис. 1.2.8) можно выбрать алгоритм с инклюзивной медианой. Очевидно, исключение медианы всегда ведет к увеличению размеров «ящика», однако при анализе рядов данных, содержащих достаточно много значений, разницей между двумя алгоритмами определения квартилей можно пренебречь, и на практике обычно медиану исключают.

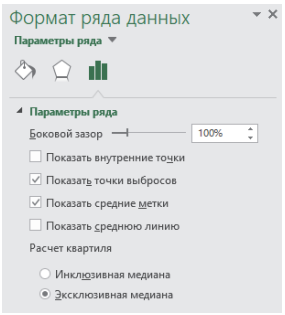
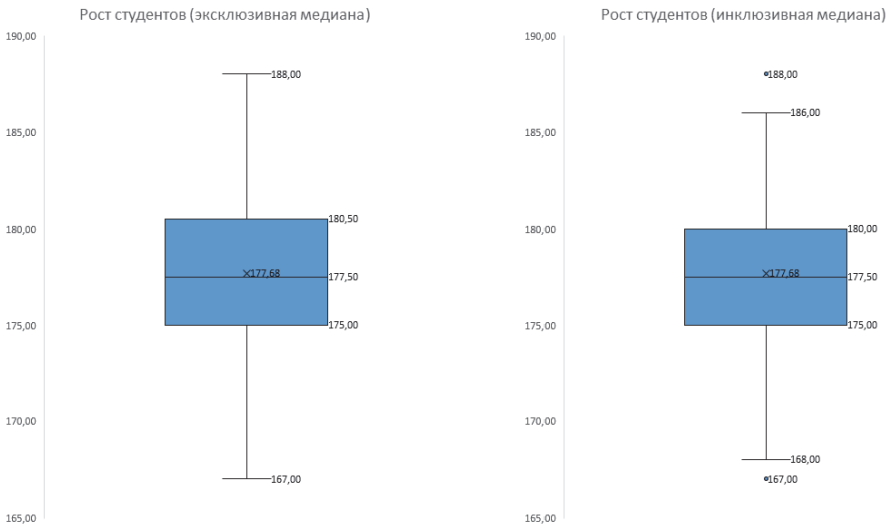


Рис. 1.2.8. *Параметры формата диаграммы размаха в Microsoft Excel*

Пример 1.2.12. Построим диаграмму размаха по данным о росте студентов из примеров 1.1.2, 1.2.9.

Для этого выделим столбец E, в котором находится рост студентов, на вкладке «Данные» нажмем кнопку «Вставить статистическую диаграмму», далее выберем тип статистической диаграммы – «Ящик с усами». Затем на вкладке «Конструктор» нажмем кнопку «Добавить элемент диаграммы», выберем «Подписи данных» — «Справа». Полученная диаграмма размаха роста студентов представлена на рис. 1.2.9, *а*. По умолчанию при расчете квартилей медиана исключалась.

Теперь вызовем на диаграмме всплывающее меню правой кнопкой мыши, выберем в нем пункт «Формат ряда данных...», и в появившейся справа области формата ряда данных (рис. 1.2.8) выберем алгоритм с инклюзивной медианой. Результат представлен на рис. 1.2.9, *б*.



а) *б*)

Рис. 1.2.9. *Диаграмма размаха роста студентов в примере 1.2.12: определение квартилей с исключением медианы (а); определение квартилей с включением медианы (б)*

Видим, что квартили отличаются незначительно. Если исключать медиану, то $x_{0,25} = 175{,}00$; $x_{0,50} = 177{,}50$; $x_{0,75} = 180{,}50$, если же медиану учитывать, то $x_{0,25} = 175{,}00$; $x_{0,50} = 177{,}50$; $x_{0,75} = 180{,}00$. Нижняя и средняя квартили не изменились, а верхняя квартиль уменьшилась на 0,5 см.

Но в результате отрезок «усов» уменьшился. При расчете квартилей с исключением медианы выбросами признаются значения роста, меньшие 167 см или большие 188 см (таких значений в рассматриваемом наборе данных нет), а при расчете с включением медианы выбросами признаются значения, меньшие 168 см или большие 186 см (в нашем примере у одного студента рост равен 167 см, у другого 188 см, эти значения отмечены точками на рис. 1.2.9, *б*).

По диаграмме размаха видно, что медиана проходит примерно посередине «ящика» и близка к среднему, а «усы» имеют приблизительно равную длину. Это говорит в пользу предположения о симметричности распределения роста студентов.

В файле **«Рост студентов.xlsx»** диаграммы размаха сохранены на рабочем листе **«Form1»**. ❑

Пример 1.2.13. Построим в Microsoft Excel диаграммы размаха для данных о котировках акций российских компаний из примеров 1.1.3, 1.2.10.

Результаты представлены на рис. 1.2.10. Поскольку акции номинированы в разных масштабах, для цен и объемов торгов по каждой акции целесообразнее всего строить индивидуальные диаграммы размаха. А так как доходность — безразмерная величина, то для доходностей всех трех акций лучше построить оьбщую диаграмму размаха. Для этого проще всего построить диаграмму размаха для доходности одной из акций (например, ВТБ), а далее правой кнопкой мыши вызвать на этой диаграмме всплывающее меню, затем нажать **«Выбрать данные...»** и далее с помощью кнопки **«Добавить»** включить в диаграмму данные о доходностях двух других акций (рис. 1.2.11).

Замечание. Если в Microsoft Excel сразу выделить данные о доходностях по всем трем акциям и вставить диаграмму «Ящик с усами», то правильная диаграмма размаха автоматически не построится!

При визуальном анализе гистограмм цен в примере 1.1.3 мы заметили: «гистограмма цены акций Газпрома похожа на симметричный колокол, в отличие от гистограмм цен акций ВТБ и Сбербанка, которые вообще ни на какой колокол не похожи». В примере 1.2.10 были вычислены коэффициенты асимметрии цен всех трех акций, которые оказались меньше единицы по модулю, что говорит в пользу незначительной скошенности распределений цен всех трех акций.

Однако на диаграмме размаха цены акций ВТБ медиана сильно смещена к верхней квартили, на диаграмме размаха цены акций Сбербанка верхний «ус» значительно длиннее нижнего, поэтому представляется, что все-таки распределения цен акций ВТБ и Сбербанка скошены (в отличие от цены акций Газпрома, диаграмма размаха которой почти симметрична). Выбросы наблюдаются только у цены акций Газпрома: 109,1 руб. в июне 2013 г.; 116,1 руб. в июле 2017 г.; 168,47 в апреле 2016 г. Даже минимальное значение цены акций Сбербанка [54,90 руб. в декабре 2014 г., когда инвесторы обеспокоились (не подтвердившимися впоследствии) слухами о том, что «Сбербанк прекратил с 16 декабря 2014 г. кредитование физических лиц»[1]] не является выбросом.

[1] Источник — Сбербанк: https://sberbank.ru/ru/press_center/all/article?newsID=200005901-1-1&blockID=&regionID=&lang=ru.

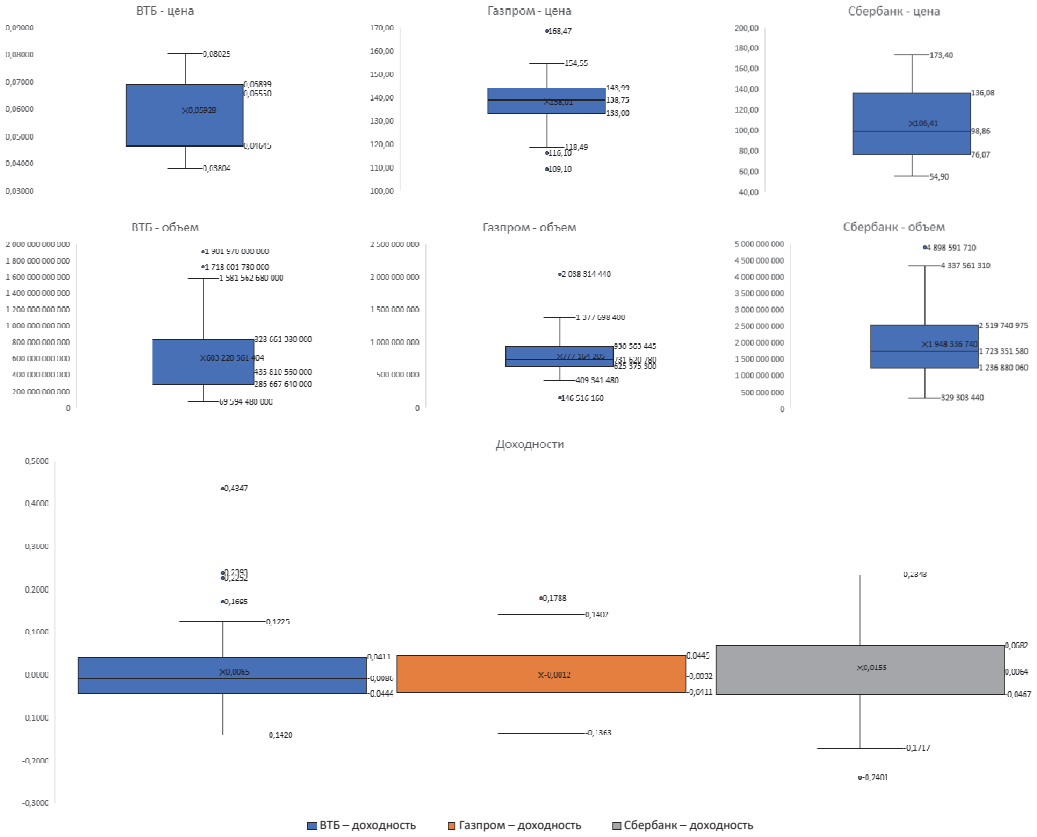


Рис. 1.2.10. *Диаграммы размаха цен, объемов торгов и доходностей акций ВТБ, Газпрома и Сбербанка*

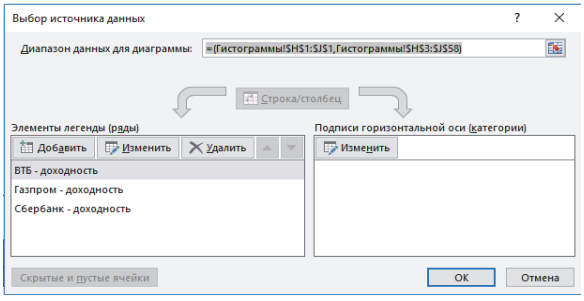


Рис. 1.2.11. *Выбор источников данных для диаграммы размаха доходностей акций ВТБ, Газпрома и Сбербанка*

На основе анализа гистограмм объемов торгов (в примере 1.1.3) и коэффициентов асимметрии (в примере 1.2.10) было сделано предположение о том, что распределения объемов торгов по всем трем акциям скошены вправо, при этом наиболее сильный скос у объемов торгов акциями ВТБ. Обэтом же говорят и диаграммы размаха объемов торгов: на всех трех диаграммах медианы сильно смещены к нижним квартилям и значительно отличаются от средних, при этом верхние «усы» на всех трех диаграммах заметно длиннее нижних.

Особенно ярко это проявляется на диаграмме размаха объема торгов по акциям ВТБ. Кроме того, на всех трех диаграммах есть выбросы. В марте 2014 г. объем торгов акциями ВТБ оказался равен 1 718 001 780 000 шт., объем торгов акциями Газпрома составил 2 038 314 440 шт., а объем торгов акциями Сбербанка — 4 898 591 710 шт. Напомним, что в марте 2014 г. произошло присоединение Крыма к России, и значительная часть инвесторов, опасаясь резкого падения цен акций, закрывала свои позиции. Как мы видим, впоследствии цены всех трех акций выросли по сравнению с 2012 г., и объемы торгов в марте 2014 г. действительно можно считать выбросами, нетипичными значениями.

В примере 1.1.3 мы заметили, что гистограмма доходности акций ВТБ скошена вправо, а гистограммы доходностей акций Газпрома и Сбербанка симметричны, в примере 1.2.10 анализ коэффициентов асимметрии (2,15 у ВТБ, 0,16 и 0,14 соответственно у Газпрома и Сбербанка) подтвердил этот вывод. Анализ диаграмм размаха показывает, что, скорее всего, дело не в скошенности распределения доходности акций ВТБ, а в выбросах: 0,2393 в мае 2014 г., 0,1695 в ноябре 2014 г., 0,4347 в декабре 2014 г. и 0,2252 в мае 2015 г. — это нетипично высокие значения.

Также по диаграмме размаха доходностей видно, что типичная доходность по акциям Сбербанка выше, чем по акциям ВТБ и Газпрома, но при этом размах доходности по акциям Сбербанка также выше, чем размах доходности по акциям ВТБ и Газпрома.

Результаты сохранены на рабочих листах «Диаграммы размаха цен, объемов» и «Диаграмма размаха доходностей» в файле «Акции.xlsx». ❑

Задачи

22. Постройте диаграмму размаха по данным о температуре в Москве из задачи 12.

23. Постройте диаграмму размаха по данным о потребительских кредитах из задачи 3.

24. Постройте диаграмму размаха по данным о росте студентов, полученным в задаче 4.

25. Постройте диаграммы размаха по данным о ценах, объемах торгов и доходностях акций из задачи 5.

§ 1.3. Визуализация качественных признаков в Microsoft Excel

1.3.1. Сводные таблицы и сводные диаграммы

При наличии большого количества данных часто совершенно непонятно, с чего начать их анализ. В этом случае может оказаться полезно создать сводную таблицу или построить сводную диаграмму.

Сводная таблица — это инструмент обработки данных, упрощающий их обобщение и визуализацию. Если имеются исходные данные в форме таблицы, столбцы которой содержат однотипные значения, то *сводная таблица* — это таблица, связанная с исходной и обобщающая исходные данные путем вычисления сумм, средних, долей и т. п.

Сводная таблица создается при помощи нескольких действий, потом ее можно быстро настроить, чтобы выбрать наиболее информативное представ-

ление результатов, и создать на основе сводной таблицы сводную диаграмму, автоматически обновляющуюся при изменении данных.

Рассмотрим создание сводных таблиц и сводных диаграмм на примерах.

Пример 1.3.1. В файле **«Кредиты по видам.xlsx»** содержатся данные о 500 кредитах, выданных банком в прошлом месяце (рис. 1.3.1). Вряд ли кто-то может сделать какой-нибудь полезный вывод, просто посмотрев на эти данные. Обобщим эти данные с помощью сводной таблицы и сводных диаграмм.

Вначале отформатируем исходные данные как таблицу (для этого нажмем на кнопку **«Форматировать как таблицу»**, расположенную на вкладке **«Главная»**).

Теперь, находясь в одной из ячеек таблицы, на вкладке **«Вставка»** нажмем кнопку **«Сводная таблица»**, в появившемся окне параметров (рис. 1.3.2) убедимся, что в качестве **таблицы или диапазона** указана таблица **«Кредиты_по_видам»**, укажем, что сводную таблицу нужно поместить **на существующий лист** в ячейку **$C$1** — справа от исходных данных.

В появившейся справа области выбора полей сводной таблицы (рис. 1.3.3) с помощью мыши перенесем поле **«Вид кредита»** в ячейку **«Строки»**, а также в ячейку **«Значения»** (при этом в ячейке **«Значения»** появится элемент **«Количество по полю Вид кредита»**). Построенная сводная таблица (на рис. 1.3.4 она находится справа от исходных данных) имеет два столбца:

- **«Названия строк»**, в котором содержатся все виды выданных кредитов;
- **«Количество по полю Вид кредита»**, в котором указано, сколько кредитов данного вида было выдано.

В любой из ячеек столбца **«Количество по полю Вид кредита»** вызовем правой кнопкой мыши всплывающее меню и выберем в нем пункт **«Сортировка»** — **«Сортировка по убыванию»**, чтобы упорядочить кредиты от наиболее популярных к наименее популярным.

	A	B
1	Номер договора	Вид кредита
2	1	Кредит для приобретения подержанного автомобиля
3	2	Потребительский кредит без обеспечения
4	3	Кредит для строительства жилого дома
5	4	Кредит для приобретения подержанного автомобиля
6	5	Потребительский кредит без обеспечения
7	6	Кредит для приобретения нового автомобиля
8	7	Кредит для приобретения подержанного автомобиля
...	...	...
501	500	Потребительский кредит без обеспечения

Рис. 1.3.1. *Данные о выданных кредитах*

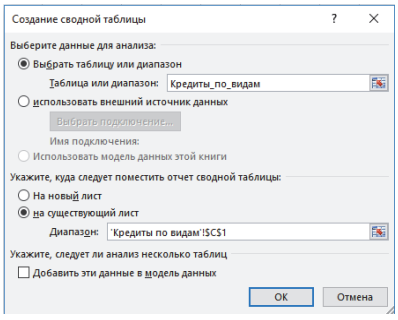


Рис. 1.3.2. *Параметры сводной таблицы*

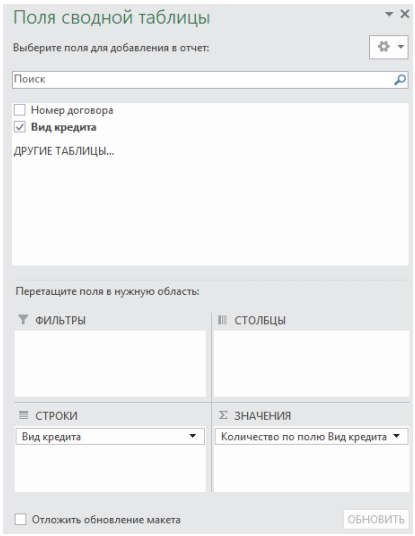


Рис. 1.3.3. *Выбор полей в сводной таблице в примере 1.3.1*

	A	B	C	D
1	Номер договора	Вид кредита	Названия строк	Количество по полю Вид кредита
2	1	Кредит для приобретения подержанного автомобиля	Кредит для приобретения нового автомобиля	125
3	2	Потребительский кредит без обеспечения	Потребительский кредит без обеспечения	105
4	3	Кредит для строительства жилого дома	Кредит для приобретения квартиры в строящемся доме	93
5	4	Кредит для приобретения подержанного автомобиля	Кредит для приобретения подержанного автомобиля	62
6	5	Потребительский кредит без обеспечения	Кредит для строительства жилого дома	31
7	6	Кредит для приобретения нового автомобиля	Кредит для приобретения готового жилья	31
8	7	Кредит для приобретения подержанного автомобиля	Нецелевой кредит под залог недвижимости	27
9	8	Кредит для приобретения нового автомобиля	Потребительский кредит под поручительство физических лиц	26
10	9	Кредит для строительства жилого дома	**Общий итог**	**500**
...	...	...		

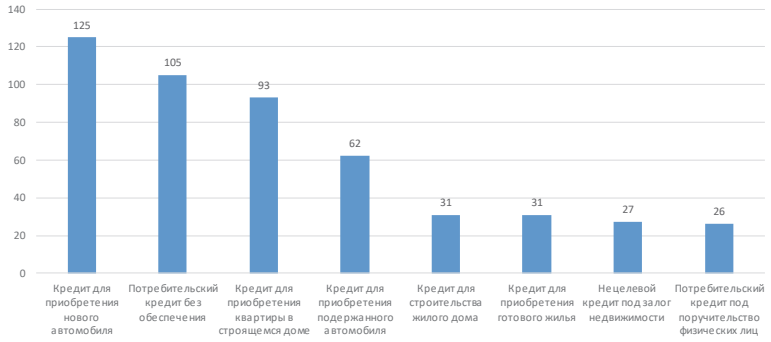
Рис. 1.3.4. *Сводная таблица о выданных кредитах*

Теперь мы видим, какие кредиты в данном банке выдаются чаще, а какие реже. Представим эту информацию в визуальной форме с помощью сводной диаграммы. Для этого на вкладке «Анализ» нажмем на кнопку «Сводная диаграмма» и выберем в качестве типа диаграммы «Гистограмма» — «Гистограмма с группировкой». Получим результат, представленный на рис. 1.3.5, *а*.

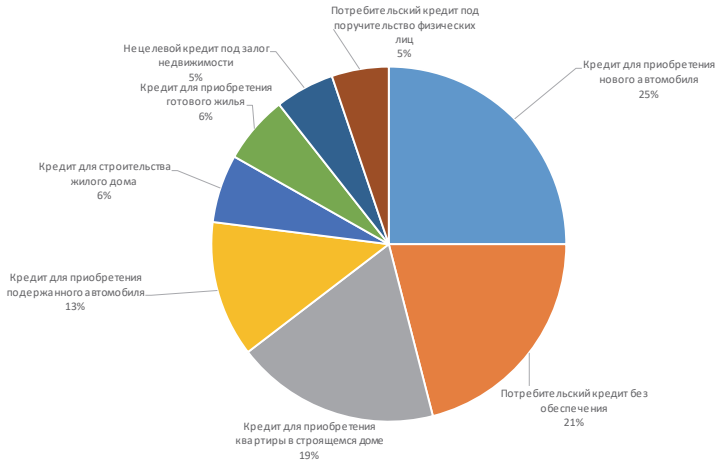
Чтобы визуализировать процентные соотношения, удобно использовать круговые диаграммы. На вкладке «Анализ» нажмем на кнопку «Сводная диаграмма» и выберем в качестве типа диаграммы «Круговая». Теперь, находясь на построенной круговой диаграмме, нажмем на вкладке «Конструктор» кнопку «Добавить элемент диаграммы», в появившемся меню выберем «Подписи данных» — «Дополнительные параметры подписей данных...» и укажем, что нужно добавить имя категории, доли и линии выноски (рис. 1.3.6).

В результате получим круговую диаграмму, визуализирующую процентное соотношение между количеством выданных кредитов разных видов (рис. 1.3.5, *б*). ❑

Для изучения структуры какого-либо множества объектов, характеризуемых количественным признаком, бывает удобно, как в следующем примере, с помощью сводных таблиц провести группировку количественных признаков и перейти от количественного признака к порядковому.



а)



б)

Рис. 1.3.5. *Визуализация данных о выданных кредитах с помощью сводных диаграмм: гистограмма с группировкой* (*а*); *круговая диаграмма* (*б*)

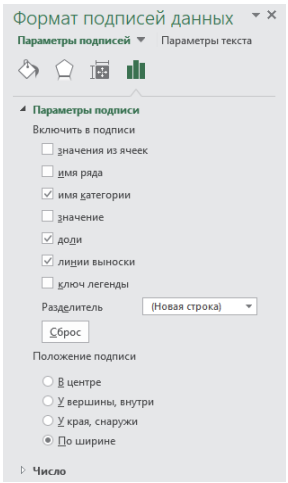


Рис. 1.3.6. *Настройка подписей данных в круговой диаграмме*

Пример 1.3.2. В табл. 1.3.1 и в файле «Население России по регионам.xlsx» представлены данные о численности населения субъектов Российской Федерации по состоянию на 1 января 2017 г.[1] Всего в России 82 региона,[2] и сразу понять, как они отличаются по численности населения, при взгляде на таблицу сложно.

Таблица 1.3.1

Численность населения регионов России по состоянию на 1 января 2017 г.

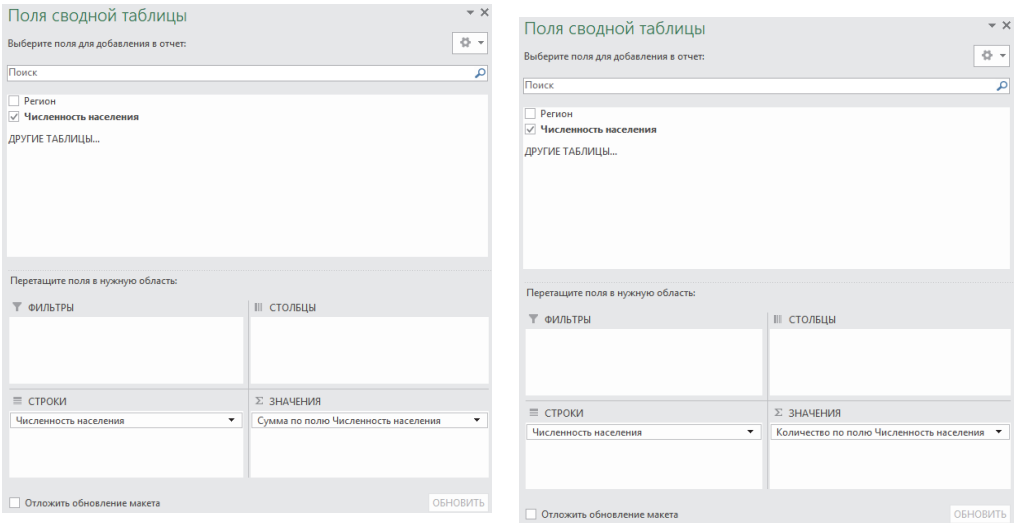
Регион	*Численность населения*	*Регион*	*Численность населения*
Алтайский край	2 365 700	Пермский край	2 632 100
Амурская область	801 800	Приморский край	1 923 100
Архангельская область, включая Ненецкий авт.округ	1 165 700	Псковская область	642 200
Астраханская область	1 018 900	Республика Адыгея	453 400
Белгородская область	1 552 900	Республика Алтай	217 000
Брянская область	1 220 500	Республика Башкортостан	4 067 000
Владимирская область	1 389 600	Республика Бурятия	984 100
Волгоградская область	2 535 200	Республика Дагестан	3 041 900
Вологодская область	1 183 900	Республика Ингушетия	480 500
Воронежская область	2 335 400	Республика Калмыкия	277 800
г. Москва	12 380 700	Республика Карелия	627 100
г. Санкт-Петербург	5 281 600	Республика Коми	850 500
г. Севастополь	428 800	Республика Крым	1 912 200
Еврейская авт. область	164 200	Республика Марий Эл	684 700
Забайкальский край	1 079 000	Республика Мордовия	808 500
Ивановская область	1 023 200	Республика Саха (Якутия)	962 800
Иркутская область	2 408 900	Республика Северная Осетия-Алания	703 300
Кабардино-Балкарская Республика	864 400	Республика Татарстан	3 885 200
Калининградская область	986 300	Республика Тыва	318 600
Калужская область	1 014 600	Республика Хакасия	537 700
Камчатский край	314 700	Ростовская область	4 231 300
Карачаево-Черкесская Республика	466 400	Рязанская область	1 126 700
Кемеровская область	2 708 800	Самарская область	3 203 700
Кировская область	1 291 700	Саратовская область	2 479 200
Костромская область	648 200	Сахалинская область	487 400
Краснодарский край	5 570 900	Свердловская область	4 329 400
Красноярский край	2 875 300	Смоленская область	953 200
Курганская область	854 100	Ставропольский край	2 804 400
Курская область	1 122 900	Тамбовская область	1 040 300
Ленинградская область	1 791 900	Тверская область	1 296 800
Липецкая область	1 156 200	Томская область	1 078 900
Магаданская область	145 600	Тульская область	1 499 400
Московская область	7 423 500	Тюменская область, включая авт. округа	3 660 000
Мурманская область	757 600	Удмуртская Республика	1 516 800
Нижегородская область	3 247 700	Ульяновская область	1 252 900
Новгородская область	612 500	Хабаровский край	1 333 300
Новосибирская область	2 779 500	Челябинская область	3 502 300
Омская область	1 972 700	Чеченская Республика	1 414 900
Оренбургская область	1 989 600	Чувашская Республика	1 235 900
Орловская область	754 800	Чукотский авт. округ	49 800
Пензенская область	1 341 500	Ярославская область	1 270 700

Отформатируем исходные данные как таблицу (напомним, что для этого нужно нажать кнопку «Форматировать как таблицу» на вкладке «Главная») и создадим сводную таблицу (нажав кнопку «Сводная таблица» на вкладке «Вставка»).

[1] Источник — Росстат: http://gks.ru/bgd/regl/b17_107.

[2] Без учета Ханты-Мансийского и Ямало-Ненецкого автономных округов, которые входят в состав Тюменской области, а также Ненецкого автономного округа, входящего в состав Архангельской области.

Из списка полей перенесем поле «Численность населения» и в ячейку «Строки», и в ячейку «∑ Значения». При этом в ячейке «∑ Значения» появится элемент «Сумма по полю Численность населения» (рис. 1.3.7, *а*).



а) *б*)

Рис. 1.3.7. *Выбор полей сводной таблицы в примере 1.3.2: суммирование* (*а*); *подсчет количества значений* (*б*)

В результате в сводной таблице (рис. 1.3.8, *а*) будет 82 строки и два одинаковых столбца, в каждом из которых содержится отсортированная по возрастанию численность населения в регионах — пока ничего нового (все регионы разные, поэтому они разделены на 82 группы — по одному региону в группе; суммарная численность населения в группе из одного региона — это просто численность населения в регионе).

Нажмем на стрелку справа от элемента «Сумма по полю Численность населения» в ячейке «∑ Значения», в появившемся меню выберем «Параметры полей значений...» и в качестве операции вместо суммы выберем количество (рис. 1.3.7, *б*).

Теперь правый столбец сводной таблицы состоит из одних единиц (поскольку нет регионов с совпадающей численностью населения).

Чтобы сделать эту сводную таблицу полезной, находясь на столбце «Численность населения» сводной таблицы, вызовем всплывающее меню, выберем в нем пункт «Группировать» и зададим подходящие параметры групп, например, от 0 до 13 000 000 с шагом в 1 000 000.

В получившейся сводной таблице регионы сгруппированы по численности населения (рис. 1.3.8, *б*). Видно, что 30 из 82 регионов (37%) имеют численность до 1 млн человек; 28 регионов (34%) имеют численность от 1 до 2 млн человек; 12 регионов имеют численность от 2 до 3 млн человек; пять регионов имеют численность от 3 до 4 млн человек; три региона имеют численность от 4 до 5 млн человек; два региона имеют численность от 5 до 6 млн человек (г. Санкт-Петербург и Краснодарский край) и по одному региону имеют численность от 7 до 8 млн человек (это Московская область) и от 12 до 13 млн человек (это Москва).

Еще лучше структуру численности населения в регионах видно на сводных диаграммах (рис. 1.3.9). ❑

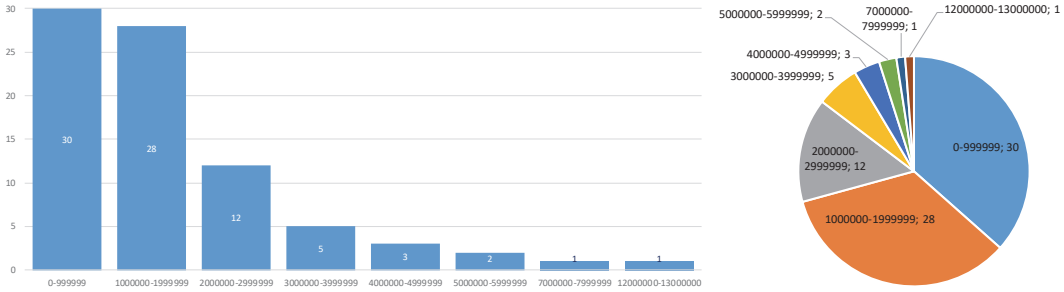
	A	B	C	D
1	Регион	Численность населения		
2	Алтайский край	2 365 700		
3	Амурская область	801 800	Сумма по полю Численность населения	
4	Архангельская область, включая Ненецкий авт. округ	1 165 700	Численность населения	Итог
5	Астраханская область	1 018 900	50 759	50 759
6	Белгородская область	1 552 900	148 105	148 105
7	Брянская область	1 220 500	168 408	168 408
8	Владимирская область	1 389 600	213 544	213 544
9	Волгоградская область	2 535 200	280 581	280 581
10	Вологодская область	1 183 900	313 612	313 612
11	Воронежская область	2 335 400	317 206	317 206
12	г. Москва	12 380 700	400 865	400 865
13	г. Санкт-Петербург	5 281 600	449 160	449 160
14	г. Севастополь	428 800	463 408	463 408
15	Еврейская авт. область	164 200	469 253	469 253
16	Забайкальский край	1 079 000	488 308	488 308
17	Ивановская область	1 023 200	535 647	535 647
18	Иркутская область	2 408 900	618 696	618 696
...	...	...		

а)

	A	B	C	D
1	Регион	Численность населения		
2	Алтайский край	2 365 700		
3	Амурская область	801 800	Сумма по полю Численность населения	
4	Архангельская область, включая Ненецкий авт. округ	1 165 700	Численность населения	Итог
5	Астраханская область	1 018 900	0–999999	30
6	Белгородская область	1 552 900	100000–1999999	28
7	Брянская область	1 220 500	2000000–2999999	12
8	Владимирская область	1 389 600	3000000–3999999	5
9	Волгоградская область	2 535 200	4000000–4999999	3
10	Вологодская область	1 183 900	5000000–5999999	2
11	Воронежская область	2 335 400	7000000–7999999	1
12	г. Москва	12 380 700	12000000–13000000	1
13	г. Санкт-Петербург	5 281 600	Общий итог	82
14	г. Севастополь	428 800		
15	Еврейская авт. область	164 200		
16	Забайкальский край	1 079 000		
17	Ивановская область	1 023 200		
18	Иркутская область	2 408 900		
...	...	...		

б)

Рис. 1.3.8. *Сводная таблица численности населения регионов России: по несгруппированным данным* (*а*); *по сгруппированным данным* (*б*)



а) *б*)

Рис. 1.3.9. *Сводные диаграммы численности населения регионов России: гистограмма с группировкой* (*а*); *круговая диаграмма* (*б*)

Задачи

26. В файле **«Продажи автомобилей.xlsx»** содержатся данные о марках автомобилей, проданных за последний год мультибрендовым автодилером. Постройте сводную таблицу, выбрав марки автомобилей в качестве строк и

количество проданных автомобилей по маркам в качестве значений. Визуализируйте результаты с помощью гистограммы с группировкой и круговой диаграммы.

27. В файле «Расходы на путешествия.xlsx» содержатся данные о возрасте и поле клиентов интернет-сервиса бронирования апартаментов, сделавших заказы на бронирование жилья, и стоимости бронирования жилья. Постройте сводную таблицу, выбрав возраст и пол в качестве строк, а среднее по полю «Стоимость бронирования апартаментов» в качестве значений. Объедините клиентов в группы по возрасту: от 25 до 34 лет, от 35 до 44 лет и т д. Визуализируйте результаты с помощью гистограммы с группировкой.

1.3.2. Таблицы сопряженности и парадокс Симпсона

Для анализа взаимосвязей между качественными признаками бывает полезно использовать таблицы сопряженности. *Таблица сопряженности* — это таблица, строки которой соответствуют значениям одного признака, столбцы соответствуют значениям другого признака, а в каждой ячейке указывается, сколько раз в наборе данных встречается соответствующая комбинация значений двух признаков.

Чтобы по исходным данным построить таблицу сопряженности, проще всего воспользоваться инструментарием создания сводных таблиц.

Пример 1.3.3. В файле «Возврат кредитов.xlsx» содержатся данные о 403 потребительских кредитах и кредитах для приобретения автомобилей, выданных двумя банками (рис. 1.3.10, *а*). Для каждого кредитного договора указан банк, выдавший кредит, назначение кредита (потребительский кредит или автокредит) и статус договора (погашен или не погашен). Чтобы визуализировать такую таблицу, в которой нет ни одного числа, построим таблицу сопряженности. Для этого отформатируем исходные данные как таблицу, затем вставим, начиная с ячейки $D$1, сводную таблицу, выбрав в качестве строк сводной таблицы признаки «Банк» и «Тип кредита», в качестве столбцов признак «Статус», а в качестве значений — «Количество по полю Статус». Результат представлен на рис. 1.3.10, *б* (ячейки D1:G9). Визуализируем эту таблицу сопряженности с помощью сводной диаграммы (рис. 1.3.10, *в*).

Вставим еще одну точно такую же сводную таблицу, начиная с ячейки $D$11, а затем вызовем всплывающее меню, находясь на одном из числовых элементов новой сводной таблицы, и выберем «Дополнительные вычисления» — «% от суммы по столбцу». Теперь во второй сводной таблице (она представлена в ячейках D11:G19 на рис. 1.3.10, *б*) указано, какой процент каждого типа кредитов в каждом банке был возвращен, а какой не был возвращен. Соответствующая сводная диаграмма представлена на рис. 1.3.10, *г*).

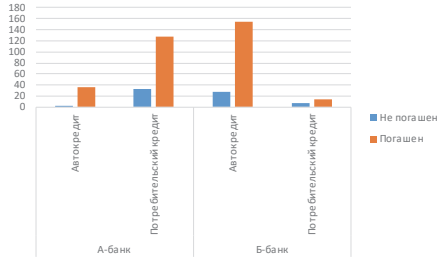
Выясним, в каком из банков выше риск невозврата кредита. В А-банке не было возвращено 7,50% автокредитов, а в Б-банке не было возвращено 14,92% автокредитов — риск невозврата автокредита в А-банке существенно ниже, чем в Б-банке. Риск невозврата потребительских кредитов в А-банке также существенно ниже, чем в Б-банке: в А-банке не было возвращено 20,00% потребительских кредитов, а в Б-банке не было возвращено 31,82% потребительских кредитов. А следующее утверждение, следующее из анализа таблицы сопряженности, на первый взгляд, противоречит здравому смыслу: в целом риск невозврата кредита в А-банке (17,50%) выше, чем в Б-банке (16,75%)!

	A	B	C
1	Банк	Вид кредита	Статус
2	А-банк	Автокредит	Погашен
3	А-банк	Автокредит	Погашен
4	А-банк	Автокредит	Погашен
5	А-банк	Автокредит	Не погашен
6	А-банк	Автокредит	Погашен
7	А-банк	Автокредит	Погашен
8	А-банк	Автокредит	Погашен
9	А-банк	Автокредит	Погашен
10	А-банк	Автокредит	Погашен
11	А-банк	Автокредит	Погашен
12	А-банк	Автокредит	Погашен
13	А-банк	Автокредит	Погашен
14	А-банк	Автокредит	Погашен
15	А-банк	Автокредит	Погашен
16	А-банк	Автокредит	Погашен
17	А-банк	Автокредит	Погашен
18	А-банк	Автокредит	Погашен
19	А-банк	Автокредит	Погашен
…	…	…	…
402	Б-банк	Потребительский кредит	Погашен
403	Б-банк	Потребительский кредит	Не погашен
404	Б-банк	Потребительский кредит	Погашен

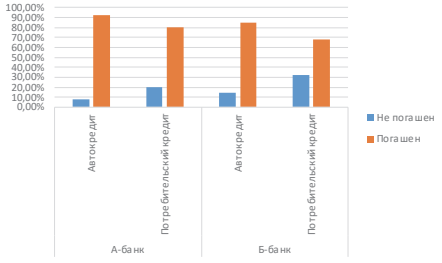
а)

	A	B	C	D	E	F	G
1	Банк	Вид кредита	Статус	Количество по полю Статус	Названия столбцов		
2	А-банк	Автокредит	Погашен	Названия строк	Не погашен	Погашен	Общий итог
3	А-банк	Автокредит	Погашен	**А-банк**	**35**	**165**	**200**
4	А-банк	Автокредит	Погашен	Автокредит	3	37	40
5	А-банк	Автокредит	Не погашен	Потребительский кредит	32	128	160
6	А-банк	Автокредит	Погашен	**Б-банк**	**34**	**169**	**203**
7	А-банк	Автокредит	Погашен	Автокредит	27	154	181
8	А-банк	Автокредит	Погашен	Потребительский кредит	7	15	22
9	А-банк	Автокредит	Погашен	**Общий итог**	**69**	**334**	**403**
10	А-банк	Автокредит	Погашен				
11	А-банк	Автокредит	Погашен	Количество по полю Статус	Названия столбцов		
12	А-банк	Автокредит	Погашен	Названия строк	Не погашен	Погашен	Общий итог
13	А-банк	Автокредит	Погашен	**А-банк**	**17,50%**	**82,50%**	**100,00%**
14	А-банк	Автокредит	Погашен	Автокредит	7,50%	92,50%	100,00%
15	А-банк	Автокредит	Погашен	Потребительский кредит	20,00%	80,00%	100,00%
16	А-банк	Автокредит	Погашен	**Б-банк**	**16,75%**	**83,25%**	**100,00%**
17	А-банк	Автокредит	Погашен	Автокредит	14,92%	85,08%	100,00%
18	А-банк	Автокредит	Погашен	Потребительский кредит	31,82%	68,18%	100,00%
19	А-банк	Автокредит	Погашен	**Общий итог**	**17,12%**	**82,88%**	**100,00%**
…	…	…	…				
402	Б-банк	Потребительский кредит	Погашен				
403	Б-банк	Потребительский кредит	Не погашен				
404	Б-банк	Потребительский кредит	Погашен				

б)



в)



г)

Рис. 1.3.10. *Данные о погашенных и непогашенных кредитах: исходные танные (а); таблицы сопряженности (б); сводная диаграмма для абсолютных значений (в); сводная диаграмма для относительных значений (г)*

Данное явление называется **парадоксом Симпсона**.[1] «Здравый смысл» нас здесь подводит, предлагая считать, что раз и автокредиты, и потребительские кредиты в А-банке возвращаются чаще, чем в Б-банке, то и в целом кредиты в А-банке возвращаются чаще, чем в Б-банке. На самом деле, если посмотреть не на относительные, а на абсолютные значения (первая сводная таблица и сводная диаграмма на рис. 1.3.10, *в*), то можно заметить, что в А-банке выдается больше потребительских кредитов, тогда как Б-банк больше специализируется на автокредитах.

Причина парадокса Симпсона заключается в некорректном усреднении. Интуитивно предполагается, что при усреднении найденных зависимостей соотношение количества выданных автокредитов и потребительских кредитов в двух банка одинаково, поэтому можно просто вычислить средние проценты невозврата кредитов в двух банках:

$$\frac{20{,}00\% + 7{,}5\%}{2} = 13{,}75\% \text{ для А-банка и } \frac{31{,}82\% + 14{,}92\%}{2} = 23{,}37\% \text{ для Б-банка.}$$

Но так делать нельзя, поскольку А-банк выдал 160 потребительских кредитов и 40 автокредитов, а Б-банк — 22 потребительских кредита и 181 автокредит. Чтобы это учесть, нужно домножить долю невозвращенных кредитов каждого типа на долю кредитов данного типа в общем числе кредитов, выданных данным банком. Для А-банка:

$$20{,}00\% \cdot \frac{160}{160+40} + 7{,}50\% \cdot \frac{40}{160+40} = 20{,}00\% \cdot 0{,}8 + 7{,}50\% \cdot 0{,}2 = 17{,}50\%,$$

для Б-банка:

$$31{,}82\% \cdot \frac{22}{22+181} + 14{,}92\% \cdot \frac{181}{22+181} = 31{,}82\% \cdot 0{,}108 + 14{,}92\% \cdot 0{,}892 = 16{,}75\%.$$

Теперь мы усреднили корректно. Именно так произведено усреднение в сводной таблице: риск невозврата кредита для А-банка и Б-банка соответственно равен 17,50% и 16,75% (ячейки E13 и E16 на рис. 1.3.10, *б*). В целом же в двух банках риск невозврата кредита равен 17,12% (ячейка E19). ❑

Задачи

28. В файле **«Приобретение автомобилей.xlsx»** содержатся данные о 344 семьях, собирающихся приобрести автомобили. Для каждой семьи указано, предпочитает ли она внедорожник, сколько в ней детей (больше двух или не больше двух), какие у семьи доходы (высокие или низкие). С помощью сводной таблицы ответьте на следующие вопросы: *а*) какие семьи с двумя и более детьми больше предпочитают внедорожники — с высокими или низкими доходами; *б*) какие семьи, имеющие меньше двух детей больше предпочитают внедорожники — с высокими или низкими доходами; *в*) какие семьи в целом больше предпочитают внедорожники — с высокими или низкими доходами?

29. В условиях задачи 28 выясните, какие семьи больше предпочитают внедорожники — менее чем с двумя детьми или с двумя и более детьми?

[1] Подробнее см. статью «Парадокс Симпсона» в Википедии (https://ru.wikipedia.org/wiki/Парадокс_Симпсона) и ссылки в этой статье.

1.3.3. Иерархия признаков

Часто в данных имеется определенная иерархия признаков, когда все множество объектов разделается на какие-либо категории, каждая категория, в свою очередь, делится на подкатегории и т. д. Распознать такую иерархию можно с помощью визуализации данных.

Пример 1.3.4. В табл. 1.3.2 и в файле «Население России по округам.xlsx» представлены данные о численности населения субъектов Российской Федерации по состоянию на 1 января 2017 г.[1]

Таблица 1.3.2

Численность городского и сельского населения федеральных округов России по состоянию на 1 января 2017 г.

Федеральный округ	*Численность городского населения*	*Численность сельского населения*	*Суммарная численность населения*
Центральный федеральный округ	32 189 400	7 020 200	39 209 600
Северо-Западный федеральный округ	11 712 000	2 187 300	13 899 300
Южный федеральный округ	10 252 700	6 175 800	16 428 500
Северо-Кавказский федеральный округ	4 803 600	4 972 200	9 775 800
Приволжский федеральный округ	21 264 000	8 372 500	29 636 500
Уральский федеральный округ	10 028 400	2 317 400	12 345 800
Сибирский федеральный округ	14 103 500	5 222 700	19 326 200
Дальневосточный федеральный округ	4 678 800	1 503 900	6 182 700

Визуализируем данные. Для этого выделим диапазон A1:C9, содержащий по федеральным округам данные о численности городского и сельского населения, на вкладке «Вставить» нажмем на кнопку «Вставить гистограмму или линейчатую диаграмму», далее выберем тип диаграммы — «Гистограмма с накоплением». Получится диаграмма, как на рис. 1.3.11, *а*.

По этой диаграмме можно сделать много выводов, например:

- во всех округах, кроме Северо-Кавказского, городского населения больше, чем сельского;
- Дальневосточный и Северо-Кавказский округа — самые маленькие по численности;
- в Северо-Западном, Уральском и Дальневосточном федеральных округах процент сельского населения существенно меньше, чем в остальных округах.

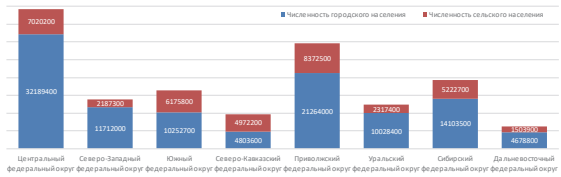
Существуют диаграммы, которые представляют иерархические данные в более эффектной форме.

Чтобы построить диаграмму «Дерево» (рис. 1.3.11, *б*), вначале данные нужно представить в виде трех столбцов, определяющих иерархию признаков (рис. 1.3.12):

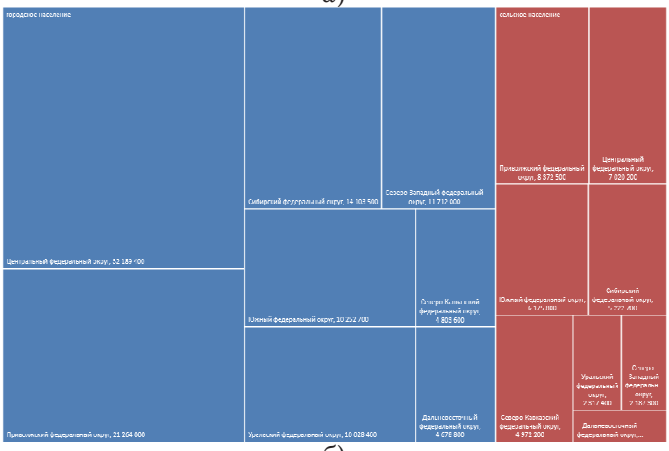
- тип населения — городское или сельское;
- федеральный округ;
- численность.

Теперь на вкладке «Вставка» нужно нажать кнопку «Рекомендуемые диаграммы», далее перейти на вкладку «Все диаграммы» и выбрать диаграмму «Дерево».

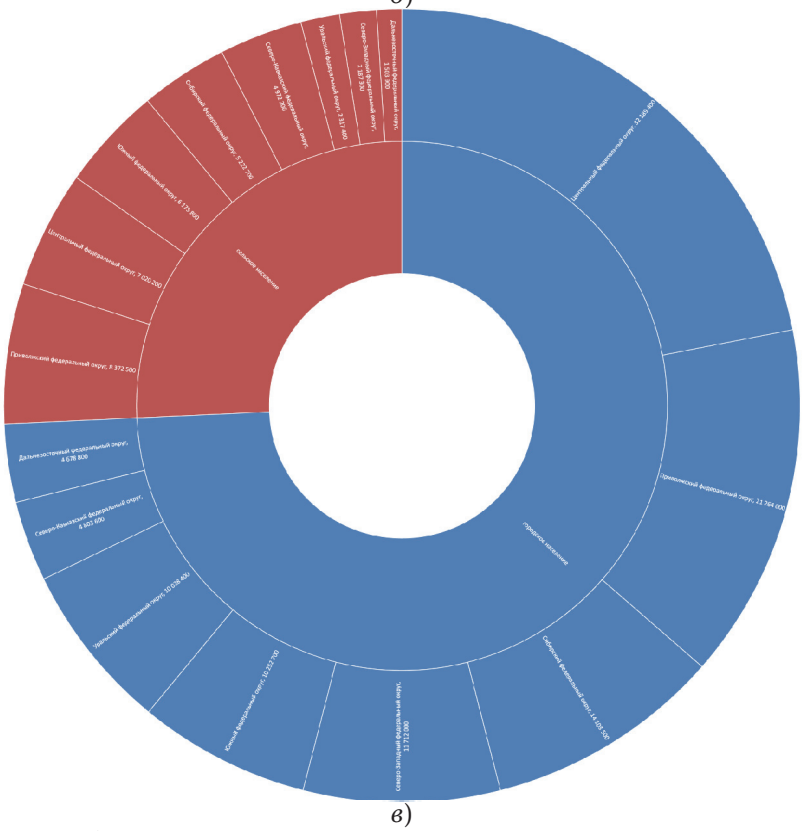
[1] Источник — Росстат: http://gks.ru/bgd/regl/b17_107.



а)



б)



в)

Рис. 1.3.11. *Визуализация данных о численности населения федеральных округов России: гистограмма с накоплением* (*а*); *диаграмма «Дерево»* (*б*); *диаграмма «Солнечные лучи»* (*в*)

	A	B	C
1	Население	Федеральный округ	Численость
2	городское население	Центральный федеральный округ	32 189 400
3	городское население	Северо-Западный федеральный округ	11 712 000
4	городское население	Южный федеральный округ	10 252 700
5	городское население	Северо-Кавказский федеральный округ	4 803 600
6	городское население	Приволжский федеральный округ	21 264 000
7	городское население	Уральский федеральный округ	10 028 400
8	городское население	Сибирский федеральный округ	14 103 500
9	городское население	Дальневосточный федеральный округ	4 678 800
10	сельское население	Центральный федеральный округ	7 020 200
11	сельское население	Северо-Западный федеральный округ	2 187 300
12	сельское население	Южный федеральный округ	6 175 800
13	сельское население	Северо-Кавказский федеральный округ	4 972 200
14	сельское население	Приволжский федеральный округ	8 372 500
15	сельское население	Уральский федеральный округ	2 317 400
16	сельское население	Сибирский федеральный округ	5 222 700
17	сельское население	Дальневосточный федеральный округ	1 503 900

Рис. 1.3.12. *Иерархическое представление данных о численности городского и сельского населения федеральных округов России*

Для того, чтобы получить диаграмму **«Солнечные лучи»** (рис. 1.3.11, *в*), на вкладке **«Вставка»** нужно нажать кнопку **«Рекомендуемые диаграммы»**, а далее на вкладке **«Все диаграммы»** выбрать диаграмму **«Солнечные лучи»**. ❑

Какой инструмент визуализации выбрать, зависит от конкретной ситуации, конкретной задачи, конкретного набора данных, и конечно, от личных предпочтений исследователя. Например, круговые диаграммы для визуализации распределения целого на части кому-то кажутся превосходным инструментом визуализации, кому-то — не очень информативным. Как правило, имеет смысл попробовать несколько вариантов визуализации и выбрать из них наиболее наглядный.

Задачи

30. В файле **«Абитуриенты магистратуры.xlsx»** содержатся данные о количестве поданных в 2017 г. заявлений на магистерские программы четырех факультетов Финансового университета. Визуализируйте эти данные с помощью диаграмм «Дерево» и «Солнечные лучи».

31. В файле **«Объем кредитов.xlsx»** содержатся данные о суммарном объеме кредитов, выданных банком по двум типам кредитных карт и по трем типов некарточных кредитов. Визуализируйте эти данные с помощью диаграмм «Дерево» и «Солнечные лучи».

§ 1.4. Предварительная обработка данных

1.4.1. Выбросы

Значительная часть работы по анализу данных связана с их предварительной обработкой. Данные, с которыми мы имеем дело на практике, не идеальны, какие-то значения могут быть зафиксированы с ошибками, какие-то значения могут быть пропущены, некоторые строки могут повторяться, в данных могут быть выбросы, сильно осложняющие их анализ и т. д.

Предварительная обработка данных в реальных задачах, как правило, занимает столько же времени, сколько их анализ с помощью различных интеллектуальных технологий, а часто даже больше.

Одной из самых типичных и самых важных проблем, возникающих при анализе реальных наборов данных, является проблема выбросов, которые искажают характеристики признаков и их взаимосвязей.

Напомним, что выбросами называются значения признака, не попадающие в отрезок $[x_{0,25} - 1{,}5IQR;\ x_{0,75} + 1{,}5IQR]$.

Первым шагом при поиске выбросов является визуализация данных с помощью диаграмм размаха и диаграмм рассеяния.

При анализе выбросов необходимо серьезно задумываться над каждым значением, являющимся «кандидатом» в выбросы — действительно ли оно является выбросом, либо в данных есть важные специальные подмножества, которые нужно рассматривать отдельно.

Часто выбросы связаны с предсказуемыми ошибками людей, участвовавших в подготовке наборов данных. Например, человек мог ввести данные не в тот столбец или при вводе числа типа 25,00 пропустить десятичную запятую, так что в результате в набор данных о признаке, принимающем типичные значения порядка нескольких десятков, попадает число 2500.

В некоторых случаях выбросы следует отбросить, чтобы избежать искажения результатов, в других ситуациях выбросы являются важнейшим предметом исследования. Например, в задаче обнаружения мошеннических транзакций по банковским картам именно выбросы — необычные, нетипичные транзакции — представляют основной интерес. Если отбросить выбросы или заменить их типичными транзакциями, то, очевидно, что задача определения мошеннических транзакций решена не будет.

В общем случае возможны следующие **варианты работы с выбросами:**

- *замена выброса соответствующей границей отрезка*

$$[x_{0,25} - 1{,}5IQR;\ x_{0,75} + 1{,}5IQR];$$

- *обработка выброса как пропущенного значения.*

Замечание. Как правило, для каждого признака X целесообразно добавить в набор данных специальный признак $X_{\text{выбр.}}$, значение которого равно единице, в случае, если значение признака X в данной строке исходного набора данных было классифицировано как выброс.

Пример 1.4.1. В файле **«Кофеварки.xlsx»** представлены данные о продажах кофеварок интернет-магазином в июле 2017 г. (рис. 1.4.1, *а*). Магазин продает кофеварки двух видов: красные и серебристые. Известно, сколько в каждый день было продано кофеварок каждого вида, на какой платформе («ВК» или *Facebook*) проводилась в соответствующий день рекламная кампания, каковы были расходы на рекламу, какая в этот день устанавливалась цена на кофеварки и каким был курс доллара.

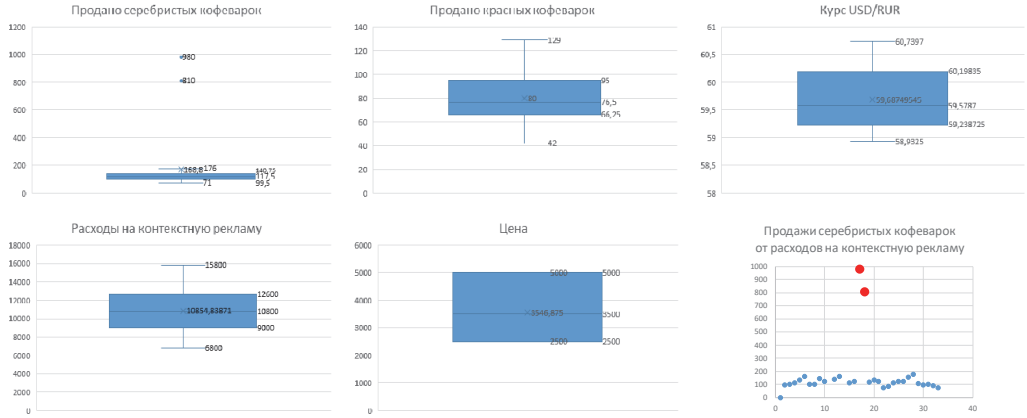
Построим в Microsoft Excel диаграммы размаха для каждого из числовых признаков, а также диаграммы рассеяния для пар признаков (рис. 1.4.1, *б*).

Видно, что выбросы присутствуют только в данных о продажах серебристых кофеварок: каждый день продавалось порядка сотни кофеварок, однако 15.07.2017 продано 980 кофеварок, а 16.07.2017 продано 810 кофеварок. Если в эти дни не проводилось никаких специальных акций и распродаж (а такие акции, судя по значению цены в эти дни, не проводились — наоборот, 15.07.2017 произошло повышение цены с 2500 руб. до 5000 руб. за кофеварку), то логично предположить, что в данном случае причиной выброса является человеческая ошибка: скорее всего, в эти дни к объемам продаж серебристых кофеварок

ошибочно были приписаны лишние нули. Поэтому целесообразно просто удалить лишние нули и заменить значения объемов продаж в эти дни на 98 и 81. ❑

	A	B	C	D	E	F	G
1	Дата	Рекламная платформа	Серебр. кофеварки	Красн. кофеварки	Курс RUR/USD	Расходы на рекламу	Цена
2	01.07.2017	VK	97	67	59,3862	9 000	2 500
3	02.07.2017	VK	98	67		9 000	2 500
4	03.07.2017	VK	110	77		10 400	2 500
5	04.07.2017	Facebook	134	99	58,9695	9 800	2 500
6	05.07.2017	Facebook	159	118	59,2295	13 500	2 500
7	06.07.2017	Facebook	103	69	59,5787	9 000	2 500
8	06.07.2017	Facebook	103	69	59,5787	9 000	2 500
9	07.07.2017	Facebook	143	101	60,2426	13 500	2 500
10		Facebook	123	86	60,3792	11 300	2 500
11	09.07.2017	Facebook		95		12 600	2 500
12	10.07.2017	Facebook	140	98		13 100	2 500
13	11.07.2017	Facebook	162	120	60,3014	13 500	2 500
14	12.07.2017	Facebook		95	60,7397	9 900	2 500
15	13.07.2017	Facebook	109	75	60,6227	9 900	2 500
16	14.07.2017	Facebook	122	85	60,1836	11 300	2 500
17	15.07.2017	Facebook	980	62	59,8806	10 800	5 000
18	16.07.2017	Facebook	810	50		9 000	5 000
19	17.07.2017	Facebook	115	76		12 600	5 000
20	18.07.2017	VK	131	92	59,0657	12 200	5 000
21	19.07.2017	VK	122	85	59,3705	11 300	5 000
22	20.07.2017	VK	71	42	59,2418		5 000
23	21.07.2017	VK	83	50	59,0823	9 000	5 000
24	22.07.2017	VK	112	75	58,9325	10 800	5 000
25	23.07.2017	VK	120	82		11 700	5 000
26	24.07.2017	VK	121	82		11 700	5 000
27	25.07.2017	VK	156	113	59,6572	13 500	5 000
28	26.07.2017	VK	176	129	59,8185	15 800	3 500
29	27.07.2017	VK	104	68	59,9102	9 900	3 500
30	28.07.2017	VK	96	63	59,4102	9 000	3 500
31	29.07.2017	VK	100	66	59,5436	9 500	3 500
32	30.07.2017	Facebook	88	57		8 100	3 500
33	31.07.2017	Facebook	76	47		6 800	3 500

а)



б)

Рис. 1.4.1. *Поиск выбросов в данных о продажах кофеварок: исходные данные (а); диаграммы размаха и диаграммы рассеяния (б)*

Задачи

32. В файле «Кредиты.xlsx» из примера 1.1.1 (табл. 1.1.1) представлены данные о выданных некоторым банком кредитах. Проведите анализ выбросов в этих данных.

33. В примере 1.2.13 были определены выбросы в данных о котировках акций российских компаний. Предложите варианты обработки этих выбросов.

1.4.2. Пропущенные значения

Важной проблемой анализа данных является отсутствие в наборах данных некоторых значений. Они могут быть неизвестны по какой-либо причине, могут быть утрачены, могут появиться в результате обработки выбросов.

При этом многие методы анализа данных невозможно использовать, если таблица с данными содержит пустые ячейки.

Пример 1.4.2. Рассмотрим табл. 1.4.1, в которой представлены данные о количестве кредитных и депозитных договоров, заключенных банком в августе 2017 г.

Таблица 1.4.1

Пример набора данных с пропущенными значениями и повторяющимися строками

Дата	*Число заключенных кредитных договоров*	*Число заключенных депозитных договоров*
01.08.2017	126	75
02.08.2017	123	0
02.08.2017	123	80
	136	84
04.08.2017		75
05.08.2017	128	73
05.08.2017	128	73

Видим, что некоторые значения в таблице пропущены, т. е. неизвестны.

Иногда можно догадаться, какое значение должно быть в соответствующей ячейке. В данном примере столбец «Дата» содержит даты августа 2017 г, начиная с 01.08.2017. При этом перед пропущенным значением идет 02.08.2017, а послпе него — 04.08.2017. Скорее всего, предположение о том, что пропущено значение 03.08.2017. Таким образом, мы провели интерполяцию пропущенного значения — заменили пропущенное значение средним из двух ближайших строк.

В рассматриваемом наборе данных отсутстует число заключенных кредитных договоров за 04.08.2017. Догадаться, сколько кредитов было выдано в этот день, невозможно. В таком случае можно, например, заполнить пропущенное значение средним, медианой или модой. В данном случае среднее равно 127,3333, медиана равна 127, а моды — 123 и 128. Поэтому целесообразнее всего поставить в эту ячейку число 127.

В реальных наборах данных пропущенным значениям не всегда соответствуют пустые ячейки. В некоторых информационных системах вместо пропущенных чисел вставляются специальные коды. Например, в Microsoft Excel используются следующие коды ошибок: «#ДЕЛ/0!», «#Н/Д», «#ИМЯ?», «#ПУСТО!», «#ЧИСЛО!», «#ССЫЛКА!», «#ЗНАЧ!». В других системах могут использоваться коды «NA» (*Not Available*), «NaN» (*Not a Number*) и др. Некоторые системы для обозначения ошибок и отсутствующих значений числовых признаков вставляют нули или специальные числовые коды, например, «9999» или «99999999».

Так, в первой из двух строк, соответствующих дате 02.08.2017, значение числа заключенных депозитных договоров равно нулю. В такой ситуации достаточно сложно понять, откуда взялся этот нуль. Теоретически возможно, что ни один депозит в определенный день не был открыт, но если посмотреть на значения этого признака в другие даты, мы видим, что типичные значения

этого признака находятся в границах [73; 84], и нулевое значение очень нетипично. Поэтому может быть целесообразным считать, что нуль соответствует пропущенному значению, и заменить его, например, медианой или средним. В данном же примере возможен и другой вариант: дате 02.08.2017 соответствуют две строки в наборе данных, и возможно, что на самом деле в этот день было заключено 80 депозитных договоров. Остановимся на этом варианте. ❑

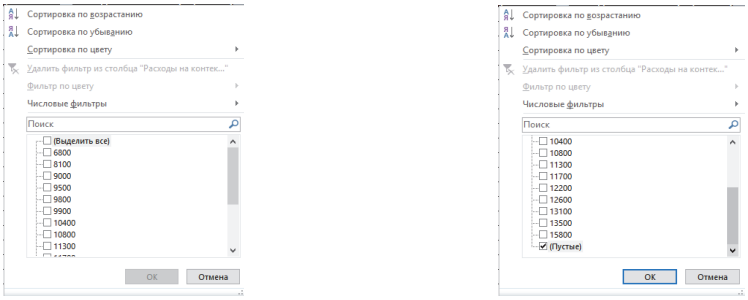
В общем случае возможны следующие **варианты работы с пропущенными значениями:**

- *замена специальным значением* (например, пропущенные значения признака «Число задержек платежей по кредиту» в наборе данных о заемщиках банка целесообразно заменить нулем, так как если банку неизвестно, сколько раз клиент задерживал платеж, то это, скорее всего, означает, что он всегда платил вовремя);
- *замена средним или медианой* (это два наиболее часто используемых метода, но они работают только для количественных признаков);
- *замена с помощью метода ближайших соседей* (когда вычисляется среднее, но не по всему столбцу, а по нескольким близким строкам — ближайшим соседям);
- *замена модой* (как правило, замена модой менее эффективна, чем замена медианой или средним, но у качественных признаков не бывает среднего и медианы, поэтому пропущенные значения таких признаков чаще всего заменяют модой);
- *удаление целой строки* (поскольку удаление строки приводит к потере информации не только о том признаке, в котором пропущено значение, но и о других признаках, этим методом пользуются редко, когда никакой другой способ не подходит);
- *замена с помощью специальных методов* (в последнее время, наряду с приведенными выше, популярны также методы заполнения пропущенных значений с помощью сингулярного разложения, линейной регрессии, случайного леса, метода *K*-средних, алгоритмов *EM*, *ZET* и др. — эти методы рассматриваются в дисциплине «Машинное обучение»).

Замечание. Как правило, при замене пропущенных значений для каждого признака X целесообразно добавить в набор данных специальный признак $X_{\text{пропущ}}$, значение которого равно единице, в случае, если значение признака X в данной строке исходного набора данных отсутствовало и было каким-либо образом синтезировано.

Пример 1.4.3. Продолжим рассмотрение примера 1.4.1. Видим, что в данных о продажах кофеварок присутствуют пропущенные значения (на рис. 1.4.1, *а* соответствующие ячейки выделены желтым).

Замечание. Рассматриваемый набор данных умещается на одном экране, поэтому пропущенные значения найти легко. Если данных много, то поиск пропущенных значений — это отдельная задача, которая в Microsoft Excel может быть решена с помощью средств фильтрации. Например, для того чтобы найти пропущенные значения в поле «Расходы на контекстную рекламу», нужно вначале отформатировать набор данных как таблицу, затем нажать на кнопку со стрелкой справа от поля «Расходы на контекстную рекламу», в появившемся окне напройки параметров сортировки и фильтрации снять выделение со строки «Выделить все» (рис. 1.4.2, *а*), прокрутить список значений вниз до конца и выделить строку «Пустые» (рис. 1.4.2, *б*). В результате будут показаны только строки, в которые значение поля «Расходы на контекстную рекламу» отсутствует.



а) *б*)

Рис. 1.4.2. *Поиск пропущенных значений признака «Расходы на контекстную рекламу»*

В столбце **«Дата»** пропущено значение между 07.07.2017 и 09.07.2017. Судя по остальным значениям данного признака, в данном случае пропущена дата 08.07.2017.

В столбце **«Курс RUR/USD»** после каждых пяти значений курса доллара идут два пропущенных значениях. Это совершенно естественно: Банк России устанавливает обменные курсы в рабочие дни, при этом в пятницу устанавливается курс на субботу, а в понедельник устанавливается курс на вторник. Поэтому пропущенные значения данного признака заполнить легко: в субботу 02.07.2017 и воскресенье 03.07.2017 нужно поставить значение признака **«Курс RUR/USD»** равным 59,3862 (как в пятницу 01.07.2017); 09.07.2017 и 10.07.2017 нужно поставить **«Курс RUR/USD»** равным 60,3792 (как 08.07.2017) и т. д.

Пропущенные значения в столбцах **«Серебр. кофеварки»** и **«Расходы на рекламу»** угадать не получится, поэтому их целесообразно заменить, например, медианами, значения которых — соответственно 111 и 10 800 — вычислены с помощью функции **МЕДИАНА**). Альтернативный вариант — вставить пропущенные значения с помощью метода ближайших соседей, например, как средние между значениями в предыдущий и последующий дни, т. е. $[(123 + 1430)/2] = [131{,}5] = 131$ для пропущенного числа проданных серебряных кофеварок и $(11\,300 + 9000)/2 = 10\,150$ для пропущенных расходов на рекламу. Возможны и другие решения.

Введем дополнительные признаки **«Выброс – серебр.»**, **«Пропущ – серебр.»**, **«Пропущ – реклама»**, установив их равными единице в тех строках, где обрабатывались соответствующие выбросы и пропущенные значения, и равными нулю в остальных строках. Результаты представлены на рис. 1.4.3 и сохранены в файле **«Кофеварки.xlsx»** на вкладке **«Обраб. выбросов и проп. знач.»**. ❑

Задачи

34. В файле **«Кредиты.xlsx»** из примера 1.1.1 (табл. 1.1.1) представлены данные о выданных некоторым банком кредитах. Найдите в данных ошибки и пропущенные значения.

35. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Кредитный рейтинг»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**). Насколько надежным следует считать заемщика, который никогда не брал кредитов?

	A	B	C	D	E	F	G	H	I	J
1	Дата	Рекламная платформа	Серебр. кофеварки	Красн. кофеварки	Курс RUR/USD	Расходы на рекламу	Цена	Выброс - серебр.	Пропущ. - серебр.	Пропущ. – реклама
2	01.07.2017	VK	97	67	59,3862	9 000	2 500	0	0	0
3	02.07.2017	VK	98	67	59,6832	9 000	2 500	0	0	0
4	03.07.2017	VK	110	77	59,6832	10 400	2 500	0	0	0
5	04.07.2017	Facebook	134	99	58,9695	9 800	2 500	0	0	0
6	05.07.2017	Facebook	159	118	59,2295	13 500	2 500	0	0	0
7	06.07.2017	Facebook	103	69	59,5787	9 000	2 500	0	0	0
8	06.07.2017	Facebook	103	69	59,5787	9 000	2 500	0	0	0
9	07.07.2017	Facebook	143	101	60,2426	13 500	2 500	0	0	0
10	08.07.2017	Facebook	123	86	60,3792	11 300	2 500	0	0	0
11	09.07.2017	Facebook	111	95	60,3792	12 600	2 500	0	1	0
12	10.07.2017	Facebook	140	98	60,3792	13 100	2 500	0	0	0
13	11.07.2017	Facebook	162	120	60,3014	13 500	2 500	0	0	0
14	12.07.2017	Facebook	111	95	60,7397	9 900	2 500	0	1	0
15	13.07.2017	Facebook	109	75	60,6227	9 900	2 500	0	0	0
16	14.07.2017	Facebook	122	85	60,1836	11 300	2 500	0	0	0
17	15.07.2017	Facebook	98	62	59,8806	10 800	5 000	1	0	0
18	16.07.2017	Facebook	81	50	59,8806	9 000	5 000	1	0	0
19	17.07.2017	Facebook	115	76	59,8806	12 600	5 000	0	0	0
20	18.07.2017	VK	131	92	59,0657	12 200	5 000	0	0	0
21	19.07.2017	VK	122	85	59,3705	11 300	5 000	0	0	0
22	20.07.2017	VK	71	42	59,2418	10 800	5 000	0	0	1
23	21.07.2017	VK	83	50	59,0823	9 000	5 000	0	0	0
24	22.07.2017	VK	112	75	58,9325	10 800	5 000	0	0	0
25	23.07.2017	VK	120	82	58,9325	11 700	5 000	0	0	0
26	24.07.2017	VK	121	82	58,9325	11 700	5 000	0	0	0
27	25.07.2017	VK	156	113	59,6572	13 500	5 000	0	0	0
28	26.07.2017	VK	176	129	59,8185	15 800	3 500	0	0	0
29	27.07.2017	VK	104	68	59,9102	9 900	3 500	0	0	0
30	28.07.2017	VK	96	63	59,4102	9 000	3 500	0	0	0
31	29.07.2017	VK	100	66	59,5436	9 500	3 500	0	0	0
32	30.07.2017	Facebook	88	57	59,5436	8 100	3 500	0	0	0
33	31.07.2017	Facebook	76	47	59,5436	6 800	3 500	0	0	0
32	**Медианы**		**111**			**10 800**				

Рис. 1.4.3. *Обработка выбросов, пропущенных значений и повторяющихся строк в данных о продажах кофеварок*

36. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Годовой доход»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**). Насколько надежным следует считать заемщика, который никогда не брал кредитов?

37. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Стаж на текущем месте работы»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**).

38. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Цель кредита»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**).

39. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Размер ежемесячного платежа по кредиту»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**).

40. Предложите и обоснуйте, чем заполнять пропущенные значения поля **«Срок с последнего нарушения условий кредитного договора»** в наборе данных о заемщиках (файл **«Кредиты.xlsx»**).

1.4.3. Повторяющиеся строки

Повторяющиеся в наборе данных строки могут исказить результаты анализа данных, поэтому перед проведением анализа данных из нескольких дублирующихся строк оставляют обычно одну.

Пример 1.4.4. Предположим, что на сайте отзывов путешественников о посещенных местах некоторому ресторану поставили оценки пять человек: первые четверо постявили по «пять звездочек», а пятый — «ни одной звездочки». В такой ситуации средняя оценка, которая должна отображаться на сайте, очевидно, равна $(5+5+5+5+0)/5=20/5=4$.

Предположим теперь, что пятый посетитель, которому не понравилось в ресторане, проголосовал на сайте 196 раз, каждый раз выставляя «ни одной звездочки». В этом случае средняя оценка будет равна $20/200=0{,}1$. Какую оценку отобразить на сайте — вопрос, на который нет однозначного ответа. Как правило, если один и тот же человек сделал оценки в разные дни, сопроводив их конкретными отзывами о качестве работы ресторана, то рекомендательный сайт будет их учитывать, если же человек проголосовал 196 раз в течение 10 минут, то дублирующиеся оценки будут исключены. ❑

Пример 1.4.5. В наборе данных о количестве заключенных кредитных и депозитных договоров, который рассматривался в примере 1.4.2, мы заметили, что дате 02.08.2017 соответствуют две строки, тогда как всем остальным датам соответствует по одной строке. При этом и в первой, и во второй строке, соответствующей дате 02.08.2017, зафиксировано значение количества заключенных кредитных договоров, равное 123, а число заключенных депозитных договоров в первой строке равно нулю, а во второй строке равно 80. В этом случае при рассмотрении примера 1.4.2 мы предположили, что вторая из строк является более полной, и заменили в первой из этих строк нуль, который мы посчитали признаком пропущенного значения, на 80. В итоге у нас появилось две совершенно идентичных строки, соответствующих дате 02.08.2017.

Логично предположить, что на самом деле одна из строк в наборе данных является дубликатом другой, и ее следует удалить. ❑

В примере 1.4.4 однозначно понятно, что строка данных об оценке ресторана однозначно идентифицируется именем пользователя сайта или *IP*-адресом, с которого был произведен вход на сайт, или *IP*-адресом и временем, в которое была сделана оценка.

В примере 1.4.5 таким *ключевым столбцом*, по которому можно идентифицировать уникальные и повторяющиеся строки, является дата. Но что делать, если две строки содержат одно и тоже значение ключевого столбца, но разные значения в других столбцах?

Бывает так, что набор данных не содержит ключевых столбцов, которые позволяют однозначно идентифицировать каждую строку. Если окажется, что в таком наборе данных есть несколько таких строк, что значения каждого из признаков в этих строках совпадают, то логично предположить, что эти строки являются дубликатами. Однако в такой ситуации следует быть осторожными: может оказаться и так, что эти строки на самом деле соответствуют разным объектам.

В общем случае возможны такие **варианты обработки повторяющихся строк:**

- сохранение первой строки с удалением оставшихся;
- сохранение последней строки с удалением оставшихся;
- объединение повторяющихся строк в одну с вычислением средних значений для каждого количественного признака и мод для каждого качественного признака.

В любом случае и при анализе выбросов, и при обработке пропущенных значений, и при удалении повторяющихся строк очень важно разбираться в предметной области, соответствующей исследуемому набору данных, и принимать решения об очистке и трансформации данных с учетом специфики предметной области, содержательного смысла конкретной ситуации и задачи.

Пример 1.4.6. На рис 1.4.3 зеленым цветом выделены повторяющиеся строки в данных о продажах кофеварок из примеров 1.4.1, 1.4.3. Одну из этих двух строк следует удалить (соответствующие результаты сохранены в файле «Кофеварки.xlsx» на рабочем листе «Удаленные дубликаты». ❑

Замечание. Для удаления повторяющихся строк в Microsoft Excel нужно выделить диапазон или таблицу, содержащую данные, нажать кнопку «Удалить дубликаты» на вкладке «Данные» и выбрать, совпадение каких признаков определяет, что данные строки являются повторяющимися.

Задачи

41. В условиях примера 1.4.2 предложите и обоснуйте варианты работы с посторяющимися строками в случае, если бы при анализе пропущенных значений в наборе данных о количестве заключенных договоров мы заменили в первой из строк, соответствующих дате 02.08.2017, нулевое значение признака «Количество заключенных депозитных договоров» не на 80 (содержащееся во второй из строк, соответствующих той же дате), а на 111 (медиану)?

42. В файле «Кредиты.xlsx» из примера 1.1.1 (табл. 1.1.1) представлены данные о выданных банком кредитах. Идентифицируйте и удалите дублирующиеся строки в этом наборе данных.

1.4.4. Синтетические признаки

Одним из важнейших этапов решения задач анализа данных является поиск *синтетических признаков*, являющихся определенными функциями от исходных признаков: удачно построенные синтетические признаки очень помогают находить в данных закономерности, определять общие свойства объектов и различия между ними.

Пример 1.4.7. Если в задаче анализа данных о заемщиках известен размер ежемесячного платежа по кредиту и средний размер месячного дохода, то, как правило, полезно рассмотреть дополнительный признак — долю ежемесячного платежа по кредиту в месячном доходе. ❑

Пример 1.4.8. При анализе данных о продажах кофеварок из примеров 1.4.1, 1.4.3, 1.4.6 полезно ввести в качестве дополнительных признаков общее количество кофеварок (=[@[Серебристые кофеварки]] + [@[Красные кофеварки]]), а также выручку (=[@[Кофеварки – всего]]*[@Цена]).

Если дополнительно известна полная себестоимость одной кофеварки для магазина (например, 2300 руб.), то, конечно, нужно добавить признак «Прибыль» (=[@[Кофеварки – всего]]*([@Цена] – 2300) – [@[Расходы на рекламу]]).

Результаты расчетов представлены на рис. 1.4.4. Построим на рис. 1.4.5 графики изменения числа проданных серебристых кофеварок, числа проданных красных кофеварок, общего числа проданных кофеварок и прибыли, а также диаграммы рассеяния числа кофеварок от расходов на рекламу.

Видим, что общее число проданных в день кофеварок, как и число проданных кофеварок каждого цвета каждый день меняется, при этом тенденции к росту или снижению числа проданных кофеварок не наблюдаются. Что касается прибыли, то по соответствующему графику мы видим ее резкий рост 15.07.2017 (когда цена была увеличена с 2500 до 5000 руб.), а затем резкое падение 27.07.2017 (при уменьшении цены до 3500 руб.).

Диаграммы рассеяния при этом показывают, что рост расходов на рекламу сопровождается ростом числа проданных кофеварок, увеличение цены не всегда ведет к уменьшению числа проданных кофеварок, рост расходов на рекламу не обязательно приводит к росту прибыли, а увеличение цены, как правило, сопровождается увеличением прибыли. ❑

Во многих ситуациях синтез новых признаков связан не столько с содержательным смыслом задачи, сколько с необходимостью использования статистических методов, основанных на применении следствий из нормального закона распределения данных, в той ситуации, когда законы распределения исходных признаков отличаются от нормального.

Во всех случаях бывает обычно полезно рассмотреть синтетические признаки, являющиеся квадратами, квадратными корнями, логарифмами исходных признаков, произведениями и частными пар признаков и т. д.

Пример 1.4.9. В примере 1.1.3 при анализе данных об объемах торгов по акциям мы заметили, что гистограммы объемов торгов по всем трем акциям (в особенности по акциям ВТБ) скошены вправо, т. е. распределения объемов торгов заметно отличаются от нормального распределения.

В том же примере мы ввели три дополнительных признака, равных натуральным логарифмам объемов торгов акциями ВТБ, Газпрома и Сбербанка, и увидели, что гистограммы для новых признаков стали более симметричными (рис. 1.1.3). ❑

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Дата	Рекламная платформа	Серебр. кофеварки	Красн. кофеварки	Курс RUR/USD	Расходы на рекламу	Цена	Выброс - серебр.	Пропущ. - серебр.	Пропущ. – реклама	Кофеварки – всего	Выручка	Прибыль
2	01.07.2017	VK	97	67	59,3862	9 000	2 500	0	0	0	164	410 000	23 800
3	02.07.2017	VK	98	67	59,6832	9 000	2 500	0	0	0	165	412 500	7 500
4	03.07.2017	VK	110	77	59,6832	10 400	2 500	0	0	0	187	467 500	8 300
5	04.07.2017	Facebook	134	99	58,9695	9 800	2 500	0	0	0	233	582 500	13 500
6	05.07.2017	Facebook	159	118	59,2295	13 500	2 500	0	0	0	277	692 500	14 200
7	06.07.2017	Facebook	103	69	59,5787	9 000	2 500	0	0	0	172	430 000	8 200
8	06.07.2017	Facebook	103	69	59,5787	9 000	2 500	0	0	0	244	610 000	10 900
9	07.07.2017	Facebook	143	101	60,2426	13 500	2 500	0	0	0	209	522 500	9 600
10	08.07.2017	Facebook	123	86	60,3792	11 300	2 500	0	0	0	206	515 000	8 000
11	09.07.2017	Facebook	111	95	60,3792	12 600	2 500	0	1	0	238	595 000	10 700
12	10.07.2017	Facebook	140	98	60,3792	13 100	2 500	0	0	0	282	705 000	14 700
13	11.07.2017	Facebook	162	120	60,3014	13 500	2 500	0	0	0	206	515 000	10 700
14	12.07.2017	Facebook	111	95	60,7397	9 900	2 500	0	1	0	184	460 000	8 500
15	13.07.2017	Facebook	109	75	60,6227	9 900	2 500	0	0	0	207	517 500	9 400
16	14.07.2017	Facebook	122	85	60,1836	11 300	2 500	0	0	0	160	800 000	405 200
17	15.07.2017	Facebook		62	59,8806	10 800	5 000	1	0	0	131	655 000	331 600
18	16.07.2017	Facebook		50	59,8806	9 000	5 000	1	0	0	191	955 000	484 000
19	17.07.2017	Facebook	115	76	59,8806	12 600	5 000	0	0	0	223	1 115 000	567 600
20	18.07.2017	VK	131	92	59,0657	12 200	5 000	0	0	0	207	1 035 000	526 900
21	19.07.2017	VK	122	85	59,3705	11 300	5 000	0	0	0	113	565 000	283 000
22	20.07.2017	VK	71	42	59,2418	10 800	5 000	0	0	1	133	665 000	336 800
23	21.07.2017	VK	83	50	59,0823	9 000	5 000	0	0	0	187	935 000	475 400
24	22.07.2017	VK	112	75	58,9325	10 800	5 000	0	0	0	202	1 010 000	513 500
25	23.07.2017	VK	120	82	58,9325	11 700	5 000	0	0	0	203	1 015 000	516 100
26	24.07.2017	VK	121	82	58,9325	11 700	5 000	0	0	0	269	1 345 000	685 900
27	25.07.2017	VK	156	113	59,6572	13 500	5 000	0	0	0	305	1 067 500	319 700
28	26.07.2017	VK	176	129	59,8185	15 800	3 500	0	0	0	172	602 000	179 300
29	27.07.2017	VK	104	68	59,9102	9 900	3 500	0	0	0	159	556 500	165 900
30	28.07.2017	VK	96	63	59,4102	9 000	3 500	0	0	0	166	581 000	173 100
31	29.07.2017	VK	100	66	59,5436	9 500	3 500	0	0	0	145	507 500	151 400
32	30.07.2017	Facebook	88	57	59,5436	8 100	3 500	0	0	0	123	430 500	128 500
33	31.07.2017	Facebook	76	47	59,5436	6 800	3 500	0	0	0	164	410 000	23 800
32	**Медианы**		**111**			**10 800**							

Рис. 1.4.4. *Синтетические признаки в данных о продажах кофеварок*



Рис. 1.4.5. *Графики и гистограммы синтетических признаков в наборе данных о продажах кофеварок*

Может быть так, что все признаки в исследуемом наборе данных имеют одинаковую природу и измеряются в одинаковых единицах. Например, можно получить данные опроса о том, сколько времени в месяц люди тратят на работу, домашние дела и развлечения (все три признака имеют одинаковый масштаб и измеряются в одинаковых единицах времени).

Или можно на основании анализа транзакций по пластиковым картам получить набор данных, содержащий структуру затрат держателей пластиковых карт — сколько люди тратят на еду, на путешествия, на развлечения и т. д., здесь все признаки будут иметь сходный масштаб и измеряться в тыс. руб.

Но так бывает редко. Большинство реальных наборов данных, с которыми приходится иметь дело в экономике, управлении, финансах, состоят из нескольких признаков, имеющих принципиально разную природу и измеряющихся в разных единицах.

При анализе данных при этом часто бывают важны относительные взаимосвязи между признаками, причем не обязательно между их абсолютными значениями. Кроме того, многие методы анализа данных работают только в том случае, когда участвующие в анализе признаки имеют одинаковый масштаб.

Поэтому бывает нужно приводить различные признаки к единому масштабу. Чтобы это сделать, наиболее часто используются два способа преобразования признаков:

- переход от признака X к признаку

$$\frac{X - \bar{x}}{s_X}$$

  приводит к тому, что у нового признака будет нулевое среднее и стандартное отклонение, равное единице (такая манипуляция называется *z-преобразованием*);

- преобразование признака X в признак

$$\frac{X - x_{(\min)}}{x_{(\max)} - x_{(\min)}}$$

  гарантирует, что новый признак будет принимать значения из отрезка [0; 1] (эта трансформация называется *min-max-преобразованием*).

Пример 1.4.10. Введем еще несколько синтетических признаков в наборе данных о котировках акций российских компаний из примеров 1.1.3, 1.2.10, 1.2.13, 1.4.10.

Поскольку акции номинированы в разных масштабах, целесообразно рассмотреть объемы торгов не в натуральном выражении (в количестве проданных акций), а в денежном выражении. Объем торгов в денежном выражении — это объем торгов в натуральном выражении, умноженный на цену акции.

Добавим в таблицу столбец **«ВТБ – объем в д. в.»** и введем в этот столбец формулу **«=[@[ВТБ – объем]]*[@[ВТБ – цена]]»**. Аналогичным образом добавим столбцы, соответствующие объемам торгов в денежном выражении для двух других акций.

На соответствующем графике (рис. 1.4.6, *а*) видно, что объемы торгов по акциям Сбербанка и Газпрома демонстрируют схожие тенденции (растут и снижаются практически одновременно). При этом в денежном выражении объемы торгов акциями Сбербанка в среднем примерно в два раза больше, чем объемы торгов акциями Газпрома, а объемы торгов акциями ВТБ в среднем примерно в четыре раза меньше, чем объемы торгов акциями Газпрома.



а)



б)



в)

Рис. 1.4.6. *Графики объемов торгов в денежном выражении* (*а*), *z-преобразования цен* (*б*) *и min-max-преобразования цен* (*в*) *в наборе данных о котировках акций*

Чтобы подвергнуть цену акций ВТБ z-преобразованию и min-max-преобразованию, добавим новые столбцы «Z_ВТБ – цена» и «MinMax_ВТБ - цена». Введем в столбец «Z_ВТБ – цена» формулу

«=([@[ВТБ – цена]] – СРЗНАЧ([ВТБ – цена]))/СТАНДОТКЛОН.В([ВТБ – цена])»,

а в столбец «MinMax_ВТБ - цена» — формулу

«=([@[ВТБ – цена]] – МИН([ВТБ – цена]))/(МАКС([ВТБ – цена]) – МИН([ВТБ – цена]))».

В этих формулах наличие знака «@» означает обращение к конкретному значению признака (к конкретной ячейке), а отсутствие этого знака — к целому столбцу.

Аналогичным образом добавим столбцы для z-преобразования и min-max-преобразования цен двух других акций.

Графики z-преобразования и min-max-преобразования цен акций построены соответственно на рис. 1.4.6, *б*, *в*. Видно, что цены, подвергнутые z-преобразованию, меняются в границах от -3 до 3, а результаты min-max-преобразования цен изменяются в границах от 0 до 1.

По каждому из этих двух графиков (рис. 1.4.6, *б*, *в*) видно, что в общем цены меняются несогласованно. Минимальные и максимальные значения цен различных акций соответствуют разным датам.

При этом примерно до июня 2014 г. Цены акций ВТБ и Сбербанка ведут себя похожим образом, увеличиваясь и уменьшаясь приблизительно в одни и те же моменты времени. После июня 2014 г. динамика цен всех трех акций различается довольно существенно.

В примере 1.1.5 мы умножили цену акций ВТБ на 1000, чтобы получить привести цены к единому масштабу (рис. 1.1.9). В общем случае приведение признаков к единому масштабу гораздо удобнее проводить с помощью z-преобразования или min-max-преобразования.

Рассчитанные значения синтетических признаков сохранены на рабочем листе «Данные», а построенные графики — на рабочем листе «Преобразования признаков» в файле «Акции.xlsx». ❑

Задачи

43. Рассматривается задача прогнозирования цены земельных участков, и в наборе данных для каждого из большого количества домов известна цена, длина и ширина. Предложите дополнительный синтетический признак, который может в большей степени влиять на цену, чем длина и ширина по отдельности.

44. В файле «Кредиты.xlsx» из примера 1.1.1 (табл. 1.1.1) представлены данные о выданных некоторым банком кредитах. Предложите синтетические признаки, позволяющие лучше понять структуру данных о заемщиках, и вычислите их значения.

45. Докажите, что каковы бы ни были значения признака X, минимальное значение признака

$$\frac{X - x_{(\min)}}{x_{(\max)} - x_{(\min)}}$$

будет равно нулю, а максимальное — единице.

Глава 2. Случайные события

§ 2.1. Основы комбинаторики

2.1.1. Правила суммы и произведения

В анализе данных часто приходится иметь дело с задачами, в которых необходимо подсчитывать число возможных способов совершения каких-либо действий. Задачи такого типа называются комбинато́рными, а раздел математики, занимающийся решением таких задач, — комбинато́рикой. Вообще, *комбинаторика* — это раздел математики, изучающий расположения объектов в соответствии со специальными правилами и методы подсчета числа всех возможных способов, которыми эти расположения могут быть совершены.

Первые теоретические исследования проблем комбинаторики были проведены в XVII в. Б. Паскалем, П. Ферма и Г. Лейбницем, в XVIII в. Я. Бернулли и Л. Эйлером. Именно тогда сложилась принятая и сегодня в комбинаторике терминология размещений, перестановок и сочетаний. К началу XX в. комбинаторика считалась законченным разделом математики, лежащим вне русла ее основного развития. В XX в. комбинаторику стали рассматривать как раздел теории множеств, изучающий вопросы, возникающие при изучении конечных множеств, что привело к более естественной и последовательной классификации основных задач комбинаторики и комбинаторных понятий. В связи с развитием вычислительной техники, теории информации и кодирования в настоящее время комбинаторика является одним из наиболее интенсивно развивающихся разделов математики.

Комбинаторные методы очень важны при вычислении вероятностей событий, связанных с экспериментами, имеющими конечное число исходов.

Сформулируем два универсальных правила, применяемых при решении комбинаторных задач.

Правило суммы. *Пусть требуется выполнить одно из каких-либо m действий, взаимно исключающих друг друга. Если первое действие можно выполнить n_1 способами, второе действие — n_2 способами и так до m-го действия, которое можно выполнить n_m способами, то выполнить одно из этих m действий можно $n_1 + n_2 + \cdots + n_m$ способами.*

Правило произведения. *Пусть требуется выполнить какие-либо m действий. Если первое действие можно выполнить n_1 способами, второе действие — n_2 способами и так до m-го действия, которое можно выполнить n_m способами, то все m действий могут быть выполнены $n_1 n_2 \cdots n_m$ способами.*

Доказательство этих правил очевидно. Для случая одного и двух действий их справедливость не вызывает сомнений, а на случай произвольного количества действий правила обобщаются при помощи метода математической индукции. Оставляем строгое доказательство читателю в задаче **48**.

Пример 2.1.1. В информационно-технологическом управлении банка работают три аналитика, десять программистов и 20 инженеров. Для сверхурочной работы в праздничный день начальник управления должен выделить одного сотрудника. Определим, сколько способов существует у начальника управления.

Начальник управления может отобрать одного аналитика $n_1 = 3$ способами, одного программиста — $n_2 = 10$ способами, а одного инженера — $n_3 = 20$ способами. Поскольку по условию задачи начальник управления может выделить любого из своих сотрудников, то согласно правилу суммы у него существует $n_1 + n_2 + n_3 = 3 + 10 + 20 = 33$ различных способа выбрать сотрудника для сверхурочной работы. ❑

Пример 2.1.2. Маша поссорилась с Петей и не хочет ехать с ним в одном автобусе. От общежития до университета с 7 до 8 ч отправляется пять автобусов. Не успевший на последний из этих автобусов опаздывает на лекцию. Рассчитаем, сколькими способами Маша и Петя могут доехать до университета в разных автобусах и не опоздать на лекцию.

Петя может доехать до университета $n_1 = 5$ различными способами (на одном из пяти автобусов), при этом Маше остается только $n_2 = 4$ способа (так как один из автобусов занят Петей). Таким образом, по правилу произведения у Пети и Маши есть $n_1 n_2 = 5 \cdot 4 = 20$ различных способов добраться до университета в разных автобусах и не опоздать на лекцию. ❑

Задачи

46. Из города А в город Б ведет m различных путей, а из города Б в город В — n различных путей. Определите, сколькими различными способами можно доехать из города А в город В через город Б.

47. В розыгрыше первенства университета по футболу принимают участие 16 команд. Определите, сколькими способами могут быть распределены золотая, серебряная и бронзовая медали.

48. Докажите с помощью метода математической индукции правило суммы и правило произведения.

2.1.2. Комбинации без повторений

Напомним понятие факториала, активно используемое в комбинаторике. *Факториалом натурального числа n* называется число

$$n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1. \qquad (2.1.1)$$

Факториалом нуля по определению является единица:

$$0! = 1. \qquad (2.1.2)$$

Рассмотрим некоторое множество S, состоящее из n различных элементов. Пусть $1 \leqslant k \leqslant n$. Назовем множество, состоящее из k элементов, *упорядоченным*, если каждому элементу этого множества поставлено в соответствие число от 1 до k, причем различным элементам множества соответствуют разные числа.

Размещениями из n элементов по k называются упорядоченные подмножества множества S, состоящие из k различных элементов и отличающиеся друг от друга составом элементов или порядком их расположения.

Число размещений из n элементов по k равно

$$A_n^k = \frac{n!}{(n-k)!} = n(n-1)(n-2)\cdots(n-k+1). \qquad (2.1.3)$$

Перестановками из n элементов называются размещения из n элементов по n, т. е. упорядоченные подмножества множества S, состоящие из всех элементов данного множества и отличающиеся друг от друга только порядком их расположения.

Число перестановок из n элементов равно

$$P_n = n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1. \qquad (2.1.4)$$

Сочетаниями из n элементов по k называются подмножества множества S, состоящие из k различных элементов и отличающиеся друг от друга только составом элементов.

Число сочетаний из n элементов по k равно

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 3\cdot 2\cdot 1}. \qquad (2.1.5)$$

При этом

$$C_n^k = C_n^{n-k} \quad (0 \leqslant k \leqslant n).$$

Доказательство. Действительно, $C_n^k = \frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)![n-(n-k)]!} = C_n^{n-k}$. ❑

В пакете Microsoft Excel существуют функции для вычисления факториалов, числа перестановок и числа сочетаний:

$$P_n = n! = \text{ФАКТР}(<n>);$$

$$A_n^k = \text{ПЕРЕСТ}(<n>; <k>);$$

$$C_n^k = \text{ЧИСЛКОМБ}(<n>; <k>).$$

Проиллюстрируем на примерах идеи доказательства формул (2.1.3)—(2.1.5).

Пример 2.1.3. Начальник службы безопасности банка должен ежедневно расставлять десятерых охранников по десяти постам. В целях усиления безопасности одна и та же комбинация расстановки охранников по постам не может повторяться чаще одного раза в месяц. Чтобы оценить, возможно ли это, найдем число различных комбинаций расстановки охранников.

Первый способ. На первый пост начальник службы безопасности может назначить любого из $n_1 = 10$ охранников, на второй пост — любого из оставшихся $n_2 = 9$ охранников и так до девятого поста, на который можно назначить любого из оставшихся $n_9 = 2$ охранников, при этом оставшийся $n_{10} = 1$ охранник будет назначен на 10-й пост. Поэтому согласно правилу произведения у начальника службы безопасности есть $n_1 n_2 \cdots n_{10} = 10\cdot 9\cdots 2\cdot 1 = 10! = 3\,628\,800$ способов расстановки охранников по постам. Поскольку количество дней в месяце не превышает 31, у начальника службы безопасности заведомо существует достаточное число способов расстановки своих подчиненных по постам.

Второй способ. Число способов расстановки десяти охранников по десяти постам, существующих у начальника службы безопасности, описывается числом перестановок из 10 элементов, т. е. $P_{10} = 10! = 3\,628\,800$.

Это число можно получить либо путем непосредственного перемножения чисел от единицы до десяти, либо воспользовавшись пакетом Microsoft Excel: результатом вычисления функции ФАКТР(10) будет число 3 628 800. ❑

Пример 2.1.4. Новый президент банка должен назначить двух новых вице-президентов из числа десяти директоров. Посмотрим, сколько способов существует у президента, если: *а*) один из вице-президентов (первый) выше другого по должности; *б*) вице-президенты по должности равны между собой.

Первый способ. *а*) Первого вице-президента можно выбрать из $n_1 = 10$ претендентов, при этом на пост второго вице-президента будут претендовать $n_2 = 9$ оставшихся директоров. Поэтому согласно правилу произведения у нового президента банка есть $n_1 n_2 = 10 \cdot 9 = 90$ способов назначения двух вице-президентов, один из которых подчиняется другому, из числа десяти директоров. *б*) Пусть первое действие заключается в том, что президент отбирает двух человек из 10 на должности вице-президентов, а второе действие — в том, что президент говорит отобранным людям, кто из них является первым вице-президентом, а кто — вторым. Пусть первое действие можно выполнить n_1 способами, второе действие, очевидно, можно выполнить $n_2 = 2$ способами, и по правилу произведения число способов назначения двух вице-президентов, один из которых подчиняется другому, из числа 10 директоров составляет $n_1 n_2 = 2n_1$. С другой стороны, в пункте а) мы нашли это число, и оно оказалось равным 90, поэтому $n_1 = 90/2 = 45$.

Второй способ. *а*) Число способов выбора двух кандидатов на две различные должности из 10 претендентов описывается числом размещений из 10 элементов по 2, т. е. $A_{10}^2 = 10!/8! = 90$. *б*) Число способов выбора двух кандидатов на две одинаковые должности из 10 претендентов описывается числом сочетаний из 10 элементов по 2, т. е. $C_{10}^2 = 10!/(2!8!) = 45$.

Конкретные числа в ответе можно получить либо непосредственной подстановкой в формулы, как это было сделано выше, либо воспользовавшись пакетом Microsoft Excel: результатом вычисления функции ПЕРЕСТ(10; 2) будет число 90, а результатом вычисления функции ЧИСЛКОМБ(10; 2) будет число 45. ❑

Пример 2.1.5. У банка $n = 9$ учредителей. Регистрационные документы хранятся в сейфе. Определим, сколько замков должен иметь сейф и сколько ключей к ним нужно изготовить, чтобы доступ к содержимому сейфа был возможен только тогда, когда соберется не менее $m = 6$ учредителей?

Какие бы $(m - 1)$ учредителей ни собрались, должен найтись замок, который они не смогут открыть (т. е. ни у одного из них не найдется ключа от этого замка), но при этом ключ от этого замка должен быть у каждого из оставшихся $(n - m + 1)$ учредителей (появление любого из них дает возможность открыть сейф). Поэтому минимальное число замков равно $C_n^{m-1} = C_n^{n-(m-1)}$, а ключей к ним должно быть $(n-m+1)C_n^{n-m+1}$. В условиях задачи $n = 9$, $m = 6$, поэтому число замков равно $C_n^{n-m+1} = C_9^{9-6+1} = C_9^4 = 9!/(4!5!) = 126$, а число ключей $(n-m+1)C_n^{n-m+1} = 4 \cdot 126 = 504$. ❑

Задачи

49. Сколькими способами можно разместить на шахматной доске восемь ладей так, чтобы они не били друг друга?

50. Петя и Маша коллекционируют сувениры из разных городов. У Пети есть 30 магнитов, 80 тарелок и 7 колокольчиков, у Маши — 20 магнитов, 5 тарелок и 90 колокольчиков. Сколькими способами Петя и Маша могут обменяться одновременно тремя магнитами, двумя тарелками и одним колокольчиком?

51. В кредитном отделе банка работают восемь человек. Определить, сколько существует способов распределить между ними три премии: *а*) одинакового размера; *б*) разных размеров, известных заранее?

52. В сессию в течение 20 дней студенты одной группы должны сдать пять экзаменов. Сколькими способами можно составить расписание экзаменов, если запрещается сдавать два экзамена в один день?

53. Сколько существует вариантов опроса группы из десяти студентов на одном занятии по анализу данных, если ни один из студентов не будет подвергнут опросу дважды, и на занятии может быть опрошено любое число студентов (в том числе, ни одного)?

2.1.3. Комбинации с повторениями

Размещениями с повторениями из n элементов по k называются упорядоченные подмножества множества S, состоящие из k элементов, среди которых могут оказаться одинаковые, и отличающиеся друг от друга составом элементов или порядком их расположения.

Число размещений с повторениями из n элементов по k равно

$$\tilde{A}_n^k = n^k. \tag{2.1.6}$$

Пример 2.1.6. В группе 20 студентов. Выясним, сколько существует у преподавателя способов вызвать к доске студентов для решения трех задач.

Для решения первой задачи можно вызвать любого из $n = 20$ студентов. Для решения второй задачи также можно вызвать любого из $n = 20$ студентов (в том числе и того, кто решал первую задачу). И для решения третьей задачи также можно вызвать любого из $n = 20$ студентов. Таким образом, по правилу произведения общее количество способов вызвать студентов для решения трех задач равно $n^3 = 20^3 = 8000$. ❑

Сочетаниями с повторениями из n элементов по k называются подмножества множества S, состоящие из k элементов, среди которых могут оказаться одинаковые, и отличающиеся друг от друга только составом элементов.

Число сочетаний с повторениями из n элементов по k равно

$$\tilde{C}_n^k = C_{n+k-1}^k = \frac{(n+k-1)!}{k!(n-1)!} = \frac{(n+k-1)(n+k-2)\cdots n}{k(k-1)(k-2)\cdots 3\cdot 2\cdot 1}. \tag{2.1.7}$$

Отметим, что формулы (2.1.3)—(2.1.7) сохраняют смысл и остаются справедливыми и при $k = 0$.

Пример 2.1.7. Маша очень любит пирожные и ежедневно в булочной рядом с университетом покупает пять пирожных (одинаковых или разных). Всего в булочной продается 11 сортов пирожных. Выясним, сколькими способами Маша может выбрать из них пять штук.

Каждому набору пирожных, которые выберет Маша, будем ставить в соответствие последовательность нулей и единиц, определяемую по следующему правилу. Напишем подряд столько единиц, сколько пирожных первого вида выбрала Маша, далее поставим нуль и после него запишем количество отобранных пирожных второго вида и т. д. Например, комбинации «одно пирожное второго вида, три пирожных пятого вида и одно пирожное 8-го вида» соответствует такая последовательность: «010001110001000» (нули отделяют виды пирожных друг от друга, поэтому нуль после 11-го вида не нужен). При этом каждому набору пирожных взаимно однозначным образом соответствует последовательность, построенная по описанному правилу. Все такие последовательности состоят, очевидно, из $11+5-1=15$ знаков, причем пять из них — единицы, которые могут находиться на любых местах. Поэтому количество способов вы-

бора пирожных равно количеству всех таких последовательностей, т. е. числу размещений 5 единиц по 15 местам: $C_{15}^{5} = C_{11+5-1}^{5} =$ ЧИСЛКОМБ(15; 5) = 3003. ❑

Если во множестве S, состоящем из n элементов, есть только m р а з л и ч н ы х элементов, то *перестановками с повторениями из n элементов* называются упорядоченные подмножества множества S, в которые первый элемент множества S входит n_1 раз, второй элемент — n_2 раз и так до m-го элемента, который входит n_m раз ($n_1 + n_2 + \cdots + n_m = n$).

Ч и с л о п е р е с т а н о в о к с п о в т о р е н и я м и из n элементов, в которые первый элемент множества S входит n_1 раз, второй элемент — n_2 раз и так до m-го элемента, который входит n_m раз ($n_1 + n_2 + \cdots + n_m = n$), равно

$$\tilde{P}_n(n_1, n_2, \ldots, n_m) = \frac{n!}{n_1! n_2! \cdots n_m!}. \qquad (2.1.8)$$

Пример 2.1.8. Определим, сколько различных слов можно составить, переставляя буквы в слове «мама».

Число различных слов, которые можно составить, переставляя буквы в слове «мама», описывается числом перестановок с повторениями из $n = 4$ элементов (букв в слове «мама»), в которые первый элемент (буква «м») входит $n_1 = 2$ раза, а второй элемент (буква «а») — $n_2 = 2$ раза ($n_1 + n_2 = 4 = n$). Это число равно $\tilde{P}_4(2, 2) = 4!/(2!2!) = 6$. Шесть различных «слов», получающиеся перестановками букв в слове «мама», таковы: «ммаа», «мама», «маам», «амма», «амам», «аамм». ❑

Задачи

54. Маша решила помириться с Петей и позвонить ему, но забыла две последних цифры его телефона и набирает их наудачу. Найдите наибольшее возможное число неудачных попыток, которые сделает Маша, прежде чем дозвонится до Пети.

55. Сколько автомобилей в одном регионе можно обеспечить государственными регистрационными знаками, если каждый регистрационный знак состоит из кода региона, трех букв, имеющих одинаковое начертание как в русском, так и в латинском алфавите («А», «В», «Е», «К», «М», «Н», «О», «Р», «С», «Т», «У», «Х»), и трех цифр?

56. В сессию в течение 20 дней студенты одной группы должны сдать пять экзаменов. Сколькими способами можно составить расписание экзаменов, если между двумя экзаменами должен пройти хотя бы один день для подготовки?

57. В конкурсе по трем номинациям участвуют десять кинофильмов. Вычислите число вариантов распределения призов, если призы в разных номинациях: *а*) различны; *б*) одинаковы.

58. Определите количество различных слов, которые можно составить, переставляя буквы в слове «математика».

§ 2.2. Определение вероятности

2.2.1. Исчисление событий

Случайное событие A, связанное с опытом S, — это такое событие, которое может произойти или не произойти в результате опыта S, причем заранее, до проведения опыта, неизвестно, произойдет оно или нет. В дальнейшем при рассмотрении случайных событий мы будем обычно опускать слово «случайное».

Случайные события мы будем обозначать заглавными латинскими буквами из начала алфавита, иногда с индексами, например, A, B, C, D_1, D_2 и т. п. С ка-

ждым опытом обязательно связаны два события — достоверное и невозможное, которые обозначаются особенно.

Достоверным событием, связанным с опытом S, называется такое событие Ω, которое обязательно произойдет в результате опыта S.

Невозможным событием, связанным с опытом S, называется такое событие $\varnothing$, которое обязательно не произойдет в результате опыта S.

Простейшим примером опыта является подбрасывание монеты. Монету будем считать симметричной, очень тонкой (теоретически бесконечно тонкой), однородной по плотности, на одной стороне монеты помещен герб («орел») на другой — надпись («решка»). Тогда у этого опыта есть два исхода: выпадение «орла» и выпадение «решки». Примерами событий, связанных с этим опытом являются следующие: A — выпадение «орла», B — выпадение «решки», Ω — выпадение «орла» или «решки» (достоверное событие), $\varnothing$ — невыпадение ни «орла», ни «решки» (невозможное событие).

Еще одним примером опыта является подбрасывание игральной кости. Игральная кость представляет собой геометрически правильный куб, однородный по плотности, на гранях которого отмечено число очков от одного до шести. Тогда у этого опыта есть шесть исходов: выпадение различного числа очков от одного до шести. В этом опыте можно рассмотреть больше событий, чем в опыте с монетой: A_1, A_2, ..., A_6 — выпадение одного очка, двух и так до шести; $A_{\text{чет}}$ — выпадение четного числа очков; $A_{\text{нечет}}$ — выпадение нечетного числа очков; $A_{<3}$ — выпадение числа очков, меньшего трех; Ω — выпадение числа очков, меньшего семи (достоверное событие); $\varnothing$ — выпадение числа очков, меньшего нуля (невозможное событие) и т. д.

Обратимся к интуитивному смыслу словосочетаний «более возможно», «одинаково возможно» и т. п. Из симметрии монеты следует, что события A и B одинаково возможны. События A_1, A_2, ..., A_6 в опыте с игральной костью также одинаково возможны (это следует из симметрии игральной кости). В то же время очевидно, что любое из событий $A_{\text{чет}}$, $A_{\text{нечет}}$ и $A_{<3}$ более возможно, чем любое событие из A_1, A_2, ..., A_6.

Вероятность наступления события, о которой мы будем говорить далее в этой главе, характеризует меру возможности наступления этого события при проведении опыта. Но сначала еще немного поговорим о самих событиях.

Над событиями A и B, связанными с одним и тем же опытом S, определены следующие **операции.**

Событие A *влечет за собой* событие B (или событие A *вложено* в событие B), если каждое появление события A сопровождается появлением события B. Это обозначается как $A \subseteq B$. События A и B называются *эквивалентными*, если $A \subseteq B$ и $B \subseteq A$. Эквивалентность обозначается так: $A = B$.

Объединением событий A и B называется событие $A \cup B$, которое наступает всегда, когда наступает либо событие A, либо событие B (либо одновременно и A, и B).

Пересечением событий A и B называется событие $A \cap B$, которое наступает всегда, когда события A и B наступают одновременно.

Дополнением события B до события A (или *разностью событий A и B*) называется событие $A \setminus B$, которое наступает всегда, когда наступает событие A и при этом не наступает событие B.

Противоположным событию A называется событие $\bar{A} = \Omega \setminus A$ (читается «не A»), которое наступает всегда, когда событие A не наступает.

События A и B называются *несовместными*, если $A \cap B = \varnothing$, т. е. если в результате опыта события A и B не могут наступить одновременно.

Говорят, что события $H_1, H_2, \ldots, H_n$ образуют *полную группу*, если они попарно несовместны ($H_i \cap H_j = \varnothing$, $i \neq j$), и их объединение эквивалентно достоверному событию ($H_1 \cup H_2 \cup \text{L} \cup H_n = \Omega$).

Замечание. Иногда полную группу событий называют *разбиением.*

В условиях опыта с подбрасыванием игральной кости событие $A_1 \cup A_2$ состоит в выпадении либо единицы, либо двойки; событие $A_{\text{чет}} \cap A_{<3}$ состоит в выпадении четного числа, меньшего трех (т. е. двойки); дополнением события A_1 до события $A_{<3}$ является событие A_2; примерами противоположных случайных событий являются, например, события $A_{\text{чет}}$ и $A_{\text{нечет}}$; примерами несовместных событий являются, например, события $A_{<3}$ и A_5; примерами совместных событий являются, например, события $A_{<3}$ и A_1; полную группу образуют события $A_1, A_2, \ldots, A_6$.

Случайное событие ω, связанное с опытом S, которое невозможно представить как объединение более простых событий, связанных с тем же опытом, называется *элементарным событием*. Очевидно, достоверное событие $\Omega = \{\omega\}$ — это множество всех элементарных событий (поэтому Ω называют еще *пространством элементарных событий*), а невозможное событие $\varnothing$ — это пустое множество. Любое событие, связанное с опытом S, можно представить как некоторое подмножество достоверного события Ω, т. е. как множество некоторых элементарных событий.

Для наглядного представления событий, операций над событиями и отношений между ними используются диаграммы Венна — Эйлера (рис. 2.2.1).

На этих диаграммах достоверное событие Ω изображается в виде некоторой области на плоскости, элементарные события ω_i — точками внутри области, соответствующей Ω. При этом любому случайному событию A будет соответствовать некоторая геометрическая фигура внутри области, соответствующей Ω (заштрихованная фигура на рис. 2.2.1, *а*). Невозможному событию $\varnothing$ соответствует пустое множество (т. е. не соответствует никакая фигура на диаграмме). Событие A влечет за собой событие B, если все элементарные события, входящие в A, входят и в B (рис. 2.2.1, *б*). Объединение $A \cup B$ событий A и B состоит из всех элементарных событий, влекущих, по крайней мере, одно из событий A или B (заштрихованная фигура на рис. 2.2.1, *в*). Пересечение $A \cap B$ событий A и B состоит из всех элементарных событий, влекущих одновременно оба события A и B (рис. 2.2.1, *г*). Дополнение $A \setminus B$ события B до события A состоит из всех элементарных событий, влекущих событие A и при этом не влекущих событие B (рис. 2.2.1, *д*). Событие $\bar{A}$, противоположное событию A, состоит из всех элементарных событий, не влекущих событие A (рис. 2.2.1, *е*). Несовместные события не имеют общих элементарных событий (рис. 2.2.1, *ж*). Полная группа событий представлена на рис. 2.2.1, *з*.

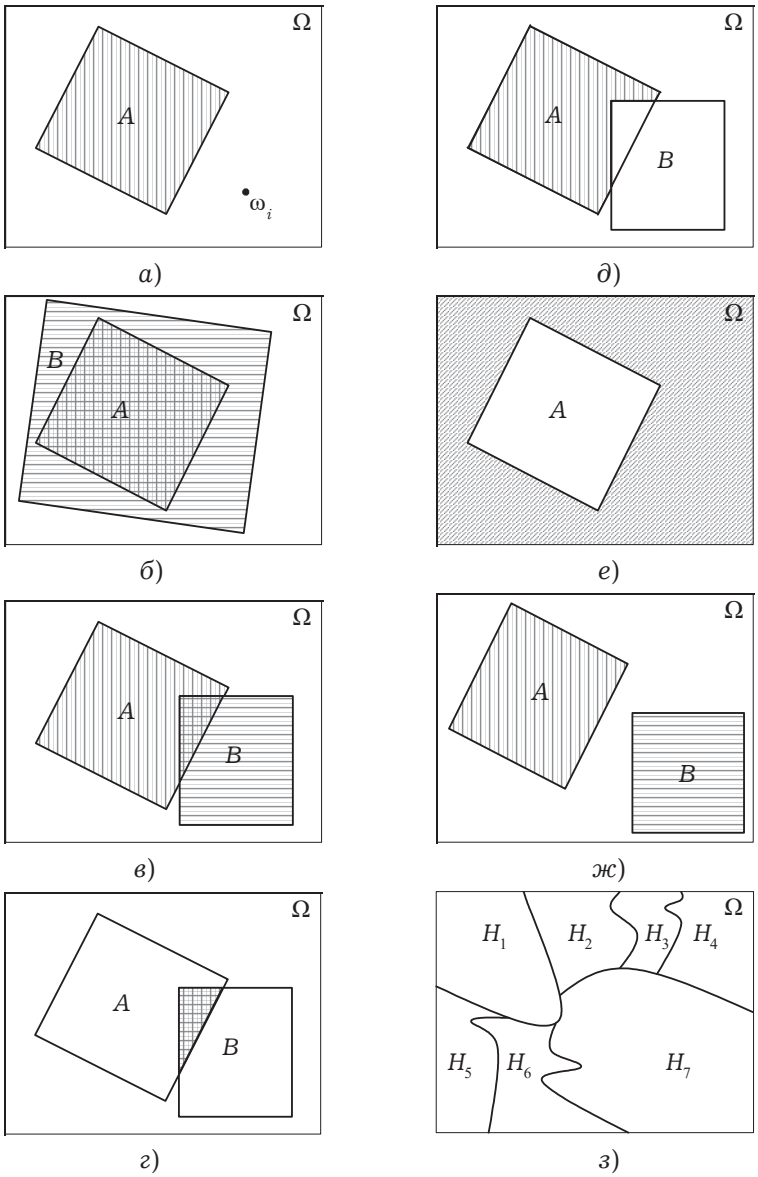


Рис. 2.2.1. *Диаграммы Венна — Эйлера*

Операции над событиями обладают следующими легко проверяемыми **свойствами.**

Коммутативность объединения событий:

$$A \cup B = B \cup A. \tag{2.2.1}$$

Ассоциативность объединения событий:

$$(A \cup B) \cup C = A \cup (B \cup C). \tag{2.2.2}$$

Другие свойства объединения событий:

$$A \cup A = A, \quad A \cup \overline{A} = \Omega, \quad A \cup \varnothing = A, \quad A \cup \Omega = \Omega; \tag{2.2.3}$$

Коммутативность пересечения событий:

$$A \cap B = B \cap A. \tag{2.2.4}$$

Ассоциативность пересечения событий:

$$(A \cap B) \cap C = A \cap (B \cap C). \tag{2.2.5}$$

Другие свойства пересечения событий:

$$A \cap A = A, \quad A \cap \overline{A} = \varnothing, \quad A \cap \varnothing = \varnothing, \quad A \cap \Omega = A; \tag{2.2.6}$$

Противоположность достоверного и невозможного событий:

$$\overline{\Omega} = \varnothing, \quad \overline{\varnothing} = \Omega; \tag{2.2.7}$$

Выражение для дополнения события:

$$A \setminus B = A \cap \overline{B}; \tag{2.2.8}$$

Дистрибутивность пересечения событий относительно объединения:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \tag{2.2.9}$$

Правила де Моргана:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}. \tag{2.2.10}$$

Доказательство. Справедливость свойств (2.2.1)—(2.2.8) следует непосредственно из определений соответствующих операций над событиями. Например, в левой части формулы (2.2.1) стоит событие $A \cup B$, т. е. событие которое происходит тогда, когда происходит либо событие A, либо событие B (либо оба события одновременно). В правой части этого равенства стоит событие $B \cup A$, т. е. событие которое происходит тогда, когда происходит либо событие B, либо событие A (либо оба события одновременно). Очевидно, события в левой и правой частях равенства (2.2.1) эквивалентны, что и доказывает свойство коммутативности объединения событий.

Предоставляем читателю самостоятельно убедиться в справедливости свойств (2.2.2)—(2.2.8).

Докажем свойство (2.2.9). Для этого представим событие $A \cap (B \cup C)$ в виде объединения трех непересекающихся событий: $A \cap (B \cup C) = [A \cap B \cap C] \cup [A \cap (B \setminus C)] \cup [A \cap (C \setminus B)]$ (рис. 2.2.2).

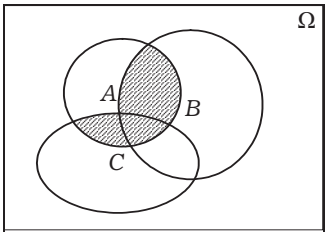


Рис. 2.2.2. *Диаграмма Венна — Эйлера для события* $A \cap (B \cup C)$

Теперь, учитывая, что $A \cap B \cap C = [A \cap B \cap C] \cup [A \cap B \cap C]$ (2.2.3), можно записать: $A \cap (B \cup C) = [A \cap B \cap C] \cup [A \cap B \cap C] \cup [A \cap (B \setminus C)] \cup [A \cap (C \setminus B)]$, а затем воспользоваться свойством коммутативности объединения событий (2.2.1):

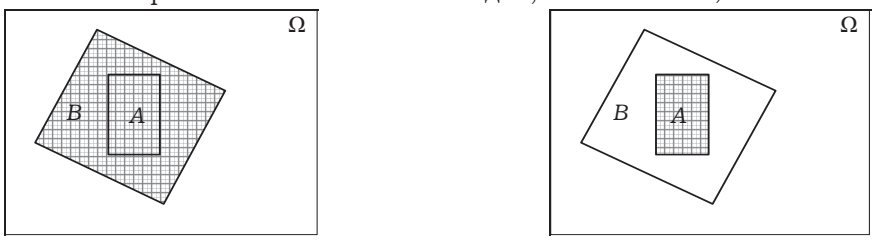
$$A \cap (B \cup C) = [A \cap B \cap C] \cup [A \cap B \cap C] \cup [A \cap (B \setminus C)] \cup [A \cap (C \setminus B)] =$$
$$= \{[A \cap (B \cap C)] \cup [A \cap (B \setminus C)]\} \cup \{[A \cap (B \cap C)] \cup [A \cap (C \setminus B)]\} = (A \cap B) \cup (A \cap C),$$

где в последней строке мы учли, что $[A \cap B \cap C] \cup [A \cap (B \setminus C)] = A \cap B$ и $[A \cap B \cap C] \cup \cup [A \cap (C \setminus B)] = A \cap C$, в чем читателю предлагаем убедиться самостоятельно.

Диаграмма Венна — Эйлера для события $A \cap (B \cup C)$, представленная на рис. 2.2.2, иллюстрирует свойство (2.2.9).

Доказательство правил де Моргана (2.2.10) оставляем читателю в виде задачи **64**. ❑

Пример 2.2.1. Пусть известно, что $A \subseteq B$. Найдем $A \cup B$ и $A \cap B$. Событие $A \cup B$ заштриховано на рис. 2.2.3, *а*, событие $A \cap B$ — на рис. 2.2.3, *б*. Из определений объединения и пересечения событий очевидно, что $A \cup B = B$, $A \cap B = A$. ❑



а) *б*)

Рис. 2.2.3. *События* $A \cup B$ (*а*) *и* $A \cap B$ (*б*), *когда* $A \subseteq B$

Пример 2.2.2. Пусть A и B — произвольные события. Убедимся, что события A и $\overline{A \cup B}$ являются несовместными.

Действительно, $A \cap (\overline{A \cup B}) = \{$по правилу де Моргана$\} = A \cap (\overline{A} \cap \overline{B}) = \{$по свойству ассоциативности пересечения событий$\} = (A \cap \overline{A}) \cap \overline{B} = \varnothing \cap \overline{B} = \varnothing$, значит, события A и $\overline{A \cup B}$ являются несовместными. ❑

Задачи

59. Пусть A, B, C — произвольные события. Запишите выражения для событий, состоящих в том, что: *а*) произошло только A; *б*) произошли A и B, но C не произошло; *в*) все три события произошли; *г*) произошло хотя бы одно из этих событий; *д*) произошло хотя бы два события; *е*) ни одно из событий A, B и C не произошло; *ж*) произошло не более двух из событий A, B и C; *з*) произошло ровно одно из этих событий; *и*) произошло ровно два из этих событий.

60. Пусть A, B, C — некоторые события, причем $A \subseteq B$. Упростите выражения: *а*) $A \cap B \cap C$; *б*) $A \cup B \cup C$.

61. Установите, при каких условиях события A и $A \cap B$ являются эквивалентными.

62. Проверьте справедливость следующих утверждений, сравнивая диаграммы Венна — Эйлера для событий, стоящих в левых и в правых частях следующих выражений: *а*) $(A \cup B) \setminus C = A \cup (B \setminus C)$; *б*) $A \cap B \cap C = A \cap B \cap (B \cup C)$; *в*) $A \cup B \cup C = A \cup [B \setminus (A \cap B)] \cup [C \setminus (A \cap C)]$; *г*) $A \cup B = [A \setminus (A \cap B)] \cup B$; *д*) $(A \cup B) \setminus A = B$; *е*) $(\overline{A \cup B \cup C}) = \overline{A} \cap \overline{B} \cap \overline{C}$; *ж*) $(\overline{A \cup B}) \cap C = (\overline{A} \cap C) \cup (\overline{B} \cap C)$; *з*) $(\overline{A \cup B}) \cap C = \overline{A} \cap \overline{B} \cap C$; *и*) $(\overline{A \cup B}) \cap C = C \setminus (C \cap (A \cup B))$; *к*) $A \cap B \cap C \subseteq (A \cap C) \cup (B \cap C)$; *л*) $(A \cap B) \cup (B \cap C) \cup (A \cap C) \subseteq A \cup B \cup C$; *м*) $A \cap \overline{B} \cap C \subseteq A \cup B$.

63. Пусть A, B — произвольные события. Проверьте, образуют ли события A, $\overline{A} \cap B$, $\overline{A \cup B}$ полную группу.

64. Докажите правила де Моргана (2.2.10).

2.2.2. Классическая вероятностная схема

Вероятность наступления события характеризует меру возможности наступления этого события при проведении некоторого опыта. Если множество элементарных событий (исходов опыта) $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ конечно,

и все элементарные события одинаково возможны, то такая вероятностная схема называется *классической*. В этом случае *вероятность* $\mathbf{P}(A)$ наступления события A, состоящего из M элементарных событий, входящих в Ω, определяется как отношение числа M элементарных событий, благоприятствующих наступлению события A, к общему числу N элементарных событий. Эта формула носит название *классической формулы вероятности*.

Классическая формула вероятности. *Если множество элементарных событий конечно, и все элементарные события одинаково возможны, то вероятность произвольного события A вычисляется по формуле*

$$\mathbf{P}(A) = \frac{M}{N}. \tag{2.2.11}$$

Введенная таким образом вероятность обладает следующими **свойствами.**

Неотрицательность вероятности. *Для всех событий A*

$$\mathbf{P}(A) \geqslant 0. \tag{2.2.12}$$

Нормированность вероятности:

$$\mathbf{P}(\Omega) = 1. \tag{2.2.13}$$

Аддитивность вероятности. *Для всех событий A и B, таких что* $A \cap B = \varnothing$,

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B). \tag{2.2.14}$$

Доказательство. Поскольку число M элементарных событий, благоприятствующих наступлению события A, и общее число N элементарных событий неотрицательны, то и вероятность $\mathbf{P}(A) = M/N \geqslant 0$, что доказывает справедливость свойства неотрицательности вероятности (2.2.12).

Так как достоверному событию благоприятствуют все N элементарных событий, то вероятность достоверного события $\mathbf{P}(\Omega) = N/N = 1$, что доказывает справедливость свойства нормированности вероятности (2.2.13).

Если события A и B несовместны, т. е. $A \cap B = \varnothing$, т. е. нет элементарных событий, благоприятствующих одновременному наступлению событий A и B, обозначим M_A число элементарных событий, благоприятствующих наступлению события A, а M_B — число элементарных событий, благоприятствующих наступлению события B (конечно, $0 \leqslant M_A + M_B \leqslant N$). Тогда число элементарных событий, благоприятствующих наступлению события $A \cup B$, будет равно $M_A + M_B$.

Поэтому согласно классической формуле вероятности $\mathbf{P}(A \cup B) = (M_A + M_B)/N = M_A/N + M_B/N = \mathbf{P}(A) + \mathbf{P}(B)$, что доказывает справедливость свойства аддитивности вероятности (2.2.14). ❑

Идея вычисления вероятности как отношения числа исходов, благоприятствующих наступления события, к общему числу исходов опыта в частных случаях появилась в рукописи Дж. Кардано «Книга об игре в кости», датированной самим Дж. Кардано 1526 г., но изданной только в 1563 г. Однако сама классическая формула вероятности в современном виде была впервые опубликована лишь в 1713 г. в трактате Я. Бернулли «Искусство предположений».

Пример 2.2.3. В корзине три красных и семь зеленых яблок. Из корзины вынимают одно яблоко. Найдем вероятность того, что оно будет красным.

Пусть опытом будет извлечение яблока из корзины, а событие A состоит в том, что извлеченное из корзины яблоко окажется красным. Тогда общее число элементарных событий $N = 10$, из которых $M = 3$ элементарных события

благоприятствуют наступлению события A. Согласно классической формуле вероятности (2.2.11) $\mathbf{P}(A) = M/N = 3/10 = 0{,}3$. ❑

Пример 2.2.4. В колоде всего 32 карты, по восемь каждой масти. Каждому из троих игроков сдано по десять карт и две оставлены в прикупе. Один из игроков видит, что у него на руках шесть карт бубновой масти, а четыре — других мастей. Он сбрасывает две карты из этих четырех и берет себе прикуп. Найдем вероятность того, что в прикупе окажутся две бубновые карты.

Пусть событие A состоит в том, что в прикупе окажутся две бубновые карты. Из 32 карт игроку известны десять, а остальные 22 неизвестны. Взять две карты из прикупа — это то же самое, что взять их из 22 неизвестных карт, среди которых две бубновые. Поэтому общее число элементарных событий $N = C_{22}^2$, из которых лишь $M = 1$ элементарное событие благоприятствует наступлению события A. Согласно классической формуле вероятности (2.2.11) $\mathbf{P}(A) = M/N = 1/C_{22}^2 = 1/231$ ($C_{22}^2 = $ ПЕРЕСТ(22; 2) =231). ❑

Пример 2.2.5. Двери лифта закрылись на первом этаже прямо перед Петей, который успел только заметить, что в лифт вошли пять человек. В здании банка семь этажей, и лифт, если откроет на каком-либо из них двери, стоит там целую минуту. Петя опаздывает на собеседование и размышляет, каковы вероятности следующих событий: а) все пятеро выйдут на одном этаже; б) все пятеро выйдут на разных этажах. Найдем эти вероятности.

а) Опыт, за которым наблюдает Петя, состоит в том, что люди выходят из лифта произвольным образом: каждый из пяти человек может выйти на любом из шести этажей (со второго по седьмой). Общее число исходов этого опыта $N = \bar{A}_6^5 = 6^5$ (первый человек может выйти на любом из пяти этажей, второй — также на любом из пяти этажей, и так до пятого человека; все эти пять шестерок перемножаются по правилу произведения). Пусть событие A состоит в том, что все пять человек выйдут на одном и том же этаже, тогда число исходов описанного опыта, благоприятствующих наступлению события A, равно $M_A = 6$ (всего есть шесть этажей). Поэтому согласно классической формуле вероятности (2.2.11) $\mathbf{P}(A) = M_A/N = 6/6^5 = 1/6^4 \approx 0{,}00077$.

б) Рассмотрим тот же опыт, что и в п. *а*). Пусть событие B состоит в том, что все пять человек выйдут на разных этажах, тогда число исходов опыта, благоприятствующих наступлению события B, равно $M_B = A_6^5 = 6!$, и по классической формуле вероятности (2.2.11)

$$\mathbf{P}(B) = \frac{M_B}{N} = \frac{6!}{6^5} = \frac{\overset{1}{\cancel{6}}}{\underset{1}{\cancel{6}}} \cdot \frac{5}{6} \cdot \frac{\overset{1}{\cancel{4}}}{\underset{3}{\cancel{6}}} \cdot \frac{\overset{1}{\cancel{3}}}{\underset{1}{\cancel{6}}} \cdot \frac{\overset{1}{\cancel{2}}}{\underset{3}{\cancel{6}}} = \frac{5}{54} \approx 0{,}0926.$$

В Microsoft Excel вычисления проведем так: $\mathbf{P}(A) =$ 1/6^4 = 0,000771605; $\mathbf{P}(B) =$ = ФАКТР(6)/6^5 = 0,092592593. ❑

Частным случаем классической вероятностной схемы является **урновая схема**: в урне содержится L шаров, среди которых K белых и $(L - K)$ черных. Из урны наугад извлекаются l шаров, и требуется вычислить вероятности $\mathbf{P}_{K,L}(k, l)$ того, что в выборке содержится ровно k белых шаров (и $(l - k)$ черных).

Если при этом шары из урны извлекаются б е з в о з в р а щ е н и я, то для вычисления вероятности $\mathbf{P}_{K,\,L}(k,\,l)$ того, что в выборке содержится ровно k белых шаров (и $(l-k)$ черных), применяется

Формула гипергеометрической вероятности:

$$\mathbf{P}_{K,L}(k,l)=\frac{C_K^k C_{L-K}^{l-k}}{C_L^l}\quad (k=0,\,1,\,2,\,\ldots,\,\min\{K;\,l\}). \tag{2.2.15}$$

Доказательство. Опыт состоит в том, что из общей совокупности L шаров извлекается l шаров. Общее число исходов этого опыта равно, очевидно, $N=C_L^l$, причем все исходы одинаково возможны. Пусть событие A состоит в том, что в результате данного опыта было извлечено ровно k белых шаров и $(l-k)$ черных. Число способов, которыми можно из K белых шаров извлечь k штук, равно C_K^k, а число способов выбрать оставшиеся $(l-k)$ шаров из числа $(L-K)$ черных, равно C_{L-K}^{l-k}. Поэтому по правилу произведения число исходов, благоприятствующих наступлению события A, равно $M=C_L^l C_{L-K}^{l-k}$, и искомую вероятность можно рассчитать по классической формуле: $\mathbf{P}_{K,L}(k,l)=M/N=C_K^k C_{L-K}^{l-k}/C_L^l$, что и требовалось доказать. ❑

Если же шары из урны извлекаются с в о з в р а щ е н и е м, то для вычисления вероятности $\mathbf{P}_{K,\,L}(k,\,l)$ того, что в выборке содержится ровно k белых шаров [и $(l-k)$ черных], используется

Формула Бернулли:

$$\mathbf{P}_{K,L}(k,l)=C_l^k\left(\frac{K}{L}\right)^k\left(\frac{L-K}{L}\right)^{l-k}\quad (k=0,\,1,\,2,\,\ldots,\,l). \tag{2.2.16}$$

Замечание. Доказательство этой формулы с помощью классической формулы вероятности оставляем любознательному читателю в задаче **76**. Более простое доказательство будет приведено в **п. 2.4.1.**

В пакете Microsoft Excel для вычисления вероятности $\mathbf{P}_{K,\,L}(k,\,l)$ по формулам (2.2.5) и (2.2.6) можно пользоваться соответствующими функциями

$$\mathbf{P}_{K,\,L}(k,\,l)=\text{ГИПЕРГЕОМ.РАСП}(\langle k\rangle;\,\langle l\rangle;\,\langle K\rangle;\,\langle L\rangle;\,\text{ЛОЖЬ})$$

и

$$\mathbf{P}_{K,\,L}(k,\,l)=\text{БИНОМ.РАСП}(\langle k\rangle;\,\langle l\rangle;\,\langle K\rangle\,/\,\langle L\rangle;\,\text{ЛОЖЬ}).$$

Пример 2.2.6. Из прошлого опыта известно, что на тысячу транзакций по картам клиентов некоторого банка приходится четыре мошеннических. Для детальной проверки отбираются 100 транзакций. Найдем вероятность того, что среди отобранных транзакций не окажется мошеннических.

По формуле гипергеометрической вероятности (2.2.15) получаем, что

$$\mathbf{P}_{4;1000}(0;100)=\frac{C_4^0 C_{996}^{100}}{C_{1000}^{100}}=\frac{\dfrac{4!}{0!4!}\cdot\dfrac{996!}{100!896!}}{\dfrac{1000!}{100!900!}}=\frac{\dfrac{996!}{100!896!}}{\dfrac{1000\cdot999\cdot998\cdot997\cdot996!}{100!900\cdot899\cdot898\cdot897\cdot896!}}=$$

$$=\frac{996!100!900\cdot899\cdot898\cdot897\cdot896!}{100!896!1000\cdot999\cdot998\cdot997\cdot996!}=\frac{\cancel{996!}\cdot\cancel{100!}\cdot900\cdot899\cdot898\cdot897\cdot\cancel{896!}}{\cancel{100!}\cdot\cancel{896!}\cdot1000\cdot999\cdot998\cdot997\cdot\cancel{996!}}=$$

$$=\frac{899\cdot449\cdot897}{10\cdot111\cdot499\cdot997}\approx0{,}656.$$

В Microsoft Excel тот же результат получим так:

$$\mathbf{P}_{4;1000}(0;100)=\text{ГИПЕРГЕОМ.РАСП}(0;\,100;\,4;\,1000;\,\text{ЛОЖЬ})=0{,}655661666.\ ❑$$

Задачи

65. В ежедневной тиражной лотерее «Гослото «5 из 36» играющий отмечает на карточке пять чисел от 1 до 36 и получает призы различного достоинства, если он угадал два, три, четыре и пять чисел, объявленных тиражной комиссией. Найдите вероятности следующих событий: не угадать ни одного числа из 36, угадать одно, два, три, четыре и пять чисел из 36.

66. Среди 12 переданных ревизору договоров семь оформлены с ошибками. Найдите вероятность того, что среди пяти договоров, произвольно отобранных ревизором для проверки, окажутся неправильно оформленными: *а*) ровно три договора; *б*) не менее трех договоров.

67. Петя и Маша приглашены на день рождения в компанию из десяти человек, включая их, но приходят на него порознь, причем, как и остальные гости, в случайное время. Найдите вероятность того, что они будут сидеть за праздничным столом рядом, если хозяин рассаживает гостей случайным образом, а стол, имеющий прямоугольную форму: *а*) стоит в середине комнаты; *б*) придвинут к стене.

68. Найдите вероятность того, что в группе из шести человек: *а*) ни у кого нет дней рождений в ноябре и декабре; *б*) хотя бы два человека родились в один месяц.

69. Найдите вероятность того, что в группе из 30 человек ни у кого нет общих дней рождений.

70. Финтех-стартап создан десятью семейными парами. Чтобы никому не было обидно, на ежегодном собрании акционеров совет директоров, состоящий из восьми человек, выбирается случайным образом. Найдите вероятности следующих событий: *а*) в совете директоров отсутствуют семейные пары; *б*) в совете директоров есть ровно одна семейная пара; *в*) в совете директоров есть ровно две семейных пары?

71. В студенческой группе 12 девушек и 15 молодых людей. Для участия в психологическом эксперименте отбирают четверых студентов группы. Найдите вероятность того, что будут обобраны две девушки и два молодых человека.

72. Студенческую группу, состоящую из $2m$ юношей и $2m$ девушек, разделили случайным образом на две подгруппы одинакового размера. Найдите вероятность того, что в каждой из подгрупп окажется поровну девушек и юношей.

73. Найдите вероятность того, что при раздаче колоды в 52 карты четырем игрокам первый из них получит ровно n пар «туз и король одной масти» ($n = 0, 1, 2, 3, 4$).

74. На избирательном участке голосовало семь человек: четверо проголосовали за кандидата в депутаты Иванова, трое — за Петрова. Найдите вероятность того, что в ходе подсчета числа голосов по бюллетеням число голосов, поданных за Иванова, было всегда выше, чем число голосов, поданных за Петрова.

75. Группа студентов, состоящая из шести молодых людей (двое из которых являются братьями) и шести девушек (двое из которых являются сестрами), селится в общежитие. Каждая комната в общежитии рассчитана на троих человек, причем в комнате могут жить или только девушки, или только молодые люди. Найдите вероятность того, что сестры будут жить в одной комнате, а братья в разных, если распределение студентов по комнатам происходит случайным образом.

76. Докажите формулу Бернулли для схемы выборки с возвращением (2.2.16), используя классическую формулу вероятности.

2.2.3. Схема геометрических вероятностей

В случае, когда множество элементарных событий бесконечно и даже несчетно (но эти события являются одинаково возможными), вероятность наступления события можно рассчитать, пользуясь геометрическим подходом, который состоит в следующем. Пусть множество элементарных событий Ω представляет собой некоторую область в d-мерном пространстве, имеющую ненулевой объем $V(\Omega)$: $\Omega \subseteq \breve{Y}^d$, $0 < V(\Omega) < +\infty$ [здесь под объемом $V(S)$ области S понимается ее длина в одномерном случае ($d = 1$), площадь в двумерном случае ($d = 2$), объем — в трехмерном случае ($d = 3$) и т. д.]. При этом каждому элементарному событию соответствует точка $\omega \in \Omega \subseteq \breve{Y}^d$, а каждое событие A представляет собой область, вложенную в Ω и имеющую объем $V(A)$: $A \subseteq \Omega \subseteq \breve{Y}^d$, $0 \leqslant V(A) < +\infty$. Тогда *вероятность* $\mathbf{P}(A)$ *наступления события* A можно рассчитать как отношение объема $V(A)$ области A к объему $V(\Omega)$ области Ω:

$$\mathbf{P}(A) = \frac{V(A)}{V(\Omega)} = \frac{\int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d}. \qquad (2.2.17)$$

Введенная таким образом вероятность (2.2.17) обладает свойствами **неотрицательности** (2.2.12), **нормированности** (2.2.13) и **аддитивности** (2.2.14).

Доказательство. Поскольку $\int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d \geqslant 0$, $\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d \geqslant 0$, то и

$$\mathbf{P}(A) = \frac{\int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} \geqslant 0,$$

что доказывает справедливость свойства неотрицательности вероятности (2.2.12).

Вероятность достоверного события

$$\mathbf{P}(\Omega) = \frac{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} = 1,$$

что доказывает справедливость свойства нормированности вероятности (2.2.13)

Если события A и B несовместны ($A \cap B = \varnothing$), т. е. области A и B не имеют общих точек, то

$$\int\int\cdots\int\limits_{A\cup B} dx_1 dx_2 \cdots dx_d = \int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d + \int\int\cdots\int\limits_{B} dx_1 dx_2 \cdots dx_d,$$

и тогда

$$\mathbf{P}(A\cup B) = \frac{\int\int\cdots\int\limits_{A\cup B} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} = \frac{\int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d + \int\int\cdots\int\limits_{B} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} =$$

$$= \frac{\int\int\cdots\int\limits_{A} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} + \frac{\int\int\cdots\int\limits_{B} dx_1 dx_2 \cdots dx_d}{\int\int\cdots\int\limits_{\Omega} dx_1 dx_2 \cdots dx_d} = \mathbf{P}(A) + \mathbf{P}(B),$$

что доказывает справедливость свойства аддитивности вероятности (2.2.14). ❑

Пример 2.2.7. Маша тратит на дорогу в университет от 40 до 50 мин, причем любое время в этом промежутке является одинаково возможным. Найдем вероятность того, что в день экзамена она потратит на дорогу от 45 до 50 мин.

Выберем в качестве множества элементарных событий отрезок длиной 10 мин (от 40 до 50 мин). Событию A, состоящему в том, что Маша потратит на дорогу от 45 до 50 мин, соответствуют элементарные события на отрезке длиной 5 мин. Поэтому в соответствии с геометрическим подходом к определению вероятности по формуле (2.2.17) получаем: $\mathbf{P}(A) = V(A) / V(\Omega) = 5/10 = 0{,}5$. ❑

Пример 2.2.8. Чтобы добраться в университет, Петя может воспользоваться автобусом одного из двух маршрутов. Автобусы первого маршрута следуют с интервалом в 18 мин, второго маршрута — с интервалом в 15 мин. Найдем вероятность того, что Петя будет ждать автобуса не более 10 мин.

Выберем в качестве множества элементарных событий Ω прямоугольник со сторонами $T_1 = 18$ мин и $T_2 = 15$ мин (рис. 2.2.4). Событие A состоит в том, что время, которое Петя будет ждать автобуса, меньше $t = 10$ мин. Элементарные события, благоприятствующие наступлению события A, на рис. 2.2.4 заштрихованы.

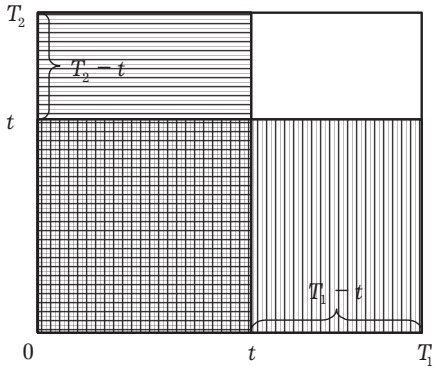


Рис. 2.2.4. *Множество элементарных событий в примере 2.2.8*

По формуле (2.2.17)

$$\mathbf{P}(A) = \frac{V(A)}{V(\Omega)} = \frac{T_1T_2 - (T_1 - t)(T_2 - t)}{T_1T_2} = 1 - \left(1 - \frac{t}{T_1}\right)\left(1 - \frac{t}{T_2}\right).$$

При $T_1 = 18$, $T_2 = 15$, $t = 10$ имеем: $\mathbf{P}(A) = 1 - (1 - 10/18)(1 - 10/15) = 23/27$. ❑

Пример 2.2.9 (Задача о встрече). Петя и Маша договорились встретиться с 12 до 13 ч на станции метро «ВДНХ» у последнего вагона поезда, идущего в центр города, однако ни один из них не смог точно указать время своего прихода. Они договорились ждать друг друга не более 15 мин.

Найдем вероятности следующих событий: *а*) Петя придет раньше Маши; *б*) Петя придет раньше Маши более чем на 15 мин; *в*) Петя и Маша встретятся; *г*) время, проведенное Петей и Машей на станции метро, окажется не меньше 10 мин.

а) Выберем в качестве множества элементарных событий квадрат со сторонами $T_1 = T_2 = 1$ ч (рис. 2.2.5). Абсцисса точки из этого квадрата соответствует времени прихода Пети к месту встречи, а ордината — времени прихода Маши. Диагональ этого квадрата соответствует тому, что Петя и Маша приходят к месту встречи одновременно. Пусть событие A состоит в том, что Петя придет раньше Маши, тогда элементарные события, благоприятствующие наступлению события A, заштрихованы на рис. 2.2.5, *а*, и по формуле (2.2.17)

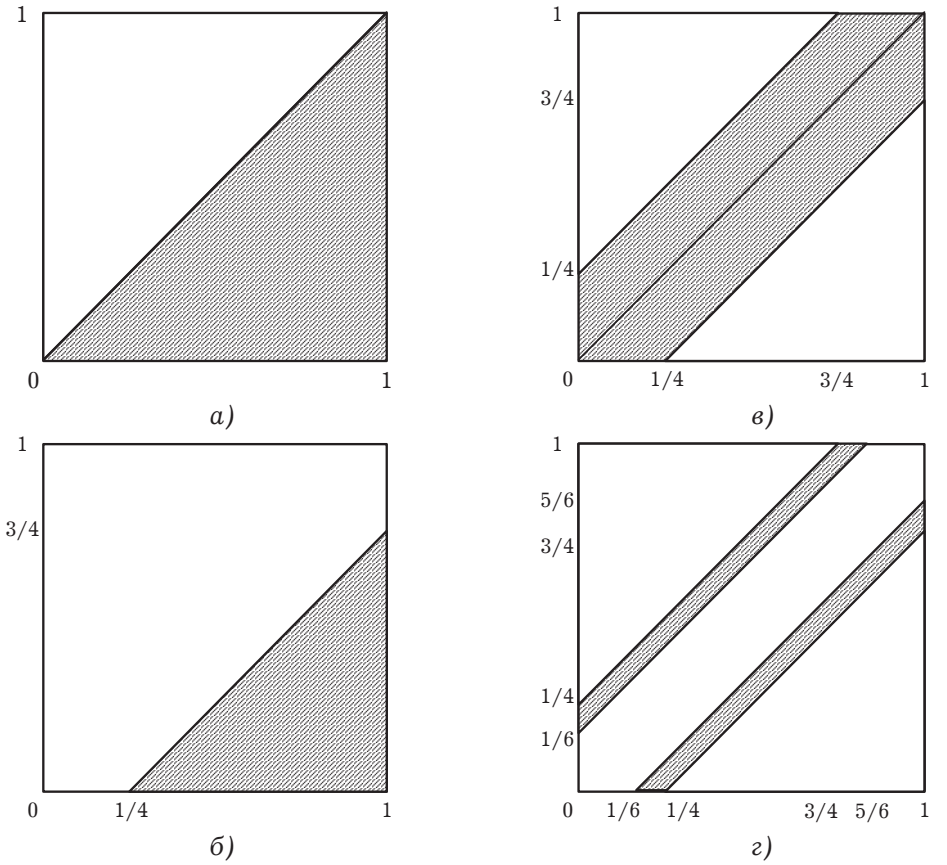


Рис. 2.2.5. *Множество элементарных событий в примере 2.2.9*

$$\mathbf{P}(A)=\frac{V(A)}{V(\Omega)}=\frac{1/2}{1}=\frac{1}{2}.$$

б) Пусть событие B состоит в том, что Петя придет раньше Маши более чем на 15 мин, тогда элементарные события, благоприятствующие наступлению события B, заштрихованы на рис. 2.2.5, *б*, и по формуле (2.2.17)

$$\mathbf{P}(B)=\frac{V(B)}{V(\Omega)}=\frac{\frac{1}{2}\left(\frac{3}{4}\right)^2}{1}=\frac{9}{32}.$$

в) Пусть событие C состоит в том, что Петя и Маша встретятся, тогда элементарные события, благоприятствующие наступлению события C, заштрихованы на рис. 2.2.5, *в*, и по формуле (2.2.17)

$$\mathbf{P}(C)=\frac{V(C)}{V(\Omega)}=\frac{1-2\cdot\frac{1}{2}\left(\frac{3}{4}\right)^2}{1}=1-\frac{9}{16}=\frac{7}{16}.$$

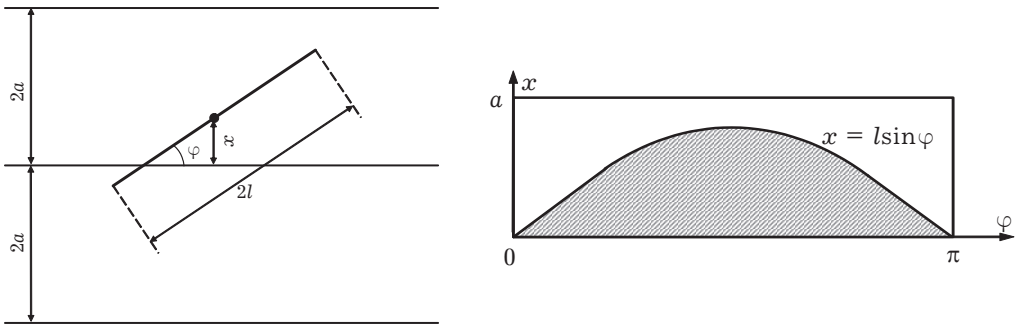
г) Пусть событие D состоит в том, что время, проведенное Петей и Машей на станции метро, окажется не меньше 10 мин, тогда элементарные события, бла-

гоприятствующие наступлению события D, заштрихованы на рис. 2.2.5, *г*. По формуле (2.2.17)

$$\mathbf{P}(D)=\frac{2\left[\frac{1}{2}\left(\frac{5}{6}\right)^2-\frac{1}{2}\left(\frac{3}{4}\right)^2\right]}{1}=\frac{25}{36}-\frac{9}{16}=\frac{19}{144}. \blacksquare$$

Пример 2.2.10 (Задача Бюффона). Плоскость разграфлена параллельными прямыми, отстоящими друг от друга на расстоянии $2a$. На плоскость случайным образом бросают иглу длиной $2l$ ($l < a$), как показано на рис. 2.2.6, а. Найдем вероятность того, что игла пересечет какую-нибудь прямую.

Пусть x — расстояние от центра иглы до ближайшей прямой, а φ — угол, который образует игла с этой прямой.



а) *б*)

Рис. 2.2.6. *Задача Бюффона: бросание иглы* (*а*); *множество элементарных событий* (*б*)

Очевидно, $0 \leqslant x \leqslant a$, $0 \leqslant \varphi \leqslant \pi$, и две величины x и φ однозначно определяют положение иглы. Поэтому в качестве множества элементарных событий Ω выберем прямоугольник со сторонами a и π (рис. 2.2.6, *а*). Каждая точка этого прямоугольника одинаково возможна, поскольку игла бросается случайным образом. Из рис. 2.2.6 ясно, что игла пересечет какую-либо прямую тогда и только тогда, когда выполнено условие $x \leqslant l\sin\varphi$.

Элементарные события, благоприятствующие наступлению события A, состоящему в том, что игла пересечет какую-либо прямую, заштрихованы на рис. 2.2.6, *б*.

По формуле (2.2.9)

$$\mathbf{P}(A)=\frac{V(A)}{V(\Omega)}=\frac{\int_0^{\pi} l\sin\varphi d\varphi}{a\pi}=\frac{2l}{a\pi},$$

где площадь области, соответствующей событию A, удобно рассчитать как интеграл. $\blacksquare$

Геометрические вероятности впервые были рассмотрены французским естествоиспытателем Ж. Бюффоном (одна из задач Ж. Бюффона рассмотрена в примере 2.2.10). Геометрическая вероятностная схема, являясь естественным обобщением классических вероятностей, существенно расширила область применения теории вероятностей, помимо математического интереса геометрические вероятности приобрели важное значение в физике, биологии, медицине, инженерном деле, производственном менеджменте и др.

В 1899 г. Ж. Бертран на хорошо подобранных примерах показал, что логически понятие геометрической вероятности не выдерживает никакой критики. Играя на неопределенности терминологии, ему удалось для (казалось бы) одной и той

же задачи получить несколько различных ответов. Эта критика заставила математиков задуматься над логическим обоснованием теории вероятностей, формулировать аксиомы, из которых бы вытекали остальные положения этой науки. Такая система аксиом была построена (и не одна). В **п. 2.2.5** мы познакомимся с наиболее удобной системой аксиом, предложенной А.Н. Колмогоровым.

Задачи

77. На 200-километровом участке газопровода между двумя компрессорными станциями происходит утечка газа, которая одинаково возможна в любой точке газопровода. Найдите вероятности следующих событий: *а*) утечка расположена не далее 20 км от одной из станций; *б*) утечка расположена ближе к первой станции, чем ко второй.

78. Интервал движения автобуса равен 40 мин. Расстояние между остановками А и Б автобус проходит за 4 мин, а пешеход за 20 мин. Петя приходит на остановку А в случайный момент времени и сразу идет пешком к остановке Б. Найдите вероятность того, что автобус не обгонит Петю.

79. Петя и Маша должны прийти к преподавателю для сдачи зачета с 9 до 10 ч. Они подходят в случайное время в указанном промежутке. Пете для сдачи зачета потребуется 10 мин, а Маше — 20 мин. Найдите вероятность того, что ни одному из этих студентов не придется ждать, пока другой закончит отвечать.

80. Два студента — Коля и Дима — по вечерам дежурят у подъезда Маши, которая им очень понравилась. Не зная, в какое время Маша выйдет из дома (и выйдет ли вообще), каждый из них в произвольный момент между 18 и 20 ч приходит к Машиному подъезду и ждет, когда Маша выйдет, в течение часа (независимо от соперника). Если за это время Маша не вышла, студент уходит. Сегодня Маша хочет пойти на дискотеку с Петей, не встречаясь со своими безуспешными воздыхателями Колей и Димой. Чтобы не опоздать, ей нужно выйти из дома не позднее 20 ч. К 19 ч. она готова выйти. Найдите вероятность того, что Маше удастся выбрать момент выхода так, чтобы никого из поклонников не было во дворе, если двор из ее окна просматривается целиком, и она в любой момент видит, кто ее ждет.

81. Петя, Маша и Вася договорились встретиться в большой перерыв, который длится час, около входа в библиотеку. Никто из них не смог точно указать время своего прихода, поэтому они договорились ждать друг друга не более 10 мин. Найдите вероятности следующих событий: *а*) все они встретятся; *б*) по крайней мере, двое из них встретятся.

2.2.4. Статистическая вероятность

Теория вероятностей и математическая статистика — дисциплины, являющиеся одновременно и строго математическими (поскольку все положения этих наук следуют из небольшого числа аксиом), и экспериментальными (поскольку случайные события, которые изучаются в данных дисциплинах, можно непосредственно наблюдать в реальной жизни).

Теория вероятностей и математическая статистика **применимы лишь к изучению таких экспериментов**, когда есть практически осуществимая или хотя бы мысленно представляемая возможность многократного воспроизведения комплекса условий опыта. Такие эксперименты называются *статистически однородными*.

Пусть в результате n-кратного проведения опыта S событие A наступило m_n раз. *Относительной частотой* $\hat{p}_n(A)$ *появления события* A *в серии из* n

опытов S называется отношение числа m_n наступлений события A к общему числу n проведенных опытов:

$$\hat{p}_n(A) = \frac{m_n}{n}. \qquad (2.2.18)$$

Простейшим примером статистически однородного опыта является опыт с подбрасыванием монеты. Действительно, одну и ту же монету можно многократно подбросить в одних и тех же условиях.

На рис. 2.2.7 представлены результаты компьютерного моделирования четырех серий из $n = 1000$ опытов по подбрасыванию монеты (на графиках отмечено, как с ростом числа проведенных опытов менялась относительная частота, результаты первой серии приведены в табл. 2.2.1).

Таблица 2.2.1

Зависимость частоты выпадения герба от числа опытов

N	$\hat{p}_n = m_n / n$	n	$\hat{p}_n = m_n / n$	N	$\hat{p}_n = m_n / n$	N	$\hat{p}_n = m_n / n$	N	$\hat{p}_n = m_n / n$
10	0,4000	210	0,4667	410	0,5293	610	0,4951	810	0,5358
20	0,5500	220	0,5591	420	0,4786	620	0,5323	820	0,4988
30	0,5000	230	0,4565	430	0,4744	630	0,5175	830	0,5554
40	0,5000	240	0,5042	440	0,5068	640	0,4859	840	0,5131
50	0,4800	250	0,5240	450	0,5222	650	0,5123	850	0,4847
60	0,5000	260	0,5077	460	0,5304	660	0,5030	860	0,5105
70	0,3857	270	0,4852	470	0,4979	670	0,5060	870	0,5069
80	0,4375	280	0,5429	480	0,5000	680	0,5103	880	0,5102
90	0,4778	290	0,5138	490	0,5224	690	0,5014	890	0,5112
100	0,6000	300	0,5367	500	0,5280	700	0,4914	900	0,5144
110	0,5818	310	0,5258	510	0,5196	710	0,4761	910	0,4802
120	0,5417	320	0,4938	520	0,5212	720	0,5028	920	0,5163
130	0,4846	330	0,4515	530	0,4849	730	0,4397	930	0,5204
140	0,4857	340	0,5618	540	0,5148	740	0,4689	940	0,5011
150	0,5000	350	0,4714	550	0,5255	750	0,5280	950	0,4832
160	0,5250	360	0,4889	560	0,4875	760	0,4987	960	0,4885
170	0,5294	370	0,5162	570	0,5123	770	0,5104	970	0,5072
180	0,5611	380	0,5079	580	0,5017	780	0,4962	980	0,5051
190	0,5421	390	0,4615	590	0,4814	790	0,5051	990	0,4859
200	0,4750	400	0,5050	600	0,5383	800	0,5113	1000	0,4850

Анализ табл. 2.2.1 и рис. 2.2.7 показывает, что по мере увеличения числа опытов n относительная частота $\hat{p}_n(A) = m_n / n$ имеет тенденцию приближаться к вероятности 0,5 [уже известной из классической формулы вероятности (2.2.11)]. Это приближение прослеживается явно, хотя и происходит достаточно медленно. Если бы было проведено еще 1000 подбрасываний монеты, то кривая зависимости относительной частоты и числа опытов выглядела бы несколько иначе, чем на графиках, представленных на рис. 2.2.7, но общая тенденция приближения относительной частоты к 0,5 сохранилась бы.

Говорят, что случайное явление обладает свойством *статистической устойчивости*, если некоторая функция результатов многократных наблюдений предсказуема с большой степенью надежности (например, в смысле сходимости по вероятности к некоторому числу), тогда как результат единичного наблюдения непредсказуем.

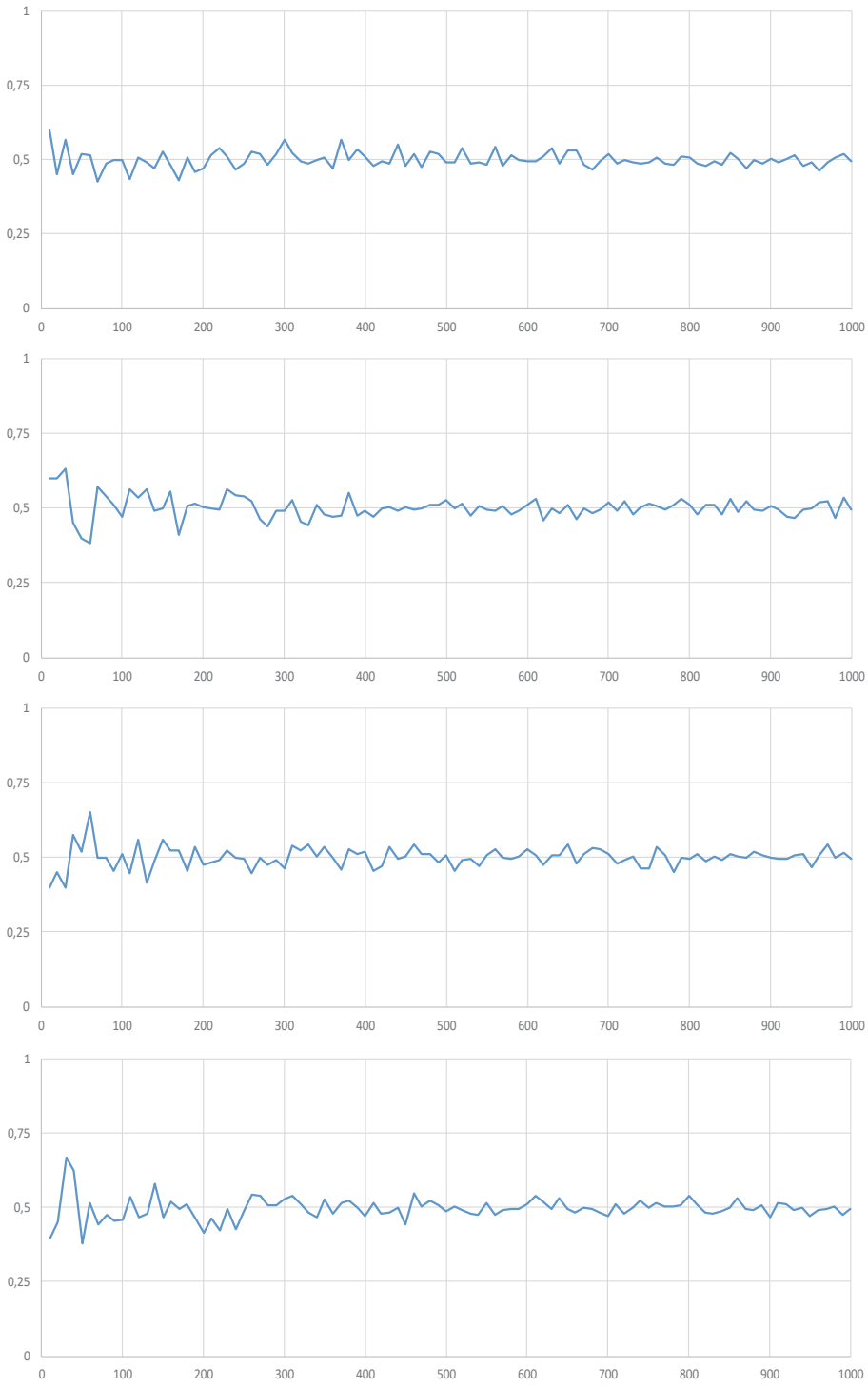


Рис. 2.2.7. *График зависимости частоты выпадения герба от числа бросаний монеты (4 серии по 1000 опытов)*

Оказывается, что *вероятность* $\mathbf{P}(A)$ *наступления события* A равна пределу относительной частоты наступления этого события в серии из n опытов при неограниченном увеличении числа опытов:

$$\frac{m_n}{n} \xrightarrow[n\to\infty]{\mathbf{P}} \mathbf{P}(A). \qquad (2.2.19)$$

Замечание. Конечно, вероятность $\mathbf{P}(A)$ не является пределом относительной частоты $\hat{p}_n(A)$ в обычном смысле предела, принятом в математическом анализе. Запись «$\xrightarrow{\mathbf{P}}$» в формуле (2.2.19) означает «сходимость по вероятности», т. е. для любого $\varepsilon > 0$ при увеличении n вероятность того, что $\hat{p}_n(A)$ отклонится от $\mathbf{P}(A)$ больше, чем на ε, стремится к нулю:

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{m_n}{n} - \mathbf{P}(A)\right| > \varepsilon\right\} = 0.$$

На практике, конечно, невозможно провести бесконечно много опытов, поэтому вероятность рассчитывают при помощи приближенного равенства

$$\mathbf{P}(A) \approx \hat{p}_n(A) = \frac{m_n}{n}. \qquad (2.2.20)$$

Статистическая устойчивость относится к реальному явлению, а не к его математической модели, поэтому принцип статистической устойчивости не является аксиомой. Статистическая устойчивость относительных частот установлена в результате многовековой практической деятельности.

В табл. 2.2.2 представлены результаты опытов с бросанием монеты. Видно, что ни в одном из опытов относительная частота не оказалась в точности равна истинному значению вероятности (0,5), но была достаточно близка к этому значению.

Таблица 2.2.2

Результаты опытов с бросанием монеты для определения статистической вероятности выпадения герба

Экспериментатор	*Число испытаний* n	*Количество появлений герба* m_n	*Относительная частота появлений герба* $\hat{p} = m_n / n$
Дж. Керрих	10 000	5 087	0,5087
Ж. Бюффон	4 040	2 048	0,5069
К. Пирсон	12 000	6 019	0,5016
К. Пирсон	24 000	12 012	0,5005

Следует помнить, что теория вероятностей и математическая статистика изучает лишь случайные события в статистически однородных опытах, и о вероятностях случайных событий можно вести речь лишь при выполнении трех **условий:**

- рассматриваемые события должны быть исходами только таких опытов, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий;
- события должны обладать статистической устойчивостью относительных частот;
- число испытаний, по результатам которых вычисляется вероятность, должно быть достаточно велико.

Пример 2.2.11. Ихтиологи выловили в пруду 100 рыб, окольцевали их и выпустили назад в воду. На следующий день они выловили 120 рыб, из которых 10 оказались окольцованными. Найдем приближенные значения вероятности того, что выловленная рыба окольцована, и количества рыб в пруду.

Пусть N — число рыб в пруду, $M = 100$ — число окольцованных рыб в пруду, событие A состоит в том, что выловленная рыба окольцована. Тогда вероятность события A равна согласно классической формуле вероятности $\mathbf{P}(A) = 100/N$. С другой стороны, поскольку из $n = 120$ выловленных рыб $m_n = m_{120} = 10$ оказались окольцованы, вероятность события A приближенно равна его относительной частоте: $\mathbf{P}(A) \approx \hat{p}_{120}(A) = m_{120}/120 = 10/120 = 1/12$. Отсюда $100/N \approx 1/12$ или $N \approx 1200$. ❑

Вопрос о соблюдении принципа статистической устойчивости в каждом конкретном случае решается на интуитивном уровне с учетом опыта практической работы. В простейшем случае статистически однородных и независимых опытов статистическая устойчивость относительных частот следует из теоремы Бернулли (см. **п. 4.1.2**). Однако проверить на практике независимость опытов не всегда оказывается возможным.

Все мыслимые эксперименты можно разбить на три класса. К первому классу отнесем такие эксперименты, в которых обеспечивается полная устойчивость опыта. Например, если при нормальном атмосферном давлении воду нагреть до 100 °С, то она закипит; если на Земле подбросить вверх камень, то он упадет. В этом случае теория вероятностей и математическая статистика не дадут никаких новых знаний об исследуемом явлении.

Ко второму классу отнесем такие эксперименты, в которых полной устойчивости опыта нет, но есть статистическая устойчивость. Простейшими примерами таких экспериментов являются опыты с бросанием монеты или игральной кости. В экономике и общественных науках таких экспериментов много. Прежде всего, это задачи, связанные с изучением поведения объекта как представителя большой однородной совокупности подобных объектов. Такие задачи часто возникают в управлении производством, управлении качеством, при изучении производителей и потребителей, в страховании, демографии, социологии и других общественных науках (не будем здесь останавливаться на широте применений вероятностно-статистических методов в естественных науках: физике, биологии, медицине и др., в этих науках проще обеспечить статистическую устойчивость опыта, так как условия опыта создаются, как правило, самим экспериментатором). Именно в таких экспериментах теория вероятностей и математическая статистика могут приносить ощутимые результаты. Если нарушено требование неизменности условий опыта, то вероятностно-статистические методы могут применяться, но при интерпретации и практическом использовании результатов следует четко оговаривать приближенность, неточность выводов.

И, наконец, к третьему классу отнесем такие опыты, в которых нет ни полной, ни статистической устойчивости. Это опыты, массовое повторение которых невозможно. Так не имеет смысла говорить о вероятностях уникальных событий, которые невозможно повторить в тех же самых условиях: возникновении войн, появлении научных открытий и гениальных произведений искусства и т. п. Бессмысленны с вероятностной точки зрения и высказывания о том, что «из-за данного куста медведь может выскочить с вероятностью 0,06» или что «студент Иванов сдаст экзамен по анализу данных на “отлично”», поскольку невозможно многократное повторение соответствующих опытов.

Как и классическая формула вероятности (2.2.11), статистический подход к определению вероятности впервые был четко сформулирован в работах Я. Бернулли.

Понятие относительной частоты и определение вероятности как предела относительных частот при неограниченном увеличении числа повторений опыта являются чрезвычайно важными в теории вероятностей. Все результаты теории вероятностей и математической статистики могут быть выведены из аксиом, которым удовлетворяют относительные частоты (такое построение теории вероятностей было проведено в начале XX в. Р. Мизесом, однако практически более удобной оказалась аксиоматика А. Н. Колмогорова, к изучению которой мы перейдем в следующем пункте).

Поговорим теперь еще об одной области применения вероятностно-статистических методов. Вспомним задачу Бюффона: вероятность того, что игла пересечет какую-нибудь прямую, оказалась равна $\mathbf{P}(A)=2l/(a\pi)$. Будем проводить опыт с бросанием иглы достаточно много раз в одинаковых условиях, тогда относительная частота пересечений $\hat{p}_n(A)=m_n/n$ (где n — число проведенных опытов, а m_n — число пересечений иглы с какой-либо прямой) будет близка к вероятности $\mathbf{P}(A)$. Поэтому можно на основании статистического подхода к определению вероятности приближенно определить значение числа π по результатам опытов: $2l/(a\pi)\approx m_n/n$, откуда $\pi\approx 2ln/(am_n)$. В докомпьютерную эру для уточнения числа π такие серии опытов проводили многие исследователи (результаты некоторых из них приведены в табл. 2.2.3).

Таблица 2.2.3

Результаты экспериментов по определению числа π в опытах с бросанием иглы

Экспериментатор	*Отношение l/a*	*Число испытаний n*	*Число пересечений m_n*	*Экспериментальное значение числа π*
Р. Вольф (1850 г.)	4/5	5000	2532	3,1596
Г. Смит (1855 г.)	2/3	3204	1218	3,1553
К. Лаццирини (1901 г.)	5/6	3408	1808	3,1415929

Задачи

82. Известно, что в среднем из 1000 выданных кредитов примерно 12 не возвращаются в срок. В текущем году банк выдал 3000 кредитов. Найдите примерное количество кредитов, которые не будут возвращены в срок.

83. За последние 100 лет дождь 15 октября наблюдался 40 раз. Найдите приближенное значение вероятности дождя 15 октября следующего года.

84. Докажите, что относительная частота (2.2.18) (а значит, и вероятность как предел относительной частоты) обладает свойствами неотрицательности (2.2.12), нормированности (2.2.13) и аддитивности (2.2.14).

2.2.5. Аксиоматическое построение теории вероятностей

Долгое время теория вероятностей представляла собой не до конца сложившуюся математическую науку, основные понятия которой были определены недостаточно четко. Эти нечеткости приводили порой к парадоксальным выводам (например, упомянутый в **п. 2.2.3** парадокс Бертрана давал три различных ответа к одной и той же задаче исключительно из-за нечеткости формулировок). Несмотря на это, неформализованный подход к теории вероятностей привел к началу XX в. к ряду весьма крупных теоретических и практических результатов. В начале XX в. развитие, прежде всего, естественных на-

ук (физики, биологии и др.) предъявило к теории вероятностей гораздо более высокие требования. Возникла необходимость в выяснении условий, при которых возможно использование результатов теории вероятностей. Поэтому формальное логическое построение теории вероятностей приобрело особенно важное значение. В основу любой математической науки, как мы знаем, должно быть положено минимальное число аксиом — предпосылок, являющихся обобщением многовекового человеческого опыта, а дальнейшее развитие такой науки должно строиться из этих аксиом посредством дедукции, без обращения к наглядным представлениям и «здравому смыслу».

В начале XX в. системы аксиом теории вероятностей были построены в работах Г. Больмана, С. Н. Бернштейна и Р. Мизеса, однако эти системы аксиом были весьма громоздкими и представляли, прежде всего, теоретический интерес. В 1933 г. А. Н. Колмогоров опубликовал монографию «Основные понятия теории вероятностей», в которой предложил аксиоматику, поразительную своей естественностью и концептуальной простотой. Эта аксиоматика превратила теорию вероятностей в строгую математическую науку и в то же время удовлетворила повышенным требованиям самых разнообразных областей естественных, технических и гуманитарных наук, открыв возможность широкого применения теории вероятностей в этих науках.

Отправным пунктом аксиоматики Колмогорова является множество элементарных событий Ω. Класс $\mathcal{S}$ подмножеств множества элементарных событий Ω называется *алгеброй событий*, если он замкнут относительно операций объединения (т. е. если $A, B \in \mathcal{S}$, то $A \cup B \in \mathcal{S}$) и перехода к противоположному событию (т. е. если $A \in \mathcal{S}$, то $\overline{A} \in \mathcal{S}$).

Следующая теорема говорит о правомерности совершения над событиями (как множествами из алгебры) любых операций, известных нам из **п. 2.2.1**, а не только операций объединения и взятия противоположного события.

Теорема о замкнутости алгебры относительно пересечения и дополнения. *Любая алгебра замкнута относительно операций пересечения и дополнения.*

Доказательство. Действительно, пусть $\mathcal{S}$ — алгебра, $A, B \in \mathcal{S}$, тогда $\overline{A}, \overline{B} \in \mathcal{S}$ (так как $\mathcal{S}$ замкнута относительно операции взятия противоположного события). Так как $\overline{A}, \overline{B} \in \mathcal{S}$, то $\overline{A} \cup \overline{B} \in \mathcal{S}$ (так как $\mathcal{S}$ замкнута относительно операции объединения). Так как $\overline{A} \cup \overline{B} \in \mathcal{S}$, то $\overline{\overline{A} \cup \overline{B}} \in \mathcal{S}$ (так как $\mathcal{S}$ замкнута относительно операции взятия противоположного события). Но согласно правилу де Моргана (2.2.10) $A \cap B = \overline{\overline{A} \cup \overline{B}}$, поэтому $A \cap B = \overline{\overline{A} \cup \overline{B}} \in \mathcal{S}$, т. е. $\mathcal{S}$ замкнута относительно операции пересечения.

Если $A, B \in \mathcal{S}$, тогда $\overline{B} \in \mathcal{S}$ (так как $\mathcal{S}$ замкнута относительно операции взятия противоположного события). Так как $A, \overline{B} \in \mathcal{S}$, то $A \cap \overline{B} \in \mathcal{S}$ (поскольку замкнута относительно операции пересечения, как только что показано). Но $A \setminus B = A \cap \overline{B}$ [по формуле (2.2.8)], поэтому $A \setminus B \in \mathcal{S}$, т. е. $\mathcal{S}$ замкнута относительно операции дополнения. Справедливость теоремы доказана. ❑

Любая непустая алгебра содержит невозможное событие ($\varnothing \in \mathcal{S}$, поскольку если некоторое событие $A \in \mathcal{S}$, то $\overline{A} \in \mathcal{S}$, значит, $\varnothing = A \cap \overline{A} \in \mathcal{S}$), а также достоверное событие ($\Omega \in \mathcal{S}$, поскольку $\Omega = \overline{\varnothing}$).

Алгебра событий $\mathcal{S}$ называется σ-*алгеброй*, если объединение и пересечение бесконечной (счетной) последовательности событий $A_1, A_2, \ldots \in \mathcal{S}$ также принадлежат $\mathcal{S}$:

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{S}, \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{S}.$$

Страшных слов «алгебра» и «σ-алгебра» бояться не нужно, наличие алгебры означает просто, что над событиями можно проводить операции объединения и пересечения (а «сигма» значит, что можно проводить бесконечно много пересечений и объединений).

Класс событий $\{\varnothing;\Omega\}$, очевидно, является σ-алгеброй. Такая σ-алгебра называется *минимальной*. Множество всех подмножеств Ω также, очевидно, образует σ-алгебру. Эта σ-алгебра называется *максимальной*.

При аксиоматическом построении теории вероятностей в каждом конкретном пространстве элементарных событий Ω выделяется σ-алгебра событий $\mathcal{S}$, и для каждого события $A\in\mathcal{S}$ задается *вероятность* (или *вероятностная мера*) — числовая функция, определенная на σ-алгебре событий $\mathcal{S}$ и удовлетворяющая следующим **аксиомам.**

Аксиома неотрицательности вероятности:

$$\textit{для всех } A\in\mathcal{S}: \quad \mathbf{P}(A)\geqslant 0. \tag{2.2.21}$$

Аксиома нормированности вероятности:

$$\mathbf{P}(\Omega)=1. \tag{2.2.22}$$

Аксиома аддитивности вероятности:

$$\textit{для всех } A,B\in\mathcal{S},\ \textit{таких что } A\cap B=\varnothing:$$
$$\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B). \tag{2.2.23}$$

Замечание. Для доказательства ряда теоретических положений вместо аксиомы аддитивности вероятности требуется ее расширенный вариант, **аксиома счетной аддитивности вероятности:** для всех $A_1, A_2, \ldots\in\mathcal{S}$, таких что при $i\neq j$ $A_i\cap A_j=\varnothing$,

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mathbf{P}(A_i). \tag{2.2.24}$$

Каждая определенная теоретико-вероятностная схема задается тройкой $\{\Omega, \mathcal{S}, \mathbf{P}\}$, где Ω — конкретное пространство элементарных событий, $\mathcal{S}$ — σ-алгебра событий, определенная на Ω, $\mathbf{P}$ — вероятность, заданная на σ-алгебре $\mathcal{S}$. Тройка $\{\Omega, \mathcal{S}, \mathbf{P}\}$ называется *вероятностным пространством.*

В предыдущих пунктах мы убедились в справедливости аксиом вероятности (2.2.21)—(2.2.23) для классической вероятностной схемы, дискретной вероятностной схемы с неодинаково возможными исходами, схемы геометрических вероятностей, а также для схемы статистических вероятностей.

Из аксиом (2.2.21)—(2.2.23) следуют следующие **свойства вероятности.**

Теорема о вероятности противоположного события. *Для всех* $A\in\mathcal{S}$

$$\mathbf{P}(\overline{A})=1-\mathbf{P}(A). \tag{2.2.25}$$

Доказательство. Поскольку $\Omega=A\cup\overline{A}$ и $A\cap\overline{A}=\varnothing$, из аксиомы аддитивности (2.2.23) следует, что $\mathbf{P}(\Omega)=\mathbf{P}(A)+\mathbf{P}(\overline{A})$. Но по аксиоме нормированности (2.2.22) $\mathbf{P}(\Omega)=1$, откуда $\mathbf{P}\{\overline{A}\}=1-\mathbf{P}\{A\}$, что и требовалось доказать. ❑

Из только что доказанной теоремы следует, в частности,

Теорема о вероятности невозможного события.

$$\mathbf{P}(\varnothing)=0. \tag{2.2.26}$$

Теорема о монотонности вероятности. *Для всех* $A,B\in\mathcal{S}$, *таких что* $A\subseteq B$,

$$\mathbf{P}(B)=\mathbf{P}(A)+\mathbf{P}(B\backslash A); \qquad \mathbf{P}(A)\leqslant\mathbf{P}(B). \tag{2.2.27}$$

Доказательство. Поскольку $A \subseteq B$, событие B можно представить в виде $B = A \cup (B \setminus A)$, причем $A \cap (B \setminus A) = \varnothing$, поэтому по аксиоме аддитивности (2.2.23) $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A)$. Но по аксиоме неотрицательности (2.2.21) $\mathbf{P}(B \setminus A) \geqslant 0$, откуда следует, что $\mathbf{P}(B) \geqslant \mathbf{P}(A)$ или $\mathbf{P}(A) \leqslant \mathbf{P}(B)$, что и требовалось доказать. ❑

Теорема об ограниченности вероятности. *Для всех* $A \in \mathcal{S}$

$$0 \leqslant \mathbf{P}(A) \leqslant 1. \tag{2.2.28}$$

Доказательство. Поскольку произвольное событие $A \in \mathcal{S}$ влечет достоверное событие (т. е. $A \in \Omega$), то по теореме о монотонности (2.2.27) $\mathbf{P}(A) \leqslant \mathbf{P}(\Omega)$ или $\mathbf{P}(A) \leqslant 1$ (так как $\mathbf{P}(\Omega) = 1$ по аксиоме нормированности). По аксиоме неотрицательности (2.2.21) $\mathbf{P}(A) \geqslant 0$. Таким образом, $0 \leqslant \mathbf{P}(A) \leqslant 1$, что и требовалось доказать. ❑

В случае произвольных событий A и B (в том числе, совместных) для вычисления вероятности объединения событий A и B нельзя пользоваться аксиомой аддитивности вероятности (2.2.23). В этом случае применяется

Теорема сложения вероятностей. *Для всех* $A, B \in \mathcal{S}$

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B). \tag{2.2.29}$$

Доказательство. Событие $B = (B \setminus A) \cup (A \cap B)$, при этом события $B \setminus A$ и $A \cap B$ несовместны (так как $(B \setminus A) \cap (A \cap B) = \varnothing$), поэтому по аксиоме аддитивности (2.2.23) $\mathbf{P}(B) = \mathbf{P}(B \setminus A) + \mathbf{P}(A \cap B)$, значит, $\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A \cap B)$. Далее, $A \cup B = A \cup (B \setminus A)$, при этом события A и $B \setminus A$ также несовместны (так как $A \cap (B \setminus A) = \varnothing$), поэтому по аксиоме аддитивности (2.2.23) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A)$. Подставляя в последнюю формулу выражение для $\mathbf{P}(B \setminus A)$, получаем окончательно: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, что и требовалось доказать. ❑

Первая четкая формулировка теоремы сложения вероятностей принадлежит А. де Муавру, хотя в частных случаях эта теорема была известна и его предшественникам.

Неравенство треугольника. *Для всех* $A, B \in \mathcal{S}$

$$\mathbf{P}(A \cup B) \leqslant \mathbf{P}(A) + \mathbf{P}(B). \tag{2.2.30}$$

Доказательство. По теореме сложения вероятностей (2.2.29) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. Но по аксиоме неотрицательности (2.2.21) $\mathbf{P}(A \cap B) \geqslant 0$, поэтому $\mathbf{P}(A \cup B) \leqslant \mathbf{P}(A) + \mathbf{P}(B)$, что и требовалось доказать. ❑

Пример 2.2.12. Известно, что курс евро к рублю может возрасти с вероятностью 0,55, а курс доллара к рублю может возрасти с вероятностью 0,35. Вероятность того, что возрастут оба курса, составляет 0,3. Найдем вероятность того, что курс евро или доллара по отношению к рублю возрастет.

Пусть событие $A_€$ состоит в том, курс евро к рублю возрастет, а событие $A_\$$ — в том, что курс доллара к рублю возрастет. Тогда по условию $\mathbf{P}(A_€) = 0{,}55$, $\mathbf{P}(A_\$) = 0{,}35$, $\mathbf{P}(A_€ \cap A_\$) = 0{,}3$. Вероятность того, что курс евро или доллара по отношению к рублю возрастет, по теореме сложения вероятностей составляет $\mathbf{P}(A_€ \cup A_\$) = \mathbf{P}(A_€) + \mathbf{P}(A_\$) - \mathbf{P}(A_€ \cap A_\$) = 0{,}55 + 0{,}35 - 0{,}3 = 0{,}6$. ❑

Обобщенная теорема сложения вероятностей. *Для всех* $A_1, A_2, \text{K}, A_n \in \mathcal{S}$

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \tag{2.2.31}$$

$$= \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{1 \leqslant i < j \leqslant n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \leqslant i < j < k \leqslant n} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right).$$

Доказательство. Будем проводить доказательство по методу математической индукции. Справедливость теоремы при $n = 1$ очевидна, а при $n = 2$ доказана в теореме сложения веро-

ятностей (2.2.28). Будем считать, что теорема верна для $n = m - 1$. Докажем справедливость теоремы для $n = m$. Поскольку для любых $(m-1)$ событий теорема верна, то

$$\mathbf{P}\left(\bigcup_{i=2}^{m} A_i\right) = \sum_{i=2}^{m} \mathbf{P}(A_i) - \sum_{2\leqslant i<j\leqslant m} \mathbf{P}(A_i \cap A_j) + \sum_{2\leqslant i<j<k\leqslant m} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{m-2} \mathbf{P}\left(\bigcap_{i=2}^{m} A_i\right),$$

$$\mathbf{P}\left(\bigcup_{i=2}^{m} (A_1 \cap A_i)\right) = \sum_{i=2}^{m} \mathbf{P}(A_1 \cap A_i) - \sum_{2\leqslant i<j\leqslant m} \mathbf{P}\big((A_1 \cap A_i) \cap (A_1 \cap A_j)\big) +$$

$$+ \sum_{2\leqslant i<j<k\leqslant m} \mathbf{P}\big((A_1 \cap A_i) \cap (A_1 \cap A_j) \cap (A_1 \cap A_k)\big) - \cdots + (-1)^{n-2} \mathbf{P}\left(\bigcap_{i=2}^{m} (A_1 \cap A_i)\right) = \sum_{i=2}^{m} \mathbf{P}(A_1 \cap A_i) -$$

$$- \sum_{2\leqslant i<j\leqslant m} \mathbf{P}(A_1 \cap A_i \cap A_j) + \sum_{2\leqslant i<j<k\leqslant m} \mathbf{P}(A_1 \cap A_i \cap A_j \cap A_k) - \cdots + (-1)^{m-2} \mathbf{P}\left(\bigcap_{i=1}^{m} A_i\right).$$

Далее, по теореме сложения вероятностей (2.2.29) $\mathbf{P}\left(\bigcup_{i=1}^{m} A_i\right) = \mathbf{P}(A_1) + \mathbf{P}\left(\bigcup_{i=2}^{m} A_i\right) - \mathbf{P}\left(A_1 \cap \left(\bigcup_{i=2}^{m} A_i\right)\right)$, но по свойству дистрибутивности пересечения событий относительно объединения (1.2.9) $A_1 \cap \left(\bigcup_{i=2}^{m} A_i\right) = \bigcup_{i=2}^{m} (A_1 \cap A_i)$, поэтому $\mathbf{P}\left(\bigcup_{i=1}^{m} A_i\right) = \mathbf{P}(A_1) + \mathbf{P}\left(\bigcup_{i=2}^{m} A_i\right) - \mathbf{P}\left(\bigcup_{i=2}^{m} (A_1 \cap A_i)\right)$. Подставляя сюда выражения для $\mathbf{P}\left(\bigcup_{i=2}^{m} A_i\right)$ и $\mathbf{P}\left(\bigcup_{i=2}^{m} (A_1 \cap A_i)\right)$, получим:

$$\mathbf{P}\left(\bigcup_{i=1}^{m} A_i\right) = \mathbf{P}\{A_1\} + \mathbf{P}\left(\bigcup_{i=2}^{m} A_i\right) - \mathbf{P}\left(\bigcup_{i=2}^{m} (A_1 \cap A_i)\right) = \mathbf{P}(A_1) + \sum_{i=2}^{m} \mathbf{P}(A_i) - \sum_{2\leqslant i<j\leqslant m} \mathbf{P}(A_i \cap A_j) +$$

$$+ \sum_{2\leqslant i<j<k\leqslant m} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{m-2} \mathbf{P}\left(\bigcap_{i=2}^{m} A_i\right) - \sum_{i=2}^{m} \mathbf{P}(A_1 \cap A_i) + \sum_{2\leqslant i<j\leqslant m} \mathbf{P}(A_1 \cap A_i \cap A_j) -$$

$$- \sum_{2\leqslant i<j<k\leqslant m} \mathbf{P}(A_1 \cap A_i \cap A_j \cap A_k) + \cdots - (-1)^{m-2} \mathbf{P}\left(\bigcap_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} \mathbf{P}(A_i) - \sum_{1\leqslant i<j\leqslant m} \mathbf{P}(A_i \cap A_j) +$$

$$+ \sum_{1\leqslant i<j<k\leqslant m} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{m-1} \mathbf{P}\left(\bigcap_{i=1}^{m} A_i\right),$$

значит, формула (2.2.30) справедлива для любого $n \in \mathbb{N}$, что и требовалось доказать. ❑

В случае $n = 3$ обобщенная теорема сложения вероятностей принимает следующий вид:

$$\textit{для всех } A, B, C \in \mathcal{S} \quad \mathbf{P}(A \cup B \cup C) = \tag{2.2.32}$$
$$= \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C).$$

Пример 2.2.13. Результаты опроса 1000 случайно отобранных молодых людей таковы:

работают	811 чел.;
учатся	518 чел.;
работают и учатся одновременно	356 чел.;
проживают в Москве	752 чел.;
из москвичей:	
работают	570 чел.;
учатся	348 чел.;
работают и учатся одновременно	297 чел.

Покажем, эта информация не может быть верной.

Пусть событие A состоит в том, что случайно выбранный молодой человек работает, событие B — в том, что случайно выбранный молодой человек про-

живает в Москве, событие C — в том, что случайно выбранный молодой человек учится. Так как $n = 1000$ достаточно велико, можно воспользоваться статистическим подходом к определению вероятности (1.3.20): $\mathbf{P}(A) \approx 0{,}811$, $\mathbf{P}(B) \approx 0{,}752$, $\mathbf{P}(C) \approx 0{,}518$, $\mathbf{P}(A \cap B) \approx 0{,}570$, $\mathbf{P}(B \cap C) \approx 0{,}348$, $\mathbf{P}(A \cap C) \approx 0{,}356$, $\mathbf{P}(A \cap B \cap C) \approx 0{,}297$. По формуле (2.2.31) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C) \approx 1{,}104 > 1$. Но вероятность не может быть больше единицы, следовательно, в данной информации содержится ошибка. ❑

Пример 2.2.14. Петя написал n электронных писем n своим друзьям, и в поле «Кому» в каждом письме вставил электронный адрес, который случайным образом выбрав из списка адресов этих n друзей. Найдем вероятность того, что хотя бы один из адресатов получит то письмо, которое предназначалось ему.

Опыт можно интерпретировать, как перестановки чисел 1, 2, ..., n. Общее число исходов опыта $N = P_n = n!$ (по формуле 2.1.4). Пусть событие A_i состоит в том, что на i-м месте стоит число i ($i = 1, 2, ..., n$), а остальные числа расположены в произвольном порядке. Тогда число исходов опыта, благоприятствующих наступлению события A_i, равно $M_{A_i} = (n-1)!$ (числу перестановок $(n-1)$ чисел). Поэтому согласно классической формуле вероятности вероятность события A_i равна $\mathbf{P}(A_i) = M_{A_i} / N = (n-1)!/n! = 1/n$.

Событие $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$ состоит в том, что на местах с номерами i_1, i_2, ..., i_k стоят числа i_1, i_2, ..., i_k, а остальные числа расположены на произвольных местах. Согласно классической формуле вероятности $\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = (n-k)!/n!$ (естественно, при $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$).

Пусть $B_k = \bigcup\limits_{1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n} \left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)$, тогда все объединяемые события, очевидно, несовместны, и число таких событий равно числу сочетаний из n по k.

Поэтому вероятность события B_k равна

$$\mathbf{P}(B_k) = \mathbf{P}\left(\bigcup_{1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n} \left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)\right) =$$

$$= \sum_{1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = C_n^k \frac{(n-k)!}{n!} = \frac{\cancel{n!}}{k!\,\cancel{(n-k)!}} \cdot \frac{\cancel{(n-k)!}}{\cancel{n!}} = \frac{1}{k!}.$$

Событие A, состоящее в том, что хотя бы одно число находятся на своем месте, эквивалентно объединению событий $A_1 \cup A_2 \cup \cdots \cup A_n$, поэтому по обобщенной теореме сложения вероятностей

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{1 \leqslant i_1 \leqslant n} \mathbf{P}(A_{i_1}) - \sum_{1 \leqslant i_1 < i_2 \leqslant n} \mathbf{P}(A_{i_1} \cap A_{i_2}) + \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots +$$

$$+ (-1)^{n-1} \mathbf{P}\left(\bigcap_{k=1}^{n} A_{i_k}\right) = (-1)^0 \mathbf{P}(B_1) + (-1)^1 \mathbf{P}(B_2) + (-1)^2 \mathbf{P}(B_3) + \cdots + (-1)^{n-1} \mathbf{P}(B_n) =$$

$$= (-1)^0 \frac{1}{1!} + (-1)^1 \frac{1}{2!} + (-1)^2 \frac{1}{3!} + \cdots + (-1)^{n-1} \frac{1}{n!} = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!}.$$

Найдем теперь предел, к которому стремится эта вероятность при $n \to \infty$.

Вспомнив разложение функции e^x в ряд Маклорена:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots,$$

замечаем, что

$$e^{-1} = 1 + \frac{-1}{1!} + \frac{(-1)^2}{2!} + \frac{(-1)^3}{3!} + \cdots + \frac{(-1)^n}{n!} + \cdots = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} + \cdots,$$

откуда искомый предел равен

$$\mathbf{P}(A^{(\infty)}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!} + \cdots = 1 - e^{-1} = 1 - \frac{1}{e} \approx 1 - \frac{1}{2{,}718} = 0{,}632. \text{ ❑}$$

Задачи

85. В партии 100 изделий, из которых шесть имеют дефекты. Партия произвольно разделена на две равные части, которые отправлены двум потребителям. Найдите вероятности следующих событий: *а*) все бракованные изделия достанутся одному потребителю; *б*) бракованные изделия достанутся обоим потребителям поровну.

86. Петя ищет работу. Он побывал на собеседованиях в банке и страховой компании. Вероятность найти работу в банке Петя оценивает в 0,5, а в страховой компании — в 0,6. Кроме того, он рассчитывает, что с вероятностью 0,3 ему поступят предложения от двух организаций сразу. Найдите вероятность того, что Петя получит хотя бы одно предложение работы.

87. Менеджер по кадрам разместил на кадровом портале объявление о том, что банку требуется начальник отдела долговых обязательств, и получил 300 резюме. Из прошлого опыта известно, что вероятность того, что претендент имеет бизнес-образование, равна 0,3, вероятность того, что претендент имеет опыт руководящей работы в банке, — 0,7, а вероятность того, что претендент имеет и бизнес-образование, и опыт руководящей работы, — 0,2. Оцените количество претендентов, имеющих опыт руководящей работы или бизнес-образование.

88. Известны вероятности дополнительной потребности банка в специалистах по анализу данных на предстоящие два года:

Число инженеров	<100	$100-199$	$200-299$	$300-399$	$400-499$	$\geqslant 500$
Вероятность	0,10	0,15	0,30	0,30	0,10	0,05

Найдите вероятности следующих событий: *а*) на протяжении предстоящих двух лет банку дополнительно потребуется не менее 400 специалистов по анализу данных; *б*) на протяжении предстоящих двух лет банку дополнительно потребуется по меньшей мере 200, но не более 399 специалистов по анализу данных.

89. Событие A состоит в том, что потенциальный покупатель увидел рекламу товара на платформе *VK*, а событие B — в том, что он увидел рекламу на платформе *Facebook*. Известно, что $\mathbf{P}(A) > 0{,}8$, $\mathbf{P}(B) > 0{,}4$. Проверьте справедливость следующих утверждений: *а*) A и B несовместны; *б*) A и B противоположны; *в*) $\mathbf{P}(A \cap B) > 0{,}2$.

90. Петя — староста группы. Когда деканат попросил его подать сведения о студентах своей группы, Петя по памяти составил следующую записку: «*В группе* 45 *студентов, в том числе* 25 *юношей;* 30 *студентов учатся на* "*хорошо*" *и* "*отлично*", *в том числе* 16 *юношей;* 28 *студентов занимаются спортом, в их числе* 18 *юношей и* 17 *студентов, успевающих на* "*хорошо*" *и* "*отлично*"*;* 15 *юношей учатся на* "*хорошо*" *и* "*отлично*" *и занимаются спортом*». Сотрудники деканата сразу определили, что Петя ошибся, и попросили его более аккуратно подойти к выполнению поручения. Определить, как сотрудникам деканата удалось «поймать» Петю?

91. Пусть A, B, C — произвольные события. Расположите следующие события в порядке возрастания их вероятностей: $A \cup C$, $A \backslash B$, $A \backslash (B \backslash C)$, Ω, $A \cup B \cup C$, $A \backslash B \backslash C$, $\varnothing$.

92. В условиях задачи **65** найдите вероятности следующих событий: *а*) угадать не менее трех чисел; *б*) угадать хотя бы одно число.

93. Пусть $A \subseteq B \cup C$ и $B \cap C = \varnothing$. Проверьте справедливость следующих утверждений: *а*) $\mathbf{P}(A \cup B) = \mathbf{P}(B)$; *б*) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(B) + \mathbf{P}(C)$; *в*) $\mathbf{P}(\bar{A}) \geqslant \mathbf{P}(\bar{B})$; *г*) $\mathbf{P}(\bar{A}) \geqslant \mathbf{P}(\bar{B}) + \mathbf{P}(\bar{C})$.

94. В условиях примера 2.2.14 найдите вероятность того, что по крайней мере m адресатов получат письма, предназначенные им ($0 \leqslant m \leqslant n$).

§ 2.3. Условные вероятности

2.3.1. Условная вероятность. Независимость событий

Как уже было отмечено, говорить о вероятности наступления какого-либо события как о мере возможности наступления этого события можно лишь при выполнении определенного комплекса условий опыта. Так, если к комплексу условий, при которых изучалась вероятность наступления события A, добавить условие наступления события B, получим другое значение вероятности $\mathbf{P}(A|B)$ — *вероятность наступления события A при условии, что событие B произошло*, или *условную вероятность события A при условии B*, которая равна по определению

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}. \tag{2.3.1}$$

Замечание. Для корректности определения (2.3.1) необходимо, чтобы событие B имело ненулевую вероятность: $\mathbf{P}(B) > 0$. Если $\mathbf{P}(B) = 0$, то условная вероятность $\mathbf{P}(A|B)$ не определена.

Вероятность $\mathbf{P}(A)$, в отличие от условной вероятности $\mathbf{P}(A|B)$, называется *безусловной*.

Из определения условной вероятности (2.3.1) следуют ее **свойства.**

Неотрицательность условной вероятности. *Для всех событий A и любого события B, такого что* $\mathbf{P}(B) > 0$,

$$\mathbf{P}(A|B) \geqslant 0. \tag{2.3.2}$$

Нормированность условной вероятности. *Для любого события B, такого что* $\mathbf{P}(B) > 0$,

$$\mathbf{P}(\Omega|B) = 1. \tag{2.3.3}$$

Аддитивность условной вероятности. *Для всех событий A и C, таких что $A \cap C = \varnothing$, и любого события B, такого что* $\mathbf{P}(B) > 0$,

$$\mathbf{P}(A \cup C|B) = \mathbf{P}(A|B) + \mathbf{P}(C|B). \tag{2.3.4}$$

Доказательство. Поскольку для любых событий A и B вероятность $\mathbf{P}(A \cap B) \geqslant 0$ по аксиоме неотрицательности вероятности (2.2.21), а $\mathbf{P}(B) > 0$ по условию, то

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \geqslant 0,$$

что доказывает справедливость свойства неотрицательности условной вероятности (2.3.2).

По свойству пересечения событий (2.2.6) $\Omega \cap B = B$, поэтому $\mathbf{P}(\Omega \cap B) = \mathbf{P}(B)$ и

$$\mathbf{P}(\Omega|B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1,$$

что доказывает справедливость свойства нормированности условной вероятности (2.3.3).

Условная вероятность $\mathbf{P}((A \cup C)|B) = \frac{\mathbf{P}((A \cup C) \cap B)}{\mathbf{P}(B)}$. По свойству коммутативности пересечения событий (2.2.4) $(A \cup C) \cap B = B \cap (A \cup C)$, а по свойству дистрибутивности пересечения событий относительно объединения(2.2.9) $B \cap (A \cup C) = (B \cap A) \cup (B \cap C)$. Так как $A \cap C = \varnothing$, очевидно, что

$(B\cap A)\cap(B\cap C)=\varnothing$, поэтому по аксиоме аддитивности вероятности (2.2.23) $\mathbf{P}((B\cap A)\cup(B\cap C))=$ $=\mathbf{P}(B\cap A)+\mathbf{P}(B\cap C)$. Окончательно получим, еще раз воспользовавшись свойством коммутативности пересечения событий (2.2.4), что $\mathbf{P}((A\cup C)\cap B)=\mathbf{P}(A\cap B)+\mathbf{P}(C\cap B)$, поэтому

$$\mathbf{P}\big((A\cup C)\,|\,B\big)=\frac{\mathbf{P}((A\cup C)\cap B)}{\mathbf{P}(B)}=\frac{\mathbf{P}(A\cap B)+\mathbf{P}(C\cap B)}{\mathbf{P}(B)}=\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)}+\frac{\mathbf{P}(C\cap B)}{\mathbf{P}(B)}=\mathbf{P}(A\,|\,B)+\mathbf{P}(C\,|\,B),$$

что доказывает справедливость свойства аддитивности условной вероятности (2.3.4). ❑

Замечание. Пусть $\mathbf{P}(B)>0$. Если σ-алгебру всех множеств вида $A\cap B$, где $A\in\mathcal{S}$, обозначить $\mathcal{S}_B$, а условную вероятность $\mathbf{P}(A|B)$ обозначить $\mathbf{P}_B(A)=\mathbf{P}(A|B)$, можно заметить, что $\mathbf{P}_B(B)=\mathbf{P}(B|B)=1$, и поскольку условная вероятность $\mathbf{P}_B(\cdot)=\mathbf{P}(\cdot|B)$ является числовой функцией, определенной на σ-алгебре событий $\mathcal{S}_B$ и удовлетворяющей аксиомам неотрицательности, нормированности и аддитивности, эта функция $\mathbf{P}_B(\cdot)$ определяет на σ-алгебре событий $\mathcal{S}_B$ вероятностную меру. Поэтому $\{B,\mathcal{S}_B,\mathbf{P}_B\}$ также является вероятностным пространством (рис. 2.3.1). Очевидно, $\{\Omega,\mathcal{S},\mathbf{P}_B\}$ также является вероятностным пространством.



а) *б*)

Рис. 2.3.1. *Вычисление вероятностей в вероятностном пространстве* $\{\Omega,\mathcal{S},\mathbf{P}\}$ *(а) и в вероятностном пространстве* $\{B,\mathcal{S}_B,\mathbf{P}_B\}$ *(б)*

Из определения условной вероятности (2.3.1) непосредственно следует также

Формула умножения вероятностей. *Если* $\mathbf{P}(A)>0$, $\mathbf{P}(B)>0$, *то*

$$\mathbf{P}(A\cap B)=\mathbf{P}(A)\mathbf{P}(B\,|\,A)=\mathbf{P}(B)\mathbf{P}(A\,|\,B). \tag{2.3.5}$$

Формулу умножения вероятностей (2.3.5) легко обобщить на случай произвольного конечного числа событий:

$$\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1\cap A_2)\cdots\mathbf{P}(A_n|A_1\cap A_2\cap\cdots\cap A_{n-1}). \tag{2.3.6}$$

Доказательство. Будем проводить доказательство по методу математической индукции. Справедливость теоремы при $n=1$ очевидна, а при $n=2$ доказана в формуле умножения вероятностей (2.3.5). Будем считать, что формула (2.3.6) верна для $n=m-1$, докажем ее справедливость для $n=m$. Поскольку для любых $m-1$ событий теорема верна, то $\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_{m-1})=$ $=\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1\cap A_2)\cdots\mathbf{P}(A_{m-1}|A_1\cap A_2\cap\cdots\cap A_{m-2})$. По формуле умножения вероятностей (2.3.6) $\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_m)=\mathbf{P}((A_1\cap A_2\cap\cdots\cap A_{m-1})\cap A_m)=\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_{m-1})\mathbf{P}(A_m|A_1\cap A_2\cap\cdots\cap A_{m-1})$. Подставляя сюда выражение для $\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_{m-1})$, получим: $\mathbf{P}(A_1\cap A_2\cap\cdots\cap A_m)=$ $=\mathbf{P}((A_1\cap A_2\cap\cdots\cap A_{m-1})\cap A_m)=\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1\cap A_2)\cdots\mathbf{P}(A_{m-1}|A_1\cap A_2\cap\cdots\cap A_{m-2})\mathbf{P}(A_m|A_1\cap A_2\cap\cdots\cap A_{m-1})$, что и требовалось доказать. ❑

Говорят, что события A и B являются *независимыми*, если

$$\mathbf{P}(A\cap B)=\mathbf{P}(A)\mathbf{P}(B). \tag{2.3.7}$$

Очевидно, при $\mathbf{P}(B)>0$ формула (2.3.7) означает, что условная вероятность события A при условии B совпадает с безусловной вероятностью события A:

$$\mathbf{P}(A\,|\,B)=\mathbf{P}(A). \tag{2.3.8}$$

Понятия условной вероятности, зависимых и независимых событий впервые сформулировал А. де Муавр, а формула для вычисления условной вероятности (2.3.1) была предложена Т. Байесом.

Пример 2.3.1. Известно, что $\mathbf{P}(A) = 0{,}8$, $\mathbf{P}(B) = 0{,}6$, $\mathbf{P}(A \cup B) = 0{,}9$.

Найдем $\mathbf{P}(A \cap B)$, $\mathbf{P}(A|B)$, $\mathbf{P}(B|A)$. По теореме сложения вероятностей (2.2.29) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, откуда $\mathbf{P}(A \cap B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cup B) = 0{,}8 + 0{,}6 - 0{,}9 = 0{,}5$.

По определению условной вероятности $\mathbf{P}(A|B) = \mathbf{P}(A \cap B) / \mathbf{P}(B) = 0{,}5 / 0{,}6 = 5/6$, $\mathbf{P}(B|A) = \mathbf{P}(A \cap B) / \mathbf{P}(A) = 0{,}5 / 0{,}8 = 5 / 8$.

Выясним теперь, зависимы ли события A и B. Для того, чтобы события A и B были независимыми, необходимо выполнение равенства (2.3.7): $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. В нашем случае $\mathbf{P}(A \cap B) = 0{,}5$, $\mathbf{P}(A)\mathbf{P}(B) = 0{,}8 \cdot 0{,}6 = 0{,}48 \neq 0{,}5$, т. е. равенство (2.3.7) не выполняется, значит, события A и B являются зависимыми. ❑

Пример 2.3.2. Среди сотрудников некоторого банка 23% получают высокую заработную плату. При этом 40% сотрудников банка — женщины, а 8% всех сотрудников — женщины, получающие высокую заработную плату. Проверим, существует ли в этом банке дискриминация женщин в оплате труда.

По условию вероятность события A, состоящего в том, что случайно выбранный сотрудник получает высокую заработную плату, равна (согласно статистическому подходу к определению вероятности) $\mathbf{P}(A) = 0{,}23$. Пусть событие B состоит в том, что случайно выбранный сотрудник банка является женщиной, тогда если $\mathbf{P}(A|B)$ окажется меньше безусловной вероятности $\mathbf{P}(A)$, то это и будет означать дискриминацию женщин в оплате труда.

По определению условной вероятности (2.3.1) $\mathbf{P}(A|B) = \mathbf{P}(A \cap B) / \mathbf{P}(B) = 008/0{,}40 = 0{,}20 < 0{,}23 = \mathbf{P}(A)$, т. е. действительно можно утверждать, что женщины, работающие в данном банке, имеют несколько меньшую вероятность получать высокую заработную плату, чем мужчины. ❑

Теорема о независимости пар противоположных событий. *Если события A и B независимы, то пары событий $\overline{A}$ и B, A и $\overline{B}$, $\overline{A}$ и $\overline{B}$ также независимы.*

Доказательство. Достаточно показать независимость событий $\overline{A}$ и B. Любое событие B можно представить в виде $B = B \cap \Omega = B \cap (A \cup \overline{A}) = (B \cap A) \cup (B \cap \overline{A}) = (A \cap B) \cup (\overline{A} \cap B)$ [здесь мы последовательно воспользовались свойствами (2.2.6), (2.2.3), (2.2.9), (2.2.4)], причем $(B \cap A) \cap (B \cap \overline{A}) = \varnothing$, поэтому по аксиоме аддитивности вероятности (2.2.23) $\mathbf{P}(B) = \mathbf{P}((A \cap B) \cup (\overline{A} \cap B)) = \mathbf{P}(A \cap B) + \mathbf{P}(\overline{A} \cap B)$, откуда $\mathbf{P}(\overline{A} \cap B) = \mathbf{P}(B) - \mathbf{P}(A \cap B)$. События A и B независимы, значит, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, поэтому $\mathbf{P}(\overline{A} \cap B) = \mathbf{P}(B) - \mathbf{P}(A)\mathbf{P}(B) = (1 - \mathbf{P}(A))\mathbf{P}(B) = \mathbf{P}(\overline{A})\mathbf{P}(B)$ [в последнем равенстве мы воспользовались теоремой о вероятности противоположного события (2.2.25)]. Итак, $\mathbf{P}(\overline{A} \cap B) = \mathbf{P}(\overline{A})\mathbf{P}(B)$, значит, события $\overline{A}$ и B независимы по определению независимости событий (2.3.7), что и требовалось доказать. ❑

События $A_1, A_2, \ldots, A_n$ называются *независимыми в совокупности*, если для любого их подмножества $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ вероятность одновременного наступления событий $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ равна произведению безусловных вероятностей этих событий:

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_k}) \quad (k = 1, 2, \ldots, n). \tag{2.3.9}$$

Если условие (2.3.9) выполняется только при $k = 2$, то события $A_1, A_2, \ldots, A_n$ называются *попарно независимыми* (естественно, предполагается справедливость этого условия и при $k = 1$, т. е. выполнение тривиального равенства $\mathbf{P}(A_i) = \mathbf{P}(A_i)$ для всех $i = 1, 2, \ldots, n$).

Замечание. Следует обратить внимание на следующие факты:

- из условия независимости не следует несовместность;
- из условия несовместности следует зависимость;
- из условия независимости в совокупности следует попарная независимость, но из попарной независимости не следует независимость в совокупности.

Приведем соответствующий

Пример 2.3.3. Пусть в корзине находится четыре шара, на ощупь неразличимых: красный, желтый, зеленый и синий, и из этой корзины наудачу извлекается один шар. Пусть событие A состоит в том, что извлечен красный или желтый шар, B — красный или зеленый, C — красный или синий. Тогда $\mathbf{P}(A)=\mathbf{P}(B)=\mathbf{P}(C)=1/2$, $\mathbf{P}(A\cap B)=\mathbf{P}(B\cap C)=\mathbf{P}(A\cap C)=1/4=(1/2)^2$. Очевидно, события A и B совместны, но при этом независимы.

Пусть теперь в корзине находится два шара, на ощупь неразличимых: красный и желтый, и из этой корзины наудачу извлекается один шар. Пусть событие A состоит в том, что извлечен красный шар, B — желтый шар. Тогда $\mathbf{P}(A)=\mathbf{P}(B)=1/2$, но, очевидно, $\mathbf{P}(A\cap B)=0\neq(1/2)^2$. Таким образом, события A и B несовместны, но при этом зависимы.

Приведем, наконец, пример попарно независимых событий, не являющихся при этом независимыми в совокупности. Пусть в корзине находится четыре шара, на ощупь неразличимых: красный, желтый, зеленый и синий, и из этой корзины наудачу извлекается один шар. Пусть событие A состоит в том, что извлечен красный или желтый шар, B — красный или зеленый, C — красный или синий. Тогда $\mathbf{P}(A)=\mathbf{P}(B)=\mathbf{P}(C)=1/2$, $\mathbf{P}(A\cap B)=\mathbf{P}(B\cap C)=$ $=\mathbf{P}(A\cap C)=1/4=(1/2)^2$, но $\mathbf{P}(A\cap B\cap C)=1/4\neq(1/2)^3$. Очевидно, события A, B и C являются попарно независимыми, но при этом не являются независимыми в совокупности. ❑

Если события $A_1, A_2, \ldots, A_n$ независимы в совокупности, то

$$\mathbf{P}\left(\bigcup_{i=1}^{n}A_i\right)=\mathbf{P}(A_1\cup A_2\cup\cdots\cup A_n)= \tag{2.3.10}$$

$$=1-(1-\mathbf{P}(A_1))(1-\mathbf{P}(A_2))\cdots(1-\mathbf{P}(A_n))=1-\prod_{i=1}^{n}(1-\mathbf{P}(A_i)).$$

Доказательство. Поскольку события $A_1, A_2, \ldots, A_n$ независимы в совокупности, то по теореме о независимости противоположных событий события $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$ также независимы в совокупности, значит,

$$\mathbf{P}\left(\bigcap_{i=1}^{n}\overline{A}_i\right)=\mathbf{P}(\overline{A}_1\cap\overline{A}_2\cap\cdots\cap\overline{A}_n)=\mathbf{P}(\overline{A}_1)\mathbf{P}(\overline{A}_2)\cdots\mathbf{P}(\overline{A}_n)=\prod_{i=1}^{n}\mathbf{P}(\overline{A}_i).$$

Но по теореме о вероятности противоположного события (2.2.25) $\mathbf{P}(\overline{A}_i)=1-\mathbf{P}(A_i)$, поэтому $\mathbf{P}\left(\bigcap_{i=1}^{n}\overline{A}_i\right)=\prod_{i=1}^{n}(1-\mathbf{P}(A_i))$. Событие $\bigcap_{i=1}^{n}\overline{A}_i$ противоположно событию $\bigcup_{i=1}^{n}A_i$ [по правилу де Моргана (2.2.10)], поэтому по теореме о вероятности противоположного события (2.2.25) $\mathbf{P}\left(\bigcup_{i=1}^{n}A_i\right)=1-\mathbf{P}\left(\bigcap_{i=1}^{n}\overline{A}_i\right)=1-\prod_{i=1}^{n}(1-\mathbf{P}(A_i))$, что и требовалось доказать. ❑

Пример 2.3.4. Пусть события A, B и C попарно независимы, причем каждое из них имеет вероятность, отличную от нуля и единицы. Проверим, могут ли события $A\cap B$, $B\cap C$ и $A\cap C$ быть: *а*) попарно независимыми; *б*) независимыми в совокупности.

а) Рассмотрим следующие события в условиях задачи **77**: событие A, заключающееся в том, что утечка газа происходит ближе к станции А, чем к В, событие B, состоящее в том, что утечка расположена между 50-м и 150-м километрами, и событие C, состоящее в том, что утечка расположена между отметками 12,5 км и 62,5 км либо между 112,5 км и 162,5 км. Тогда согласно геометрическому подходу к определению вероятности $\mathbf{P}(A)=\mathbf{P}(B)=\mathbf{P}(C)=1/2$, $\mathbf{P}(A\cap B)=\mathbf{P}(B\cap C)=$ $=\mathbf{P}(A\cap C)=1/4$, значит, события A, B и C попарно независимы. При этом, очевидно, $\mathbf{P}(A\cap B\cap C)=$ $=1/16$, $\mathbf{P}((A\cap C)\cap(B\cap C))=\mathbf{P}(A\cap B\cap C)=1/16=\mathbf{P}(A)[\mathbf{P}(B)]^2\mathbf{P}(C)=\mathbf{P}((A\cap B)\cap(B\cap C))$. Аналогично получаем равенства $\mathbf{P}((A\cap B)\cap(A\cap C))=\mathbf{P}(A\cap B)\mathbf{P}(A\cap C)$ и $\mathbf{P}((B\cap C)\cap(A\cap C))=$ $=\mathbf{P}(B\cap C)\mathbf{P}(A\cap C)$, т. е. события A, B и C являются независимыми в совокупности.

б) Независимость в совокупности означает, во-первых, что $\mathbf{P}(A\cap B\cap C)=\mathbf{P}((A\cap B)\cap(B\cap C))=$ $=\mathbf{P}(A)[\mathbf{P}(B)]^2\mathbf{P}(C)$ и, во-вторых, что $\mathbf{P}(A\cap B\cap C)=\mathbf{P}((A\cap B)\cap(B\cap C)\cap(A\cap C))=$ $=[\mathbf{P}(A)]^2[\mathbf{P}(B)]^2[\mathbf{P}(C)]^2$. Поэтому $\mathbf{P}(A)[\mathbf{P}(B)]^2\mathbf{P}(C)=[\mathbf{P}(A)]^2[\mathbf{P}(B)]^2[\mathbf{P}(C)]^2$, но это равенство невозможно, так как вероятности событий A, B и C отличны от нуля и единицы. Поэтому события $A\cap B$, $B\cap C$ и $A\cap C$ не могут быть независимы в совокупности. ❑

Задачи

95. В группе из 1000 человек 452 имеют текущие счета, 336 — депозитные счета, а 302 — и текущие, и депозитные. Определите, являются ли события «обладание текущим счетом» и «обладание депозитным счетом» независимыми.

96. В тесте десять задач, пронумерованных от 1 до 10. Вероятность правильного решения задачи испытуемым одна и та же у всех задач. Какова вероятность того, что испытуемый верно решил задачу № 2 при условии, что он верно решил восемь задач из десяти?

97. Производитель автомобилей перед разработкой новой модели провел маркетинговый опрос потенциальных покупателей, в результате которого выяснилось, что 30% покупателей в основном оценивают автомобиль по его техническим характеристикам, 50% — по его дизайну, а 28% принимают во внимание и технические характеристики, и дизайн. Определите, являются ли предпочтения покупателей зависимыми или независимыми.

98. Злоумышленник решил угнать автомобиль, на котором установлены электронная сигнализация и механическая блокировка рычага переключения передач. Вероятность того, что угонщик справится с сигнализацией, составляет 0,2, а вероятность того, что он сломает блокировку рычага переключения передач, равна 0,1. Найдите вероятности следующих событий: *а*) автомобиль будет угнан; *б*) угонщик справится только с одной системой защиты.

99. Игральная кость бросается два раза. Найдите вероятность того, что оба раза появится одно и то же число очков.

100. Из корзины, содержащей три красных яблока и семь зеленых, вынимают по очереди два яблока. Найдите вероятность того, что оба они будут красными.

101. Жюри состоит из трех судей, выносящих решение независимо друг от друга: двое из них, каждый с вероятностью 0,8, принимают правильное решение, а третий для вынесения решения подбрасывает монету. Окончательное решение принимается большинством голосов. Найдите вероятность вынесения правильного решения.

102. Нефтедобывающая компания проводит буровые работы в трех различных местах А, Б и В. Вероятности успешного бурения в А, Б и В равны соответственно 0,5, 0,4 и 0,1. Предположив, что события, заключающиеся в успешности бурения в местах А, Б и В, независимы, вычислите вероятности следующих событий: *а*) хотя бы одно бурение окажется успешным; *б*) ровно одно бурение окажется успешным.

103. Маша пришла на устный экзамен, зная ответы на 20 вопросов программы из 25. Профессор задает три вопроса. Найти вероятности следующих событий: *а*) Маша ответит на все три вопроса; *б*) Маша ответит ровно на два вопроса; *в*) Маша ответит ровно на один вопрос; *г*) Маша ответит хотя бы на один вопрос; *д*) Маша не ответит ни на один вопрос.

104. Вероятность того, что кредитная карта находится в письменном столе, равна p, причем с равной вероятностью карта может находиться в любом из восьми ящиков стола. Ее владелец осмотрел семь ящиков и пока не нашел свою кредитную карту. Найдите вероятность того, что она находится в восьмом ящике.

105. В вершине A пятиугольника $ABCDE$ находится яблоко, а в вершине C сидит червяк. Ежедневно червяк переползает в одну из соседних вершин с равной вероятностью. Он не помнит своих предыдущих положений и может оказаться в той вершине, где он уже был. Достигнув вершины A, червяк располагается пообедать. Найдите вероятность того, что червяк доберется до яблока не более чем за семь дней.

106. Докажите, что формула $\mathbf{P}(A\,|\,B)+\mathbf{P}(\overline{A}\,|\,B)=1$ является верной, а равенства $\mathbf{P}(A\,|\,B)+\mathbf{P}(A\,|\,\overline{B})=1$ и $\mathbf{P}(A\,|\,B)+\mathbf{P}(\overline{A}\,|\,\overline{B})=1$ неверны.

107. Подбрасываются три игральные кости. Событие A состоит в том, что на первой и второй костях выпало одинаковое число очков, событие B — в том, что на второй и третьей костях выпало одинаковое число очков, событие C — в том, что на первой и третьей костях выпало одинаковое число очков. Проверьте, являются ли события A, B и C: *а*) попарно независимыми; *б*) независимыми в совокупности.

108. Пусть события A, B и C независимы в совокупности, причем каждое из них имеет вероятность, отличную от нуля и единицы. Проверьте, могут ли события $A\cap B$, $B\cap C$ и $A\cap C$ быть: *а*) попарно независимыми; *б*) независимыми в совокупности.

2.3.2. Формула полной вероятности

Если события $H_1, H_2, \ldots, H_n$ образуют полную группу, то для вычисления вероятности произвольного события A применяется сформулированная впервые П. Лапласом

Формула полной вероятности. *Если события* $H_1, H_2, \ldots, H_s \in \mathcal{S}$ *образуют полную группу и имеют положительные вероятности, то для любого события* $A \in \mathcal{S}$

$$\mathbf{P}(A)=\mathbf{P}(A\,|\,H_1)\mathbf{P}(H_1)+\mathbf{P}(A\,|\,H_2)\mathbf{P}(H_2)+\cdots+\mathbf{P}(A\,|\,H_n)\mathbf{P}(H_n)= \qquad (2.3.11)$$

$$=\sum_{i=1}^{n}\mathbf{P}(A\,|\,H_i)\mathbf{P}(H_i).$$

Доказательство. Очевидно, $A=(A\cap H_1)\cup(A\cap H_2)\cup\cdots\cup(A\cap H_n)$, причем события $(A\cap H_1)$, $(A\cap H_2)$, ..., $(A\cap H_n)$ несовместны (поскольку несовместны события H_1, H_2, ..., H_n). По аксиоме аддитивности вероятности (2.2.23) $\mathbf{P}(A)=\mathbf{P}(A\cap H_1)+\mathbf{P}(A\cap H_2)+\cdots+\mathbf{P}(A\cap H_n)$. Каждую из вероятностей $\mathbf{P}(A\cap H_i)$ раскроем по формуле умножения вероятностей (2.3.5): $\mathbf{P}(A\cap H_i)=\mathbf{P}(A\,|\,H_i)\mathbf{P}(H_i)$ ($i=1, 2, \ldots, n$). Подставляя, получаем: $\mathbf{P}(A)=\mathbf{P}(A\,|\,H_1)\mathbf{P}(H_1)+\mathbf{P}(A\,|\,H_2)\mathbf{P}(H_2)+\cdots+\mathbf{P}(A\,|\,H_n)\mathbf{P}(H_n)$, что и требовалось доказать. ❑

В соответствии с формулой (2.3.11) вероятность наступления события A может быть представлена как сумма произведений условных вероятностей события A при условии наступления событий H_i на безусловные вероятности этих событий H_i. Поскольку среди событий $H_1, H_2, \ldots, H_n$, образующих полную группу, в результате опыта должно наступить одно и только одно, эти события H_i называют *гипотезами* ($i=1, 2, \ldots, n$).

Замечание. Формула полной вероятности (2.3.11) остается справедливой и в случае, если условие, состоящее в том, что события $H_1, H_2, \ldots, H_n$ образуют полную группу, заменить более слабым: гипотезы $H_1, H_2, \ldots, H_n$ попарно несовместны ($H_i\cap H_j=\varnothing$ при $i\neq j$), а событие A влечет за собой объединение этих гипотез ($A\subseteq H_1\cup H_2\cup\cdots\cup H_n$).

В пакете Microsoft Excel существует функция СУММПРОИЗВ для вычисления суммы попарных произведений элементов массивов. Эту функцию можно применять при расчетах по формуле полной вероятности:

$$\mathbf{P}(A)=\sum_{i=1}^{n}\mathbf{P}(A|H_i)\mathbf{P}(H_i)=\text{СУММПРОИЗВ}(\langle\text{массив }\mathbf{P}(A|H_i)\rangle;\ \langle\text{массив }\mathbf{P}(H_i)\rangle).$$

Пример 2.3.4. Брокерская компания проводит операции с ценными бумагами: 10% сделок заключается с инвестиционными банками, 20% — с другими брокерскими компаниями, остальные — с физическими лицами. Вероятности того, что контрагент не выполнит условия сделки, составляют для указанных групп контрагентов 0,01, 0,05 и 0,2 соответственно. Определим, какая доля сделок в среднем не исполняется по вине контрагентов.

Пусть событие A состоит в том, что условия сделки не выполняются контрагентом, гипотеза H_1 — в том, что контрагентом по этой сделке выступает инвестиционный банк, гипотеза H_2 — в том, что контрагентом является другая

брокерская компания, гипотеза H_3 — в том, что контрагентом является физическое лицо. По условию задачи имеем: вероятности гипотез составляют $\mathbf{P}(H_1) = 0{,}1$, $\mathbf{P}(H_2) = 0{,}2$, $\mathbf{P}(H_3) = 1 - \mathbf{P}(H_1) - \mathbf{P}(H_2) = 0{,}7$, а условные вероятности события A, в свою очередь, равны $\mathbf{P}(A|H_1) = 0{,}01$, $\mathbf{P}(A|H_2) = 0{,}05$, $\mathbf{P}(A|H_3) = 0{,}2$. По формуле полной вероятности (2.3.11) имеем: $\mathbf{P}(A) = \mathbf{P}(A|H_1)\mathbf{P}(H_1) + \mathbf{P}(A|H_2)\mathbf{P}(H_2) + \mathbf{P}(A|H_3)\mathbf{P}(H_3) = 0{,}01 \cdot 0{,}1 + 0{,}05 \cdot 0{,}2 + 0{,}2 \cdot 0{,}7 = 0{,}151$.

Вычисления в Microsoft Excel иллюстрирует рис. 2.3.2 (справа от ячеек, в которых есть формулы, приведены эти формулы). ❑

	A	B	C	D	E
1	i	P(H_i)		P(A|H_i)	
2	1	0,1		0,01	
3	2	0,2		0,05	
4	3	0,7	=1-B2-B3	0,2	
5			P(A)=	0,151	=СУММПРОИЗВ(D2:D4;B2:B4)

Рис. 2.3.2. *Вычисления по формуле полной вероятности*

Пример 2.3.5 (Задача о разорении).[1] Петя с папой играют в следующую игру. Петя бросает монету, предварительно сообщив папе, какая сторона, по его мнению, выпадет: «герб» или «решетка». Если Петя угадал, то папа платит Пете 1 руб., в противном случае Петя платит папе 1 руб. Начальный капитал папы безграничен, а начальный капитал Пети составляет $x = 100$ руб. Игра продолжается до тех пор, пока Петя не наберет заранее определенную сумму s, либо пока он не разорится, проиграв весь имеющийся капитал x. Найдем вероятность того, что Петя разорится, так и не набрав желаемую сумму.

Пусть $p(x)$ — вероятность того, что, имея x руб., Петя все-таки разорится, событие A состоит в Петином разорении, гипотеза H_1 — в том, что Петя выиграл на первом шаге игры, гипотеза H_2 — в том, что Петя проиграл на первом шаге игры. При этом, очевидно, вероятность разорения при условии выигрыша на первом шаге составит $\mathbf{P}(A|H_1) = p(x+1)$, а вероятность разорения при условии проигрыша на первом шаге будет равна $\mathbf{P}(A|H_2) = p(x-1)$.

Согласно классической формуле вероятности $\mathbf{P}(H_1) = \mathbf{P}(H_2) = 0{,}5$. По формуле полной вероятности

$$p(x) = \mathbf{P}(A) = \mathbf{P}(A|H_1)\mathbf{P}(H_1) + \mathbf{P}(A|H_2)\mathbf{P}(H_2) = 0{,}5p(x+1) + 0{,}5p(x-1) = 0{,}5[p(x+1) + p(x-1)],$$

откуда

$$0{,}5[p(x+1) - p(x)] = 0{,}5[p(x) - p(x-1)]$$

или

$$p(x+1) - p(x) = p(x) - p(x-1) = p(x-1) - p(x-2) = \cdots = p(1) - p(0) = c,$$

где c — некоторая константа. Из приведенной цепочки равенств следует, что $p(x) = p(0) + cx$, при этом, очевидно, $p(0) = 1$, а $p(s) = 1 + cs = 0$. Окончательно получаем:

$$p(x) = 1 - x/s.$$

Если, например, Петя собирался набрать 110 руб., то $x = 100$, $s = 110$, поэтому $p(100) = 1/11 \approx 0{,}091$; если же Петя хотел закончить игру с выигрышем, равным 1000 руб., то $x = 100$, $s = 1000$, поэтому $p(100) = 0{,}9$.

Предположим теперь, что папа играет нечестно: он дал Пете монету со смещенным центром тяжести, так что Петя (считая монету «правильной» и выбирая в среднем в половине случаев «герб» и в половине случаев — «решетку») выигрывает с вероятностью 0,4 и проигрывает с вероятностью 0,6. Начальный капитал Пети составляет, как и раньше, $x = 100$ руб., игра продолжается до тех пор, пока Петя не наберет заранее определенную сумму s, либо пока он не разорится, проиграв весь имеющийся капитал x.

Аналогично предыдущим выкладкам можно составить уравнение

$$p(x) = 0{,}4p(x+1) + 0{,}6p(x-1),$$

[1] Практическая интерпретация задачи о разорении состоит в следующем. Инвестиционная компания «играет» с рынком. Возможности рынка безграничны (чего нельзя сказать о возможностях компании). Компания пытается угадать, какие финансовые терпит убытки. Если убытков становится много, то рано или поздно компания разоряется.

откуда

$$0{,}4[p(x+1)-p(x)]=0{,}6[p(x)-p(x-1)].$$

Поэтому

$$(0{,}4)^x\prod_{k=1}^{x}[p(k+1)-p(k)]=(0{,}6)^x\prod_{k=1}^{x}[p(k)-p(k-1)]=(0{,}6)^x\prod_{k=1}^{x}[p(k)-p(k-1)]$$

или

$$(0{,}4)^x\prod_{k=2}^{x+1}[p(k)-p(k-1)]=(0{,}6)^x\prod_{k=1}^{x}[p(k)-p(k-1)].$$

Поделив левую и правую части последнего уравнения на

$$(0{,}4)^x\prod_{k=2}^{x}[p(k)-p(k-1)],$$

получим:

$$p(x+1)-p(x)=\left(\frac{0{,}6}{0{,}4}\right)^x[p(1)-p(0)]$$

или (поскольку $p(0)=1$)

$$p(x+1)-p(x)=(1{,}5)^x[p(1)-1].$$

Вычислим теперь

$$p(s)-p(x)=\sum_{k=x}^{s-1}[p(k+1)-p(k)]=\sum_{k=x}^{s-1}\{(1{,}5)^k[p(1)-1]\}=[p(1)-1]\sum_{k=x}^{s-1}(1{,}5)^k=[p(1)-1]\frac{(1{,}5)^s-(1{,}5)^x}{1{,}5-1}=$$
$$=2[p(1)-1][(1{,}5)^s-(1{,}5)^x]$$

и учтем в равенстве

$$p(s)-p(x)=2[p(1)-1][(1{,}5)^s-(1{,}5)^x]$$

условие $p(s)=0$:

$$p(x)=2[1-p(1)][(1{,}5)^s-(1{,}5)^x].$$

Подставляя теперь $x=0$ и учитывая, что $p(0)=1$, получим:

$$2[1-p(1)][(1{,}5)^s-1]=1 \quad \text{или} \quad 2[1-p(1)]=\frac{1}{(1{,}5)^s-1}.$$

Окончательно получаем:

$$p(x)=\frac{(1{,}5)^s-(1{,}5)^x}{(1{,}5)^s-1}.$$

Вспомнив, откуда взялось число 1,5 $\left(1{,}5=(1-q)/q\right)$, преобразуем эту формулу к общему виду:

$$p(x)=\frac{\left(\frac{1-q}{q}\right)^s-\left(\frac{1-q}{q}\right)^x}{\left(\frac{1-q}{q}\right)^s-1}.$$

Легко видеть, что при $s=110$ $p(x)\approx 0{,}983$ (т. е. у Пети есть очень небольшие шансы на выигрыш), а при $s=1000$ вероятность разорения Пети $p(x)$ практически равна единице (с точностью более тридцати знаков после запятой) — у Пети нет совершенно никаких шансов выиграть! ❑

Пример 2.3.6 (Задача о разборчивой невесте).[1] У одной из Машиных подруг есть достаточно большое число женихов. Заранее она ничего о своих женихах не знает, кроме их числа n. Расположившись в очередь в случайном порядке, женихи представляются разборчивой не-

[1] Приведем финансовую интерпретацию задачи о разборчивой невесте. Пусть инвестор владеет некоторым активом, цена которого может изменяться в течение некоторого срока, а по окончании этого срока станет фиксированной (таким активом может быть облигация, рыночная котировка которой может изменяться до момента погашения, или какой-либо инвестиционный проект, пока он еще не завершен). Стоимости актива в предыдущие дни известны, а в последующие — неизвестны. К предыдущему дню вернуться уже нельзя. Требуется определить момент продажи актива с наибольшей выгодой.

весте один за другим, так что встречая очередного жениха, она знает всех предшествующих. Представленный и отвергнутый жених больше не возвращается. Невеста решила избрать следующую стратегию выбора: она просматривает первых m женихов, никого из них не выбирая, а затем останавливает свой выбор на первом из оставшихся $(n - m)$ женихов, который окажется лучше, чем любой из первых m женихов. Найдем вероятность $p_m(A)$ сделать наилучший выбор при такой стратегии.

Формализуем задачу, считая, что Машина подруга приписывает каждому жениху некоторую функцию качества — число, причем понятно, что все женихи обладают различным качеством — совсем одинаковых не бывает. При этом число всевозможных расположений женихов по порядку равно, очевидно, $P_n = n!$ (после того, как первый жених был осмотрен, у второго есть две возможности: оказаться хуже первого или оказаться лучше первого; у третьего есть уже три возможности: оказаться хуже первого и второго, оказаться лучше первого и второго, оказаться между ними по качеству и т. д., после того как осмотрен k-й жених, есть $(k + 1)$ промежуток, куда может попасть $(k + 1)$-й). Будем считать все $n!$ расположений женихов одинаково возможными.

Рассмотрим те исходы опыта (по выбору жениха), при которых избранная стратегия предлагает выбрать k-го жениха, соответствующее событие (выбор k-го по счету жениха) обозначим H_k $(k = m + 1, m + 2, \ldots, n)$, а событие, состоящее в том, что выбран жених с максимальным значением функции качества (т. е. наилучший), обозначим A.

Абстрагируемся теперь от женихов и будем говорить о точках на числовой прямой — значениях этой функции качества: имеется n точек, m из них мы знаем, и теперь рассматриваем по очереди точки с $(m + 1)$-й по n-ю — выбираем первую из них, которая окажется больше, чем самая правая из точек с первой по m-ю. Тогда число исходов, благоприятствующих появлению события H_k, рассчитаем так. Выбор k-й точки означает, что все точки с $(m + 1)$-й по $(k - 1)$-ю оказались левее, чем самая правая из первых m пробных точек, а k-я точка — правее всех пробных точек. На расположение первых m точек это никаких ограничений не налагает, но $(m + 1)$-я точка может теперь попасть не в любой из $(m + 1)$ промежутка, а только в m из них (она не может быть крайней правой), $(m + 2)$-я точка может попасть только в один из $(m + 1)$ промежутка, и так до $(k - 1)$-й точки, которая может попасть в любой из $(k - 2)$ промежутков; k-я точка может оказаться только самой правой — есть только один промежуток, куда она может попасть, а на все остальные точки никаких ограничений не накладывается: $(k + 1)$-я может попасть в любой из $(k + 1)$ промежутков, $(k + 2)$-я — в любой из $(k + 2)$ и так до n-й, которая может попасть в любой из n промежутков.

В результате получаем:

$$\mathbf{P}(H_k) = \frac{1\cdot 2\cdots m\cdot m(m+1)\cdots(k-2)\cdot(k+1)(k+2)\cdots n}{n!} =$$

$$= \frac{1\cdot 2\cdots m\cdot m(m+1)\cdots(k-2)\cdot(k+1)(k+2)\cdots n}{1\cdot 2\cdots m(m+1)\cdots(k-2)(k-1)k(k+1)(k+2)\cdots n} = \frac{m}{(k-1)k}.$$

Аналогично находим вероятность события $A \cap H_k$ — сделать наилучший выбор именно на k-м шаге: в этом случает отличие от предыдущей формулы будет в последних множителях числителя — точки с $(k + 1)$-й по n-ю теперь не могут оказаться самыми правыми, т. е. $(k + 1)$-я может попасть в любой из k промежутков, $(k + 2)$-я — в любой из $(k + 1)$ и так до n-й, которая может попасть в любой из $(n + 1)$ промежутков:

$$\mathbf{P}(A\cap H_k) = \frac{1\cdot 2\cdots m\cdot m(m+1)\cdots(k-2)\cdot k(k+1)\cdots(n-1)}{n!} =$$

$$= \frac{1\cdot 2\cdots m\cdot m(m+1)\cdots(k-2)\cdot k(k+1)\cdots(n-1)}{1\cdot 2\cdots m(m+1)\cdots(k-2)(k-1)k(k+1)(k+2)\cdots(n-1)n} = \frac{m}{(k-1)n}.$$

При этом

$$\mathbf{P}(A\,|\,H_k) = \frac{\mathbf{P}(A\cap H_k)}{\mathbf{P}(H_k)} = \frac{\dfrac{m}{(k-1)n}}{\dfrac{m}{(k-1)k}} = \frac{k}{n},$$

и по формуле полной вероятности получаем:

$$\mathbf{P}(A)=\sum_{k=m+1}^{n}\mathbf{P}(A\,|\,H_k)\mathbf{P}(H_k)=\sum_{k=m+1}^{n}\frac{km}{n(k-1)k}=\frac{m}{n}\sum_{k=m+1}^{n}\frac{1}{k-1}=\frac{m}{n}\sum_{k=m}^{n-1}\frac{1}{k}.$$

Таким образом,

$$p_m(A)=\frac{m}{n}\sum_{k=m}^{n-1}\frac{1}{k}.$$

Теперь определим такое число m_n, чтобы вероятность $p_{m_n}(A)$ была максимальной среди всех $p_m(A)$, $m = 0, 1, 2, ..., n$.

Воспользовавшись соотношением

$$\frac{m}{n}\sum_{k=m}^{n-1}\frac{1}{k+1}\leqslant\frac{m}{n}\int_m^n\frac{dx}{x}=-\frac{m}{n}\ln\frac{m}{n}\leqslant\frac{m}{n}\sum_{k=m}^{n-1}\frac{1}{k}$$

и тем, что $1/\mathrm{e}$ — точка максимума функции $-x\ln x$ на интервале $(0; 1)$, получим, что максимум $p_m(A)$ достигается при

$$m_n = n/\mathrm{e}+\mathrm{o}(n)\approx 0{,}37n\,,$$

при этом

$$p_{m_n}(A)=\max_{0\leqslant m\leqslant n}p_m(A)=1/\mathrm{e}+\mathrm{o}(1)\approx 0{,}37\,.$$

Итак, при большом числе женихов оптимальная стратегия Машиной подруги состоит в том, что она должна просмотреть примерно 35—40% всех своих женихов, никого из них не выбирая, а затем выбрать того из оставшихся, который окажется лучше любого из m первоначально осмотренных женихов. При этом вероятность наилучшего выбора при такой стратегии окажется приблизительно равна 0,37. ❑

Задачи

109. Из корзины, содержащей три красных яблока и семь зеленых, вынимают по очереди все яблоки. Найдите вероятность того, что вторым по счету будет вынуто красное яблоко.

110. Из корзины, содержащей три красных яблока и семь зеленых, вынимают одно за другим все яблоки, кроме одного. Найдите вероятность того, что последнее оставшееся в корзине яблоко будет зеленым.

111. Проводится тестирование остаточных знаний студента по математике: 10% вопросов теста проверяют знание линейной алгебры, 20% — знание математического анализа, остальные — знание теории вероятностей. Студент считает, что вероятности того, что он не ответит на вопрос теста, составляют для указанных групп вопросов 0,01, 0,05 и 0,2 соответственно. Найдите ожидаемый процент вопросов теста, на которые не ответит студент.

112. Петя знает не все вопросы программы. В каком случае вероятность вытащить неизвестный билет будет меньше: когда он тянет билет первым или последним?

113. Студенты считают, что из 50 экзаменационных билетов 10 являются «хорошими». Петя и Маша по очереди тянут по одному билету. Найдите вероятности следующих событий: *а*) Пете достался «хороший» билет; *б*) Маше достался «хороший» билет; *в*) им обоим достались «хорошие» билеты.

114. Вероятность того, что дневной оборот продавца мороженого превысит 20 000 руб., при солнечной погоде равна 80%, при переменной облачности — 50%, а при пасмурной погоде — 10%. Найдите вероятность того, что на следующий день оборот превысит 20 000 руб., если вероятность солнечной погоды в данное время года составляет 20%, вероятность переменной облачности и вероятность пасмурной погоды — по 40%.

115. В каждой из трех корзин находится по семь красных яблок и по три зеленых. Из первой корзины наудачу достали одно яблоко и переложили во вто-

рую, затем из второй корзины наудачу достали яблоко и переложили в третью. Найдите вероятность того, что яблоко, наудачу извлеченное после этих манипуляций из третьей корзины, окажется красным.

116. В магазине проходит бесплатная дегустация нового газированного напитка. Вероятность того, что участвовать в дегустации согласится посетитель — мужчина, составляет 0,1, а вероятность того, что участвовать в дегустации согласится посетитель — женщина, равна 0,2. Сейчас в магазине находятся пятнадцать женщин и десять мужчин. Девушка — промоутер обратилась с предложением принять участие в дегустации к двум посетителям магазина. Найдите вероятность того, что один из посетителей примет это предложение, а другой откажется.

117. В салон самолета входит старушка, выбирает себе произвольное место и садится. Все остальные пассажиры рассаживаются по следующему правилу: если место пассажира свободно, то он занимает свое место, иначе садится на любое свободное место. Найдите вероятность того, что последний пассажир займет свое место (всего в самолете, в котором 64 места, летят 64 пассажира, включая старушку).

118. Предположим, что в условиях примера 2.3.5 нечестно играет Петя: он предложил играть монетой со смещенным центром тяжести и теперь выигрывает с вероятностью 0,6 и проигрывает с вероятностью 0,4. Начальный капитал Пети составляет, как и в предыдущей задаче, $x = 100$ руб. Игра продолжается до тех пор, пока Петя не наберет заранее определенную сумму s, либо пока он не разорится, проиграв весь имеющийся капитал x. Найдите вероятность разорения Пети в общем случае (для произвольной суммы s) и в конкретных случаях: *а*) $s = 110$ руб.; *б*) $s = 1000$ руб.

119. Как изменится вероятность разорения Пети в условиях примера 2.3.5, если ставка на каждом ходе будет равна не 1 руб., как раньше, а 2 руб.?

120. В условиях примера 2.3.5 начальный капитал папы уже не безграничен, а равен $y = 1100$ руб., начальный капитал Пети по прежнему составляет $x = 100$ руб., игра продолжается до тех пор, пока кто-нибудь из игроков не разорится. Найдите вероятность того, что разорится Петя в случае, когда вероятность угадывания монеты равна: *а*) 0,5; *б*) 0,4; *в*)0,6.

2.3.3. Формула Байеса

Из формулы полной вероятности следует

Формула Байеса. *Если события* $H_1, H_2, \ldots, H_s \in \mathcal{S}$ *образуют полную группу и имеют положительные вероятности, событие* $A \in \mathcal{S}$ *также имеет положительную вероятность, то*

$$\mathbf{P}(H_k \mid A) = \frac{\mathbf{P}(A \mid H_k)\mathbf{P}(H_k)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \mid H_k)\mathbf{P}(H_k)}{\sum_{i=1}^{n} \mathbf{P}(A \mid H_i)\mathbf{P}(H_i)} \quad (k = 1, 2, \ldots, n). \qquad (2.3.12)$$

Доказательство. По формуле умножения вероятностей (2.3.6) $\mathbf{P}(A \cap H_i) = \mathbf{P}(A \mid H_k)\mathbf{P}(H_k)$. С другой стороны, по той же формуле умножения вероятностей $\mathbf{P}(A \cap H_i) = \mathbf{P}(H_k \mid A)\mathbf{P}(A)$. Таким образом, $\mathbf{P}(A \mid H_k)\mathbf{P}(H_k) = \mathbf{P}(H_k \mid A)\mathbf{P}(A)$, откуда $\mathbf{P}(H_k \mid A) = \dfrac{\mathbf{P}(A \mid H_k)\mathbf{P}(H_k)}{\mathbf{P}(A)}$. Вероятность в знаменателе можно раскрыть по формуле полной вероятности (2.3.11). Справедливость формулы Байеса доказана. ❑

Общая схема применения формулы Байеса при решении практических задач такова. Пусть событие A может происходить в различных условиях, о характере которых можно выдвинуть n гипотез: $H_1, H_2, \ldots, H_n$. Из некоторых соображений известны вероятности $\mathbf{P}(H_i)$ гипотез H_i (*априорные вероятности* — вероятности гипотез H_i до проведения опыта), известны также условные вероятности $\mathbf{P}(A \mid H_1)$, $\mathbf{P}(A \mid H_2)$, ..., $\mathbf{P}(A \mid H_n)$. Предположим, что произведен опыт, в результате которого обнаружено наступление события A. Это должно вызвать

переоценку вероятностей гипотез H_i; формула Байеса и дает выражения для условных вероятностей гипотез H_i, уточненных в результате опыта, исходом которого стало событие A, — *апостериорных вероятностей* $\mathbf{P}(H_i|A)$.

Замечание. Формула (2.3.12), носящее имя Т. Байеса, самому Т. Байесу была неизвестна, поскольку в то время еще не была известна формула полной вероятности (2.3.11). На самом деле формула (2.3.12), как и формула полной вероятности (2.3.11), была впервые сформулирована П. Лапласом. Томас Байес (*Thomas Bayes*) был английским священником и при жизни опубликовал всего одну математическую работу. Его фамилия в российских учебниках чаще всего пишется через «а»: Байес. Однако существуют учебники, в которых утверждается, что правильно писать эту фамилию через «е» или «э»: Бейес или Бэйес. На самом деле эта фамилия правильно произносится как Бэйз.

Пример 2.3.7. В условиях примера 2.3.4 начальнику отдела управления рисками доложили, что по одной из сделок контрагент не выполняет своих обязательств. Найдем вероятность того, что таким контрагентом является брокерская компания.

Пусть, как и в примере 2.3.4, событие A состоит в том, что условия сделки не выполняются контрагентом, гипотеза H_1 — в том, что контрагентом по этой сделке выступает инвестиционный банк, гипотеза H_2 — в том, что контрагентом является другая брокерская компания, гипотеза H_3 — в том, что контрагентом является физическое лицо. По условию задачи имеем: вероятности гипотез составляют $\mathbf{P}(H_1) = 0{,}1$, $\mathbf{P}(H_2) = 0{,}2$, $\mathbf{P}(H_3) = 1 - \mathbf{P}(H_1) - \mathbf{P}(H_2) = 0{,}7$, а условные вероятности события A, в свою очередь, равны $\mathbf{P}(A|H_1) = 0{,}01$, $\mathbf{P}(A|H_2) = 0{,}05$, $\mathbf{P}(A|H_3) = 0{,}2$. По формуле Байеса (2.3.12) $\mathbf{P}(H_2|A) = \mathbf{P}(A|H_2)\mathbf{P}(H_2) / \mathbf{P}(A) = 0{,}05 \cdot 0{,}2 / 0{,}151 = 10/151 \approx 0{,}066$, где вероятность $\mathbf{P}(A) = 0{,}151$ была рассчитана по формуле полной вероятности (2.3.11) в примере 2.3.4. ❑

Пример 2.3.8. Если проектом руководит квалифицированный менеджер, то такой проект может быть выполнен в срок с вероятностью 0,8, если же менеджер проекта недостаточно квалифицирован, то своевременное выполнение проекта возможно лишь с вероятностью 0,4. Достаточной квалификацией обладают 70% менеджеров проектов на рынке труда. Для руководства срочным проектом компания наняла нового менеджера и заключила с ним контракт на руководство только данным проектом.

Определим условную вероятность того, что менеджер проекта обладает достаточной квалификацией (и его можно принять в штат компании) при условии, что работы по проекту оказались завершены в срок;

Пусть событие A состоит в том, что проект завершится в срок, гипотеза H_1 — в том, что проектом руководит квалифицированный менеджер, а гипотеза H_2 — в том, что менеджер проекта недостаточно компетентен. По условию задачи имеем: вероятности гипотез составляют $\mathbf{P}(H_1) = 0{,}7$, $\mathbf{P}(H_2) = 0{,}3$, а условные вероятности события A, в свою очередь, равны $\mathbf{P}(A|H_1) = 0{,}8$, $\mathbf{P}(A|H_2) = 0{,}4$. Безусловная вероятность своевременного завершения проекта по формуле полной вероятности (2.3.11) равна $\mathbf{P}(A) = \mathbf{P}(A|H_1)\mathbf{P}(H_1) + \mathbf{P}(A|H_2)\mathbf{P}(H_2) = 0{,}8 \cdot 0{,}7 + 0{,}4 \cdot 0{,}3 = 0{,}68$.

По формуле Байеса (2.3.12) $\mathbf{P}(H_1|A) = \mathbf{P}(A|H_1)\mathbf{P}(H_1) / \mathbf{P}(A) = 0{,}8 \cdot 0{,}7 / 0{,}68 = 0{,}56 / 0{,}68 = 14/17 \approx 0{,}82$, т. е. достаточно высока вероятность того, что принятый на работу менеджер обладает требуемой квалификацией, если известно, что проект завершен в срок.

Теперь найдем условную вероятность того, что менеджер проекта не обладает достаточной квалификацией при условии, что работы по проекту не были

завершены в срок. Вероятность задержки сроков завершения проекта $\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A) = 1 - 0{,}68 = 0{,}32$, при этом условные вероятности $\mathbf{P}(\bar{A}\,|\,H_1) = 1 - \mathbf{P}(A\,|\,H_1) = 1 - 0{,}8 = 0{,}2$, $\mathbf{P}(\bar{A}\,|\,H_2) = 1 - \mathbf{P}(A\,|\,H_2) = 1 - 0{,}4 = 0{,}6$, и по формуле Байеса (2.3.12) $\mathbf{P}(H_2\,|\,\bar{A}) = \mathbf{P}(\bar{A}\,|\,H_2)\mathbf{P}(H_2) / \mathbf{P}(\bar{A}) = 0{,}6 \cdot 0{,}3 / 0{,}32 = 0{,}18 / 0{,}32 = 9 / 16 \approx \approx 0{,}56$. При условии, что проект не был завершен в срок, достаточно сложно сделать выводы о компетентности менеджера проекта: вероятности того, что он компетентен и некомпетентен, примерно равны 0,5, но все же вероятность того, что менеджер имеет недостаточную квалификацию, несколько выше, чем вероятность того, что менеджер достаточно компетентен. ❑

Пример 2.3.9 (Парадокс Монти Холла). В декабре 1963 г. на американском телеканале *NBC* впервые вышла программа *Let' s Make a Deal* («Заключим сделку!»), в которой участники, выбранные из зрителей в студии, играли в небольшие игры или просто угадывали ответы на вопросы. В конце передачи участники могли сыграть в «сделку дня». Перед ними было три двери, про которые было известно, что за одной из них — автомобиль, а за двумя другими — живые козы). После того, как игрок делал свой выбор, ведущий программы Монти Холл открывал одну из двух оставшихся дверей, показывая, что приза за ней нет, и давая участнику порадоваться тому, что он сохраняет шансы на выигрыш.

При этом автомобиль может с одинаковой вероятностью находиться за каждой дверью, ведущий всегда обязан из двух невыбранных дверей открывать дверь с козой и предлагать игроку возможность изменения выбора. Если же за обеими невыбранными дверями находятся козы, ведущий выбирает любую с одинаковой вероятностью.

Предположим, что мы оказались на шоу Монти Холла в роли участника, и в заключительный момент, открыв дверь с козой, ведущий предложил нам поменять свой выбор. Имеет ли смысл это делать? Повлияет ли наше решение — согласиться или нет — на вероятность выигрыша?

Большинство людей в такой ситуации полагают, что после открытия одной из дверей вероятность нахождения автомобиля за любой из двух закрытых дверей равна 1/2, т. е. нет никакой разницы, соглашаться или не соглашаться изменить свое решение. Более того, многие люди с трудом осознают ответ, отличный от этого, даже после того, как им было рассказано подробное решение.

Давайте посмотрим на эту ситуацию внимательно. Пронумеруем двери. Ту дверь, которую мы выбрали изначально, зазовем первой, ту, которую открыл ведущий, назовем второй, а оставшуюся дверь — третьей.

Пусть событие H_i состоит в том, что автомобиль находится за дверью i ($i = 1, 2, 3$), а событие A — в том, что ведущий открыл вторую дверь, за которой оказалась коза. До того, как ведущий открыл дверь с козой, т. е. до того, как наступило событие A, вероятности всех трех гипотез были равны 1/3: $\mathbf{P}(H_1) = \mathbf{P}(H_2) = \mathbf{P}(H_3) = 1/3$.

Если справедлива гипотеза H_1, то вероятность того, что ведущий откроет вторую дверь, равна $\mathbf{P}(A\,|\,H_1) = 1/2$ (так как если за обеими невыбранными дверями находятся козы, ведущий выбирает любую с одинаковой вероятностью). При справедливости гипотезы H_2 ведущий не откроет вторую дверь (так как автомобиль находится именно за ней), т. е. $\mathbf{P}(A\,|\,H_2) = 0$. Если же верна гипотеза

H_3, то вероятность того, что ведущий откроет вторую дверь, равна $\mathbf{P}(A|H_3)=1$ (ведь за второй дверью находится коза, а за третьей — автомобиль).

Теперь по формуле полной вероятности

$$\mathbf{P}(A)=\mathbf{P}(A|H_1)\mathbf{P}(H_1)+\mathbf{P}(A|H_2)\mathbf{P}(H_2)+\mathbf{P}(A|H_3)\mathbf{P}(H_3)=$$
$$=(1/2)\cdot(1/3)+0\cdot(1/3)+1\cdot(1/3)=1/6+0+1/3=1/2,$$

и по формуле Байеса мы можем вычислить

$$\mathbf{P}(H_1|A)=\frac{\mathbf{P}(A|H_1)\mathbf{P}(H_1)}{\mathbf{P}(A)}=\frac{(1/2)\cdot(1/3)}{1/2}=\frac{1}{3};$$

$$\mathbf{P}(H_2|A)=\frac{\mathbf{P}(A|H_2)\mathbf{P}(H_2)}{\mathbf{P}(A)}=\frac{0\cdot(1/3)}{1/2}=0;\qquad \mathbf{P}(H_3|A)=\frac{\mathbf{P}(A|H_3)\mathbf{P}(H_3)}{\mathbf{P}(A)}=\frac{1\cdot(1/3)}{1/2}=\frac{2}{3}.$$

Таким образом, вероятность того, что автомобиль находится за первой дверью, по прежнему равна 1/3, за второй дверью автомобиля точно нет, а за третьей дверью он находится с вероятностью 2/3, т. е. изменение выбранной двери вдвое увеличивает наши шансы на выигрыш! ❑

Задачи

121. В корзине три красных и семь зеленых яблок. Из корзины вынули одно яблоко и не глядя отложили в сторону. После этого из корзины достали еще одно яблоко, которое оказалось зеленым. Найдите вероятность того, что яблоко, отложенное в сторону, также было зеленым.

122. В условиях задачи **111** студент при получении очередного вопроса теста даже не понял, к какому из трех разделов математики относится этот вопрос. Найдите вероятность того, что вопрос относится к теории вероятностей.

123. Магазин получает однотипный товар от трех поставщиков: 55% товара поступает от первого поставщика, 20% от второго и 25% от третьего. Продукция, поступающая от первого поставщика, содержит 5% брака, поступающая от второго поставщика — 6% брака, а поступающая от третьего поставщика — 8% брака. Покупатель оставил в книге пожеланий покупателей жалобу о неудовлетворительном качестве приобретенного товара. Найдите вероятность того, что плохой товар, вызвавший нарекания покупателя, поступил от второго поставщика.

124. В некоторой страховой компании среди держателей полисов страхования автогражданской ответственности 20% относятся к группе высокого риска (у таких страхователей вероятность наступления страхового случая в течение года равна 0,1), а 80% — к группе низкого риска (у них вероятность наступления страхового случая в течение года равна 0,01). С какой вероятностью случайно выбранный страхователь, у которого в прошлом году был страховой случай, относится к группе: *а*) высокого риска; *б*) низкого риска.

125. Опыт отдела маркетинга фармацевтической фирмы показывает, что после проведения рекламной кампании нового вида зубной пасты 5% мужчин и 10% захотят приобрести новую пасту, а остальные будут продолжать пользоваться прежними видами зубных паст. Число мужчин и число женщин в городе соотносятся как 2 : 3 и все они покупают зубную пасту. Найдите вероятность того, что покупатель, случайно отобранный среди купивших новый вид пасты, окажется женщиной.

126. Для принятия решений об участии в инвестиционных проектах предлагается некоторая компьютерная система анализа бизнес-планов проектов. Из данных за прошлые периоды известно, что 5% всех рассматриваемых проектов являются «плохими» — не подходящими для инвестирования. Предложенная

система определяет 98% «плохих» проектов как «потенциально плохие», но при этом 15% проектов, пригодных для инвестиций, также определяет как «потенциально плохие». Найдите: *а*) вероятность того, что проект не подходит для инвестирования, при условии, что данной системой анализа бизнес-планов он был определен как «потенциально хороший»; *б*) вероятность того, что проект подходит для инвестирования, при условии, что данной системой он был определен как «потенциально плохой». На основе полученных результатов прокомментируйте пригодность системы для принятия инвестиционных решений.

127. Чтобы подготовить качественную документацию для участия в конкурсе на право выполнения государственного контракта, необходимы значительные инвестиции в предварительные исследования и разработки. Если фирма А произведет такие инвестиции, а ее основной конкурент этого не сделает, то вероятность заключения договора с фирмой составит 0,8. Однако если конкурент также проведет предварительные исследования и разработки, то вероятность заключения договора с фирмой уменьшается до 0,4. Аналитическая служба фирмы оценивает вероятность проведения конкурентом изысканий по предстоящему проекту в 0,3. Вычислите вероятности следующих событий: *а*) контракт будет заключен с фирмой А при отсутствии информации о решении конкурента; *б*) контракт с фирмой А не будет заключен при условии, что конкурент проведет предварительные исследования; *в*) конкурент проведет предварительные исследования при условии, что контракт будет заключен с фирмой А; *г*) конкурент проведет предварительные исследования при условии, что контракт с фирмой А не будет заключен.

128. В некотором городе у 55 взрослых жителей из 100 есть страницы в социальных сетях. При этом часть взрослых жителей города работает в государственных организациях, а часть — в коммерческих компаниях. Также известно, что страницы в социальных сетях есть у 60% жителей, работающих в государственных организациях, и у 40% жителей, работающих в коммерческих компаниях. Найдите вероятность того, что данный житель города работает в государственной организации, если известно, что у него есть страница в социальной сети.

§ 2.4. Последовательности испытаний

2.4.1. Биномиальная схема. Формула Бернулли

Пусть проводится конечное число n последовательных испытаний, в каждом из которых некоторое событие A может либо наступить (такую ситуацию назовем *успехом*) либо не наступить (такую ситуацию назовем *неудачей*), причем эти испытания удовлетворяют следующим **условиям:**

- *событие A случайно относительно каждого испытания*, т. е. до проведения испытания нельзя сказать, закончится ли оно успехом или неудачей;
- *испытания проводятся в одинаковых, с вероятностной точки зрения, условиях*, т. е. вероятность успеха в каждом отдельно взятом испытании равна одному и тому же числу p;
- *испытания независимы*, т. е. события A_1, A_2, ..., A_n, где A_i состоит в успехе на i-м испытании ($i = 1, 2, ..., n$), независимы в совокупности.

Такая последовательность испытаний называются *схемой Бернулли* или *биномиальной схемой*, а сами испытания — *испытаниями Бернулли*.

Вероятность $\mathbf{P}_n(k)$ того, что в серии из n испытаний Бернулли окажется ровно k успешных, рассчитывается по формуле, полученной впервые Я. Бернулли.

Формула Бернулли:

$$\mathbf{P}_n(k) = C_n^k p^k (1-p)^{n-k} \quad (k = 0, 1, 2, ..., n). \tag{2.4.1}$$

Доказательство. Пусть проведено n испытаний Бернулли, в каждом из которых событие A могло наступить или не наступить. Если в результате испытаний событие A наступило k раз, то это означает, что совместно наступили k событий A и $(n-k)$ событий $\overline{A}$, вероятности которых в каждом конкретном испытании равны соответственно p и $(1-p)$. Поскольку все n событий независимы, то по теореме умножения вероятностей вероятность появления в определенной последовательности k событий A и $(n-k)$ событий $\overline{A}$ равна $p^k(1-p)^{n-k}$. Но число таких различных последовательностей, в которых событие A наступило ровно k раз, равно C_n^k. Эти последовательности, очевидно, несовместны, поэтому по аксиоме аддитивности вероятности $\mathbf{P}_n(k)=C_n^k p^k(1-p)^{n-k}$, что и требовалось доказать. ❑

Сумма всех биномиальных вероятностей (2.4.1) равна единице:

$$\sum_{k=0}^{n}\mathbf{P}_n(k)=\sum_{k=0}^{n}C_n^k p^k(1-p)^{n-k}=[p+(1-p)]^n=1^n=1.$$

В пакете Microsoft Excel для вычисления вероятности $\mathbf{P}_n(k)$ можно пользоваться функцией

$$\mathbf{P}_n(k) = \text{БИНОМ.РАСП(<}k\text{>; <}n\text{>; <}p\text{>; ЛОЖЬ)}.$$

Формула Бернулли для схемы выборки с возвращением (2.2.16) теперь непосредственно следует из только что доказанной теоремы.

Доказательство формулы (2.2.16). Действительно, выбор l шаров из L представляет собой последовательность l независимых испытаний, в которых успехом считается извлечение белого шара. По классической формуле вероятности вероятность успеха $p = K/L$, вероятность неудачи $p = (L-K)/L$, и по формуле Бернулли (2.4.1) вероятность k успехов (т. е. вероятность того, что будет извлечено ровно k белых шаров) равна $\mathbf{P}_l(k)=C_l^k p^k(1-p)^{l-k}=$ $=C_l^k(K/L)^k[(L-K)/L]^{l-k}$, что и требовалось доказать. ❑

Замечание. Если в схеме схемы выборки без возвращения $K\to\infty$, $L\to\infty$ так, что $K/L\to p=\text{const}$, то формула гипергеометрической вероятности (2.2.15) переходит в пределе в формулу Бернулли (2.4.1).

Доказательство. При $L\to\infty$, $K/L\to p$ все пределы

$$\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{K}{L}=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{K-1}{L-1}=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{K-2}{L-2}=\cdots=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{K-k+1}{L-k+1}=p,$$

а все пределы

$$\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{L-K}{L-k}=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{L-K-1}{L-k-1}=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{L-K-2}{L-k-2}=\cdots=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{L-K-l+k+1}{L-l+1}=1-p,$$

поэтому

$$\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\mathbf{P}_{K,L}(k,l)=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\frac{C_K^k C_{L-K}^{l-k}}{C_L^l}=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\left(\frac{K!}{k!(K-k)!}\frac{(L-K)!}{(l-k)!(L-K-l+k)!}\frac{l!(L-l)!}{L!}\right)=$$

$$=\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\left(\frac{l!}{k!(l-k)!}\frac{K(K-1)(K-2)\cdots(K-k+1)}{L(L-1)(L-2)\cdots(L-k+1)}\frac{(L-K)(L-K-1)(L-K-2)\cdots(L-K-l+k+1)}{(L-k)(L-k-1)(L-k-2)\cdots(L-l+1)}\right)=$$

$$=C_l^k\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\underbrace{\left(\frac{K}{L}\frac{K-1}{L-1}\frac{K-2}{L-2}\cdots\frac{K-(k-1)}{L-(k-1)}\right)}_{k\text{ множителей}}\lim_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\underbrace{\left(\frac{L-K}{L-k}\frac{L-K-1}{L-k-1}\frac{L-K-2}{L-k-2}\cdots\frac{L-K-(l-k-1)}{L-k-(l-k-1)}\right)}_{(l-k)\text{ множителей}}=C_l^k p^k(1-p)^{l-k},$$

т. е. $\lim\limits_{\substack{L\to\infty\\ \frac{K}{L}\to p}}\mathbf{P}_{K,L}(k,l)=C_l^k p^k(1-p)^{l-k}$, что и требовалось доказать. ❑

Пример 2.4.1. Известно, что 70% из числа подписчиков определенного видеоблога смотрят рекламные блоки. Группу, состоящую из трех наугад выбранных под-

писчиков, опрашивают относительно содержания рекламного блока. Рассчитаем вероятности того, что рекламные блоки смотрели 0, 1, 2 и 3 человека из группы.

Вероятность того, что наугад выбранный подписчик видеоблога смотрит рекламные блоки, по классической формуле вероятности (2.2.11) равна $p = 0{,}7$. Интерпретируя опрос трех подписчиков, как три испытания Бернулли, и считая успехом ситуацию, когда подписчик смотрит рекламные блоки, найдем искомые вероятности по формуле Бернулли (2.4.2), в которой $n = 3$, $p = 0{,}7$: $\mathbf{P}_3(k) = C_3^k 0{,}7^k 0{,}3^{3-k}$ $(k = 0,1,2,3)$: $\mathbf{P}_3(0) = C_3^0 0{,}7^0 0{,}3^3 = 0{,}027$; $\mathbf{P}_3(1) = C_3^1 0{,}7^1 0{,}3^2 = 0{,}189$; $\mathbf{P}_3(2) = C_3^2 0{,}7^2 0{,}3^1 = 0{,}441$; $\mathbf{P}_3(3) = C_3^3 0{,}7^3 0{,}3^0 = 0{,}343$. ❑

Вычисления в Microsoft Excel иллюстрирует рис. 2.4.1 (справа от ячеек, в которых есть формулы, приведены эти формулы). ❑

Пример 2.4.2. Из 1000 опрошенных 700 человек поддерживают некоторую муниципальную программу. Для участия в телевизионной передаче необходимо собрать группу, в которую бы вошли и сторонники, и противники данной программы. Определим минимальную численность группы, в которой с вероятностью, не меньшей 0,9, хотя бы один респондент не поддерживает эту программу.

	A	B	C
1	k	P_3(k)	
2	0	0,027	=БИНОМ.РАСП(A2;3;0,7;ЛОЖЬ)
3	1	0,189	=БИНОМ.РАСП(A3;3;0,7;ЛОЖЬ)
4	2	0,441	=БИНОМ.РАСП(A4;3;0,7;ЛОЖЬ)
5	3	0,343	=БИНОМ.РАСП(A5;3;0,7;ЛОЖЬ)

Рис. 2.4.1. *Вычисления по формуле полной вероятности*

Пусть численность группы равна n. Будем интерпретировать опрос группы из n человек как испытания Бернулли, считая успехом то, что случайно выбранный респондент поддерживает муниципальную программу. Согласно статистическому подходу к определению вероятности вероятность успеха равна $p = 700/1000 = 0{,}7$. Пусть событие A состоит в том, что в группе из n человек хотя бы один не поддерживает муниципальную программу, тогда событие $\overline{A}$ означает, что в группе из n человек все n поддерживают эту программу. $\mathbf{P}(A) = 1 - \mathbf{P}\{\overline{A}\} = 1 - \mathbf{P}_n(n) =$ {по формуле Бернулли} $= 1 - C_n^n p^n (1-p)^0 = 1 - p^n =$ $= 1 - 0{,}7^n$. По условию вероятность $\mathbf{P}(A)$ должна быть не меньше 0,9, поэтому $1 - (0{,}7)^n \geqslant 0{,}9$ или $(0{,}7)^n \leqslant 0{,}1$. Чтобы найти минимальное значение n, при котором выполняется это неравенство, будем последовательно подставлять в него числа 1, 2, 3 и т. д., пока неравенство не удовлетворится: $(0{,}7)^1 = 0{,}7$, $(0{,}7)^2 = 0{,}49$, $(0{,}7)^3 = 0{,}343$; $(0{,}7)^4 = 0{,}240$, $(0{,}7)^5 = 0{,}168$, $(0{,}7)^6 \approx 0{,}118$, $(0{,}7)^7 \approx$ $\approx 0{,}082$. Видно, что неравенство $(0{,}7)^n \leqslant 0{,}1$ не выполняется при $n = 1, 2, \ldots, 6$, но выполняется при $n = 7$, поэтому минимальная численность группы, в которой с вероятностью, не меньшей 0,9, хотя бы один респондент не поддерживает эту программу, равна 7 чел.

Отметим, что неравенство $(0{,}7)^n \leqslant 0{,}1$ можно было бы решить намного проще, прологарифмировав обе его части: $n \ln 0{,}7 \geqslant \ln 0{,}1$, откуда $n \geqslant \ln 0{,}1 / \ln 0{,}7 \approx$ $\approx -2{,}3026 / (-0{,}3567) \approx 6{,}46$, и наименьшее натуральное n, удовлетворяющее последнему неравенству, равно 7. ❑

Пример 2.4.3 (Задача о разделе ставки).[1] Петя и Маша часто играют друг с другом на бильярде, причем Петя выигрывает в два раза чаще, чем Маша. Исходя из этого, они оценили свои

[1] Подобные задачи возникают при определении доли инвестора, который хочет выйти из незавершенного проекта. Задача эта интересна тем, что с ее решения в XVI в Б. Паскалем ведет свою историю теория вероятностей. Кстати, на шахматном матче между А. Карповым и Г. Каспаровым в 1985 г. случилась ситуация, подобная данной.

вероятности победить как 2/3 для Пети и 1/3 для Маши и начали турнир на следующих условиях: каждый выигрыш приносит одно очко, Петя для победы должен набрать двенадцать очков, а Маша — шесть. После того, как Петя набрал восемь очков, а Маша — четыре, игру пришлось прекратить, и победу решили присудить тому, у кого вероятность окончательного выигрыша больше. Кому в такой ситуации присудить победу?

Очевидно, максимальное количество партий, которое осталось сыграть Пете и Маше, равно пяти (либо Петя выиграет три раза, а Маша — два раза, либо Маша выиграет один раз, а Петя — четыре раза). Поэтому событие, заключающееся в выигрыше Пети (а значит, проигрыше Маши), состоит в том, что Маша из пяти партий не выиграет ни одной или выиграет всего одну. Поэтому вероятность выигрыша Пети равна вероятности того, что в пяти испытаниях Бернулли, в каждом из которых успех интерпретируется как выигрыш Машей очередной партии (т. е. вероятность успеха в каждом испытании составляет $p = 1/3$), наступит 0 или 1 успех:

$$\mathbf{P}(\text{выигрыш Пети}) = \mathbf{P}(0 \text{ или } 1 \text{ выигрыш Маши из } 5 \text{ партий}) = \mathbf{P}_5(0) + \mathbf{P}_5(1) =$$
$$= C_5^0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^5 + C_5^1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^4 = \frac{112}{243}.$$

При этом

$$\mathbf{P}(\text{выигрыш Маши}) = \mathbf{P}\left(\overline{\text{выигрыш Пети}}\right) = 1 - \mathbf{P}(\text{выигрыш Пети}) =$$
$$= 1 - \frac{112}{243} = \frac{131}{243} > \frac{112}{243}.$$

Поэтому в данном случае победу должны присудить Маше. ❑

Пример 2.4.4 (Задача Банаха). Известный математик Стефан Банах всегда носил с собой две коробки спичек, в каждой из которых первоначально было по r спичек. Каждый раз, когда он хотел зажечь спичку, Банах доставал наугад одну из коробок. Найдем вероятность того, что когда он в первый раз вынимал пустую коробку, в другой коробке оказывалось ровно k спичек, где $k = 1, 2, \ldots, r$.

Спички брались всего $(2r - k)$ раз, причем r раз из коробки, оказавшейся пустой. Это соответствует r успехам в $(2r - k)$ независимых испытаниях (вероятность успеха в этих испытаниях равна $p = 1/2$ в каждом испытании), поэтому вероятность того, что когда Банах в первый раз вынимал пустую коробку, в другой коробке оказывалось ровно k спичек, равна

$$\mathbf{P}(A) = \mathbf{P}_{2r-k}(r) = C_{2r-k}^r\left(\frac{1}{2}\right)^r\left(1-\frac{1}{2}\right)^{r-k} = \frac{C_{2r-k}^r}{2^{2r-k}}. \text{ ❑}$$

В ряде задач представляет интерес наиболее вероятное число успехов в последовательности испытаний, т. е. наибольшая из вероятностей (2.4.1). Наиболее вероятное число k^* успехов в последовательности из n испытаний Бернулли удовлетворяет неравенствам

$$np - (1 - p) \leqslant k^* \leqslant np + p. \tag{2.4.2}$$

Доказательство. Рассмотрим отношение

$$\frac{\mathbf{P}_n(k+1)}{\mathbf{P}_n(k)} = \frac{C_n^{k+1}p^{k+1}(1-p)^{n-k-1}}{C_n^k p^k (1-p)^{n-k}} = \frac{\dfrac{n!}{(k+1)!(n-k-1)!}p^{k+1}(1-p)^{n-k-1}}{\dfrac{n!}{k!(n-k)!}p^k(1-p)^{n-k}} =$$
$$= \frac{k!(n-k)!p^{k+1}(1-p)^{n-k-1}}{(k+1)!(n-k-1)!p^k(1-p)^{n-k}} = \frac{(n-k)p}{(k+1)(1-p)}.$$

При этом $\dfrac{\mathbf{P}_n(k)}{\mathbf{P}_n(k-1)} = \dfrac{(n-k+1)p}{k(1-p)}$. Очевидно, число k^* является наиболее вероятным числом успехов в серии из n испытаний Бернулли тогда и только тогда, когда $\mathbf{P}_n(k^*) \geqslant \mathbf{P}_n(k^* - 1)$ и $\mathbf{P}_n(k^*) \geqslant \mathbf{P}_n(k^* + 1)$. В свою очередь, $\mathbf{P}_n(k) \geqslant \mathbf{P}_n(k-1)$ тогда и только тогда, когда $\dfrac{\mathbf{P}_n(k)}{\mathbf{P}_n(k-1)} \geqslant 1$ или

$\frac{(n-k+1)p}{k(1-p)} \geqslant 1$, т. е. $(n-k+1)p \geqslant k(1-p)$ или $np-kp+p \geqslant k-kp$ или $k \leqslant np+p$. Аналогично, $\mathbf{P}_n(k) \geqslant \mathbf{P}_n(k+1)$ тогда и только тогда, когда $\frac{\mathbf{P}_n(k+1)}{\mathbf{P}_n(k)} \leqslant 1$ или $\frac{(n-k)p}{(k+1)(1-p)} \leqslant 1$, т. е. $(n-k)p \leqslant$ $\leqslant (k+1)(1-p)$ или $np-kp \leqslant k(1-p)+(1-p)$ или $k \geqslant np-(1-p)$. Окончательно получаем, что k^* должно удовлетворять неравенствам (2.4.4),что и требовалось доказать. ❑

Пример 2.4.5. В условиях примера 2.4.1 найдем наиболее вероятное k^* число лиц в группе, которые смотрят рекламные блоки. Оно подчиняется неравенствам (2.4.2): $np-(1-p) \leqslant k^* \leqslant np+p$, в которых $n=3$, $p=0{,}7$, т. е. $3 \cdot 0{,}7 - 0{,}3 \leqslant k^* \leqslant 3 \cdot 0{,}7 + 0{,}7$ или $1{,}8 \leqslant k^* \leqslant 2{,}8$, откуда $k^*=2$. Это подтверждается и результатами, полученными в примере 2.4.18. ❑

Задачи

129. Найдите вероятность появления ровно 5 гербов при 10-кратном бросании монеты.

130. Игральную кость бросают пять раз. Найдите вероятность того, что дважды появится число, кратное трем.

131. Что вероятнее: выиграть на бильярде у равносильного противника три партии из четырех или пять партий из восьми?

132. Что вероятнее: выиграть на бильярде у равносильного противника не менее трех партий из четырех или не менее пяти партий из восьми?

133. В контрольной работе десять задач. Студент знает примерно 70% материала (т. е. можно считать, что вероятность правильного решения задачи студентом равна 0,7). Студент получает оценку «зачтено», если он верно решил не менее 8 задач. Найдите вероятность того, что студент получит за эту контрольную работу оценку «зачтено».

134. В самолете *Boeing* 737-800, выполняющем рейс «Москва — Санкт-Петербург», 190 мест. Авиакомпания знает, что в среднем 5% пассажиров не приходят на регистрацию, и поэтому продает на эти 190 мест 200 билетов. Найдите вероятность того, что не всем пришедшим на регистрацию пассажирам хватит мест.

135. Из числа 100 студентов потока требуется выбрать пять старост групп. Считая, что типы темперамента распределены примерно равномерно, найдите вероятность того, что все пять старост окажутся сангвиниками.

136. В течение месяца данная акция может подорожать на 1% с вероятностью 0,7 и подешеветь на 1% с вероятностью 0,3. Предполагая ежемесячные изменения цены независимыми, рассчитайте вероятности того, что за три месяца цена акции возрастет: *а*) в $(1{,}01)^3$ раза; *б*) в $0{,}99(1{,}01)^2$ раза.

137. Работа коммерческого представителя признается удовлетворительной, если за месяц он продал программные продукты хотя бы 10 клиентам. Считая, что любая компания, в которую представитель обратится с коммерческим предложением, купит программный продукт с вероятностью 0,1, вычислите вероятность того, что работа представителя будет признана удовлетворительной, если за месяц он разошлет 40 предложений.

138. Среди билетов лотереи половина выигрышных. Найдите минимальное число билетов, которые необходимо приобрести, чтобы с вероятностью, не меньшей 0,99, быть уверенным в выигрыше хотя бы по одному билету.

139. В городе работают 1000 предприятий, из которых 330 допускают нарушения налогового законодательства. Определите число предприятий, которые должна отобрать для проверки налоговая служба, чтобы с вероятностью, не меньшей 0,99, среди них оказался хотя бы один нарушитель законодательства.

140. В условиях задачи **139** налоговая служба проводит проверку 12 предприятий, выбирая их случайным образом. Выбранные предприятия проверяются независимо друг от друга. Допущенные проверяемым предприятием нарушения могут быть выявлены налоговой службой с вероятностью 0,8. Найдите вероятность того, что в ходе этой проверки будет выявлен хотя бы один нарушитель налогового законодательства.

141. Стоимость проезда в автобусе равна 55 руб., месячный проездной билет на автобус стоит 2000 руб., а штраф за безбилетный проезд составляет 1500 руб. Петя 24 раза в месяц ездит на автобусе в университет и обратно. Он не покупает проездного билета, никогда не платит за проезд и считает (по своему опыту), что вероятность быть пойманным и заплатить штраф равна 0,05. Сравните стоимость проездного билета с наиболее вероятной величиной штрафа за 48 поездок.

142. Петя играл с Васей (равносильным противником) в шахматы на приз в 100 руб.: каждый выигрыш приносил одно очко, ничьи не считались. Игра шла до 8 очков. Когда Петя выиграл пять партий, а Вася — три, внезапно погас свет, и игру пришлось прекратить. Как им разделить приз — 100 руб.?

2.4.2. Формула Пуассона

Часто на практике встречаются ситуации, в которых число n испытаний Бернулли велико, а вероятность успеха мала. И хотя функция БИНОМ.РАСП дает верные результаты при $n \leqslant 2147\,483\,646$, при исследовании таких редких событий бывает полезна приближенная формула, полученная впервые С. Пуассоном.

Теорема Пуассона. *Если* $n \to \infty, p \to 0$ *так, что* $np \to \lambda = \text{const}$, *то*

$$\mathbf{P}_n(k) \approx \mathbf{P}(k) = \lim_{\substack{n\to\infty \\ np\to\lambda}} \mathbf{P}_n(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, 2, \ldots). \tag{2.4.3}$$

Доказательство. Вероятность

$$\mathbf{P}_n(k) = C_n^k p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k}.$$

По условию $n \to \infty, np \to \lambda = \text{const}$, значит, можно считать, что $p \to \lambda / n$. Поэтому $\lim\limits_{\substack{n\to\infty \\ np\to\lambda}} \mathbf{P}_n(k) = \dfrac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(\dfrac{\lambda}{n}\right)^k \left(1-\dfrac{\lambda}{n}\right)^{n-k}$, и $\mathbf{P}_n(k)$ можно вычислять как предел

$$\mathbf{P}_n(k) \approx \lim_{\substack{n\to\infty \\ np\to\lambda}} \mathbf{P}_n(k) = \lim_{n\to\infty}\left[\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}\right] =$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty}\left[\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdots\frac{n-k+1}{n}\cdot\left(1-\frac{\lambda}{n}\right)^{-k}\cdot\left(1-\frac{\lambda}{n}\right)^{n}\right] =$$

$$= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\cdot\lim_{n\to\infty}\left(1-\frac{2}{n}\right)\cdots\lim_{n\to\infty}\left(1-\frac{k-1}{n}\right)\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\cdot\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\lambda} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Все пределы, кроме последнего, равны единице, а последний множитель является вторым замечательным пределом, равным $e^{-\lambda}$. Таким образом, справедливость теоремы доказана. ❑

В пакете Microsoft Excel для вычисления $\mathbf{P}(k)$ можно пользоваться функцией

$$\mathbf{P}(k) = \text{ПУАССОН.РАСП}(<k>; <\lambda>; \text{ЛОЖЬ}).$$

Пример 2.4.6. На лекции по анализу данных присутствуют 200 человек. Считая вероятность того, что день рождения случайно выбранного студента приходится на определенный день года, равной $1/365$, найдем вероятность того, что один человек из присутствующих родился 1 января и два человека родились 8 марта.

Пусть событие A состоит в том, что случайно выбранный студент родился 1 января, событие B — в том, что k человек из 200 родились 1 января. Тогда по условию $p = \mathbf{P}(A) = 1/365$. Предположим, что опрос $n = 200$ студентов относительно даты их рождения удовлетворяет условиям, которые накладываются на испытания Бернулли, где успехом единичного испытания считается наступление события A.

Тогда, поскольку $n = 200$ велико, а произведение $\lambda_1 = np = 200/365 = 0{,}548 < 10$, для подсчета вероятности события B можно воспользоваться формулой Пуассона:

$$\mathbf{P}_{200}(k) = \frac{(0{,}548)^k}{k!} e^{-0{,}548} = \text{ПУАССОН.РАСП}(<k>; 0{,}548; \text{ЛОЖЬ}),$$

и при $k = 1$ получаем

$$\mathbf{P}_{200}(1) = \text{ПУАССОН.РАСП}(<k>; <\lambda>; \text{ЛОЖЬ}) = 0{,}317.$$

Пусть событие C состоит в том, что m человек из присутствующих родились 8 марта. Тогда в соответствии с формулой умножения вероятностей

$$\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C|B),$$

где $\mathbf{P}(C|B) = \mathbf{P}_{n-k}(m)$ — вероятность того, что из $(n-k)$ студентов m родились 8 марта. Так как число $n - k = 200 - 1 = 199$ велико, а $\lambda_2 = (n-k)p = 199/365 = 0{,}542 < 10$, для расчета условной вероятности $\mathbf{P}(C|B)$ можно вновь воспользоваться формулой Пуассона:

$$\mathbf{P}_{n-k}(m) = \frac{(0{,}542)^m}{m!} e^{-0{,}542} = \text{ПУАССОН.РАСП}(<k>; 0{,}542; \text{ЛОЖЬ}).$$

При $n = 200$, $k = 1$, $m = 2$ получаем:

$$\mathbf{P}(C|B) = \mathbf{P}_{199}(2) = \text{ПУАССОН.РАСП}(2; 0{,}542; \text{ЛОЖЬ}) = 0{,}086,$$

поэтому искомая вероятность $\mathbf{P}(B \cap C) = 0{,}317 \cdot 0{,}086 = 0{,}027$. ❑

При больших n и малых p погрешность от замены точной формулы Бернулли (2.4.1) приближенной формулой Пуассона (2.4.3) оказывается достаточно малой, и на практике формулой Пуассона пользуются в случае, когда число n испытаний Бернулли составляет несколько десятков или более, а произведение $\lambda = np < 10$. Результаты вычислений по формулам (2.4.1) и (2.4.3) для случаев $n = 100$, $p = 0{,}02$ и $n = 10$, $p = 0{,}2$ приводятся в табл. 2.4.1 (и в первом случае, и во втором $\lambda = np = 2 < 10$).

В случае, когда n велико, но $\lambda = np \geqslant 10$, формула Пуассона дает очень грубое приближение, и для получения приближенных значений $\mathbf{P}_n(k)$ используют интегральную теорему Муавра — Лапласа (см. **п. 4.2.2**).

Таблица 2.4.1

Сравнительные результаты вычислений по формулам Бернулли и Пуассона

k	0	1	2	3	4	5
$\mathbf{P}(k)$	0,1353	0,2707	0,2707	0,1805	0,0902	0,0361
$\mathbf{P}_{100}(k)$	0,1326	0,2707	0,2734	0,1823	0,0902	0,0353
$\mathbf{P}_{10}(k)$	0,1074	0,2684	0,3020	0,2013	0,0880	0,0264

Задачи

143. Считая вероятность того, что день рождения случайно выбранного человека приходится на определенный день года, равной 1/365, найдите вероятность того, что в группе, состоящей из 30 студентов, хотя бы двое родились в один день.

144. Владельцы кредитных карт теряют их редко — вероятность потерять кредитную карту в течение недели для случайно выбранного вкладчика составляет 0,001. Банк выдал кредитные карты 2000 клиентов. Найдите вероятность того, что за предстоящую неделю будет утеряна: *а*) ровно одна кредитная карта; *б*) хотя бы одна кредитная карта.

145. Среди транзакций по банковским картам 0,01% — мошеннические. При этом существующая система информационной безопасности распознает только 90% мошеннических транзакций, а средний ущерб от нераспознанной мошеннической транзакции составляет 14 000 руб. Каждый час банк проводит в среднем 20 000 транзакций. Определите: *а*) вероятность того, что за час пройдет хотя бы одна мошенническая транзакция (не будучи распознанной); *б*) наиболее вероятное время, за которое окупит себя новая система машинного обучения, которая стоит 40 000 000 руб. и распознает все мошеннические транзакции; *в*) вероятность того, что новая система оправдает себя за это время.

Глава 3. Случайные величины

§ 3.1. Определение случайной величины

3.1.1. Понятие случайной величины

В жизни нам часто приходится иметь дело с различными величинами; такими величинами являются, например, продолжительность суток в часах, температура кипения воды при обычных условиях, число детей, родившихся в Москве в течение суток, обменный курс евро к рублю. Первые две величины являются *детерминированными*, т. е. их значения известны. Так бывает не всегда, например, невозможно сказать заранее, сколько детей появится на свет в Москве в течение следующих суток или какой через месяц будет курс евро.

Важнейшим понятием теории вероятностей является понятие случайной величины, т. е. величины, которая в результате опыта может принять в зависимости от случая какое-либо значение из некоторого множества, однако заранее, до проведения опыта, невозможно сказать, какое именно значение она примет.

Примерами случайных величин являются число очков, выпадающих при однократном бросании игральной кости; число бракованных изделий среди наугад отобранных n изделий; время безотказной работы некоторого прибора; число посетителей некоторого интернет-сайта в определенный момент времени; суточная выручка в магазине; число застрахованных, обратившихся в страховую компанию за возмещением ущерба; число детей, родившихся в большом городе в течение суток; обменный курс двух валют в определенный день в будущем; число заемщиков, не вернувших кредит в срок, и т. д.

Случайная величина — это число, которое ставится в соответствие каждому возможному исходу опыта. Поскольку исходы опыта описываются элементарными событиями, случайную величину удобно рассматривать как функцию $X = X(\omega)$, определенную на пространстве элементарных событий Ω.

При изучении случайных величин нам часто придется искать ответ на вопрос, какова вероятность того, что значение случайной величины $X(\omega)$ принадлежит некоторому множеству B. Поэтому для достаточно широкого класса множеств мы должны быть уверены в том, что множество $\{\omega \in \Omega\colon X(\omega) \in B\}$ принадлежит σ-алгебре событий $\mathcal{S}$, и потому можно рассматривать вероятность $\mathbf{P}\{\omega \in \Omega\colon X(\omega) \in B\}$.

Оказывается, для этого достаточно, чтобы для любого $x \in \mathbb{R}$ множество $\{\omega \in \Omega\colon X(\omega) \in (-\infty; x]\} = \{\omega \in \Omega\colon X(\omega) \leqslant x\}$ являлось случайным событием, т. е. чтобы для любого $x \in \mathbb{R}$ множество событий ω, удовлетворяющих условию $X(\omega) \leqslant x$, принадлежало σ-алгебре событий $\mathcal{S}$:

$$\{\omega \in \Omega\colon X(\omega) \leqslant x\} \in \mathcal{S}.$$

Перейдем к строгим определениям.

Пусть $\{\Omega, \mathcal{S}, \mathbf{P}\}$ — некоторое вероятностное пространство.

Числовая функция $X(\omega)$, заданная на множестве Ω, называется *измеримой* относительно σ-алгебры $\mathcal{S}$, если для любого $x \in \mathbb{R}$ $\{\omega \in \Omega\colon X(\omega) \leqslant x\} \in \mathcal{S}$. *Случайной величиной* называется числовая функция $X(\omega)$, определенная на пространстве элементарных событий Ω и измеримая относительно σ-алгебры событий $\mathcal{S}$.

Далее случайные величины будут обозначаться заглавными буквами из конца латинского алфавита: X, Y, Z, U, V, W, ..., а их возможные значения — соответствующими строчными буквами: x, y, z, u, v, w, ...

Пусть $X(\omega)$ — случайная величина, т. е. функция

$$X\colon \Omega \to \mathbb{R} = (-\infty; +\infty),$$

измеримая относительно σ-алгебры $\mathcal{S}$, а $B \subseteq (-\infty; +\infty)$ — некоторое множество на числовой прямой.

В практических задачах нас часто будут интересовать множества вида

$$X^{-1}(B) = \{\omega \in \Omega\colon\ X(\omega) \in B\}$$

[$X^{-1}(B)$ — это множество таких элементарных исходов ω, что $X(\omega) \in B$].

Являются ли такие множества случайными событиями? Иначе говоря, можно ли утверждать, что $X^{-1}(B) \in \mathcal{S}$?

Если $B = (-\infty; x]$, то такое утверждение следует непосредственно из определения случайной величины, так как при этом $X^{-1}(B) = \{\omega \in \Omega\colon\ X(\omega) \in (-\infty; x]\} = \{\omega \in \Omega\colon\ X(\omega) \leqslant x\}$. Из этого следует, что множества

$$\{\omega \in \Omega\colon\ X(\omega) > x\} = \{\omega \in \Omega\colon\ X(\omega) \in (x; +\infty)\},$$

$$\{\omega \in \Omega\colon\ X(\omega) < x\} = \{\omega \in \Omega\colon\ X(\omega) \in (-\infty; x)\},$$

$$\{\omega \in \Omega\colon\ X(\omega) \geqslant x\} = \{\omega \in \Omega\colon\ X(\omega) \in [x; +\infty)\},$$

$$\{\omega \in \Omega\colon\ X(\omega) = x\} = \{\omega \in \Omega\colon\ X(\omega) \in \{x\}\},$$

$$\{\omega \in \Omega\colon\ a \leqslant X(\omega) < b\} = \{\omega \in \Omega\colon\ X(\omega) \in [a; b)\},$$

$$\{\omega \in \Omega\colon\ a < X(\omega) \leqslant b\} = \{\omega \in \Omega\colon\ X(\omega) \in (a; b]\},$$

$$\{\omega \in \Omega\colon\ a \leqslant X(\omega) \leqslant b\} = \{\omega \in \Omega\colon\ X(\omega) \in [a; b]\},$$

$$\{\omega \in \Omega\colon\ a < X(\omega) < b\} = \{\omega \in \Omega\colon\ X(\omega) \in (a; b)\},$$

где x, a, $b \in \mathbb{R}$, также являются случайными событиями, т. е. принадлежат σ-алгебре $\mathcal{S}$.

Доказательство. Для доказательства достаточно заметить, что

$$\{\omega \in \Omega\colon\ X(\omega) > x\} = \overline{\{\omega \in \Omega\colon\ X(\omega) \leqslant x\}},$$

$$\{\omega \in \Omega\colon\ X(\omega) \geqslant x\} = \bigcap_{n=1}^{\infty} \left\{\omega \in \Omega\colon\ X(\omega) > x - \frac{1}{n}\right\},$$

$$\{\omega \in \Omega\colon\ X(\omega) = x\} = \{\omega \in \Omega\colon\ X(\omega) \leqslant x\} \cap \{\omega \in \Omega\colon\ X(\omega) \geqslant x\},$$

$$\{\omega \in \Omega\colon\ X(\omega) < x\} = \{\omega \in \Omega\colon\ X(\omega) \leqslant x\} \cap \overline{\{\omega \in \Omega\colon\ X(\omega) = x\}},$$

$$\{\omega \in \Omega\colon\ a \leqslant X(\omega) < b\} = \{\omega \in \Omega\colon\ X(\omega) < b\} \cap \overline{\{\omega \in \Omega\colon\ X(\omega) < a\}},$$

$$\{\omega \in \Omega\colon\ a < X(\omega) \leqslant b\} = \{\omega \in \Omega\colon\ X(\omega) \leqslant b\} \cap \overline{\{\omega \in \Omega\colon\ X(\omega) \leqslant a\}},$$

$$\{\omega \in \Omega\colon\ a \leqslant X(\omega) \leqslant b\} = \{\omega \in \Omega\colon\ X(\omega) \leqslant b\} \cap \overline{\{\omega \in \Omega\colon\ X(\omega) < a\}},$$

$$\{\omega \in \Omega\colon\ a < X(\omega) < b\} = \{\omega \in \Omega\colon\ X(\omega) < b\} \cap \overline{\{\omega \in \Omega\colon\ X(\omega) \leqslant a\}},$$

а σ-алгебра $\mathcal{S}$ замкнута относительно операций пересечения и перехода к противоположному событию. ❑

В дальнейшем, когда мы будем говорить о случайных величинах, мы обычно будем опускать их явную зависимость от элементарных событий и вместо $X(\omega)$ писать просто X, а вместо $X^{-1}(B) = \{\omega \in \Omega\colon\ X(\omega) \in B\}$ — просто $\{X \in B\}$, например, $\{X < x\}$, $\{a \leqslant X \leqslant b\}$ и т. п.

Понятие случайной величины было впервые введено в 1837 г. С. Пуассоном, строгое формальное определение случайной величины было дано в 1933 г. А. Н. Колмогоровым.

Пример 3.1.1. Рассмотрим простейший пример случайной величины — **индикатор события**. Пусть A — некоторое событие. Случайная величина

$$I_A = I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases} \tag{3.1.1}$$

равная единице, когда событие A происходит, и нулю в противном случае, называется *индикатором события A*.

При любом $x \in \mathbb{R}$ множество $\{I_A \leqslant x\}$ можно выразить так:

$$\{I_A \leqslant x\} = \{\omega \in \Omega : I_A(\omega) \leqslant x\} = \begin{cases} \varnothing, & x < 0, \\ \overline{A}, & 0 \leqslant x < 1, \\ \Omega, & x \geqslant 1, \end{cases}$$

т. е. при любом $x \in \mathbb{R}$ множество $\{I_A \leqslant x\}$ является случайным событием (принадлежащим σ-алгебре $\mathcal{S}$), следовательно, I_A действительно является случайной величиной. ❑

Пример 3.1.2. Проверим, является ли число X очков, выпавших при бросании игральной кости, случайной величиной.

Пространство элементарных событий в этом случае имеет вид $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, где элементарное событие ω_i состоит в том, что выпало i очков ($i = 1, 2, 3, 4, 5, 6$). В качестве σ-алгебры $\mathcal{S}$ выберем максимальную σ-алгебру, соответствующую Ω. Вероятности элементарных событий $\mathbf{P}(\omega_i) = 1/6$ ($i = 1, 2, 3, 4, 5, 6$). Число выпавших очков X является функцией от элементарных событий, причем $X(\omega_i) = i$. При этом

$$\{X \leqslant x\} = \{\omega \in \Omega : X(\omega) \leqslant x\} = \begin{cases} \varnothing, & x < 1, \\ \{\omega_1\}, & 1 \leqslant x < 2, \\ \{\omega_1, \omega_2\}, & 2 \leqslant x < 3, \\ \{\omega_1, \omega_2, \omega_3\}, & 3 \leqslant x < 4, \\ \{\omega_1, \omega_2, \omega_3, \omega_4\}, & 4 \leqslant x < 5, \\ \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, & 5 \leqslant x < 6, \\ \Omega, & x \geqslant 6, \end{cases}$$

и все эти множества принадлежат максимальной σ-алгебре $\mathcal{S}$. Поэтому число выпавших очков X является случайной величиной. ❑

Если X и Y — случайные величины, заданные на вероятностном пространстве $\{\Omega, \mathcal{S}, \mathbf{P}\}$, а $c \in \mathbb{R}$ — некоторая неслучайная постоянная, то cX, $X + c$, $|X|$, $X + Y$, $X - Y$, — также случайные величины. Предлагаем любознательному читателю доказать этот факт в задаче **148**.

Задачи

146. Петя и Маша договорились встретиться с 12 до 13 ч на станции метро «ВДНХ» у последнего вагона поезда, идущего в центр города. И Петя, и Маша сказали, что точно придут в промежутке с 12 до 13 ч, но ни один из них не смог указать точное время своего прихода. Они договорились обязательно дождаться друг друга. X — это время, проведенное на станции Петей. Проверьте, является ли X случайной величиной.

147. Докажите, что если σ-алгебра $\mathcal{S}$ — минимальная σ-алгебра (т. е. $\mathcal{S} = \{\varnothing, \Omega\}$), то все случайные величины, которые могут быть заданы на вероятностном пространстве $\{\Omega, \mathcal{S}, \mathbf{P}\}$, являются константами (т. е. не зависят от ω).

148. Докажите, что если X и Y — случайные величины, заданные на вероятностном пространстве $\{\Omega, \mathcal{S}, \mathbf{P}\}$, а $c \in \mathbb{R}$ — некоторая неслучайная постоянная, то cX, $X + c$, $|X|$, $X + Y$, $X - Y$ — также случайные величины.

3.1.2. Функция распределения случайной величины

Законом распределения вероятностей случайной величины называется правило, устанавливающее соответствие между значениями этой случайной величины (или множествами значений) и вероятностями того, что случайная величина примет данное значение (или попадет в соответствующее множество).

Функцией распределения вероятностей (или, короче, *функцией распределения*) случайной величины X называется функция

$$F_X(x) = \mathbf{P}\{X \leqslant x\} = \mathbf{P}\{\omega \in \Omega\colon X(\omega) \leqslant x\}, \qquad x \in \mathbb{R}. \tag{3.1.2}$$

Если известно, о какой случайной величине идет речь, то индекс, обозначающий эту случайную величину, опускается: $F(x) = F_X(x)$.

Для любой случайной величины X и любых чисел x_1, $x_2 \in \mathbb{R}$, таких что $x_1 < x_2$, вероятность попадания случайной величины X в полуинтервал $(x_1; x_2]$ можно рассчитать по формуле

$$\mathbf{P}\{x_1 < X \leqslant x_2\} = F(x_2) - F(x_1). \tag{3.1.3}$$

Доказательство. Если $x_1 < x_2$, то $\{x_1 < X \leqslant x_2\} \subseteq \{X \leqslant x_2\}$, при этом $\{X \leqslant x_1\} = \{X \leqslant x_2\} \setminus \{x_1 < X \leqslant x_2\}$ и по теореме о монотонности вероятности (2.2.27) $\mathbf{P}\{X \leqslant x_2\} = \mathbf{P}\{X \leqslant x_1\} + \mathbf{P}\{(X \leqslant x_2) \setminus (x_1 < X \leqslant x_2)\} = \mathbf{P}\{X \leqslant x_1\} + \mathbf{P}\{x_1 < X \leqslant x_2\}$. Отсюда $\mathbf{P}\{x_1 < X \leqslant x_2\} = \mathbf{P}\{X \leqslant x_2\} - \mathbf{P}\{X \leqslant x_1\} = F(x_2) - F(x_1)$, что и требовалось доказать. ❑

Как числовая функция от числового аргумента x, функция распределения $F(x)$ произвольной случайной величины обладает следующими **свойствами.**

Монотонность функции распределения. *Функция $F(x)$ является неубывающей, т. е. для любых $x_1, x_2 \in \mathbb{R}$, таких что $x_1 < x_2$,*

$$F(x_1) \leqslant F(x_2). \tag{3.1.4}$$

Доказательство. Если $x_1 < x_2$, то по формуле (3.1.3) $\mathbf{P}\{x_1 < X \leqslant x_2\} = F(x_2) - F(x_1)$. Но по аксиоме неотрицательности вероятности (2.2.21) $\mathbf{P}\{x_1 < X \leqslant x_2\} \geqslant 0$, поэтому $F(x_2) - F(x_1) \geqslant 0$ или $F(x_1) \leqslant F(x_2)$, что и требовалось доказать. ❑

Ограниченность функции распределения. *Функция $F(x)$ является ограниченной, т. е. для любого $x \in \mathbb{R}$*

$$0 \leqslant F(x) \leqslant 1, \tag{3.1.5}$$

при этом на концах числовой прямой

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \qquad F(+\infty) = \lim_{x \to +\infty} F(x) = 1. \tag{3.1.6}$$

Доказательство. Неравенство (3.1.4) непосредственно следует из теоремы об ограниченности вероятности (2.2.27). При этом $F(-\infty) = \mathbf{P}\{X \leqslant -\infty\} = \mathbf{P}(\varnothing) = 0$, $F(+\infty) = \mathbf{P}\{X \leqslant +\infty\} = \mathbf{P}(\Omega) = 1$. ❑

Непрерывность функции распределения. *Функция $F(x)$ является непрерывной справа, т. е. для любого $x \in \mathbb{R}$*

$$F(x) = F(x + 0) = \lim_{\substack{z \to x \\ z > x}} F(z). \tag{3.1.7}$$

Доказательство. Неравенство (3.1.7) следует из того, что $F(x+0)=\lim\limits_{\substack{z\to x\\ z>x}}F(z)=$

$$=\lim_{n\to\infty}F\left(x+\frac{1}{n}\right)=\lim_{n\to\infty}\mathbf{P}\left\{X\leqslant x+\frac{1}{n}\right\}=\lim_{n\to\infty}\mathbf{P}\left\{X\leqslant x+\frac{1}{n}\right\}=\mathbf{P}\left\{\bigcap_{n=1}^{\infty}\left(X\leqslant x+\frac{1}{n}\right)\right\}=\mathbf{P}\{X\leqslant x\}=F(x).\ \square$$

Замечание. Можно доказать, что если некоторая функция $F(x)$ обладает свойствами монотонности (3.1.4), ограниченности (3.1.5)—(3.1.6) и непрерывности (3.1.7), то найдется такая случайная величина X, что $F(x)$ будет ее функцией распределения. Доказательство этого факта, а также формул (3.1.6), (3.1.7) можно найти во многих учебниках по теории вероятностей для математиков.

По аналогии с независимыми событиями определяются **независимые случайные величины.** Две случайные величины X и Y называются *независимыми*, если для всех $x, y \in \mathbb{R}$

$$\mathbf{P}\{(X\leqslant x)\cap(Y\leqslant y)\}=\mathbf{P}\{X\leqslant x\}\mathbf{P}\{Y\leqslant y\}=F_X(x)F_Y(y), \tag{3.1.8}$$

т. е. если для всех $x, y \in \mathbb{R}$ события $\{X \leqslant x\}$ и $\{Y \leqslant y\}$ независимы.

Пример 3.1.3. Найдем **функцию распределения индикатора события** A (3.1.1):

$$F_{I_A}(x)=\mathbf{P}\{I_A\leqslant x\}=\mathbf{P}\{\omega\in\Omega: I_A(\omega)\leqslant x\}=\begin{cases}P(\varnothing), & x<0,\\ P(\overline{A}), & 0\leqslant x<1,\\ P(\Omega), & x\geqslant 1\end{cases}$$

или, окончательно,

$$F_{I_A}(x)=\begin{cases}0, & x<0,\\ 1-\mathbf{P}(A), & 0\leqslant x<1,\\ 1, & x\geqslant 1.\ \square\end{cases} \tag{3.1.9}$$

Пример 3.1.4. Пусть нам известна функция распределения некоторой случайной величины X:

$$F_X(x)=\begin{cases}0, & x<0,\\ cx^2, & 0\leqslant x<2,\\ 1, & x\geqslant 2.\end{cases}$$

Определим, какие значения могут быть у параметра c.

Для того, чтобы существовала случайная величина X, которая бы имела функцию распределения $F(x)$, необходимо и достаточно, чтобы $F(x)$ обладала свойствами монотонности (3.1.4), ограниченности (3.1.5)—(3.1.6) и непрерывности (3.1.7). Из условия (3.1.5) следует, что $0 \leqslant cx^2 \leqslant 1$ при $x \in [0; 2)$, откуда $c \in [0; 0{,}25]$. Из условия (3.1.4) следует, что производная $F'(x) \geqslant 0$, значит, $2cx \geqslant 0$ при $x \in [0; 2)$, откуда $c \geqslant 0$. Условия (3.1.6), (3.1.7), очевидно, выполнены. Поэтому $c \in [0; 0{,}25]$.

Если дополнительно нам известно, что $F(x)$ непрерывна в точке $x = 2$, возможное значение постоянной c будет единственным. Из условия непрерывности функции $F(x)$ в точке $x = 2$ следует, что $cx^2\big|_{x=2} = c\cdot 2^2 = 4c = 1$, откуда $c = 0{,}25$.

Теперь найдем, например,

$$\mathbf{P}\{X>1\}=1-\mathbf{P}\{X\leqslant 1\}=1-F(1)=1-\frac{x^2}{4}\bigg|_{x=1}=\frac{3}{4}.\ \square$$

Пример 3.1.5. Найдем функцию распределения случайной величины X, определенной в примере 3.1.2 — числа очков, выпавших при бросании игральной кости.

В примере 3.1.2 мы ввели вероятностное пространство $\{\Omega, \mathcal{S}, \mathbf{P}\}$ и определили события $\{X \leqslant x\}$ для всех $x \in \mathbb{R}$. При этом $\mathbf{P}\{\omega_i\} = 1/6$, поэтому

$$F(x)=\mathbf{P}\{X\leqslant x\}=\begin{cases}\mathbf{P}(\varnothing), & x<1,\\ \mathbf{P}\{\omega_1\}, & 1\leqslant x<2,\\ \mathbf{P}\{\omega_1,\omega_2\}, & 2\leqslant x<3,\\ \mathbf{P}\{\omega_1,\omega_2,\omega_3\}, & 3\leqslant x<4,\\ \mathbf{P}\{\omega_1,\omega_2,\omega_3,\omega_4\}, & 4\leqslant x<5,\\ \mathbf{P}\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5\}, & 5\leqslant x<6,\\ \mathbf{P}(\Omega), & x\geqslant 6,\end{cases}=\begin{cases}0 & x<1,\\ 1/6, & 1\leqslant x<2,\\ 2/6, & 2\leqslant x<3,\\ 3/6, & 3\leqslant x<4,\\ 4/6, & 4\leqslant x<5,\\ 5/6, & 5\leqslant x<6,\\ 1, & x\geqslant 6.\end{cases}\ \square$$

Задачи

149. Функция распределения некоторой случайной величины X такова:

$$F(x)=\begin{cases}0, & x<2,\\ (x-a)^2, & 2\leqslant x<3,\\ 1, & x\geqslant 3.\end{cases}$$

Найдите значение постоянной a, а также вероятность $\mathbf{P}\{1<X\leqslant 2{,}5\}$.

150. Определите, может ли функция

$$F(x)=\begin{cases}0, & x<0,\\ 1-\dfrac{1}{x}, & 0\leqslant x<1,\\ 1, & x\geqslant 1\end{cases}$$

быть функцией распределения какой-либо случайной величины.

151. Найдите функцию распределения случайной величины X, определенной в задаче **146**.

152. Проверьте, являются ли независимыми случайные величины X и Y, если $X=X(\omega)=1$ для всех $\omega\in\Omega$, $Y=Y(\omega)=2$ для всех $\omega\in\Omega$.

153. Пусть $A\in\mathcal{S}$ — некоторое событие, $p=\mathbf{P}(A)$. Проверьте, являются ли независимыми случайная величина $X=I_A$ и случайная величина Y, которая на каждом элементарном событии совпадает с X [т. е. $Y(\omega)=X(\omega)$ для всех $\omega\in\Omega$] в случае, когда: *а*) $0<p<1$; *б*) $p=0$; *в*) $p=1$.

§ 3.2. Дискретные случайные величины и их важнейшие числовые характеристики

3.2.1. Способы задания дискретных случайных величин

Дискретная случайная величина X — это случайная величина, принимающая значения из конечного или счетного множества.

Рассмотрим такую случайную величину, заданную на вероятностном пространстве $\{\Omega,\mathcal{S},\mathbf{P}\}$. Пусть эта случайная величина X принимает значения $x_1, x_2, \ldots, x_n, \ldots$ Для любой случайной величины, в том числе, для X, множество $\{X=x\}=\{\omega\in\Omega\colon X(\omega)=x\}$ (как мы убедились в **п. 3.1.1**) является случайным событием, а потому определена его вероятность при любом $x\in\mathbb{R}$; в частности, определены числа $p_i=\mathbf{P}\{X=x_i\}$, $i=1, 2, \ldots$

З а к о н р а с п р е д е л е н и я дискретной случайной величины задается чаще всего не функцией распределения, а *рядом распределения* — таблицей

$$\begin{array}{c|cccccc} X & x_1 & x_2 & \cdots & x_n & \cdots \\ \hline p & p_1 & p_2 & \cdots & p_n & \cdots \end{array} \tag{3.2.1}$$

в которой $x_1, x_2, \ldots, x_n, \ldots$ — р а з л и ч н ы е значения дискретной случайной величины X, а $p_1, p_2, \ldots, p_n, \ldots$ — отвечающие этим значениям вероятности. Число столбцов в этой таблице может быть конечным (если соответствующая случайная величина принимает конечное число значений) или бесконечным (счетным).

Часто ряд распределения записывают в виде матрицы

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}. \tag{3.2.2}$$

Очевидно, все $p_i \geqslant 0$, $i = 1, 2, \ldots$ [по аксиоме неотрицательности вероятности (1.4.1)], и

$$\sum_i p_i = 1 \tag{3.2.3}$$

[так как $\Omega = (X = x_1) \cup (X = x_2) \cup \cdots \cup (X = x_n) \cup \cdots$, а события $\{X = x_i\}$ и $\{X = x_j\}$ несовместны при $i \neq j$, то по аксиоме счетной аддитивности вероятности (2.2.24) $\mathbf{P}(\Omega) = \sum_i \mathbf{P}\{X = x_i\} = \sum_i p_i$, но по аксиоме нормированности вероятности (2.2.22) $\mathbf{P}(\Omega) = 1$].

Кривой (или *многоугольником*) *распределения вероятностей* дискретной случайной величины X называется ломаная, соединяющая точки $(x_i; p_i)$ в порядке возрастания x_i.

Если значения дискретной случайной величины расположены в ряде распределения по возрастанию,[1] то по ряду распределения дискретной случайной величины можно восстановить ее функцию распределения и наоборот.

Доказательство. Пусть есть дискретная случайная величина, заданная рядом распределения (3.2.1), причем

$$-\infty < x_1 < x_2 < \cdots < x_n < \cdots < +\infty.$$

Тогда

$$\{X \leqslant x\} = \begin{cases} \varnothing, & x < x_1, \\ \{X = x_1\}, & x_1 \leqslant x < x_2, \\ \{X = x_1\} \cup \{X = x_2\}, & x_2 \leqslant x < x_3, \\ \{X = x_1\} \cup \{X = x_2\} \cup \{X = x_3\}, & x_3 \leqslant x < x_4, \\ \quad \vdots & \quad \vdots \\ \bigcup_{i=1}^{n} \{X = x_i\}, & x_n \leqslant x < x_{n+1}, \\ \quad \vdots & \quad \vdots \end{cases}$$

а

$$F(x) = \mathbf{P}\{X \leqslant x\} =$$

[1] А так бывает не всегда, например, невозможно упорядочить по возрастанию с ч е т н о е множество рациональных чисел!

$$= \begin{cases} \mathbf{P}\{\varnothing\}, & x < x_1, \\ \mathbf{P}\{X = x_1\}, & x_1 \leqslant x < x_2, \\ \mathbf{P}\{(X = x_1) \cup (X = x_2)\}, & x_2 \leqslant x < x_3, \\ \mathbf{P}\{(X = x_1) \cup (X = x_2) \cup (X = x_3)\}, & x_3 \leqslant x < x_4, \\ \quad \vdots & \quad \vdots \\ \mathbf{P}\left\{\bigcup\limits_{i=1}^{n} \{X = x_i\}\right\}, & x_n \leqslant x < x_{n+1} \\ \quad \vdots & \quad \vdots \end{cases} = \begin{cases} 0, & x < x_1, \\ p_1, & x_1 \leqslant x < x_2, \\ p_1 + p_2, & x_2 \leqslant x < x_3, \\ p_1 + p_2 + p_3, & x_3 \leqslant x < x_4, \\ \quad \vdots & \quad \vdots \\ \sum\limits_{i=1}^{n} p_i, & x_n \leqslant x < x_{n+1} \\ \quad \vdots & \quad \vdots \end{cases} \quad \square$$

Итак, мы убедились, что функция распределения любой дискретной случайной величины является кусочно-постоянной (ступенчатой) функцией со скачками в точках, соответствующих значениям случайной величины, причем величина скачка в каждой точке равна вероятности того, что случайная величина попадет в эту точку, а значение функции распределения в каждой точке равно сумме вероятностей значений случайной величины, не превосходящих данной точки).[1]

Пусть теперь нам дана некоторая ступенчатая функция

$$F(x) = \begin{cases} 0, & x < x_1, \\ q_1, & x_1 \leqslant x < x_2, \\ q_2, & x_2 \leqslant x < x_3, \\ q_3, & x_3 \leqslant x < x_4, \\ \quad \vdots & \quad \vdots \\ q_n, & x_n \leqslant x < x_{n+1}, \\ \quad \vdots & \quad \vdots \end{cases}$$

обладающая свойствами монотонности (3.1.4) ограниченности (3.1.5)—(3.1.6), и непрерывности (3.1.7).

Тогда $\mathbf{P}\{X = x_i\} = \mathbf{P}\{x_i < X \leqslant x_{i+1}\} = \{\text{по формуле (3.1.3)}\} = F(x_{i+1}) - F(x_i) = q_{i+1} - q_i$ ($i = 1, 2, \ldots$). Если же $x \neq x_i$ ($i = 1, 2, \ldots$), то $\mathbf{P}\{X = x\} = 0$ (предлагаем читателю доказать это самостоятельно).

Пример 3.2.1. Составим **ряд распределения индикатора события** (3.1.1).

Вспомним, что индикатор события может принимать одно из двух значений: нуль (когда A не наступило), или единица (когда A наступило). Поэтому ряд распределения индикатора события A имеет вид

$$\begin{array}{c|cc} I_A & 0 & 1 \\ \hline p & 1 - \mathbf{P}(A) & \mathbf{P}(A) \end{array} \quad \square \tag{3.2.4}$$

Пример 3.2.2. В лотерее на каждые 100 билетов приходится 15 выигрышей. Количество и размеры выигрышей таковы:

Размер выигрыша, руб.	2 000	500	100
Количество билетов	1	4	10

Случайная величина X описывает размер выигрыша на один случайно выбранный билет.

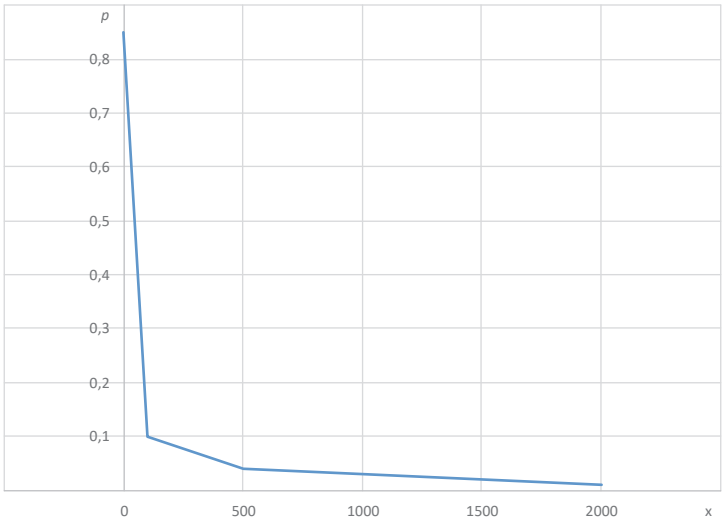
Составим ряд распределения случайной величины X.

[1] Мы убедились в этом, на самом деле, только для случая, когда $x_1 < x_2 < \cdots < x_n < \cdots$. В общем случае функция распределения дискретной случайной величины также обязательно будет кусочно-постоянной; сложность состоит не в доказательстве этого факта, а в невозможности указать способ расстановки по возрастанию произвольного счетного множества, например, невозможно указать такой способ для множества рациональных чисел.

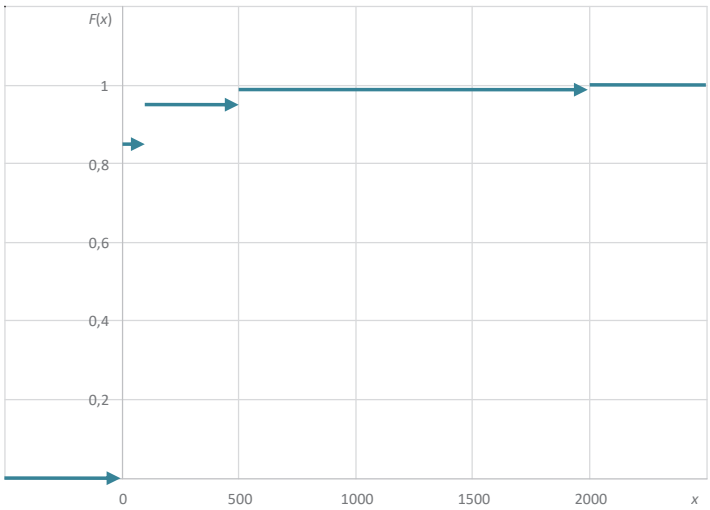
Случайная величина X может принимать значения 2000, 500, 100 и 0 руб. Вероятности этих значений можно вычислить по классической формуле вероятности (1.3.1): $1/100 = 0{,}01$, $4/100 = 0{,}04$, $10/100 = 0{,}10$, $1 - 0{,}01 - 0{,}04 - 0{,}10 = 0{,}85$ соответственно. Расположив пары $(x_i; p_i)$ в порядке возрастания значений случайной величины X, получим ее ряд распределения:

X	0	100	500	2000
p	0,85	0,10	0,04	0,01

Соединив точки $(x_i; p_i)$ отрезками, получим многоугольник распределения случайной величины X (рис. 3.2.1, *а*).



а)



б)

Рис. 3.2.1. *Многоугольник распределения* (*а*) *и график функции распределения* (*б*) *в примере 3.2.2*

Функция распределения случайной величины X имеет следующий вид:

$$F(x)=\mathrm{P}\{X\leqslant x\}=\begin{cases}0, & x<0,\\ 0{,}85, & 0\leqslant x<100,\\ 0{,}85+0{,}10, & 100\leqslant x<500,\\ 0{,}85+0{,}10+0{,}04, & 500\leqslant x<2000,\\ 0{,}85+0{,}10+0{,}04+0{,}01, & x\geqslant 2000\end{cases}=\begin{cases}0, & x<0,\\ 0{,}85, & 0\leqslant x<100,\\ 0{,}95, & 100\leqslant x<500,\\ 0{,}99, & 500\leqslant x<2000,\\ 1, & x\geqslant 2000.\end{cases}$$

Ее график представлен на рис. 3.2.1, *б*.

Вычислим теперь вероятности каких-нибудь событий. Например, вероятность выиграть больше 100 руб. равна $\mathbf{P}\{X>100\}=\mathbf{P}\{X=500\}+\mathbf{P}\{X=2000\}=$ $=0{,}04+0{,}01=0{,}05$.

А вероятность выиграть ровно 55 руб. ($\mathbf{P}\{X=55\}$) и вероятность выиграть более 2500 руб. ($\mathbf{P}\{X>2500\}$) обе равны нулю, так как выигрыш X ни при каких условиях не может оказаться равным 55 руб. или превысить 2500 руб. ❑

Условие независимости дискретных случайных величин. *Пусть дискретная случайная величина X принимает значения x_1, x_2, …, x_m, … с вероятностями p_1, p_2, …, p_m, … соответственно, а дискретная случайная величина Y принимает значения y_1, y_2, …, y_n, … с вероятностями q_1, q_2, …, q_n, … соответственно. Эти случайные величины X и Y являются независимыми тогда и только тогда, когда для всех $i=1,\ 2,\ \ldots,\ m,\ \ldots,\ j=1,\ 2,\ \ldots,\ n,\ \ldots$*

$$\mathbf{P}\{(X=x_i)\cap(Y=y_j)\}=\mathbf{P}\{X=x_i\}\mathbf{P}\{Y=y_j\}. \qquad (3.2.5)$$

Доказательство этого утверждения оставляем любознательному читателю в задаче **161**.

Пример 3.2.3. Независимые случайные величины X и Y распределены одинаково:

X	-5	0	5	10
p	0,1	0,2	0,5	0,2

Y	-5	0	5	10
p	0,1	0,2	0,5	0,2

Составим ряды распределения случайных величин $Z=X^2$ и $V=XY$.

Случайная величина $Z=X^2$ может принимать значения $(-5)^2$, 0^2, 5^2 и 10^2 с теми же вероятностями, с которыми случайная величина X принимает значения (-5), 0, 5 и 10 соответственно. Поэтому случайная величина Z имеет ряд распределения

Z	0	25	100
p	0,2	0,6	0,2

(здесь $\mathbf{P}\{Z=25\}=\mathbf{P}\{X^2=25\}=\mathbf{P}\{X=-5\}+\mathbf{P}\{X=5\}=0{,}1+0{,}5=0{,}6$).

Случайная величина $V=XY$ может иметь в качестве значений произведения любых пар чисел из множества $\{-5;\ 0;\ 5;\ 10\}$, при этом вероятности этих значений вычисляются по теореме умножения вероятностей для независимых событий:

V	$(-5)\cdot(-5)$	$(-5)\cdot 0$	$(-5)\cdot 5$	$(-5)\cdot 10$	$0\cdot(-5)$	$0\cdot 0$	$0\cdot 5$	$0\cdot 10$	$5\cdot(-5)$	$5\cdot 0$	$5\cdot 5$	$5\cdot 10$	$10\cdot(-5)$	$10\cdot 0$	$10\cdot 5$	$10\cdot 10$
p	0,01	0,02	0,05	0,02	0,02	0,04	0,10	0,04	0,05	0,10	0,25	0,10	0,02	0,04	0,10	0,04

или

V	25	0	-25	-50	0	0	0	0	-25	0	25	50	-50	0	50	100
p	0,01	0,02	0,05	0,02	0,02	0,04	0,10	0,04	0,05	0,10	0,25	0,10	0,02	0,04	0,10	0,04

Чтобы получить ряд распределения случайной величины V, сложим вероятности, соответствующие одинаковым значениям этой случайной величины:

V	–50	–25	0	25	50	100
p	0,04	0,10	0,36	0,26	0,20	0,04

❑

Задачи

154. В результате анализа счетов 400 инвесторов на фондовой бирже получена следующая информация о количестве сделок за последний месяц:

X, количество сделок	0	1	2	3	4	5	6	7	8	9	10
Количество инвесторов	146	97	73	34	23	10	6	3	4	2	2

Определите вероятности того, что случайно выбранный инвестор произвел: *а*) нуль сделок; *б*) по крайней мере, одну сделку; *в*) более пяти сделок; *г*) менее шести сделок.

155. Вечером Пете понадобилось купить хлеб. В микрорайоне, где живет Петя, есть три булочных, в каждой из которых в это время хлеб может быть с вероятностью 1/3. Составьте ряд распределения числа N булочных, которые придется посетить Пете.

156. Начальный капитал инвестора, желающего разместить средства в доверительном управлении, составляет 100 000 руб. Опытные коллеги сказали ему, что каждый месяц капитал с вероятностью 0,5 увеличивается в полтора раза, с вероятностью 0,25 остается без изменений и с вероятностью 0,25 уменьшается в полтора раза. Составьте ряд распределения капитала через два месяца.

157. Проект состоит из трех этапов. Первый и второй этапы можно выполнять параллельно, а третий этап можно начинать только по завершении первых двух. Длительности этапов (в рабочих днях) описываются дискретными случайными величинами T_i ($i = 1, 2, 3$) с рядами распределения

T_1	2	3	4
p	0,1	0,8	0,1

T_2	2	3	4
p	0,4	0,4	0,2

T_3	2	3	4
p	0,2	0,3	0,5

Составьте ряд распределения времени выполнения проекта T и найти вероятность того, что от начала работ по проекту до его завершения пройдет более шести рабочих дней.

158. 90% деталей, изготовленных первым заводом, и 80% деталей, изготовленных вторым заводом, обладают высшим качеством. Из деталей, произведенных первым заводом, случайным образом отобрано три штуки, а из деталей, произведенных вторым заводом, — две. Случайная величина X — это количество деталей высшего качества из пяти отобранных. Составьте ряд распределения случайной величины X.

159. Между Петей и Васей происходит дуэль (из-за Маши). Стреляют по очереди до первого попадания. Вероятность того, что попадет Петя, составляет 0,3, вероятность того, что попадет Вася, — 0,6. Петя стреляет первым. Найдите вероятность того, что Вася поразит противника первым.

160. Стопку из двенадцати денежных купюр, из которых ровно три фальшивых, разделили на три равные по числу купюр части. Случайная величина X — это максимальное количество фальшивых купюр, попавших в результате деления в какую-либо часть. Составьте ряд распределения случайной величины X.

161. Докажите формулу (3.2.5) для независимых дискретных случайных величин.

3.2.2. Математическое ожидание, дисперсия и среднее квадратическое отклонение дискретной случайной величины

Наиболее употребительной числовой характеристикой центра группирования значений случайной величины является математическое ожидание.

Математическим ожиданием дискретной случайной величины X называется число

$$\mathbf{E}(X) = \sum_i x_i p_i = \sum_i x_i \mathbf{P}\{X = x_i\}, \tag{3.2.6}$$

равное среднему взвешенному значению случайной величины с весами — вероятностями.

Замечание. В случае, когда дискретная случайная величина принимает счетное число значений, математическое ожидание определено, естественно, только в том случае, когда ряд $\sum_i x_i p_i$ в правой части (3.2.6) сходится абсолютно.

В пакете Microsoft Excel существует функция СУММПРОИЗВ для вычисления суммы попарных произведений элементов массивов (мы уже пользовались этой функцией при вычислениях по формуле полной вероятности в **п. 2.3.2**).

С помощью этой функции можно рассчитать математическое ожидание дискретной случайной величины, принимающей конечное число значений:

$$\mathbf{E}(X) = \sum_i x_i p_i = \text{СУММПРОИЗВ}(\text{<массив } x_i\text{>; <массив } p_i\text{>}).$$

Пример 3.2.4. Найдем **математическое ожидание индикатора события** (3.1.1).

В предыдущем пункте мы нашли ряд распределения (3.2.4) индикатора события A:

I_A	0	1
p	$1-\mathbf{P}(A)$	$\mathbf{P}(A)$

Поэтому математическое ожидание индикатора события A по определению (3.2.6) равно $\mathbf{E}(I_A) = 0[1 - \mathbf{P}(A)] + 1\mathbf{P}(A) = \mathbf{P}(A)$. Итак,

$$\mathbf{E}(I_A) = \mathbf{P}(A). \; \square \tag{3.2.7}$$

Пример 3.2.5. Оценим средний выигрыш игрока в условиях примера 3.2.2, вычислив математическое ожидание случайной величины X.

Ряд распределения случайной величины X был получен в примере 3.2.2. Оценим средний выигрыш на один билет с помощью математического ожидания выигрыша X:

$$\mathbf{E}(X) = 0{\cdot}0{,}85 + 100{\cdot}0{,}1 + 500{\cdot}0{,}04 + 2000{\cdot}0{,}01 = 0 + 10 + 20 + 20 = 50$$

(при расчете математического ожидания случайной величины X мы перемножаем возможные значения x_i случайной величины X на соответствующие этим значениям вероятности p_i и все эти произведения складываем).

Расчеты с помощью Microsoft Excel приведены на рис. 3.2.2. $\square$

	A	B	C	D	E
1	X	0	100	500	2000
2	p	0,85	0,1	0,04	0,01
3	E(X)	=СУММПРОИЗВ(B1:E1; B2:E2)			
4	[X – E(X)]^2	=(B1-$B$3)^2	=(C1-$B$3)^2	=(D1-$B$3)^2	=(E1-$B$3)^2
5	Var(X)	=СУММПРОИЗВ(B4:E4; B2:E2)			
6	sigma_x	=КОРЕНЬ(B5)			

а) формулы

	A	B	C	D	E
1	X	0	100	500	2000
2	p	0,85	0,1	0,04	0,01
3	E(X)	50			
4	[X – E(X)]^2	2500	2500	202 500	3 802 500
5	Var(X)	48 500			
6	sigma_x	220,23			

б) результаты расчетов

Рис. 3.2.2. *Расчеты в примере 3.2.5*

Математическое ожидание дискретной случайной величины обладает следующими **свойствами.**

Формула для математического ожидания константы. *Для любой неслучайной величины* $c \in \mathbb{R}$

$$\mathbf{E}(c) = c. \tag{3.2.8}$$

Доказательство. Рассмотрим случайную величину $Z = Z(\omega) = c$, где $c \in \mathbb{R}$ — некоторая неслучайная величина. Тогда, очевидно, эта случайная величина Z принимает единственное значение $z_1 = c$ с вероятностью $p_1 = 1$, и по определению математического ожидания (3.2.6) $\mathbf{E}(c) = \mathbf{E}(Z) = z_1 p_1 = c \cdot 1 = c$, что и требовалось доказать. ❑

Линейность математического ожидания. *Для любой дискретной случайной величины* X *и любой неслучайной величины* $c \in \mathbb{R}$

$$\mathbf{E}(cX) = c\mathbf{E}(X). \tag{3.2.9}$$

Доказательство. Пусть случайная величина X принимает значения $x_1, x_2, \ldots, x_n, \ldots$ с вероятностями $p_1, p_2, \ldots, p_n, \ldots$ соответственно. Рассмотрим случайную величину $V = cX$, где $c \in \mathbb{R}$ — некоторая неслучайная величина. Тогда, очевидно, эта случайная величина V принимает значения $v_1 = cx_1, v_2 = cx_2, \ldots, v_n = cx_n, \ldots$ с вероятностями $p_1, p_2, \ldots, p_n, \ldots$ соответственно и по определению математического ожидания (3.2.6) $\mathbf{E}(cX) = \mathbf{E}(V) = \sum_i v_i p_i =$

$= \sum_i (cx_i) p_i = c\sum_i x_i p_i = c\mathbf{E}(X)$, что и требовалось доказать. ❑

Аддитивность математического ожидания. *Для любых дискретных случайных величин* X *и* Y

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y). \tag{3.2.10}$$

Доказательство. Пусть случайная величина X принимает значения $x_1, x_2, \ldots, x_m, \ldots$ с вероятностями $p_1, p_2, \ldots, p_m, \ldots$ соответственно, а случайная величина Y принимает значения $y_1, y_2, \ldots, y_n, \ldots$ с вероятностями $q_1, q_2, \ldots, q_n, \ldots$ соответственно. Рассмотрим события $A_i = \{X = x_i\}$, $B_j = \{Y = y_j\}$, $C_{ij} = A_i \cap B_j = \{(X = x_i) \cap (Y = y_j)\}$ для всех возможных $i = 1, 2, \ldots, m, \ldots, j = 1, 2, \ldots, n, \ldots$ При этом C_{ij} будет множеством постоянства и случайной величины X, и случайной величины Y, т. е. множеством постоянства случайной величины $X + Y$, и по определению математического ожидания (3.2.6)

$$\mathbf{E}(X+Y) = \sum_i \sum_j (x_i + y_j)\mathbf{P}(C_{ij}) = \sum_i \left(x_i \sum_j \mathbf{P}(C_{ij}) \right) + \sum_j \left(y_j \sum_i \mathbf{P}(C_{ij}) \right) = \sum_i x_i p_i + \sum_j y_j q_j = \mathbf{E}(X) + \mathbf{E}(Y),$$

что и требовалось доказать [мы учли при доказательстве, что $A_i = \bigcup_j C_{ij}$, $C_{ij} \cap C_{ik} = \varnothing$ (при $j \neq k$),

$B_j = \bigcup_i C_{ij}$, $C_{ij} \cap C_{kj} = \varnothing$ (при $i \neq k$)]. ❑

Формула для математического ожидания произведения независимых случайных величин. *Для любых н е з а в и с и м ы х дискретных случайных величин X и Y*

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y). \tag{3.2.11}$$

Доказательство. Пусть случайная величина X принимает значения $x_1, x_2, ..., x_m, ...$ с вероятностями $p_1, p_2, ..., p_m, ...$ соответственно, а случайная величина Y принимает значения $y_1, y_2, ..., y_n, ...$ с вероятностями $q_1, q_2, ..., q_n, ...$ соответственно, и при этом X и Y независимы, т. е. для любых $i = 1, 2, ..., m, ..., j = 1, 2, ..., n, ...$ $\mathbf{P}\{(X = x_i) \cap (Y = y_j)\} = \mathbf{P}\{X = x_i\}\mathbf{P}\{Y = y_j\}$ [по формуле (3.2.4)]. Тогда по определению математического ожидания (3.2.6) $\mathbf{E}(XY) = \sum_i \sum_j (x_i y_j)\mathbf{P}\{(X = x_i) \cap (Y = y_j)\} = \sum_i \sum_j x_i y_j \mathbf{P}\{X = x_i\}\mathbf{P}\{Y = y_j\} = \sum_i x_i \mathbf{P}\{X = x_i\} \sum_j y_j \mathbf{P}\{Y = y_j\} =$ $= \mathbf{E}(X)\mathbf{E}(Y)$, что и требовалось доказать. ❑

Замечание. Когда случайные величины X и Y зависимы, условие (3.2.11) может и не выполняться. Например, рассмотрим случайные величины $X = X(\omega) = I_A(\omega)$ и $Y = Y(\omega) = X(\omega)$ для всех $\omega \in \Omega$ [где $A \in \mathcal{S}$ — некоторое событие, вероятность которого равна $p = \mathbf{P}(A)$]. Если читатель решал задачу **153**, он убедился в том, что данные случайные величины X и Y являются зависимыми при $0 < p < 1$ и независимыми при $p = 0$ или $p = 1$. При этом $\mathbf{E}(XY) = p$, $\mathbf{E}(X)\mathbf{E}(Y) = p^2$ (предлагаем читателю убедиться в этом самостоятельно), поэтому при $p = 0$ или $p = 1$ (т. е. когда X и Y независимы) $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$, а при $0 < p < 1$ (т. е. когда X и Y зависимы) $\mathbf{E}(XY) \neq \mathbf{E}(X)\mathbf{E}(Y)$.

Формула для математического ожидания неотрицательной целочисленной случайной величины. *Если случайная величина X принимает значения $0, 1, 2, ...$, и у этой случайной величины существует математическое ожидание $\mathbf{E}(X)$, то*

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} \mathbf{P}\{X > x\} = \sum_{x=1}^{\infty} \mathbf{P}\{X \geqslant x\}. \tag{3.2.12}$$

Доказательство. При $k > 1$

$$\sum_{x=0}^{k} x\mathbf{P}\{X = x\} = \sum_{x=1}^{k} x\left[\mathbf{P}\{X > x-1\} - \mathbf{P}\{X > x\}\right] = \sum_{x=1}^{k} (x-1)\mathbf{P}\{X > x-1\} +$$

$$+ \sum_{x=1}^{k} \mathbf{P}\{X > x-1\} - \sum_{x=1}^{k} x\mathbf{P}\{X > x\} = \sum_{x=0}^{k-1} \mathbf{P}\{X > x\} - k\mathbf{P}\{X > k\}.$$

Если устремить k к ∞, то $k\mathbf{P}\{X > k\} = k \sum_{x=k+1}^{\infty} \mathbf{P}\{X = x\} \leqslant \sum_{x=k}^{\infty} x\mathbf{P}\{X = x\} \to 0$ в силу конечности ряда $\mathbf{E}(X) = \sum_{x=0}^{\infty} x\mathbf{P}\{X = x\} < \infty$, и

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} x\mathbf{P}\{X = x\} = \lim_{k\to\infty} \sum_{x=1}^{k} x\left[\mathbf{P}\{X > x-1\} - \mathbf{P}\{X > x\}\right] =$$

$$= \lim_{k\to\infty} \sum_{x=0}^{k-1} \mathbf{P}\{X > x\} - \lim_{k\to\infty} \left(k\mathbf{P}\{X > k\}\right) = \sum_{x=0}^{\infty} \mathbf{P}\{X > x\},$$

что и требовалось доказать. ❑

Наиболее употребительной числовой характеристикой степени разброса значений случайной величины (произвольной, не обязательно дискретной) вокруг центра группирования является дисперсия.

Дисперсией случайной величины X называется число

$$\mathbf{Var}(X) = \mathbf{E}\{[X - \mathbf{E}(X)]^2\}, \tag{3.2.13}$$

равное математическому ожиданию квадрата отклонения случайной величины от своего математического ожидания.

В теоретических выкладках иногда проще использовать еще две формулы для вычисления дисперсии:

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2, \tag{3.2.14}$$

$$\mathbf{Var}(X) = \mathbf{E}[X(X-1)] + \mathbf{E}(X) - [\mathbf{E}(X)]^2. \tag{3.2.15}$$

Доказательство. Имеем: $\mathbf{Var}(X) = \mathbf{E}[X - \mathbf{E}(X)]^2 = \mathbf{E}\{X^2 - 2X\mathbf{E}(X) + [\mathbf{E}(X)]^2\} = \mathbf{E}(X^2) - 2\mathbf{E}[X\mathbf{E}(X)] + \mathbf{E}[\mathbf{E}(X)^2] = \mathbf{E}(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + [\mathbf{E}(X)]^2 = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2$, что доказывает справедливость формулы (3.2.14). Случайную величину X^2 представим в виде $X^2 = X(X-1) + X$. При этом $\mathbf{E}(X^2) = \mathbf{E}[X(X-1) + X] = \mathbf{E}[X(X-1)] + \mathbf{E}(X)$, и из формулы (3.2.14) получаем: $\mathbf{Var}(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 = \mathbf{E}[X(X-1)] + \mathbf{E}(X) - [\mathbf{E}(X)]^2$, что доказывает справедливость формулы (3.2.15). ❑

Замечание. Формулы (3.2.14)—(3.2.15) категорически нельзя использовать для практических расчетов, если случайная величина имеет сколь-нибудь большое число различных возможных значений, поскольку эти формулы обладают существенно большей вычислительной погрешностью, чем формула (3.2.13).

Для дискретных случайных величин формулы (3.2.13)—(3.2.15) принимают вид

$$\mathbf{Var}(X) = \sum_i [x_i - \mathbf{E}(X)]^2 p_i, \tag{3.2.16}$$

$$\mathbf{Var}(X) = \sum_i x_i^2 p_i - [\mathbf{E}(X)]^2, \tag{3.2.17}$$

$$\mathbf{Var}(X) = \sum_i x_i(x_i - 1)p_i + \mathbf{E}(X) - [\mathbf{E}(X)]^2 \tag{3.2.18}$$

соответственно.

Пример 3.2.6. Найдем **дисперсию индикатора события** (3.1.1).

Ранее мы составили ряд распределения (3.2.4) индикатора события A:

I_A	0	1
p	$1 - \mathbf{P}(A)$	$\mathbf{P}(A)$

и нашли его математическое ожидание (3.2.7): $\mathbf{E}(I_A) = \mathbf{P}(A)$. Дисперсия индикатора по формуле (3.2.16) равна $\mathbf{Var}(I_A) = [0 - \mathbf{E}(I_A)]^2[1 - \mathbf{P}(A)] + [1 - \mathbf{E}(I_A)]^2\mathbf{P}(A) = [0 - \mathbf{P}(A)]^2[1 - \mathbf{P}(A)] + [1 - \mathbf{P}(A)]^2\mathbf{P}(A) = [\mathbf{P}(A)]^2[1 - \mathbf{P}(A)] + [1 - \mathbf{P}(A)]^2\mathbf{P}(A) = \mathbf{P}(A)[1 - \mathbf{P}(A)][\mathbf{P}(A) + 1 - \mathbf{P}(A)] = \mathbf{P}(A)[1 - \mathbf{P}(A)]$. Можно вычислить дисперсию индикатора и по формуле (3.2.17): $\mathbf{Var}(I_A) = 0^2[1 - \mathbf{P}(A)] + 1^2\mathbf{P}(A) - [\mathbf{E}(I_A)]^2 = 0 + \mathbf{P}(A) - [\mathbf{P}(A)]^2 = \mathbf{P}(A)[1 - \mathbf{P}(A)]$. Естественно, обе формулы дали одинаковый роезультат:

$$\mathbf{Var}(I_A) = \mathbf{P}(A)[1 - \mathbf{P}(A)]. \text{ ❑} \tag{3.2.19}$$

Дисперсия дискретной случайной величины обладает следующими **свойствами.**

Формула для дисперсии константы. *Для любой неслучайной величины* $c \in \mathbb{R}$

$$\mathbf{Var}(c) = 0; \tag{3.2.20}$$

а если для некоторой случайной величины X *выполняется равенство* $\mathbf{Var}(X) = 0$, *то существует такое число* $c = \mathbf{E}(X)$, *что* $\mathbf{P}\{X = c\} = 1$, *т. е. с вероятностью, равной единице,*[1] *эта случайная величина равна константе.*

Доказательство. По определению дисперсии (3.2.13) имеем: $\mathbf{Var}(c) = \mathbf{E}[c - \mathbf{E}(c)]^2$. Но по свойству (3.2.8) $\mathbf{E}(c) = c$, поэтому $\mathbf{Var}(c) = \mathbf{E}[c - \mathbf{E}(c)]^2 = \mathbf{E}[(c - c)^2] = \mathbf{E}(0^2) = \mathbf{E}(0)$. Поскольку 0 — это число, вновь воспользуемся свойством (3.2.8): $\mathbf{E}(0) = 0$. Окончательно получаем: $\mathbf{Var}(c) = \mathbf{E}[(c - \mathbf{E}(c)]^2 = \mathbf{E}[(c - c)^2] = \mathbf{E}(0^2) = \mathbf{E}(0) = 0$, что доказывает первую часть утверждения. Вторую часть утверждения мы докажем в **п 4.1.1**. ❑

[1] В таких случаях говорят «почти наверное».

Квадратичная нелинейность дисперсии. *Для любой случайной величины X и любой неслучайной величины $c \in \mathbb{R}$*

$$\mathbf{Var}(cX) = c^2\mathbf{Var}(X). \qquad (3.2.21)$$

Доказательство. По определению дисперсии (3.2.13) имеем: $\mathbf{Var}(cX) = \mathbf{E}\{[cX - \mathbf{E}(cX)]^2\}$. По свойству (3.2.9) $\mathbf{E}(cX) = c\mathbf{E}(X)$, поэтому $\mathbf{Var}(cX) = \mathbf{E}\{[cX - c(\mathbf{E}\{X\})]^2\} = \mathbf{E}\{c^2[X - \mathbf{E}(X)]^2\}$. Вновь воспользуемся свойством (3.2.9): $\mathbf{E}\{c^2[X - \mathbf{E}(X)]^2\} = c^2\mathbf{E}\{[X - \mathbf{E}(X)]^2\} = c^2\mathbf{Var}(X)$, что и требовалось доказать. ❑

Аддитивность дисперсии независимых случайных величин (Равенство Бьенэме). *Для любых н е з а в и с и м ы х случайных величин X и Y*

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y). \qquad (3.2.22)$$

Доказательство. По формуле (3.2.14) имеем: $\mathbf{Var}(X + Y) = \mathbf{E}[(X + Y)^2] - [\mathbf{E}(X + Y)]^2$. Во-первых, $\mathbf{E}[(X + Y)^2] = \mathbf{E}(X^2 + 2XY + Y^2)$. Раскроем скобки в последнем выражении, воспользовавшись аддитивностью математического ожидания (3.2.10): $\mathbf{E}[(X + Y)^2] = \mathbf{E}(X^2 + 2XY + Y^2) = \mathbf{E}(X^2) + \mathbf{E}(2XY) + \mathbf{E}(Y^2)$. По свойству линейности математического ожидания (3.2.9) $\mathbf{E}(2XY) = 2\mathbf{E}(XY)$, поэтому $\mathbf{E}[(X + Y)^2] = \mathbf{E}(X^2) + 2\mathbf{E}(XY) + \mathbf{E}(Y^2)$. Во-вторых, по свойству аддитивности математического ожидания (3.2.10) $\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y)$, поэтому $[\mathbf{E}(X + Y)]^2 = [\mathbf{E}(X) + \mathbf{E}(Y)]^2 = [\mathbf{E}(X)]^2 + 2\mathbf{E}(X)\mathbf{E}(Y) + [\mathbf{E}(Y)]^2$. Собирая все вместе, получим: $\mathbf{Var}(X + Y) = \mathbf{E}[(X + Y)^2] - [\mathbf{E}(X + Y)]^2 = \mathbf{E}(X^2) + 2\mathbf{E}(XY) + \mathbf{E}(Y^2) - [\mathbf{E}(X)]^2 - 2\mathbf{E}(X)\mathbf{E}(Y) - [\mathbf{E}(Y)]^2 = \{\mathbf{E}(X^2) - [\mathbf{E}(X)]^2\} + \{\mathbf{E}(Y^2) - [\mathbf{E}(Y)]^2\} + 2[\mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)]$. Два слагаемых в фигурных скобках в последнем выражении представляют собой выражения для $\mathbf{Var}(X)$ и $\mathbf{Var}(Y)$ по формуле (3.2.14), поэтому $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2[\mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)]$. Теперь для доказательства формулы (3.2.22) достаточно заметить, что поскольку случайные величины X и Y независимы, то по формуле (3.2.11) для математического ожидания произведения независимых случайных величин $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$, значит, если X и Y независимы, то $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$, что и требовалось доказать. ❑

Средним квадратическим отклонением случайной величины X называется неотрицательное значение квадратного корня из дисперсии:

$$\sigma_X = +\sqrt{\mathbf{Var}(X)}\,. \qquad (3.2.23)$$

Пример 3.2.7. Оценим разброс выигрыша игрока в условиях примеров 3.2.2, 3.2.5, вычислив среднее квадратическое отклонение случайной величины X.

Дисперсия $\mathbf{Var}(X) = \mathbf{E}\{[X - \mathbf{E}(X)]^2\} = (0 - 50)^2 \cdot 0{,}85 + (100 - 50)^2 \cdot 0{,}1 + (500 - 50)^2 \cdot 0{,}04 + (2000 - 50)^2 \cdot 0{,}01 = 2500 \cdot 0{,}85 + 2500 \cdot 0{,}1 + 202\,500 \cdot 0{,}04 + 3802500 \cdot 0{,}01 = 48\,500$ (при расчете математического ожидания случайной величины $[X - \mathbf{E}(X)]^2$ мы перемножаем возможные значения $[x_i - \mathbf{E}(X)]^2$ случайной величины $[X - \mathbf{E}(X)]^2$ на соответствующие этим значениям вероятности p_i (ведь случайная величина $[X - \mathbf{E}(X)]^2$ принимает значения $[x_i - \mathbf{E}(X)]^2$ тогда и только тогда, когда случайная величина X принимает значения x_i, а это происходит с вероятностями p_i) и все эти произведения складываем).

Наконец, среднее квадратическое отклонение $\sigma_X = \sqrt{\mathbf{Var}(X)} = \sqrt{48\,500} \approx 220{,}23$. Расчеты с помощью Microsoft Excel приведены на рис. 3.2.2. ❑

Знание математических ожиданий и средних квадратических отклонений случайных величин очень важно при анализе группы случайных величин, оно помогает выбрать из множества случайных величин о п т и м а л ь н ы е п о П а р е т о, отбросив заведомо «плохие».

Пусть, например, есть набор случайных величин R_1, R_2, ..., R_n — случайные будущие доходности по n финансовым инструментам, и нам известны *ожидаемые доходности* $\mathbf{E}(R_1)$, $\mathbf{E}(R_2)$, ..., $\mathbf{E}(R_n)$ (математические ожидания соответствующих случайных величин) и *риски* $\sigma_1, \sigma_2, \ldots, \sigma_n$ (средние квадратические отклонения).

Говорят, что i-я операция *доминирует* j-ю, если

$$\begin{cases}\mathbf{E}(R_i) \geqslant \mathbf{E}(R_j), \\ \sigma_i < \sigma_j\end{cases} \quad \text{или} \quad \begin{cases}\mathbf{E}(R_i) > \mathbf{E}(R_j), \\ \sigma_i \leqslant \sigma_j.\end{cases} \tag{3.2.24}$$

Операция называется *оптимальной по Парето*, если не существует операций, которые бы ее доминировали. Отметим, что операции, оптимальные по Парето, не обязательно являются «самыми лучшими» (и даже просто «хорошими») — эти операции не являются худшими. Выбор операций среди оптимальных по Парето осуществляется на основе склонности лица, принимающего соответствующее решение, к риску.

Пример 3.2.8. Инвестор рассматривает четыре инвестиционные операции со случайными доходностями, описываемыми случайными величинами R_1, R_2, R_3 и R_4 с рядами распределения

R_1	2	4	5	8
p	1/6	1/6	1/2	1/6

R_2	2	3	4	12
p	1/2	1/6	1/6	1/6

R_3	3	5	8	10
p	1/6	1/6	1/2	1/6

R_4	1	2	4	8
p	1/2	1/6	1/6	1/6

Ожидаемые доходности и риски операций равны соответственно $\mathbf{E}(R_1) = 4{,}83$, $\sigma_1 = 1{,}77$, $\mathbf{E}(R_2) = 4{,}17$, $\sigma_2 = 3{,}58$, $\mathbf{E}(R_3) = 7{,}00$, $\sigma_3 = 2{,}31$, $\mathbf{E}(R_4) = 2{,}83$, $\sigma_4 = 2{,}54$. Нанесем точки $\left(\mathbf{E}(R_i), \sigma_i\right)$ на график (рис. 3.2.3).

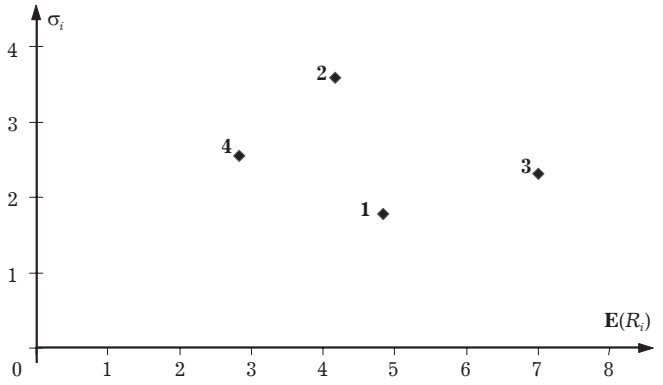


Рис. 3.2.3. *График «риск — доходность» в примере 3.2.8*

Видно, что первая операция доминирует вторую и четвертую, третья операция также доминирует вторую и четвертую. При этом первая операция не доминирует третью, а третья не доминирует первую. Первая и третья операции, таким образом, оптимальны по Парето. ❑

Пример 3.2.9. Пусть нам известно, что $\mathbf{E}(X_1) = \mathbf{E}(X_2) = \cdots = \mathbf{E}(X_n) = a$, $\mathbf{Var}(X_1) = \mathbf{Var}(X_2) = \cdots = \mathbf{Var}(X_n) = \sigma^2$.

Найдем математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины $\bar{X} = \sum_{i=1}^{n} X_i \Big/ n$.

Имеем:

$$\mathbf{E}(\overline{X}) = \mathbf{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{1}{n}(na) = a,\ \mathbf{Var}(\overline{X}) = \mathbf{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n},$$

$$\sigma_{\overline{X}} = \sqrt{\mathbf{Var}(\overline{X})} = \sqrt{\frac{\mathbf{Var}(X)}{n}} = \frac{\sqrt{\mathbf{Var}(X)}}{\sqrt{n}} = \frac{\sigma_X}{\sqrt{n}}. \square$$

Некоторые семейства однотипных случайных величин часто встречаются в различных практических ситуациях. Такие семейства распределений хорошо изучены, в частности, вычислены их основные числовые характеристики в зависимости от параметров (см. **прил. 2**). С одним из таких семейств мы уже знакомы — это семейство индикаторов событий (2.1.1).

В следующем параграфе мы познакомимся подробнее с другими дискретными случайными величинами, часто встречающимися в экономической практике.

Задачи

162. В условиях задачи **154** найдите математическое ожидание и среднее квадратическое отклонение числа сделок случайно выбранного инвестора.

163. Найдите математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины X, заданной рядом распределения

X	x_1	x_2	$\cdots$	x_n
p	$\frac{1}{n}$	$\frac{1}{n}$	$\cdots$	$\frac{1}{n}$

164. Инструментальный завод продает роботизированные линии по переработке смартфонов. Стоимость одной линии составляет 10 000 000 руб., причем все линии продаются с рассрочкой платежа. При этом прибыль завода при продаже одной линии составляет 5 000 000 руб. Из опыта известно, что примерно 10% покупателей не выполняют условия рассрочки. Поэтому для уменьшения риска завод страхует каждый контракт на случай невыполнения покупателем условий рассрочки платежа. По договору страхования завод платит страховой компании 500 000 руб., и в случае длительной просрочки оплаты покупателем страховая компания возмещает заводу 80% от долга (т. е. 8 000 000 руб.). Составьте ряды распределения следующих случайных величин: X_1 — прибыли завода от продажи одной линии в случае отказа от страхования рассрочки платежа; X_2 — прибыли завода от продажи одной линии в случае страхования рассрочки платежа; Y — прибыли страховой компании от страхования рассрочки платежа. Найдите математические ожидания и средние квадратические отклонения случайных величин X_1, X_2, Y и прокомментируйте их экономический смысл. Стоит ли заводу страховать риск невыполнения покупателями договора рассрочки?

165. Инвестор рассматривает три финансовых инструмента со случайными доходностями, описываемыми случайными величинами R_1, R_2, и R_3, которые заданы рядами распределения

R_1	−5	0	5	10
p	0,1	0,2	0,5	0,2

R_2	−5	0	5	10
p	0,3	0,2	0,1	0,4

R_3	−5	0	5	10
p	0,1	0,2	0,6	0,1

Найдите ожидаемые доходности и риски операций, нанесите точки $\left(\mathbf{E}(R_i), \sigma_i\right)$ на график, определите операции, оптимальные по Парето.

166. Инвестор рассматривает четыре инвестиционные операции со случайными эффективностями, описываемыми случайными величинами R_1, R_2, R_3 и R_4 с рядами распределения

R_1	-5	0	5	10
p	0,1	0,2	0,5	0,2

R_2	-5	0	5
p	0,1	0,4	0,5

R_3	-5	0	5	10
p	0,4	0,1	0,1	0,4

R_4	-5	0	10
p	0,1	0,7	0,2

Найдите ожидаемые доходности и риски операций, нанесите точки $\left(\mathbf{E}(R_i), \sigma_i\right)$ на график, определите операции, оптимальные по Парето.

167. Клиент должен вернуть банку кредит до сегодняшнего дня. Неделю назад он отправил денежный перевод из другого банка, который до сих пор не дошел. Время T прибытия денег оценивается клиентом так:

T	1	2	3	4	5
p	0,3	0,3	0,2	0,1	0,1

За каждый день опоздания возврата кредита клиент должен выплатить банку 3% от его суммы (проценты простые). Есть возможность обратиться к специалисту по финансовым расследованиям, который обязуется за 5% от суммы разыскать ее в течение дня. Определите, что клиенту выгоднее — обратиться к детективу или ждать прихода денег, вычислив математическое ожидание времени прибытия денег.

168. Оцените ожидаемое время T, которое Петя потратит на поход за хлебом в условиях задачи **155**, если известно, что на посещение каждого булочной уходит полчаса.

169. В условиях задачи **156** найдите математическое ожидание капитала инвестора через два месяца.

170. В условиях задачи **157** найдите математическое ожидание и дисперсию времени выполнения проекта.

171. Независимые случайные величины X и Y заданы рядами распределения

X	-1	0	1
p	0,1	0,1	?

Y	-2	2
p	?	0,7

(знаком «?» отмечены неизвестные вероятности).

Найдите $\mathbf{E}(X)$, $\mathbf{E}(Y)$, $\mathbf{Var}(X)$, $\mathbf{Var}(Y)$. Составьте ряд распределения случайной величины $Z = X + Y$, найдите $\mathbf{E}(Z)$ и $\mathbf{Var}(Z)$, убедитесь в справедливости формул (3.2.10) и (3.2.21). Составьте ряд распределения случайной величины $V = XY$, найдите $\mathbf{E}(V)$ и $\mathbf{Var}(V)$, убедитесь в справедливости формулы (3.2.11). Составьте ряд распределения случайной величины $W = \min\{0; X\}$, найдите $\mathbf{E}(W)$ и $\mathbf{Var}(W)$.

172. Случайная величина X принимает значения 7; 9; 10; 11 и 13 (каждое с вероятностью 1/5), а случайная величина Y принимает значения 22; 24; 25; 26; 28 (также каждое с вероятностью 1/5). Найдите $\mathbf{Var}(X)$ и $\mathbf{Var}(Y)$; проверьте, выполняется ли равенство $\mathbf{Var}(Y) = \mathbf{Var}(X)$.

173. Петя поехал на каникулы на n дней и решил, что будет ежедневно тратить соответствующую часть денег: в первый день — $1/n$, во второй день — $1/(n-1)$ от остатка и т. д. Пусть

X_i — часть от остатка денег, которая отделяется на расходы в i-й день ($i = 1, 2, ..., n$). Здраво понимая, что траты каждый день будут различными, Петя решил, что их можно описать независимыми случайными величинами X_i с математическими ожиданиями $\mathbf{E}(X_i) = 1/(n-i-1)$. Найдите математическое ожидание случайной величины $Y = (1 - X_1)(1 - X_2)\cdots(1 - X_n)$, равной остатку денег к последнему дню.

174. Один час занятий в 25% фитнес-центров города стоит 2000 руб, в 25% фитнес-центров — 4000 руб., а в 50% фитнес-центров час занятий стоит 6000 руб. Маша случайно выбирает четыре фитнес-центра, и идет в тот, стоимость часа занятий в котором ниже (если таких центров несколько, Маша выбирает любой из них). Оцените ожидаемую стоимость часа занятий в том фитнес-центре, который выберет Маша.

175. Выразите $\mathbf{Var}(XY)$ для независимых дискретных случайных величин X и Y через $\mathbf{E}(X)$, $\mathbf{E}(Y)$, $\mathbf{E}(X^2)$, и $\mathbf{E}(Y^2)$.

176. Докажите, что для независимых дискретных случайных величин X и Y $\mathbf{Var}(XY) \geqslant \mathbf{Var}(X)\mathbf{Var}(Y)$.

177. Пусть X — дискретная случайная величина, $\overset{\circ}{X} = \dfrac{X - \mathbf{E}(X)}{\sigma_X}$. Докажите, что $\mathbf{E}(\overset{\circ}{X}) = 0$, $\mathbf{Var}(\overset{\circ}{X}) = 1$.

178. Найдите среднюю продолжительность $m(x)$ игры, описанной в примере 2.3.5.

§ 3.3. Дискретные случайные величины, часто встречающиеся в экономической практике

3.3.1. Биномиальный закон распределения

Случайная величина X называется распределенной *по биномиальному закону* с параметрами n, p (где $n \in \mathbb{N}$, $0 < p < 1$), если она принимает значения 0, 1, 2, ..., n с вероятностями

$$\mathbf{P}\{X = x\} = C_n^x p^x (1-p)^{n-x} \qquad (x = 0, 1, 2, ..., n). \tag{3.3.1}$$

Поскольку вероятности $\mathbf{P}\{X = x\}$ рассчитываются по формуле Бернулли (2.4.1), случайную величину X, распределенную по биномиальному закону (3.3.1), можно интерпретировать как число успехов в последовательности из n испытаний Бернулли, в каждом из которых успех может произойти с вероятностью p.

Формула (3.3.1) действительно определяет дискретную случайную величину, поскольку все $\mathbf{P}\{X = x\} \geqslant 0$ и

$$\sum_{x=0}^{n} \mathbf{P}\{X = x\} = \sum_{x=0}^{n} C_n^x p^x (1-p)^{n-x} = [p + (1-p)]^n = 1^n = 1.$$

Тот факт, что случайная величина X распределена по биномиальному закону с параметрами n, p, кратко обозначается так: $X = \mathrm{Bin}(n; p)$.

Математическое ожидание случайной величины $X = \mathrm{Bin}(n; p)$ равно

$$\mathbf{E}(X) = np, \tag{3.3.2}$$

а ее дисперсия

$$\mathbf{Var}(X) = np(1 - p). \tag{3.3.3}$$

Доказательство. Пусть события A_i состоят в том, что в i-м из n испытаний Бернулли произошел успех ($i = 0, 1, 2, ..., n$). Рассмотрим n случайных величин $X_i = I_{A_i}$ — индикаторов со-

бытий A_i. При этом по формуле (3.2.7) $\mathbf{E}(X_i)=\mathbf{E}(I_{A_i})=\mathbf{P}(A_i)=p$, по формуле (3.2.19) $\mathbf{Var}(X_i)=\mathbf{Var}(I_{A_i})=\mathbf{P}(A_i)[1-\mathbf{P}(A_i)]=p(1-p)$, а случайная величина X — число успехов в последовательности из n испытаний Бернулли — очевидно, равна $X=\sum_{i=1}^{n}X_i$. По свойству аддитивности математического ожидания (2.2.10) $\mathbf{E}(X)=\sum_{i=1}^{n}\mathbf{E}(X_i)=np$, что доказывает формулу (3.3.2). По свойству аддитивности дисперсии независимых случайных величин (3.2.22) $\mathbf{Var}(X)=\sum_{i=1}^{n}\mathbf{Var}(X_i)=np(1-p)$ [а случайные величины X_i независимы, так как независимы испытания Бернулли], что доказывает формулу (3.3.3). ❑

В пакете Microsoft Excel для вычисления вероятностей $\mathbf{P}\{X=x\}$ и $\mathbf{P}\{X\leqslant x\}$ можно пользоваться функциями

$$\mathbf{P}\{X=x\}=\text{БИНОМ.РАСП(<}x\text{>; <}n\text{>; <}p\text{>; ЛОЖЬ)},$$

$$\mathbf{P}\{X\leqslant x\}=\text{БИНОМ.РАСП(<}x\text{>; <}n\text{>; <}p\text{>; ИСТИНА)}.$$

Пример 3.3.1. В пяти из 30 филиалов компании допускаются нарушения финансовой дисциплины. Тщательной проверке (которая гарантированно выявляет нарушения) подвергаются 10 случайно выбранных филиалов. Найти ожидаемое количество и среднее квадратическое отклонение количества филиалов, в которых выявлены нарушения.

Ожидаемое количество выявленных нарушений можно найти как математическое ожидание случайной величины X, распределенной по биномиальному закону с параметрами $n=10$, $p=5/30=1/6$: $\mathbf{E}(X)=np=5/3=1\frac{2}{3}$, среднее квадратическое отклонение равно при этом $\sigma_X=\sqrt{\mathbf{Var}(X)}=\sqrt{np(1-p)}=$ $=\sqrt{10\cdot\frac{1}{6}\cdot\frac{5}{6}}=\sqrt{\frac{50}{36}}=\frac{5\sqrt{2}}{6}\approx 1{,}18$. ❑

Пример 3.3.2. Построим ожидаемое распределение результатов испытаний, которое было бы получено для 256 абсолютно невежественных экзаменующихся, случайно угадывающих ответы на четыре вопроса с четырьмя возможными вариантами ответа на каждый вопрос (из которых один и только один верен).

Угадывание каждым экзаменующимся ответов на четыре вопроса можно интерпретировать как $n=4$ испытания Бернулли. При этом, поскольку экзаменующийся невежествен, для него равновероятны все четыре ответа на каждый вопрос, т. е. вероятность успеха (правильного ответа на вопрос) равна $p=1/4$.

Тогда число X угаданных одним экзаменующимся ответов на четыре вопроса представляет собой биномиальную случайную величину $X=$ $=\mathrm{Bin}(n=4;\ p=1/4)$ и $\mathbf{P}\{X=x\}=C_n^x p^x(1-p)^{n-x}=C_4^x\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{4-x}$, $x=0,\ 1,\ 2,\ 3,\ 4$, а ожидаемое распределение результатов для 256 экзаменующихся, учитывая их независимость друг от друга, будет иметь следующий вид:

Число правильных ответов, X	0	1	2	3	4	
Число экзаменующихся, $256\mathbf{P}\{X=x_i\}$	81	108	54	12	1	$\sum_{i=0}^{4}256\mathbf{P}\{X=x_i\}=256$

❑

Задачи

179. Случайная величина $X = \mathrm{Bin}(n = 5; p = 2/3)$. Составьте ряд распределения этой случайной величины, найти ее функцию распределения, математическое ожидание и дисперсию. Постройте график ее функции распределения.

180. Абитуриент при поступлении в университет сдает четыре экзамена, вероятность успешно сдать каждый экзамен равна 0,8. Случайная величина X описывает число сданных абитуриентом экзаменов (в предположении, что различные экзамены представляют собой независимые испытания). Составьте ряд распределения случайной величины X.

181. В группе из 16 человек 12 поддерживают некоторую муниципальную программу. Из этой группы наудачу отбирают троих человек. Составьте ряд распределения числа людей в выборке, поддерживающих программу, найдите среднее число таких людей и дисперсию числа таких людей.

182. Финансовый инструмент *форвард* — это сделка купли-продажи некоторого товара в определенный момент в будущем по цене, определяемой сторонами в настоящий момент времени. Фермер предполагает, что через месяц, когда он соберет урожай, цена пшеницы в каждом из десяти регионов, куда он обычно ее продает, может с вероятностью 0,9 понизиться и с вероятностью 0,1 повыситься. Поэтому он заключает с десятью мельниками в этих регионах десять форвардов на поставку им пшеницы через месяц по сегодняшней цене. Цены в регионах изменяются независимо. Найдите математическое ожидание числа форвардов, которые окажутся выгодными для фермера и вероятность того, что все десять проданных форвардов окажутся для него выгодными (форвард окажется выгодным, если в данном регионе за месяц цена понизится).

183. Две игральных кости подбрасывают до тех пор, пока хотя бы на одной из них не выпадет шестерка. Найдите вероятность того, что две кости будут брошены ровно k раз ($k = 1, 2, 3, \ldots$).

3.3.2. Биномиальная модель ценообразования финансовых инструментов

Предположим, что инвестор имеет возможность:

- размещать средства на банковском счете и брать с него в долг;
- покупать и продавать ценные бумаги.

Тогда для этого инвестора на рынке существует один безрисковый актив (банковский счет) B и n рисковых активов (ценных бумаг) $S^{(1)}, S^{(2)}, \ldots, S^{(n)}$.

Будем считать, что проценты на банковский счет начисляются по схеме сложных процентов с постоянной ставкой $i(t) = i = \mathrm{const}$, так что в конце каждого периода сумма на счете увеличивается в $(1 + i)$ раз:

$$B_t = B_{t-1}(1+i).$$

Это означает, что за время t сумма на банковском счете увеличивается с B_0 до

$$B_t = B_0(1+i)^t.$$

Отсюда следует, что процентная ставка, начисляемая за время t (не обязательно за целое число лет), составляет

$$i_{(t)} = (1+i)^t - 1,$$

например, процентная ставка за полгода равна

$$i_{(1/2)} = (1+i)^{1/2} - 1,$$

за месяц —

$$i_{(1/12)} = (1+i)^{1/12} - 1$$

и т. п.

Предположим для простоты, что на рынке обращается о д н а ценная бумага (для определенности — акция), и ее стоимость в конце периода времени t составляет S_t.

Будем предполагать, что *операционные издержки*, связанные с переводом средств между активами, отсутствуют, а также что активы являются *безгранично делимыми*, т. е. можно купить и продать любую часть ценной бумаги, положить на счет и снять с него любую его часть.

Функционирующий по таким правилам рынок называется (B, S)-*рынком*.

Рассмотрим поведение стоимости акции S_t на (B, S)-рынке в течение года, т. е. в течение периода времени $[0; 1]$.

Предположим, что в течение данного периода времени стоимость акций может увеличиться в u раз или увеличиться в d раз (уменьшиться в $1/d$ раз), причем $0 < d < 1 + i < u$ (рис. 3.3.1).

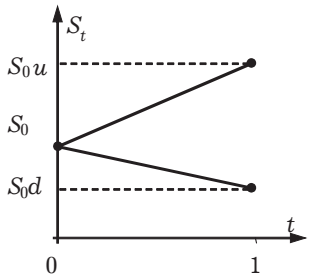


Рис. 3.3.1. *Изменение цены акции за один период времени*

На идеальном рынке о т с у т с т в у ю т а р б и т р а ж н ы е в о з м о ж н о с т и, т. е. невозможно извлечь безрисковый доход, больший чем доход по банковскому депозиту.

Предположение о том, что на рынке отсутствуют арбитражные возможности, означает, что математическое ожидание цены акции на таком рынке к концу года $\mathbf{E}(S_1)$ должно совпадать с суммой, которая оказалась бы на банковском счете к концу года, если бы сумма S_0 была в начале года положена на счет, т. е. с суммой $S_0(1 + i)$: $\mathbf{E}(S_1) = S_0(1 + i)$. Истинные вероятности того, что в течение данного периода акция подорожает и подешевеет, нам неизвестны, но в предположении отсутствия арбитражных возможностей можно с помощью только что полученного условия вычислить так называемые *вероятности, нейтральные к риску*: пусть p — вероятность того, что в начале следующего периода цена акции окажется равной $S_0 u$, тогда вероятность того, что цена акции будет равна $S_0 d$, составит $(1 - p)$; при этом $\mathbf{E}(S_1) = S_0 up + S_0 d(1 - p)$. Отсюда

$$S_0 up + S_0 d(1 - p) = S_0(1 + i).$$

Разделим обе части этого равенства на S_0:

$$up + d(1 - p) = 1 + i \quad \text{или} \quad (u - d)p + d = 1 + i,$$

поэтому

$$p = \frac{1 + i - d}{u - d}. \tag{3.3.4}$$

Теперь рассмотрим промежуток времени от 0 до T.

Предположим, что рассматриваемый промежуток $[0, T]$ разбит на n периодов, в каждом из которых стоимость акции может увеличиться в u или в d раз. Вероятность того, что стоимость акции увеличится в u или в d раз, неизвестна, однако можно вновь воспользоваться принципом вероятности, нейтральной к риску. Основное отличие состоит в том, что корректируются процентные ставки, т. е. проценты, выплачиваемые за год, естественно, больше, чем проценты, выплачиваемые за часть года, и в числителе формулы (3.3.4) вместо $(1 + i)$ будет $(1 + i)^{T/n}$, т. е.

$$p_{(T/n)} = \frac{(1 + i)^{T/n} - d}{u - d}. \tag{3.3.5}$$

Процесс изменения цены акции в течение n периодов (для случая $n = 4$) проиллюстрирован рис. 3.3.2.

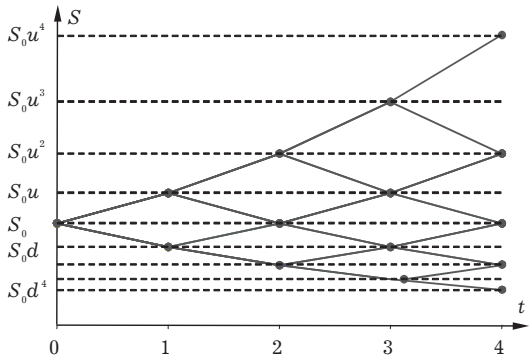


Рис. 3.3.2. *Изменение цены акции в течение четырех периодов*

При этом процесс изменения цены акции в течение n периодов можно представить как последовательность n независимых испытаний, в которых успехом считается повышение цены акции в u раз, а неудачей — ее понижение в $1/d$ раз. Если в течение n периодов цена акции поднималась k раз и опускалась $(n-k)$ раз, то ее цена к концу последнего периода составит $S_T = S_0 u^k d^{n-k}$. Вероятность наступления k повышений и $(n-k)$ понижений цены акции составит по формуле Бернулли (1.6.1) $\mathbf{P}_n(k) = C_n^k p^k (1-p)^{n-k}$. Вероятность успеха p здесь имеет смысл оценить с помощью нейтральной к риску вероятности $p_{(T/n)}$, определяемой формулой (3.3.5). Таким образом, цена акции к концу n-го периода (т. е. в момент времени T) может принимать значения $S_T = S_0 u^k d^{n-k}$ с вероятностями

$$\mathbf{P}\{S_T = S_0 u^k d^{n-k}\} = C_n^k p_{(T/n)}^k \left(1-p_{(T/n)}\right)^{n-k}, \quad k = 0, 1, 2, ..., n. \tag{3.3.6}$$

Формула (3.3.6) действительно определяет дискретную случайную величину, поскольку все вероятности неотрицательны и в сумме дают единицу, точно так же, как это было показано в предыдущем пункте для случайной величины, распределенной по биномиальному закону.

Данная модель (3.3.6), называемая *биномиальной моделью ценообразования акции*, была предложена в 1979 г. Дж.К. Коксом, С.А. Россом и М.Э. Рубинштейном.

Пример 3.3.3. Составим ряд распределения цены акции к концу года, разбив этот год на четыре периода, если текущая цена акции составляет $S_0 = 35$ ден. ед., годовая безрисковая процентная ставка составляет $i = 10\% = 0{,}1$ и известно, что в каждом периоде акция может возрасти в цене или упасть в цене в $u = 1{,}105$ раз.

Вероятность, нейтральная к риску, $p_{(1/4)} = \dfrac{(1+i)^{1/4} - d}{u - d} = \dfrac{1{,}024 - 0{,}905}{1{,}105 - 0{,}905} \approx$ $\approx 0{,}595$, $1 - p_{(1/4)} \approx 0{,}405$ (здесь $d = 1/u = 1/1{,}105 \approx 0{,}905$). Теперь мы можем составить ряд распределения цены акции к концу четвертого периода: цена принимает значения $S_0 u^k d^{4-k}$ ($k = 0, 1, 2, 3, 4$) соответственно с вероятностями $\mathbf{P}_4(k) = C_4^k p_{(1/4)}^k \left(1-p_{(1/4)}\right)^{4-k}$.

Окончательно имеем:

S_1	23,478	28,667	35,002	42,737	52,182
p	0,027	0,158	0,348	0,341	0,126

❑

Опцион — это ценная бумага, представляющая собой договор, по которому одна из сторон (*эмитент опциона*) продает опцион за определенную *премию*, а другая сторона (*держатель опциона*) при этом получает право (но не обязанность) в течение срока, оговоренного в условиях опциона, либо купить определенный актив по фиксированной цене, определяемой в момент заключения договора и называемой *терминальной стоимостью опциона* (такой опцион называется *опционом покупателя*), либо продать актив по терминальной стоимости (такой опцион называется *опционом продавца*). По срокам исполнения опционы делятся на европейские и американские. *Американский опцион* может быть предъявлен к исполнению в любое время до истечения срока опциона, *европейский опцион* может быть использован только в день истечения его срока.

Рассмотрим ценообразование опционов в рамках **биномиальной модели Кокса — Росса — Рубинштейна.**

Рациональной считается такая стоимость финансового инструмента, которая исключает возможность арбитража без риска, иными словами, доходность безрискового финансового инструмента, имеющего рациональную стоимость, должна совпадать с доходностью банковского счета.

Очевидно, рациональная стоимость опциона в момент его исполнения совпадает с прибылью, которую можно получить, исполнив опцион. Эта прибыль называется *платежной функцией опциона* и обозначается f_T. Пусть в момент исполнения T цена акции, на которую выписан европейский опцион покупателя, составляет S_T, а цена исполнения этого опциона равна X. Тогда если S_T окажется не больше X, исполнять опцион бессмысленно, т. е. его платежная функция равна нулю. Если же S_T будет больше X, то выигрыш от исполнения такого опциона составит $(S_T - X)$. Объединяя эти два случая, получаем формулу для платежной функции европейского опциона покупателя:

$$f_T = \max\{S_T - X; 0\}.$$

Формула Кокса — Росса — Рубинштейна. *Рациональная стоимость* $\hat{\mathbb{C}}_T$ *стандартного европейского опциона покупателя на биномиальном* (B, S)*-рынке определяется формулой*

$$\hat{\mathbb{C}}_T = \frac{\sum\limits_{k=0}^{n} \max\{S_0 u^k d^{n-k} - X;\ 0\} C_n^k p_{(T/n)}^k \left(1 - p_{(T/n)}\right)^{n-k}}{(1+i)^T}, \tag{3.3.7}$$

где S_0 *— стоимость акции в начальный момент,* T *— срок действия опциона (в годах),* X *— цена исполнения опциона,* i *— годовая безрисковая процентная ставка,* n *— количество периодов, на которые делится срок действия опциона (в каждый из периодов цена акции, на которую выписан опцион, может повыситься в* u *или в* d *раз),* $p_{(T/n)}$ *— нейтральная к риску вероятность, определяемая формулой* (2.3.5).

Доказательство. Опцион покупателя имеет смысл *исполнять*, т. е. пользоваться заложенным в нем правом покупки акции по цене X, лишь в том случае, когда рыночная цена S_T этой акции к моменту окончания срока действия опциона, т. е. к концу последнего периода, будет больше X. Если рыночная цена акции S_T окажется больше X, держатель опциона, исполнив его, получит доход $(S_T - X)$. Если же рыночная цена акции S_T окажется меньше X, держатель опциона просто не будет его исполнять и получит нулевой доход. Таким образом, если цена акции в момент исполнения опциона известна и равна S_T, то доход от исполнения такого опциона составит $C_T = \max\{S_T - X; 0\}$. Поскольку цена акции S_T является случайной величиной, определяемой рядом распределения (3.3.6), доход от исполнения опциона покупателя C_T также является случайной величиной, которая принимает значения $c_k = \max\{S_0 u^k d^{n-k} - X; 0\}$ $(k = 0, 1, 2, \ldots, n)$ с вероятностями $\mathbf{P}_n(k) = C_n^k p_{(n)}^k \left(1 - p_{(n)}\right)^{n-k}$.

Оценка опциона происходит перед началом первого периода, поэтому для получения его рациональной стоимости $\hat{\mathbb{C}}_T$ достаточно дисконтировать ожидаемый доход от исполнения опциона на срок его действия:

$$\hat{\mathbb{C}}_T = \frac{\mathbf{E}(C_T)}{(1+i)^T} = \frac{\sum\limits_{k=0}^{n} \max\{S_0 u^k d^{n-k} - X;\ 0\} C_n^k\, p_{(T/n)}^k \left(1 - p_{(T/n)}\right)^{n-k}}{(1+i)^T}. \ \square$$

На реальном рынке продавцу опциона выгодны цены, не меньшие рациональной стоимости опциона, а покупателю — не большие.

Пример 3.3.4. Вычислим рациональную стоимость опциона покупателя с терминальной стоимостью $X = 40$ ден. ед. и сроком исполнения 1 год, выписанного на акцию из примера 3.3.3.

Ряд распределения дохода от исполнения опциона при расчетах по четырехпериодной биномиальной модели имеет следующий вид:

C_1	0	2,737	12,182
p	$0{,}027 + 0{,}158 + 0{,}348 = 0{,}533$	0,341	0,126

Ожидаемый доход от исполнения опциона покупателя, равный математическому ожиданию случайной величины C_1, составляет

$$\mathbf{E}(C_1)=\sum_{k=0}^{4}\max\{S_0u^kd^{n-k}-X;0\}C_4^kp_{(1/4)}^k\left(1-p_{(1/4)}\right)^{4-k}=$$

$$=0\cdot0{,}027+0\cdot0{,}158+2{,}737\cdot0{,}341+12{,}182\cdot0{,}126=2{,}468,$$

и окончательно получаем: $\hat{\mathbb{C}}_1=\dfrac{2{,}468}{1+0{,}1}\approx2{,}24$. ❑

Пусть одновременно заключаются два опционных контракта (опцион покупателя и опцион продавца) с одной и той же ценой исполнения X и одним и тем же сроком исполнения T на одну и ту же акцию, стоимость которой в начальный момент равна S_0, банковская (годовая) процентная ставка равна i и пусть $\hat{\mathbb{C}}_T$ и $\hat{\mathbb{P}}_T$ — рациональные стоимости этих опционов.

Рациональность стоимости означает, что из инструментов с такой стоимостью невозможно создать безрисковый портфель, доходность которого превысит безрисковую доходность. Если

$$\hat{\mathbb{P}}_T-\hat{\mathbb{C}}_T+S_0>\frac{X}{(1+i)^T},$$

то для получения безрисковой прибыли можно использовать следующую торговую стратегию: одновременно продать акцию, продать опцион продавца на нее и купить опцион покупателя. Тогда в начальный момент времени будет получена сумма $\hat{\mathbb{P}}_T-\hat{\mathbb{C}}_T+S_0$, а в момент исполнения опционов необходимо будет выплатить сумму X, современная ценность которой составляет $X/(1+i)^T$. Таким образом, будет получен выигрыш по сравнению с безрисковыми вложениями в банковский счет, что противоречит предположению о рациональности стоимости опционов.

Аналогично, если $\hat{\mathbb{P}}_T-\hat{\mathbb{C}}_T+S_0$ будет меньше, чем $X/(1+i)^T$, то для получения безрисковой прибыли можно использовать следующую торговую стратегию: купить акцию, купить опцион продавца на нее и продать опцион покупателя. Тогда в начальный момент времени будет выплачена сумма $\hat{\mathbb{P}}_T-\hat{\mathbb{C}}_T+S_0$, а в момент исполнения опционов — получена сумма X, современная ценность которой составляет $X/(1+i)^T$, т. е. данная стратегия принесет выигрыш по сравнению с безрисковыми вложениями в банковский счет, что противоречит предположению о рациональной стоимости опционов. Полученные противоречия показывают, что для рациональной стоимости $\hat{\mathbb{C}}_T$ опциона покупателя и рациональной стоимости $\hat{\mathbb{P}}_T$ опциона продавца справедлива

Теорема о паритете опционов покупателя и продавца:

$$\hat{\mathbb{P}}_T-\hat{\mathbb{C}}_T+S_0=\frac{X}{(1+i)^T}.$$

Отметим, что данное соотношение паритета справедливо только для европейских опционов.

Теперь легко найти рациональную стоимость $\hat{\mathbb{P}}_T$ европейского опциона продавца:

$$\hat{\mathbb{P}}_T=\frac{X+\sum_{k=0}^{n}\max\{S_0u^kd^{n-k}-X;\,0\}C_n^kp_{(n)}^k(1-p_{(n)})^{n-k}}{(1+i)^T}-S_0.$$

Пример 3.3.5. Найдем рациональную стоимость опциона продавца с терминальной стоимостью $X=40$ руб. и сроком исполнения 1 год, выписанного на акцию из условия примеров 3.3.3, 3.3.4.

Исполнять опцион продавца выгодно, если цена акции на рынке окажется не больше терминальной стоимости X, поэтому ряд распределения дохода от исполнения опциона при расчетах по четырехпериодной биномиальной модели

P_1	16,252	11,333	4,998	0
p	0,027	0,158	0,348	$0{,}341+0{,}126=0{,}467$

легко получить, зная ряд распределения цены акции к концу четвертого периода (полученный в примере 3.3.3).

Ожидаемый доход от исполнения опциона покупателя, равный математическому ожиданию случайной величины P_1, составляет

$$\mathbf{E}(P_1) = \sum_{k=0}^{4} \max\{X - S_0 u^k d^{4-k}; 0\} \mathrm{C}_4^k \, p_{(4)}^k (1 - p_{(4)})^{4-k} =$$

$$= 16{,}252 \cdot 0{,}027 + 11{,}333 \cdot 0{,}158 + 4{,}998 \cdot 0{,}348 + 0 \cdot 0{,}341 + 0 \cdot 0{,}126 = 3{,}969.$$

Окончательно получаем

$$\hat{\mathbb{P}}_1 = \frac{\mathbf{E}(P_1)}{1+i} = \frac{3{,}969}{1+0{,}1} \approx 3{,}61 \text{ руб.},$$

т. е рациональная стоимость такого опциона равна 3 руб. 61 коп.

Теорема о паритете дает, естественно, тот же результат:

$$\hat{\mathbb{P}}_T = \hat{\mathbb{C}}_T - S_0 + \frac{X}{(1+i)^T} = 2{,}24 - 35 + \frac{40}{1+0{,}1} \approx 3{,}61 \text{ руб.} \ \square$$

Задачи

184. Составьте ряд распределения цены акции к концу года, разбив этот год на пять периодов, если текущая цена акции составляет $S_0 = 100$ ден. ед., годовая безрисковая процентная ставка составляет $i = 16\% = 0{,}16$ и известно, что в каждом периоде акция может возрасти в цене или упасть в цене в $u = 1{,}2$ раза.

185. Вычислите рациональную стоимость опциона покупателя с терминальной стоимостью $X = 60$ ден. ед. и сроком исполнения 1 год, выписанного на акцию из задачи **184**.

186. Вычислить рациональную стоимость опциона продавца с терминальной стоимостью $X = 60$ ден. ед. и сроком исполнения 1 год, выписанного на акцию из задачи **184** двумя способами: *а*) непосредственно по ряду распределения; *б*) с помощью теоремы о паритете.

3.3.3. Геометрический закон распределения

Случайная величина X называется распределенной *по геометрическому закону* с параметром p (где $0 < p < 1$), если она принимает значения 1, 2, 3, ... с вероятностями

$$\mathbf{P}\{X = x\} = p(1-p)^{x-1} \qquad (x = 1, 2, 3, \ldots). \tag{3.3.8}$$

Случайную величину X, распределенную по геометрическому закону (3.3.8), можно интерпретировать как номер испытания Бернулли, на котором впервые наступил успех, если вероятность успеха в единичном испытании равна p.

Формула (3.3.8) действительно определяет дискретную случайную величину, поскольку все $\mathbf{P}\{X = x\} \geqslant 0$ и

$$\sum_{x=1}^{n} \mathbf{P}\{X = x\} = \sum_{x=1}^{\infty} p(1-p)^{x-1} = \frac{p}{1-(1-p)} = 1.$$

Тот факт, что случайная величина X распределена по геометрическому закону с параметрами n и p, кратко обозначается так: $X = \mathrm{Geom}(p)$.

Математическое ожидание случайной величины $X = \mathrm{Geom}(p)$ равно

$$\mathbf{E}(X) = \frac{1}{p}, \tag{3.3.9}$$

а ее дисперсия

$$\mathbf{Var}(X)=\frac{1-p}{p^2}. \tag{3.3.10}$$

Доказательство. Имеем: $\mathbf{E}(X)=\sum_{i=1}^{\infty}x_ip_i=\sum_{x=1}^{\infty}xp(1-p)^{x-1}$. Вынесем в последнем ряде p за скобку и обозначим $q=1-p$, получим:

$$\mathbf{E}(X)=p\sum_{x=1}^{\infty}xq^{x-1}=p\sum_{x=1}^{\infty}\frac{dq^x}{dq}=p\frac{d\left(\sum_{x=1}^{\infty}q^x\right)}{dq}=p\frac{d\left(\frac{1}{1-q}\right)}{dq}=p\frac{1}{(1-q)^2}=p\frac{1}{[1-(1-p)]^2}=\frac{p}{p^2}=\frac{1}{p},$$

что доказывает формулу (3.3.9). Дисперсию будем вычислять по формуле (2.2.18):

$$\mathbf{Var}(X)=\sum_i x_i(x_i-1)p_i+\mathbf{E}(X)-[\mathbf{E}(X)]^2=\sum_{x=1}^{\infty}x(x-1)p(1-p)^{x-1}+\frac{1}{p}-\left(\frac{1}{p}\right)^2=$$

$$=\sum_{x=2}^{\infty}x(x-1)p(1-p)^{x-1}-\frac{1-p}{p^2}=\sum_{x=2}^{\infty}x(x-1)pq^{x-1}-\frac{q}{p^2}=pq\sum_{x=0}^{\infty}x(x-1)q^{x-2}-\frac{q}{p^2}=$$

$$=pq\sum_{x=0}^{\infty}x(x-1)q^{x-2}-\frac{q}{p^2}=pq\sum_{x=0}^{\infty}\frac{d^2[q^x]}{dq^2}-\frac{q}{p^2}=pq\frac{d^2\left(\sum_{x=0}^{\infty}q^x\right)}{dq^2}-\frac{q}{p^2}=pq\frac{d^2\left(\frac{1}{1-q}\right)}{dq^2}-\frac{q}{p^2}=$$

$$=pq\frac{2}{(1-q)^3}-\frac{q}{p^2}=\frac{2p(1-p)}{p^3}-\frac{1-p}{p^2}=\frac{2(1-p)}{p^2}-\frac{1-p}{p^2}=\frac{1-p}{p^2},$$

что доказывает формулу (3.3.10). ❑

В пакете Microsoft Excel для вычисления вероятностей $\mathbf{P}\{X=x\}$ можно пользоваться функцией

$$\mathbf{P}\{X=x\}=\text{ОТРБИНОМ.РАСП(<}x-1\text{>; 1; <}p\text{>)}.$$

Случайная величина X, распределенная по геометрическому закону (3.3.8), обладает **свойством отсутствия последействия**, т. е. при всех $x=1, 2, \ldots$, $y=1, 2, \ldots$

$$\mathbf{P}\{X=x+y\,|\,X>x\}=\mathbf{P}\{X=y\}. \tag{3.3.11}$$

Доказательство. Действительно,

$$\mathbf{P}\{X=x+y\,|\,X>x\}=\frac{\mathbf{P}(\{X=x+y\}\cap\{X>x\})}{\mathbf{P}\{X>x\}}=\frac{\mathbf{P}\{X=x+y\}}{\mathbf{P}\{X>x\}}=\frac{p(1-p)^{x+y-1}}{\sum_{n=x+1}^{\infty}p(1-p)^{n-1}}=$$

$$=\frac{\frac{p}{1-p}(1-p)^{x+y}}{\frac{p}{1-p}\sum_{n=x+1}^{\infty}(1-p)^n}=\frac{(1-p)^{x+y}}{\sum_{n=x+1}^{\infty}(1-p)^n}=(1-p)^{x+y}[1-(1-p)]=p(1-p)^y=\mathbf{P}\{X=y\}.\ ❑$$

Замечание. Интересен и обратный результат: если некоторая дискретная случайная величина X принимает значения 1, 2, 3, ... и обладает свойством отсутствия последействия (3.3.11), то эта случайная величина распределена по геометрическому закону (3.3.8).

Доказательство. Поскольку свойство (3.3.11) выполнено, то

$$\mathbf{P}\{X=x+y\,|\,X>x\}=\frac{\mathbf{P}(\{X=x+y\}\cap\{X>x\})}{\mathbf{P}\{X>x\}}=\frac{\mathbf{P}\{X=x+y\}}{\mathbf{P}\{X>x\}}=\mathbf{P}\{X=y\}$$

для всех $y=1, 2, 3, \ldots$

Положим $y = 1$, тогда $\dfrac{\mathbf{P}\{X = x+1\}}{\mathbf{P}\{X > 1\}} = \mathbf{P}\{X = 1\}$ или $\mathbf{P}\{X = x + 1\} = \mathbf{P}\{X = 1\}\mathbf{P}\{X > 1\}$, т. е. $\mathbf{P}\{X = x + 1\} = \mathbf{P}\{X = 1\}(1 - \mathbf{P}\{X = 1\})$ для всех $x = 1, 2, 3, \ldots$ Поэтому $\mathbf{P}\{X = x + 1\} = p(1 - p)^{x-1}$, где $p = \mathbf{P}\{X = 1\}$, т. е. X распределена по геометрическому закону (3.3.8) с параметром $p = \mathbf{P}\{X = 1\}$, что и требовалось доказать. ❑

Свойство отсутствия последействия (3.3.11) можно интерпретировать следующим образом. Пусть длительность телефонных разговоров измеряется целым числом минут, и в конце каждой минуты абонент принимает решение закончить разговор [с вероятностью p] или продолжить разговор [с вероятностью $(1 - p)$]. Тогда длительность телефонного разговора будет случайной величиной, распределенной по геометрическому закону (3.3.8) с параметром p. Свойство отсутствия последействия (3.3.11) означает, что условная вероятность того, что разговор будет продолжаться ровно $x + y$ мин, если известно, что он не закончился за x мин, равна безусловной вероятности того, что разговор будет продолжаться ровно y мин.

Пример 3.3.6. В среднем левши составляют 1% всего населения. Оценим количество людей, которых придется опросить, чтобы набрать десятерых левшей.

При интерпретации опроса как последовательности независимых испытаний с вероятностью успеха $p = 1\% = 0{,}01$ число X опрошенных до появления левши в первый раз (так же, как и число опрошенных после появления левши в i-й раз до появления левши в $(i + 1)$-й раз) — это случайная величина $X = \mathrm{Geom}(p = 0{,}01)$, и ее среднее значение оценивается математическим ожиданием $\mathbf{E}(X) = 1/p = 1/0{,}01 = 100$. Для того, чтобы отобрать десять левшей, учитывая свойство аддитивности математических ожиданий (3.2.9), в среднем нужно опросить в 10 раз больше людей, т. е. 1000 людей. ❑

Пример 3.3.7 (Санкт-Петербургский парадокс). Петя и Маша играют в следующую игру. Петя выплачивает Маше детерминированную сумму c, а Маша бросает монету до первого выпадения «орла» и выплачивает Пете 2^k руб., если «орел» выпал впервые при k-м бросании. При какой сумме c, Маше имеет смысл согласиться на такую игру?

Пусть X — выигрыш Пети, тогда $\mathbf{P}\{X = 2^k\} = 1/2^k$ и $\mathbf{E}(X) = \sum\limits_{k=1}^{\infty} 2^k \cdot \dfrac{1}{2^k} = +\infty$, т. е. кажется, что Маша не должна соглашаться на такую игру ни в коем случае, ни при каком c. Парадокс же состоит в том, что подобный вывод не реалистичен: например, при $c = 1\,000\,000\,000$ руб. Маше было бы трудно отказаться от игры.

Д. Бернулли предположил, что «удовлетворение, испытываемое игроком от дохода x» (или, в современной терминологии, полезность дохода x) возрастает не пропорционально доходу, но эту полезность можно измерить некоторой функцией $u(x)$ — функцией полезности дохода.

Так, обладатель капитала в 1 000 000 000 руб. вряд ли испытывает такое же удовлетворение от дополнительного дохода в 100 руб., как и обладатель капитала в 1 руб. Если предположить, что приращение полезности пропорционально не абсолютному, а относительному изменению дохода, т. е.

$$du(x) = \alpha \frac{dx}{x},$$

то получим дифференциальное уравнение

$$\frac{du(x)}{dx} = \frac{\alpha}{x},$$

решением которого является *логарифмическая функция полезности*

$$u(x) = \alpha \ln x + c, \tag{3.3.12}$$

где $c = \mathrm{const}$.

Если же исходить не из дохода, а из его полезности, и в качестве функции полезности взять логарифмическую полезность $u(x) = \log_2 x$, то случайная величина $u(X)$ будет принимать значения $\log_2(2^k) = k$ с вероятностями $\mathbf{P}\{X = 2^k\} = 1/2^k$, т. е. будет распределена по геометрическому закону с параметром $p = 1/2$, и ее математическое ожидание $\mathbf{E}(u(X))$ по формуле (3.3.9) будет равно $\mathbf{E}(u(X)) = \dfrac{1}{1/2} = 2$.

Таким образом, полезность Петиного выигрыша равна двум (единицам полезности — ю т и л я м), значит, детерминированный эквивалент c Петиного выигрыша определяется из уравнения $u(c) = 2$ или $\log_2 c = 2$, откуда $c = 4$ руб. Именно за такую (или за большую) сумму Маше имеет смысл соглашаться на игру. ❑

Задачи

187. Из опыта работы банка известно, что 80% заемщиков возвращают кредиты в срок, а 20% — нет. Скольким людям нужно в среднем выдать кредиты, чтобы впервые попасть в ситуацию невозврата кредита?

188. Вероятность того, что кандидат, предлагаемый агентством по подбору персонала, окажется подходящим для компании, равна 0,1. Запишите первые пять столбцов ряда распределения числа кандидатов, которые будут отвергнуты до того, как появится подходящий кандидат, найдите математическое ожидание, дисперсию и среднее квадратическое отклонение числа кандидатов, которые будут отвергнуты до того, как появится подходящий кандидат.

189. Петя захотел найти человека, день рождения которого совпадает с Петиным. Составьте ряд распределения числа N незнакомцев, которых придется опросить Пете, и найдите среднее число опрошенных незнакомцев.

190. Бизнесмен Александр Андреевич очень любит ходить в казино, и если он туда зашел, то не выходит, пока на рулетке не выпадет «зеро» (то есть число «нуль»). Каждый раз Александр Андреевич ставит пять рублей на «зеро» и по одному рублю на «двадцать девять» и на «тридцать два». После этого крупье вращает колесо рулетки, и шарик указывает на одно из чисел от 0 до 36. Если шарик указывает на число, соответствующее некоторой ставке Александра Андреевича, последний получает выигрыш, в 35 раз больший, чем эта ставка, а те ставки Александра Андреевича, которые не соответствуют выпавшему числу, теряются. Определите, сколько раз играет в среднем Александр Андреевич и каков его средний выигрыш.

3.3.4. Закон распределения Пуассона

Случайная величина X называется распределенной *по закону Пуассона* с параметром λ (где $\lambda > 0$), если она принимает значения 0, 1, 2, ... с вероятностями

$$\mathbf{P}\{X = x\} = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \qquad (x = 0, 1, 2, \ldots). \tag{3.3.13}$$

Поскольку вероятности $\mathbf{P}\{X = x\}$ рассчитываются по формуле Пуассона (2.4.3), случайную величину X, распределенную по закону (3.3.15), можно (согласно теореме Пуассона (2.4.3) — см. **п. 2.4.2**) интерпретировать как число успехов в последовательности из n испытаний Бернулли, в каждом из которых успех может произойти с вероятностью p в том случае, когда $n \to \infty$, $p \to 0$, так что $np \to \lambda = \text{const}$, и на практике приближать случайные величины, распределенные по биномиальному закону, случайными величинами, распределенными по закону Пуассона, если n велико, а $\lambda = np < 10$.

Формула (3.3.15) действительно определяет дискретную случайную величину, поскольку все $\mathbf{P}\{X = x\} \geqslant 0$ и

$$\sum_{x=0}^{\infty}\mathbf{P}\{X=x\}=\sum_{x=0}^{\infty}\left(\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\right)=\mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=\mathrm{e}^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\cdots\right)=\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda}=1.$$

Тот факт, что случайная величина X распределена по закону Пуассона с параметром λ, кратко обозначается так: $X = \mathrm{Pois}(\lambda)$.

Математическое ожидание и дисперсия случайной величины $X = \mathrm{Pois}(\lambda)$ совпадают и равны

$$\mathbf{E}(X) = \mathbf{Var}(X) = \lambda.$$

Доказательство. Математическое ожидание случайной величины X вычислим по формуле (2.2.6):

$$\mathbf{E}(X)=\sum_{i=0}^{\infty}x_i\mathbf{P}\{X=x_i\}=\sum_{x=0}^{\infty}\left(x\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\right)=\mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}\left(x\frac{\lambda^x}{x!}\right)=\mathrm{e}^{-\lambda}\sum_{x=1}^{\infty}\left(x\frac{\lambda^x}{x!}\right)=\mathrm{e}^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!}=$$

$$=\lambda\mathrm{e}^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}=\lambda\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=\lambda\mathrm{e}^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\cdots\right)=\lambda\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda}=\lambda.$$

Теперь вычислим дисперсию случайной величины X по формуле (2.2.17):

$$\mathbf{Var}(X)=\sum_{i=0}^{\infty}x_i^2p_i-[\mathbf{E}(X)]^2=\sum_{x=0}^{\infty}\left(x^2\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\right)-\lambda^2=\sum_{x=1}^{\infty}\left(x^2\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\right)-\lambda^2=\mathrm{e}^{-\lambda}\sum_{x=1}^{\infty}\left(x^2\frac{\lambda^x}{x!}\right)-\lambda^2=$$

$$=\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\left(k\frac{\lambda^k}{(k-1)!}\right)-\lambda^2=\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\left(\frac{[(k-1)+1]\lambda^k}{(k-1)!}\right)-\lambda^2=\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\left(\frac{(k-1)\lambda^k}{(k-1)!}\right)+\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\left(\frac{\lambda^k}{(k-1)!}\right)-\lambda^2=$$

$$=\lambda^2\mathrm{e}^{-\lambda}\sum_{k=2}^{\infty}\left(\frac{\lambda^{k-2}}{(k-2)!}\right)+\lambda\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\left(\frac{\lambda^{k-1}}{(k-1)!}\right)-\lambda^2=\lambda^2\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda}+\lambda\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda}-\lambda^2=\lambda^2+\lambda-\lambda^2=\lambda,$$

что и требовалось доказать. ❑

В пакете Microsoft Excel для вычисления вероятностей $\mathbf{P}\{X = x\}$ и $\mathbf{P}\{X \leqslant x\}$ можно пользоваться функциями

$$\mathbf{P}\{X = x\} = \text{ПУАССОН.РАСП(<}x\text{>; <}\lambda\text{>; ЛОЖЬ)},$$

$$\mathbf{P}\{X \leqslant x\} = \text{ПУАССОН.РАСП(<}x\text{>; <}\lambda\text{>; ИСТИНА)}.$$

Пример 3.3.8. Пусть случайная величина $X = \mathrm{Pois}(\lambda = 3)$. Найдем

$$\mathbf{P}\{X=2\}=\frac{3^2}{2!}\mathrm{e}^{-3}\approx\frac{9}{2}\cdot 0{,}05=0{,}225.$$

Теперь вычислим

$$\mathbf{P}\{X>1\}=1-\mathbf{P}\{X=0\}-\mathbf{P}\{X=1\}=1-\frac{3^0}{0!}\mathrm{e}^{-3}-\frac{3^1}{1!}\mathrm{e}^{-3}\approx 1-\frac{1}{1}\cdot 0{,}05-\frac{3}{1}\cdot 0{,}05=0{,}80.$$

Рассчитаем вероятность

$$\mathbf{P}\{0<X<3\}=\mathbf{P}\{X=1\}+\mathbf{P}\{X=2\}=\frac{3^1}{1!}\mathrm{e}^{-3}-\frac{3^2}{2!}\mathrm{e}^{-3}\approx\frac{3}{1}\cdot 0{,}05+\frac{9}{2}\cdot 0{,}05=0{,}373.$$

Найдем условную вероятность

$$\mathbf{P}\{X=1\,|\,X>0\}=$$

$$=\frac{\mathbf{P}\{(X=1)\cap(X>0)\}}{\mathbf{P}\{X>0\}}=\frac{\mathbf{P}\{X=1\}}{1-\mathbf{P}\{X=0\}}=\frac{\frac{3^1}{1!}\mathrm{e}^{-3}}{1-\frac{3^0}{0!}\mathrm{e}^{-3}}\approx\frac{\frac{3}{1}\cdot 0{,}05}{1-\frac{1}{1}\cdot 0{,}05}=\frac{0{,}15}{0{,}95}\approx 0{,}158.$$

Математическое ожидание и дисперсия случайной величины X совпадают и равны $\mathbf{E}(X) = \mathbf{Var}(X) = \lambda = 3$.

Расчеты с помощью Microsoft Excel приведены на рис. 3.3.3. Естественно, значения вероятностей, рассчитанные в Microsoft Excel, точнее, чем рассчитанные вручную.

	A	B
1	P{X = 2} =	=ПУАССОН.РАСП(2; 3; ЛОЖЬ)
2	P{X > 1} =	=1 – ПУАССОН.РАСП (1; 3; ИСТИНА)
3	P{0 < X < 3} =	=ПУАССОН.РАСП (1; 3; ЛОЖЬ) + ПУАССОН.РАСП (2; 3; ЛОЖЬ)
4	P{X = 1 | X > 0} =	=ПУАССОН.РАСП (1; 3; ЛОЖЬ)/(1 – ПУАССОН.РАСП (0; 3; ЛОЖЬ))

а) формулы

	A	B
1	P{X = 2} =	0,224042
2	P{X > 1} =	0,800852
3	P{0 < X < 3} =	0,373403
4	P{X = 1 | X > 0} =	0,157187

б) результаты расчетов

Рис. 3.3.3. *Расчеты в примере 3.3.7*

Для построения с помощью Microsoft Excel графика функции распределения достаточно воспользоваться функцией ПУАССОН.РАСП:

$$F(x) = \mathbf{P}\{X \leqslant x\} = \text{ПУАССОН.РАСП(<}x\text{>; <}\lambda\text{>; ИСТИНА)}.$$

График функции распределения представлен на рис. 3.3.4. ❑

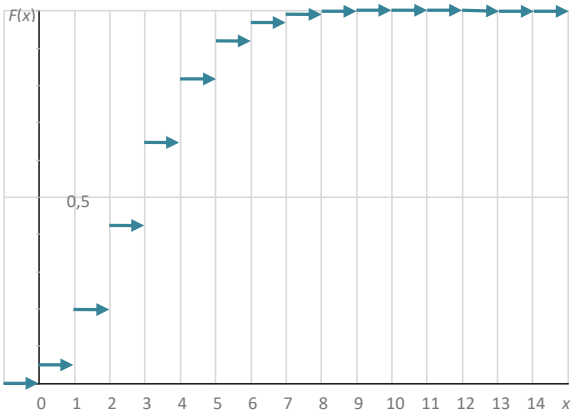


Рис. 3.3.4. *График функции распределения в примере 3.3.8*

Задачи

191. Пивоваренный завод отправил в магазин 400 ящиков пива. Вероятность того, что ящик будет разбит при транспортировке в данных условиях, равна 0,005. По приезде в магазин экспедитор, перевозивший груз, заявил, что семь ящиков с пивом были разбиты при транспортировке. Размышляя, можно ли доверять экспедитору, директор магазина хочет найти вероятность разбить семь ящиков, вероятность разбить не менее семи ящиков, математическое ожидание, дисперсию и среднее квадратическое отклонение количества ящиков, разбитых при транспортировке, чтобы оценить возможность потерь, заявленных экспедитором. Вычислите указанные величины.

192. Платежная система обработала в минуту 4000 банковских операций. Вероятность того, что операция является мошеннической, равна 0,0001. Найдите: *а*) вероятность того, что среди 4000 операций хотя бы одна является мошеннической; *б*) вероятность того, что среди 4000 операций не более трех являются мошенническими; *в*) математическое ожидание и дисперсию числа мошеннических операций из 4000.

193. Для продвижения своей продукции на рынок фирма проводит массовую рассылку объявлений по электронной почте. Прежний опыт работы фирмы показывает, что примерно в одном случае из 2000 за прочтением рекламного письма следует заказ. Найдите: *а*) вероятность того, что после рассылки 10 000 рекламных писем поступит хотя бы один заказ; *б*) среднее число поступивших заказов после рассылки 10 000 рекламных писем; *в*) дисперсию числа поступивших заказов после рассылки 10 000 рекламных писем.

194. При некотором виде страхования количество страховых случаев на один страховой договор в течение года распределено по закону Пуассона с параметром $\lambda = 0{,}2$. Найдите вероятности следующих событий: *а*) по случайно выбранному договору не наступит ни одного страхового случая; *б*) по случайно выбранному договору наступит ровно один страховой случай; *в*) по случайно выбранному договору наступит ровно два страховых случая; *г*) по случайно выбранному договору наступит ровно три страховых случая; *д*) по случайно выбранному договору наступит более трех страховых случаев.

195. При расчете страховой премии (т. е. стоимости страхового полиса) по некоторому виду страхования аналитики страховой компании предполагали, что страховые случаи происходят с вероятностью 0,005, а средний размер страховой выплаты составит 100 000 ден. ед. В прошлом году страховая компания заключила 2000 договоров по данному виду страхования, и за год произошло 15 страховых случаев (что на 50% больше ожидаемого количества $2000 \cdot 0{,}005 = 10$) со средним размером выплаты в 101 000 ден. ед. (что вполне приемлемо). Предполагая, что число страховых случаев распределено по закону Пуассона, рассчитайте вероятность того, что за год наступит не менее 15 страховых случаев, и на основании этого сделайте вывод, является ли превышение ожидаемого количества страховых случаев на 50% «невезением» по случайным причинам или результатом ошибочного предположения о вероятности страхового случая.

3.3.5. Простейший поток событий

Потоком событий называется последовательность событий, наступающих в случайные моменты времени.

Поток событий называется *стационарным*, если его вероятностные характеристики не зависят от времени (более строго, если вероятность того, что за время t наступит ровно k событий, не зависит от начала отсчета промежутка t, а зависит только от его длины).

Поток событий называется *ординарным*, если за малый промежуток времени t наступление двух или более событий маловероятно (т. е. если вероятность наступления двух или более событий за малый промежуток времени t пренебрежимо мала по сравнению с вероятностью наступления одного события за этот промежуток).

Поток событий называется *потоком с отсутствием последействия*, если будущее наступление событий не зависит от того, как они наступали в прошлом (т. е. если вероятность наступления k событий в любом промежутке времени не зависит от того, сколько событий уже наступило к началу этого промежутка, и в какие моменты времени они наступили).

Поток событий называется *простейшим*, если он обладает свойствами стационарности, ординарности и отсутствия последействия. *Интенсивность потока* μ — это среднее число событий, наступающих в единицу времени.

Число событий простейшего потока с интенсивностью μ, наступивших за время t, представляет собой случайную величину X, распределенную по закону Пуассона с параметром $\lambda = \mu t$:

$$\mathbf{P}\{X = x\} = \frac{(\mu t)^x}{x!} e^{-\mu t} \qquad (x = 0, 1, 2, \ldots). \tag{3.3.14}$$

Доказательство. Разделим отрезок времени длиной t на n частей одинаковой длины $\Delta t = t / n$. Математическое ожидание числа событий, наступающих в каждом из таких интервалов, будет равно $\mu \Delta t = \mu t / n$ (где μ — интенсивность потока). Поскольку простейший поток обладает свойством ординарности, при больших n можно пренебречь наступлением двух и более событий на каждом из участков длиной Δt и считать, что возможны два варианта для каждого участка: либо на этом участке появилось ровно одно событие из потока, либо не появилось ни одного события. Введем события A_i ($i = 1, 2, \ldots, n$), состоящие в том, что на i-м участке событие из потока появилось. При этом $\mathbf{P}(A_i) = \mathbf{E}(I_{A_i})$. С другой стороны, $\mathbf{E}(I_{A_i}) = \mu t / n$, поэтому $\mathbf{P}(A_i) = \mu t / n$. Так как простейший поток обладает свойством отсутствия последействия, то события A_i ($i = 1, 2, \ldots, n$) являются независимыми в совокупности, при вероятность каждого из этих событий равна $p = \mathbf{P}(A_i) = \mu t / n$, значит, случайную величину X (число событий простейшего потока, наступивших за время t) можно представить как число успехов в n испытаниях Бернулли (см. **п. 2.4.1**), где успехом в i-м испытании считается наступление события A_i. При этом произведение $\lambda = np = n\dfrac{\mu t}{n} = \mu t$ не зависит от n. Устремив n к бесконечности, замечаем, что $p \to 0$, а $\lambda = \mu t = \text{const}$, поэтому для вычисления вероятности того, что число событий простейшего потока, наступивших за время t, равно x, можно воспользоваться формулой Пуассона (2.4.3):

$\mathbf{P}\{X = x\} = \lim\limits_{\substack{n\to\infty \\ np\to\lambda}} \mathbf{P}_n(x) = \dfrac{\lambda^x}{x!} e^{-\lambda} = \dfrac{(\mu t)^x}{x!} e^{-\mu t}$, что и требовалось доказать. ❑

Найдем теперь функцию распределения интервала времени T между двумя последовательными наступлениями события в простейшем потоке с интенсивностью μ.

Очевидно, событие $\{T < t\}$ при $t > 0$ эквивалентно событию $\{X > 0\}$ — оба этих события состоят в том, что на промежутке времени длиной t произошло хотя бы одно событие из простейшего потока.

Вероятность того, что за время t не произойдет ни одного события из простейшего потока, равна (по формуле 3.3.16)

$$\mathbf{P}\{X = 0\} = \frac{(\mu t)^0}{0!} e^{-\mu t} = e^{-\mu t}.$$

Вероятность противоположного события равна $\mathbf{P}\{X > 0\} = 1 - e^{-\mu t}$, поэтому $F_T(t) = \mathbf{P}\{T < t\} = \mathbf{P}\{X > 0\} = 1 - e^{-\mu t}$. Итак,

$$F_T(t) = 1 - e^{-\mu t}. \tag{3.3.155}$$

При $t \leqslant 0$, очевидно,

$$F_T(t) = \mathbf{P}\{T < t\} = 0. \tag{3.3.166}$$

Объединяя формулы (3.3.17) и (3.3.18), получим окончательно:

$$F_T(t) = \begin{cases} 0, & t \leqslant 0, \\ 1 - \mathrm{e}^{-\mu t}, & t > 0. \end{cases} \tag{3.3.177}$$

Задачи

196. В диспетчерскую интернет-сервиса такси поступает простейший поток заказов такси с интенсивностью $\mu = 1{,}2$ заказа в секунду. Найдите вероятности следующих событий: *а*) за две секунды не поступит ни одного заказа; *б*) за две секунды поступит ровно один заказ; *в*) за две секунды поступит хотя бы один заказ.

197. Магазин имеет два входа, потоки покупателей на этих входах независимы и являются простейшими. Через первый вход проходит в среднем $\mu_1 = 1{,}5$ чел./мин, а через второй вход — $\mu_2 = 0{,}5$ чел./мин. Определите вероятность того, что в наугад выбранную минуту хотя бы один человек посетит магазин.

3.3.6. Гипергеометрический закон распределения

Случайная величина X называется распределенной *по гипергеометрическому закону* с параметрами $K,\ L,\ l$, если она принимает значения $0, 1, 2, \ldots, \min\{K; l\}$ с вероятностями

$$\mathbf{P}\{X = x\} = \frac{C_K^x C_{L-K}^{l-x}}{C_L^l} \qquad (x = 0, 1, 2, \ldots, \min\{K; l\}). \tag{3.3.18}$$

Случайную величину X, распределенную по гипергеометрическому закону (3.3.18), можно интерпретировать, например, так. Имеется партия из L изделий, K из которых первого сорта, а остальные $(L - K)$ — второго сорта. Из этой партии случайным образом извлекается l изделий. Тогда случайная величина X — число изделий первого сорта среди отобранных — будет распределена по гипергеометрическому закону (3.3.18).

Тот факт, что случайная величина X распределена по гипергеометрическому закону с параметрами $n,\ p$, кратко обозначается так: $X = \mathrm{HGeom}(K,\ L,\ l)$.

Математическое ожидание случайной величины $X = \mathrm{HGeom}(K,\ L,\ l)$ равно

$$\mathbf{E}(X) = l\frac{K}{L}, \tag{3.3.19}$$

а ее дисперсия

$$\mathbf{Var}(X) = l\frac{K}{L}\left(1 - \frac{K}{L}\right)\frac{1 - \dfrac{l}{L}}{1 - \dfrac{1}{L}}. \tag{3.3.20}$$

Доказательство. Рассмотрим последовательность из l (з а в и с и м ы х) испытаний. В каждом испытании из партии (состоящей из L изделий, K из которых первого сорта, а остальные $(L - K)$ — второго сорта), извлекается по одному изделию. Пусть события A_i состоят в том, что в i-м из l таких испытаний произошел успех ($i = 0, 1, 2, \ldots, l$). Вероятность каждого из

этих событий $\mathbf{P}(A_i) = K/L$ (предлагаем читателю убедиться в этом). Рассмотрим l случайных величин $X_i = I_{A_i}$ — индикаторов событий A_i. При этом по формуле (3.2.7) $\mathbf{E}(X_i) = \mathbf{E}\left(I_{A_i}\right) = \mathbf{P}(A_i) = K/L$, а случайная величина X, очевидно, равна $X = \sum_{i=1}^{l} X_i$. По свойству аддитивности математического ожидания (3.2.10) $\mathbf{E}(X) = \sum_{i=1}^{l} \mathbf{E}(X_i) = l\frac{K}{L}$, что доказывает формулу (3.3.19). Доказательство формулы (3.3.20) мы проведем позже, в п. 3.7.2. ❑

Замечание. Как было показано в **п. 2.4.1**,

$$\lim_{\substack{L\to\infty \\ \frac{K}{L}\to p}} \mathbf{P}_{K,L}(k,l) = \lim_{\substack{L\to\infty \\ \frac{K}{L}\to p}} \frac{C_K^k C_{L-K}^{l-k}}{C_L^l} = C_l^k p^k (1-p)^{l-k},$$

поэтому при $L \to \infty, K/L \to p$ случайная величина $X = \mathrm{HGeom}(K,\ L,\ l)$ переходит в случайную величину, распределенную по биномиальному закону с параметрами l, p.

В пакете Microsoft Excel для вычисления вероятностей $\mathbf{P}\{X = x\}$ и $\mathbf{P}\{X \leqslant x\}$ можно пользоваться функциями

$$\mathbf{P}\{X = x\} = \text{ГИПЕРГЕОМ.РАСП(<}x\text{>; <}l\text{>; <}K\text{>; <}L\text{>; ЛОЖЬ)},$$

$$\mathbf{P}\{X \leqslant x\} = \text{ГИПЕРГЕОМ.РАСП(<}x\text{>; <}l\text{>; <}K\text{>; <}L\text{>; ИСТИНА)}.$$

Гипергеометрический закон распределения широко используется в управлении качеством (в задачах статистического контроля качества продукции), в маркетинге и социологии (при организации выборочных опросов) и др.

Задачи

198. На автомобиле, прибывшем в авторемонтную мастерскую, требуется заменить четыре тормозных колодки. У автомеханика есть девять таких колодок, из которых пять — новые, а оставшиеся четыре — бывшие в употреблении. Из этих колодок автомеханик наугад выбирает четыре. Случайная величина X — число новых колодок среди выбранных. Составьте ряд распределения этой случайной величины.

199. Найдите математическое ожидание случайной величины X, определенной в предыдущей задаче, двумя способами: *а*) непосредственно по ряду распределения [т. е. по формулам (3.2.6), (3.2.16)]; *б*) по формулам (3.3.19)—(3.3.20).

200. Журнал *Fortune* ежегодно составляет и публикует рейтинг *Fortune Global* 500 — список 500 мировых компаний, имеющих самую большую годовую выручку.[1] В некоторый год 40 компаний из этого списка возглавляли женщины, во главе остальных компаний стояли мужчины. С какой вероятностью среди руководителей случайно выбранных десяти компаний из списка *Fortune Global* 500 не будет ни одной женщины? С какой вероятностью среди руководителей 10 случайно выбранных из этого списка компаний будет 0, 1, 2, ..., 9, 10 женщин?

[1] Источник — *Fortune Global* 500: http://fortune.com/global500.

§ 3.4. Абсолютно непрерывные случайные величины и их важнейшие числовые характеристики

3.4.1. Способы задания абсолютно непрерывных случайных величин

Случайная величина X называется *абсолютно непрерывной*, если ее функция распределения может быть представлена в виде

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz. \tag{3.4.1}$$

При этом функция $f_X(x)$ называется *плотностью распределения вероятностей* (или, короче, *плотностью распределения*) случайной величины X. График плотности распределения случайной величины X называется *кривой распределения вероятностей* (или, короче, *кривой распределения*) случайной величины X.

Как и раньше, если известно, о какой случайной величине идет речь, то индекс, обозначающий эту случайную величину, опускается: $f(x) = f_X(x)$.

Из определения абсолютно непрерывной случайной величины следует, что у функции распределения такой случайной величины существует производная во всех точках $x \in \mathbb{R}$, в которых плотность распределения $f(x)$ непрерывна, причем в этих точках

$$F'(x) = f(x). \tag{3.4.2}$$

Плотность распределения обладает следующими **свойствами.**

Неотрицательность плотности распределения. *Во всех точках $x \in \mathbb{R}$, в которых плотность распределения $f(x)$ абсолютно непрерывной случайной величины непрерывна,*

$$f(x) \geqslant 0. \tag{3.4.3}$$

Доказательство. Функция распределения случайной величины является неубывающей [по свойству монотонности функции распределения (3.1.4)], значит, в тех точках $x \in \mathbb{R}$, в которых существует производная $F'(x)$ функции распределения, $F'(x) \geqslant 0$. Но производная функции распределения существует во всех точках, где непрерывна плотность распределения $f(x)$, причем в этих точках $F'(x) = f(x)$. Таким образом, мы показали, что во всех точках $x \in \mathbb{R}$, в которых плотность распределения $f(x)$ непрерывна, $f(x) = F'(x) \geqslant 0$, что и требовалось доказать. ❑

Нормированность плотности распределения.

$$\int_{-\infty}^{+\infty} f(z)dz = 1. \tag{3.4.4}$$

Доказательство. Данное утверждение непосредственно следует из того, что по свойству (3.1.6) $F(+\infty) = \lim\limits_{x \to +\infty} F(x) = 1$, но, согласно (3.4.1), $F(+\infty) = \int\limits_{-\infty}^{+\infty} f(z)dz$. ❑

Вероятность того, что абсолютно непрерывная случайная величина X примет конкретное числовое значение, равна нулю, т. е. для всех $x \in \mathbb{R}$

$$\mathbf{P}\{X = x\} = 0. \tag{3.4.5}$$

Доказательство. Действительно, $\mathbf{P}\{X=x\} = \lim\limits_{\Delta x\to 0}\mathbf{P}\{x-\Delta x<X\leqslant x\}=\lim\limits_{\Delta x\to 0}[F(x)-F(x-\Delta x)]=$

$=\lim\limits_{\Delta x\to 0}\left(\int\limits_{-\infty}^{x} f(z)dz-\int\limits_{-\infty}^{x-\Delta x} f(z)dz\right)=\lim\limits_{\Delta x\to 0}\int\limits_{x-\Delta x}^{x} f(z)dz=0$ [здесь мы учли, что по формуле (3.1.3) $\mathbf{P}\{x_1<X\leqslant x_2\}=F(x_2)-F(x_1)$]. ❑

Вероятность попадания абсолютно непрерывной случайной величины в числовой промежуток можно рассчитать по формуле

$$\mathbf{P}\{x_1\leqslant X\leqslant x_2\}=\mathbf{P}\{x_1\leqslant X<x_2\}=\mathbf{P}\{x_1<X\leqslant x_2\}=\mathbf{P}\{x_1<X<x_2\}= \quad (3.4.6)$$

$$=F(x_2)-F(x_1)=\int\limits_{x_1}^{x_2} f(x)dx\,,$$

справедливой для всех $x_1,\ x_2\in\mathbb{R}$, таких что $x_1<x_2$.

Доказательство. Действительно, $\mathbf{P}\{x_1<X\leqslant x_2\}=F(x_2)-F(x_1)=\int\limits_{-\infty}^{x_2} f(z)dz-\int\limits_{-\infty}^{x_1} f(z)dz=\int\limits_{x_1}^{x_2} f(z)dz\,,$

$\mathbf{P}\{x_1\leqslant X\leqslant x_2\}=\mathbf{P}\{(x_1<X\leqslant x_2)\cup(X=x_1)\}=\mathbf{P}\{x_1<X\leqslant x_2\}+\mathbf{P}\{X=x_1\}=\mathbf{P}\{x_1<X\leqslant x_2\}+0=$

$=\mathbf{P}\{x_1<X\leqslant x_2\}=F(x_2)-F(x_1)$.

Справедливость оставшихся равенств доказывается аналогично. ❑

Если некоторая функция $f(x)$ обладает свойствами неотрицательности (3.4.3) и нормированности (3.4.4), то эта функция является плотностью распределения некоторой абсолютно непрерывной случайной величины с функцией распределения, определяемой формулой (3.4.1).

Доказательство следует из того, что функция $F(x)=\int\limits_{-\infty}^{x} f(z)dz$ обладает всеми свойствами монотонности (3.1.4), ограниченности (3.1.5)—(3.1.6) и непрерывности (3.1.7), а значит, как было отмечено в **п. 3.1.2**, является функцией распределения некоторой случайной величины. ❑

Поскольку функция распределения абсолютно непрерывной случайной величины представляется в виде интеграла (3.4.1), эта функция распределения, очевидно, является непрерывной, т. е. для всех $x\in\mathbb{R}$

$$F(x)=F(x\pm 0)=\lim_{z\to x}F(z)\,. \quad (3.4.7)$$

Доказательство этого утверждения предоставляем читателю в задаче **203**.

Пример 3.4.1 (Закон распределения Парето). Годовой доход случайно выбранного налогоплательщика описывается случайной величиной X с плотностью распределения

$$f(x)=\begin{cases}0, & x<1,\\ \dfrac{c}{x^{3,5}}, & x>1.\end{cases}$$

Найдем значение параметра c и функцию распределения годового дохода.

Параметр c определим из условия (3.4.3):

$$1=\int\limits_{-\infty}^{+\infty} f(x)dx=\int\limits_{-\infty}^{1} 0dx+\int\limits_{1}^{+\infty}\frac{c}{x^{3,5}}dx=-\frac{c}{2,5}\cdot\frac{1}{x^{2,5}}\bigg|_{1}^{+\infty}=\frac{c}{2,5}\,,$$

откуда $c = 2{,}5$. Таким образом, $f(x)=\begin{cases}0, & x<1,\\ \dfrac{2{,}5}{x^{3{,}5}}, & x>1.\end{cases}$

Функцию распределения найдем по формуле (3.4.1): $F(x)=\int\limits_{-\infty}^{x} f(z)dz=$

$=\int\limits_{-\infty}^{x} 0dz=0$ при $x\leqslant 1$, $F(x)=\int\limits_{-\infty}^{x} f(z)dz=\int\limits_{-\infty}^{1} f(z)dz+\int\limits_{1}^{x} f(z)dz=\int\limits_{1}^{x}\dfrac{2{,}5}{z^{3{,}5}}dt=-\dfrac{1}{t^{2{,}5}}\Big|_{1}^{x}=$

$=1-\dfrac{1}{x^{2{,}5}}$ при $x>1$. Таким образом, $F(x)=\begin{cases}0, & x\leqslant 1,\\ 1-\dfrac{1}{x^{2{,}5}}, & x>1.\end{cases}$ ❑

Задачи

201. Годовой доход случайно выбранного налогоплательщика описывается случайной величиной X с плотностью распределения

$$f(x)=\begin{cases}0, & x<1,\\ \dfrac{c}{x^{4}}, & x>1.\end{cases}$$

Найдите значение параметра c и функцию распределения годового дохода.

202. Случайная величина X является абсолютно непрерывной. Известно, что $\mathbf{P}\{-\pi/4\leqslant X\leqslant \pi/4\}=1$, а для любого отрезка $[a,b]\subseteq[-\pi/4;\pi/4]$ $\mathbf{P}\{a\leqslant X\leqslant b\}=$ $=\sin(b-a)\cos(a+b)$. Найдите плотность распределения случайной величины X.

203. Докажите, что функция распределения абсолютно непрерывной случайной величины является непрерывной, т. е. для всех $x\in\mathbb{R}$ $F(x)=F(x\pm 0)=\lim\limits_{z\to x}F(z)$.

3.4.2. Математическое ожидание, дисперсия и среднее квадратическое отклонение абсолютно непрерывной случайной величины

В **п. 3.2.2** мы для дискретной случайной величины X, заданной рядом распределения

X	x_1	x_2	$\cdots$	x_n	$\cdots$
p	p_1	p_2	$\cdots$	p_n	$\cdots$

определили математическое ожидание

$$\mathbf{E}(X)=\sum_i x_i p_i=\sum_i x_i \mathbf{P}\{X=x_i\}$$

[см. (3.2.6)].

Попробуем обобщить это определение на случай абсолютно непрерывной случайной величины X, заданной плотностью распределения $f(x)$.

Пусть вначале $f(x)$ отлична от нуля только на некотором отрезке $[c;d]$. Разобьем этот отрезок на n полуинтервалов $[x_{i-1};x_i)$ одинаковой длины Δ, где $x_0=c$, $x_{i+1}=x_i+\Delta$, $\Delta=\dfrac{d-c}{n}$, $i=0, 1, 2, \ldots, n$.

Тогда вероятность попадания случайной величины X в полуинтервал $[x_{i-1}; x_i)$ по формуле (3.4.6) равна

$$\mathbf{P}\{X\in[x_{i-1};x_i)\}=F(x_{i-1})-F(x_i)=\int_{x_i}^{x_{i-1}} f(x)dx\,.$$

По теореме о среднем существует такая точка $x_i^*\in[x_{i-1};x_i)$, что

$$\int_{x_i}^{x_{i-1}} f(x)dx=f(x_i^*)[x_i-x_{i-1}]=f(x_i^*)\Delta\,.$$

Рассмотрим теперь дискретную случайную величину

$X^{(n)}$	x_1^*	x_2^*	$\cdots$	x_n^*
p	p_1	p_2	$\cdots$	p_n

где

$$p_i=\mathbf{P}\{X^{(n)}=x_i^*\}=\mathbf{P}\{X\in[x_{i-1};x_i)\}=f(x_i^*)\Delta\,.$$

Все p_i, действительно, являются неотрицательными и в сумме дают единицу — читатель легко может в этом убедиться. Поэтому $X^{(n)}$ на самом деле является случайной величиной.

Математическое ожидание этой случайной величины $X^{(n)}$ равно по определению

$$\mathbf{E}\big(X^{(n)}\big)=\sum_i x_i^* p_i=\sum_{i=1}^{n} x_i^* f(x_i^*)\Delta\,.$$

Если мы теперь устремим n к бесконечности, то интегральная сумма в правой части будет стремиться к интегралу $\int_c^d xf(x)dx$:

$$\mathbf{E}\big(X^{(n)}\big)\xrightarrow[n\to\infty]{}\int_c^d xf(x)dx\,.$$

Переход от отрезка $[c;\,d]$ к области $(-\infty;\,+\infty)$ означает переход к несобственному интегралу

$$\int_{-\infty}^{+\infty} xf(x)dx\,.$$

Поэтому *математическим ожиданием* абсолютно непрерывной случайной величины X естественно назвать число

$$\mathbf{E}(X)=\int_{-\infty}^{+\infty} xf(x)dx\,. \tag{3.4.8}$$

Математическое ожидание абсолютно непрерывной случайной величины обладает теми же **свойствами** (3.2.8)—(3.2.11), что и математическое ожидание дискретной случайной величины.

Формулы (3.2.13)—(3.2.15) для вычисления дисперсии абсолютно непрерывных случайных величин принимают вид

$$\mathbf{Var}(X)=\int_{-\infty}^{+\infty}[x-\mathbf{E}(X)]^2 f(x)dx\,, \tag{3.4.9}$$

$$\mathbf{Var}(X)=\int_{-\infty}^{+\infty} x^2 f(x)dx-[\mathbf{E}(X)]^2\,, \tag{3.4.10}$$

$$\mathbf{Var}(X)=\int_{-\infty}^{+\infty} x(x-1)f(x)dx+\mathbf{E}(X)-[\mathbf{E}(X)]^2 \tag{3.4.11}$$

соответственно.

Дисперсия абсолютно непрерывной случайной величины обладает теми же **свойствами** (3.2.20)—(3.2.22), что и дисперсия дискретной случайной величины.

В следующем параграфе мы познакомимся с несколькими семействами однотипных абсолютно непрерывных случайных величин, часто встречающимися в экономической практике (см. также **прил. 2**).

Пример 3.4.2. В условиях примера 3.4.1 оценим средний годовой доход случайно выбранного налогоплательщика, вычислив математическое ожидание

$$\mathbf{E}(X)=\int\limits_{-\infty}^{+\infty} xf(x)dx=\int\limits_{1}^{+\infty} x\frac{2{,}5}{x^{3{,}5}}dx=-\frac{2{,}5}{1{,}5}\cdot\frac{1}{x^{1{,}5}}\Bigg|_1^{+\infty}=\frac{5}{3}.$$

При этом дисперсия годового дохода равна

$$\mathbf{Var}(X)=\int\limits_{-\infty}^{+\infty} x^2 f(x)dx-[\mathbf{E}(X)]^2=\int\limits_{1}^{+\infty} x^2\frac{2{,}5}{x^{3{,}5}}dx-\left(\frac{5}{3}\right)^2=-\frac{2{,}5}{0{,}5}\cdot\frac{1}{x^{0{,}5}}\Bigg|_1^{+\infty}-\left(\frac{5}{3}\right)^2=5-\left(\frac{5}{3}\right)^2=\frac{20}{9};$$

среднее квадратическое отклонение $\sigma_X=2\sqrt{5}/3\approx 1{,}49$. ❑

Задачи

204. В условиях задачи **201** оцените средний годовой доход и среднее квадратическое отклонение годового дохода случайно выбранного налогоплательщика.

205. Плотность распределения случайной величины X

$$f(x)=\begin{cases} ax, & x\in(0;2),\\ 0, & x\notin(0;2).\end{cases}$$

Найдите значение параметра a и функцию распределения $F(x)$, математическое ожидание $\mathbf{E}(X)$ и дисперсию $\mathbf{Var}(X)$, постройте графики функций $f(x)$ и $F(x)$. Вычислите вероятность $\mathbf{P}\{|X-\mathbf{E}(X)|<0{,}5\}$ двумя способами: используя $f(x)$ и $F(x)$, отметьте эту вероятность на обоих графиках.

206. Найдите плотность распределения, математическое ожидание и дисперсию случайной величины с функцией распределения

$$F(x)=\begin{cases} 0, & x<-2,\\ \dfrac{x+2}{4}, & -2\leqslant x<2,\\ 1, & x\geqslant 2.\end{cases}$$

207. **Треугольный закон распределения.** Плотность распределения случайной величины X изображена на рис. 3.4.1. Найдите функцию распределения, математическое ожидание и дисперсию этой случайной величины.

208. Пусть X — абсолютно непрерывная случайная величина, $\overset{\circ}{X}=\dfrac{X-\mathbf{E}(X)}{\sigma_X}$. Докажите, что $\mathbf{E}(\overset{\circ}{X})=0$, $\mathbf{Var}(\overset{\circ}{X})=1$.

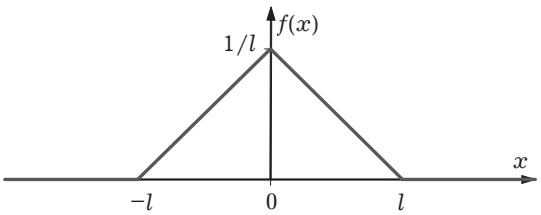


Рис. 3.4.1. *График плотности случайной величины, распределенной по треугольному закону*

§ 3.5. Абсолютно непрерывные случайные величины, часто встречающиеся в экономической практике

3.5.1. Равномерный закон распределения

Часто встречаются случайные величины, которые могут принимать значения только в строго определенных границах некоторого отрезка $[a; b]$, причем с содержательной точки зрения все значения внутри этого отрезка одинаково возможны. Например, минутная стрелка остановившихся механических часов будет показывать какое-то значение — заранее, до поломки, неизвестно, какое именно, но точно известно, что оно будет в пределах от 0 до 60. Погрешность округления также относится к подобным величинам.

Про такие величины говорят, что они распределены по равномерному закону на некотором отрезке $[a; b]$.

Более строго, случайная величина X с функцией распределения

$$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & x\in[a;b],\\ 1, & x>b\end{cases} \tag{3.5.1}$$

называется распределенной *по равномерному закону* на отрезке $[a; b]$.

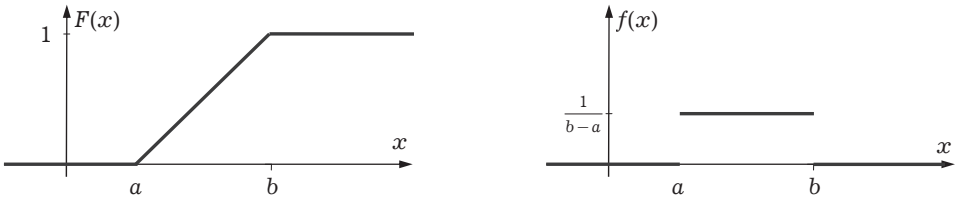
Легко убедиться, что функция (3.5.1) удовлетворяет всем свойствам (3.1.4)—(3.1.7) функции распределения, предоставляем это сделать читателю в задаче **211**.

Плотность распределения этой случайной величины получается дифференцированием:

$$f(x)=F'(x)=\begin{cases}0, & x<a.\\ \dfrac{1}{b-a}, & a<x<b,\\ 0, & x>b.\end{cases} \tag{3.5.2}$$

Тот факт, что случайная величина X распределена по равномерному закону на отрезке $[a; b]$, кратко обозначается так: $X = \mathrm{Unif}(a; b)$.

Графики функции распределения (3.5.1) и плотности распределения (3.5.2) случайной величины $X = \mathrm{Unif}(a; b)$ представлены на рис. 3.5.2.



а) *б*)

Рис. 3.5.2. *График функции* (*а*) *и плотности* (*б*) *распределения случайной величины* $X = \mathrm{Unif}(a; b)$

Математическое ожидание случайной величины $X = \mathrm{Unif}(a; b)$ равно

$$\mathbf{E}(X) = \frac{a+b}{2}, \tag{3.5.3}$$

а ее дисперсия

$$\mathbf{Var}(X) = \frac{(b-a)^2}{12}. \tag{3.5.4}$$

Доказательство. $\mathbf{E}(X) = \int\limits_{-\infty}^{+\infty} xf(x)dx = \int\limits_{a}^{b} x\frac{1}{b-a}dx = \frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{\cancel{(b-a)}(b+a)}{2\cancel{(b-a)}} = \frac{a+b}{2}$;

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 = \int\limits_{-\infty}^{+\infty} x^2 f(x)dx - \left(\frac{a+b}{2}\right)^2 = \int\limits_{a}^{b} x^2\frac{1}{b-a}dx - \frac{a^2+2ab+b^2}{4} =$$

$$= \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b - \frac{a^2+2ab+b^2}{4} = \frac{b^3-a^3}{3(b-a)} - \frac{a^2+2ab+b^2}{4} = \frac{\cancel{(b-a)}(b^2+ab+a^2)}{3\cancel{(b-a)}} - \frac{a^2+2ab+b^2}{4} =$$

$$= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}. \square$$

Пример 3.5.1. Случайная величина $X = \mathrm{Unif}(a = 0; b = 100)$. Найдем вероятности $\mathbf{P}\{X > 10\}$, $\mathbf{P}\{40 < X < 90\}$, $\mathbf{P}\{X = 50\}$ и $\mathbf{P}\{X > 50 \,|\, X < 80\}$, а также математическое ожидание и дисперсию этой случайной величины:

$$\mathbf{P}\{X > 10\} = 1 - \mathbf{P}\{X < 10\} = 1 - F(10) = 1 - \frac{10-0}{100-0} = 0{,}9;$$

$$\mathbf{P}\{40 < X < 90\} = F(90) - F(40) = \frac{90-0}{100-0} - \frac{40-0}{100-0} = 0{,}5;$$

$$\mathbf{P}\{X = 50\} = 0;\ \mathbf{P}\{X > 50 \,|\, X < 80\} = \frac{3}{8};$$

$$\mathbf{E}(X) = \frac{0+100}{2} = 50;\ \mathbf{Var}(X) = \frac{(100-0)^2}{12} = \frac{2500}{3} = 833{,}(3). \square$$

Задачи

209. Поезда метро следуют регулярно с интервалом 2 мин. Пассажир приходит на платформу станции в случайный момент времени, никак не связанный с расписанием движения поездов. Случайная величина X — время, в течение которого пассажиру придется ждать прибытия очередного поезда. Найдите: *а*) функцию и плотность распределения случайной величины X; *б*) ее математическое ожидание и дисперсию.

210. Правило трех сигм. Случайная величина $X = \mathrm{Unif}(a; b)$. Найдите $\mathbf{P}\{|X - \mathbf{E}(X)| < 3\sigma_X\}$.

211. Докажите, что функция (3.5.1) удовлетворяет всем свойствам (2.1.4)—(2.1.7) функции распределения.

212. Случайные величины $X = \mathrm{Unif}(a; b)$ и $Y = \mathrm{Unif}(c; d)$ независимы. Найдите $\mathbf{E}(XY)$ и $\mathbf{Var}(X + Y)$.

3.5.2. Показательный закон распределения

Рассмотрим некоторый простейший поток событий (см. **п. 3.3.5**). Вероятность того, что за время x произойдет хотя бы одно событие из этого простейшего потока по формуле (3.3.16) равна

$$F_X(x)=\begin{cases}0, & x<0,\\ 1-\mathrm{e}^{-\mu x}, & x\geqslant 0.\end{cases} \tag{3.5.5}$$

Случайная величина X с функцией распределения (3.5.5) называется распределенной *по показательному закону* с параметром μ.

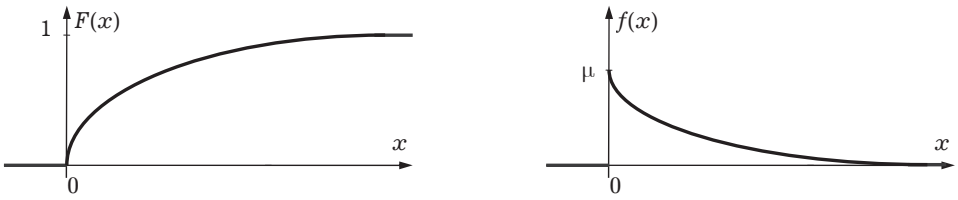
Легко убедиться, что функция (3.5.5) удовлетворяет всем свойствам (3.1.4)—(3.1.7) функции распределения, предоставляем это сделать читателю в задаче **217**.

Плотность распределения этой случайной величины получается дифференцированием:

$$f(x)=F'(x)=\begin{cases}0, & x<0,\\ \mu\,\mathrm{e}^{-\mu x}, & x>0.\end{cases} \tag{3.5.6}$$

Тот факт, что случайная величина X распределена по показательному закону с параметром μ, кратко обозначается так: $X=\mathrm{Exp}(\mu)$.

Графики функции распределения (3.5.5) и плотности распределения (3.5.6) случайной величины $X=\mathrm{Exp}(\mu)$ представлены на рис. 3.5.3.



а) *б*)

Рис. 3.5.3. *График функции* (*а*) *и плотности* (*б*) *распределения случайной величины* $X=\mathrm{Exp}(\mu)$

Интервал времени между двумя последовательными появлениями некоторого редкого события описывается случайной величиной, распределенной по показательному закону. Также по показательному закону распределено время безотказной работы какого-либо устройства (если не учитывать его старение).

Математическое ожидание и дисперсия случайной величины X, распределенной по показательному закону с параметром μ, выражаются формулами

$$\mathbf{E}(X)=\frac{1}{\mu}, \tag{3.5.7}$$

$$\mathbf{Var}(X)=\frac{1}{\mu^2}. \tag{3.5.8}$$

Доказательство. Имеем:

$$\mathbf{E}(X)=\int\limits_{-\infty}^{+\infty}xf(x)dx=\int\limits_{0}^{+\infty}x\mu\,\mathrm{e}^{-\mu x}\,dx=-\int\limits_{0}^{+\infty}xd(\mathrm{e}^{-\mu x})=-\left[(x\,\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty}-\int\limits_{0}^{+\infty}\mathrm{e}^{-\mu x}\,dx\right]=$$

$$= -\left[(x\,\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty} + \frac{1}{\mu}\int_0^{+\infty} \mathrm{e}^{-\mu x}\, d(-\mu x) \right] = -\left[(x\,\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty} + \frac{1}{\mu}(\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty} \right] = \frac{1}{\mu},$$

$$\mathbf{E}(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \int_0^{+\infty} x^2 \mu\,\mathrm{e}^{-\mu x}\,dx = -\int_0^{+\infty} x^2 d(\mathrm{e}^{-\mu x}) = -\left[(x^2\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty} - \int_0^{+\infty} \mathrm{e}^{-\mu x}\, d(x^2) \right] =$$

$$-\left[(x^2\mathrm{e}^{-\mu x})\Big|_{x=0}^{+\infty} - \int_0^{+\infty} 2x\,\mathrm{e}^{-\mu x}dx \right] = \frac{2}{\mu}\int_0^{+\infty} x\mu\,\mathrm{e}^{-\mu x}dx = \frac{2\mathbf{E}(X)}{\mu} = \frac{2}{\mu^2},$$

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{2}{\mu^2} - \frac{1}{\mu^2} = \frac{1}{\mu^2}. \square$$

В пакете Microsoft Excel плотность показательного распределения можно вычислить с помощью функции

$$f(x) = \text{ЭКСП.РАСП(<}x\text{>; <}\mu\text{>; ЛОЖЬ)},$$

а функцию показательного распределения — с помощью функции

$$F(x) = \text{ЭКСП.РАСП(<}x\text{>; <}\mu\text{>; ИСТИНА)}.$$

Случайная величина X, распределенная по показательному закону (3.5.5), обладает **свойством отсутствия последействия**, т. е. при всех $x > 0$, $y > 0$

$$\mathbf{P}\{X < x + y \mid X > x\} = \mathbf{P}\{X < y\}. \tag{3.5.9}$$

Доказательство. Действительно,

$$\mathbf{P}\{X < x+y \mid X > x\} = \frac{\mathbf{P}(\{X < x+y\} \cap \{X > x\})}{\mathbf{P}\{X > x\}} = \frac{\mathbf{P}\{x \leqslant X < x+y\}}{\mathbf{P}\{X > x\}} = \frac{F(x+y) - F(x)}{1 - F(x)} =$$

$$= \frac{1 - \mathrm{e}^{-\mu(x+y)} - (1 - \mathrm{e}^{-\mu x})}{1 - (1 - \mathrm{e}^{-\mu x})} = \frac{\mathrm{e}^{-\mu x} - \mathrm{e}^{-\mu(x+y)}}{\mathrm{e}^{-\mu x}} = \frac{\mathrm{e}^{-\mu x}(1 - \mathrm{e}^{-\mu y})}{\mathrm{e}^{-\mu x}} = 1 - \mathrm{e}^{-\mu y} = F(y) = \mathbf{P}\{X < y\}. \square$$

Важно отметить, что справедлив и обратный результат: если некоторая случайная величина X с непрерывной функцией распределения обладает свойством отсутствия последействия (3.5.9), то эта случайная величина распределена по показательному закону (3.5.5).

Доказательство. Пусть $F(x)$ — функция распределения случайной величины X. Согласно (3.5.9), $\mathbf{P}\{X < y\} = \mathbf{P}\{X < x + y \mid X > x\}$ или $F(y) = \dfrac{F(x+y) - F(x)}{1 - F(x)}$. Положим $Q(x) = 1 - F(x)$, тогда $1 - Q(y) = \dfrac{Q(x) - Q(x+y)}{Q(x)}$, откуда $Q(x) - Q(x + y) = Q(x)[1 - Q(y)]$ или $Q(x + y) = Q(x)Q(y)$. Все непрерывные ограниченные решения этого функционального уравнения имеют вид $Q(t) = \mathrm{e}^{-\mu t}$, где $\mu > 0$, значит, $F(x) = 1 - Q(x) = 1 - \mathrm{e}^{-\mu t}$, что и требовалось доказать. $\square$

Свойство отсутствия последействия (3.5.9) можно интерпретировать так же, как аналогичное свойство (2.3.11) геометрического распределения. Пусть длительность телефонных разговоров измеряется (вещественным неотрицательным) числом минут (причем разговор продолжается в среднем μ мин.), и в каждый момент времени абонент может принять решение закончить разговор или продолжить его. Тогда длительность телефонного разговора будет случайной величиной, распределенной по показательному закону (3.5.5) с параметром μ.

Свойство отсутствия последействия (3.5.9) означает, что условная вероятность того, что разговор будет продолжаться менее $x + y$ мин, если известно, что он не закончился за x мин, равна безусловной вероятности того, что разговор будет продолжаться менее y мин.

Пример 3.5.2. Обычно родители ругают Петю за принесенную «двойку» около 6 мин. На этот раз 6 мин уже прошло, а нотация все еще продолжается. Найти дисперсию длительности нотации. Определим, с какой вероятностью родители закончат читать нотацию в течение седьмой минуты.

Длительность нотации X можно считать распределенной по показательному закону. По условию обычная средняя длительность нотации (или ее математическое ожидание) составляет $\mathbf{E}(X) = 6$ мин. Но для показательного распределения $\mathbf{E}(X) = 1/\mu$, откуда $\mu = 1/\mathbf{E}(X) = 1/6$. Дисперсия длительности нотации при этом равна $\mathbf{Var}(X) = 1/\mu^2 = 1/(1/6)^2 = 36$. Вероятность того, что родители закончат читать нотацию в течение ближайшей (седьмой) минуты при условии, что нотация длится больше 6 мин, равна

$$\mathbf{P}\{X<7 \mid X>6\} = \frac{\mathbf{P}\{(X<7)\cap(X>6)\}}{\mathbf{P}\{X>6\}} = \frac{\mathbf{P}\{6<X<7\}}{\mathbf{P}\{X>6\}} = \frac{F(7)-F(6)}{1-F(6)} =$$

$$= \frac{1-\mathrm{e}^{-7/6}-(1-\mathrm{e}^{-6/6})}{1-(1-\mathrm{e}^{-6/6})} = \frac{\mathrm{e}^{-1}-\mathrm{e}^{-7/6}}{\mathrm{e}^{-1}} = 1-\mathrm{e}^{-1/6} \approx 0{,}154. \ \square$$

Задачи

213. Случайная величина $X = \mathrm{Exp}(\mu = 2)$. Определите вероятности $\mathbf{P}\{X > 1\}$, $\mathbf{P}\{X < 2\}$, $\mathbf{P}\{X > -1\}$, $\mathbf{P}\{X = 3\}$ и $\mathbf{P}\{X > 1 | X < 3\}$, математическое ожидание и дисперсию этой случайной величины.

214. Обычно брокер получает от своего клиента приказы об операциях на фондовой бирже раз в неделю. Найдите вероятность того, что сегодня поступит приказ, если последний приказ поступил два дня назад (поток приказов является простейшим).

215. Разговоры по мобильному телефону продолжаются в среднем 3 мин. Найдите: *а*) вероятность того, что очередной разговор будет продолжаться более 3 мин; *б*) долю разговоров, которые длятся менее 1 мин; *в*) вероятность того, что разговор, который длится уже больше 10 мин, закончится в течение одиннадцатой минуты; *г*) математическое ожидание и дисперсию длительности разговора.

216. Случайная величина $X = \mathrm{Exp}(\mu)$. Найдите: *а*) $\mathbf{P}\{0 \leqslant X \leqslant \tau\}$; *б*) $\mathbf{P}\{t \leqslant X \leqslant t+\tau\}$; *в*) $\mathbf{P}\{t \leqslant X \leqslant t+\tau \mid X \geqslant t\}$.

217. Докажите, что функция (3.5.5) удовлетворяет всем свойствам (3.1.4)—(3.1.7) функции распределения.

3.5.3. Нормальный закон распределения

Очень многие случайные величины имеют так называемый нормальный закон распределения; как мы увидим в **§ 4.2**, нормальный закон является предельным, к которому приближаются другие законы распределения при часто встречающихся одинаковых условиях.

Случайная величина X называется распределенной *по нормальному закону* с параметрами a и σ, если ее плотность распределения имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (3.5.10)$$

Нормальный закон распределения впервые встретился в 1838 г. в работах А. де Муавра. В конце XIX в. Ч. Пирсон предложил термин «нормальный закон

распределения». Обоснованием важности применения нормального закона распределения для решения прикладных задач занимались также П. С. Лаплас, К. Ф. Гаусс, Р.А. Эдрейн и др. Часто нормальный закон распределения называют *законом Гаусса*.

Очевидно, функция (3.5.10) неотрицательна при любых $a \in Ў$ и $\sigma > 0$, и нормирована (так как

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-a)^2}{2\sigma^2}}\, dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{[(x-a)/\sigma]^2}{2}}\, d\left(\frac{x-a}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\, dz = \frac{1}{\sqrt{2\pi}}\sqrt{2\pi} = 1$$

где последний интеграл

$$\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\, dz = \sqrt{2\pi}$$

представляет собой известный из курса математического анализа интеграл Пуассона), поэтому плотность распределения (3.5.10) действительно определяет некоторую случайную величину.

Функция распределения при этом равна

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}}\, dt. \qquad (3.5.11)$$

Для вычисления плотности и функции нормального распределения можно воспользоваться пакетом Microsoft Excel, в котором

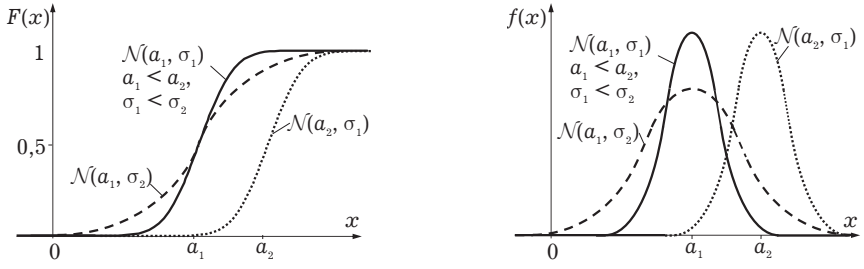
$$f(x) = \text{НОРМ.РАСП}(\langle x\rangle; \langle a\rangle; \langle\sigma\rangle; \text{ЛОЖЬ}),$$

а

$$F(x) = \text{НОРМ.РАСП}(\langle x\rangle; \langle a\rangle; \langle\sigma\rangle; \text{ИСТИНА}).$$

Тот факт, что случайная величина X распределена по нормальному закону с параметрами a и σ, кратко обозначается так: $X = \mathcal{N}(a; \sigma)$.

Графики плотностей распределения (3.5.10) и функций распределения (3.5.11) случайных величин $\mathcal{N}(a; \sigma)$ с различными значениями параметров a и σ представлены на рис. 3.5.4.



а) *б)*

Рис. 3.5.4. *График функции* (*а*) *и плотности* (*б*) *распределения случайных величин* $\mathcal{N}(a; \sigma)$ *с различными значениями параметров* a *и* σ

Рис. 3.5.4 демонстрирует, что плотность распределения достигает максимума в точке $x = a$, при увеличении a и неизменном значении σ график плотности распределения смещается вправо, а при увеличении σ и неизменном значении a график становится более низким и более растянутым по горизонтали: максимальное значение плотности увеличивается, а в удаленных от a точках значения плотности уменьшаются (поскольку площадь под кривой плотности распределения остается равной единице).

Математическое ожидание и дисперсия случайной величины, распределенной по нормальному закону, равны соответственно

$$\mathbf{E}(X) = a \tag{3.5.12}$$

и

$$\mathbf{Var}(X) = \sigma^2. \tag{3.5.13}$$

Доказательство. По определению $\mathbf{E}(X) = \int\limits_{-\infty}^{+\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} x\,\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}\,dx$. Введем замену переменной $z = \frac{x-a}{\sigma}$, тогда $x = a + \sigma z$, $dx = d(a+\sigma z) = \sigma dz$ и

$$\mathbf{E}(X) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (a+\sigma z)\mathrm{e}^{-\frac{z^2}{2}}\,dz = \frac{a}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} \mathrm{e}^{-\frac{z^2}{2}}\,dz + \frac{\sigma}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} z\,\mathrm{e}^{-\frac{z^2}{2}}\,dz = \frac{a}{\sqrt{2\pi}}\sqrt{2\pi} + 0 = a.$$

Дисперсия $\mathbf{Var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \int\limits_{-\infty}^{+\infty} (x - \mathbf{E}(X))^2 f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (x-a)^2\,\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}\,dx$.

Воспользуемся той же заменой переменных $z = \frac{x-a}{\sigma}$:

$$\mathbf{Var}(X) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (a+\sigma z - a)^2\,\mathrm{e}^{-\frac{z^2}{2}}\,\sigma dz = \frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} z^2\,\mathrm{e}^{-\frac{z^2}{2}}\,dz = -\frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} z\,d\left(\mathrm{e}^{-\frac{z^2}{2}}\right) =$$

$$= \{\text{по частям}\} = -\frac{2\sigma^2}{\sqrt{2\pi}} \left(z\,\mathrm{e}^{-\frac{z^2}{2}}\Big|_{-\infty}^{+\infty} - \int\limits_{-\infty}^{+\infty} \mathrm{e}^{-\frac{z^2}{2}}\,dz \right).$$

Уменьшаемое в скобках, как легко видеть, равно нулю, поэтому

$$\mathbf{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} \mathrm{e}^{-\frac{z^2}{2}}\,dz = \frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi} = \sigma^2. \;\square$$

Таким образом, математическое ожидание случайной величины, распределенной по нормальному закону с параметрами a и σ, равно a, а ее среднее квадратическое отклонение равно σ.

Если $X = \mathcal{N}(a; \sigma)$, а $Y = kX + b$, где $k \neq 0$, то $Y = \mathcal{N}(ka + b; k\sigma)$.

Доказательство. Пусть $k > 0$. Тогда $Y \leqslant y \Leftrightarrow kX + b \leqslant y \Leftrightarrow X \leqslant \frac{y-b}{k}$,

$$F_Y(y) = \mathbf{P}\{Y \leqslant y\} = \mathbf{P}\left\{X \leqslant \frac{y-b}{k}\right\} = F_X\left(\frac{y-b}{k}\right) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{y-b}{k}} \mathrm{e}^{-\frac{(t-a)^2}{2\sigma^2}}\,dt =$$

$$= \left\{\text{замена } w = kt + b,\ t = \frac{w-b}{k},\ dt = \frac{1}{k}dw,\ \text{при } t = \frac{y-b}{k} \text{ верхний предел } w = y\right\} =$$

$$= \frac{1}{k\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{y} \mathrm{e}^{-\frac{\left(\frac{w-b}{k}-a\right)^2}{2\sigma^2}}\,dw = \frac{1}{(k\sigma)\sqrt{2\pi}} \int\limits_{-\infty}^{y} \mathrm{e}^{-\frac{(w-(ka+b))^2}{2(k\sigma)^2}}\,dw.$$

Доказательство этого факта при $k < 0$ мы оставляем читателю в задаче **228.** $\square$

Пусть $X = \mathcal{N}(a;\sigma)$, $\overset{\circ}{X} = \overset{\circ}{X} = \dfrac{X-a}{\sigma} = \dfrac{1}{\sigma}X - \dfrac{a}{\sigma}$. Тогда $\overset{\circ}{X} = \mathcal{N}\left(\dfrac{1}{\sigma}a - \dfrac{a}{\sigma}; \dfrac{1}{\sigma}\sigma\right) = \mathcal{N}(0;1)$.

Нормальный закон распределения с параметрами $a = 0$ и $\sigma = 1$ называется *стандартным нормальным законом*. Плотность распределения и функция распределения случайной величины, распределенной по стандартному нормальному закону, равны соответственно

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$$

и

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\, dt.$$

Для вычисления плотности и функции стандартного нормального распределения в Microsoft Excel есть функции

$f(x)$ = НОРМ.СТ.РАСП(<x>; ЛОЖЬ) = НОРМ.РАСП(<x>; 0; 1; ЛОЖЬ),

$F(x)$ = НОРМ.СТ.РАСП(<x>; ИСТИНА) = НОРМ.РАСП(<x>; 0; 1; ИСТИНА).

Правило трех сигм. *Для любой случайной величины $X = \mathcal{N}(a;\sigma)$ вероятность*

$$\mathbf{P}\{|X - \mathbf{M}X| < 3\sigma_X\} = 0{,}9973. \qquad (3.5.14)$$

Доказательство. Вычислим

$$\mathbf{P}\{|X - \mathbf{M}X| < 3\sigma_X\} = \mathbf{P}\{|X - a| < 3\sigma\} = \mathbf{P}\left\{\left|\frac{X-a}{\sigma}\right| < 3\right\} = \mathbf{P}\left\{\left|\overset{\circ}{X}\right| < 3\right\} =$$

$$= \mathbf{P}\left\{-3 < \overset{\circ}{X} < 3\right\} = F_{\overset{\circ}{X}}(3) - F_{\overset{\circ}{X}}(-3).$$

Поскольку $\overset{\circ}{X} = \mathcal{N}(0;1)$, окончательно получаем:

$$\mathbf{P}\{|X - \mathbf{M}X| < 3\sigma_X\} = F_{\overset{\circ}{X}}(3) - F_{\overset{\circ}{X}}(-3) =$$

= НОРМ.СТ.РАСП(3; ИСТИНА) – НОРМ.СТ.РАСП(–3; ИСТИНА) = 0,9973. ❑

Таким образом, отклонение распределенной по нормальному закону случайной величины более, чем на три средних квадратических отклонения от своего математического ожидания, может произойти лишь с очень малой вероятностью $1 - 0{,}9973 = 0{,}0027 \approx 0{,}3\%$.

Обсудим теперь у с л о в и я, в которых возникает нормальный закон распределения и причины его распространенности.

Оказывается, что если складывается много независимых (или слабо зависимых) случайных величин $X_1, X_2, \ldots, X_n$

$$X = X_1 + X_2 + \cdots + X_n,$$

причем каждая из этих величин мала (относительно суммы $X = X_1 + X_2 + \cdots + X_n$) и никакая из них не превалирует над остальными, то сумма этих величин X распределена приблизительно по нормальному закону с некоторыми параметрами a и σ (обоснование этого факта отложим до **п. 4.2.1**).

Пример 3.5.3. Например, мы рассматриваем суточную выручку некоторого магазина. Она складывается из денежных сумм, которые оставляют в магазине покупатели.

Кто именно придет сегодня в магазин и какие покупки сделает, конечно, неизвестно, но, как правило, в магазин приходят люди с примерно одинаковым достатком, причем стоимость покупок каждого сравнительно мала по сравнению с суточной выручкой магазина, а совершают покупки люди независимо друг от друга.

Поэтому суточная выручка представляет собой случайную величину, распределенную по нормальному закону. ❑

Пример 3.5.4. Пусть страховая компания застраховала на один год большое число владельцев автомобилей от ущерба, нанесенного этим автомобилям.

С какими из этих автомобилей за год произойдут страховые случаи, каков будет в каждом из этих случаев размер ущерба, который компания должна возместить, заранее неизвестно, но суммарные выплаты, которые страховая компания произведет за год, распределены по нормальному закону, поскольку суммарные выплаты представляют собой сумму большого числа выплат по каждому из страховых случаев, произошедших за год, а выплаты по отдельным случаям, во-первых, независимы, а во-вторых, малы по сравнению с суммарными выплатами страховой компании за год. ❑

Пример 3.5.5. Значения теста IQ (коэффициента интеллекта) распределены приблизительно по нормальному закону с математическим ожиданием $a = 100$ и средним квадратическим отклонением $\sigma = 16$.

Запишем выражения для функции распределения коэффициента интеллекта:

$$F(x) = \frac{1}{16\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(z-100)^2}{512}} dz$$

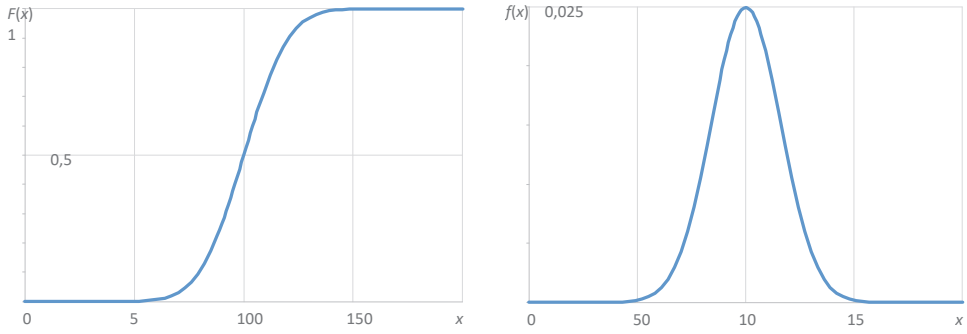
и плотности его распределения:

$$f(x) = \frac{1}{16\sqrt{2\pi}} e^{-\frac{(x-100)^2}{512}}.$$

Графики этих функций представлены на рис. 3.5.5. Для построения графиков удобно воспользоваться пакетом Microsoft Excel, где

$F(x)$ = НОРМ.РАСП(<x>; 100; 16; ИСТИНА),

$f(x)$ = НОРМ.РАСП(<x>; 100; 16; ЛОЖЬ). ❑



а *б*)

Рис. 3.5.5. *Графики функции* (*а*) *и плотности распределения* (*б*) *случайной величины, распределенной по нормальному закону*

Пример 3.5.6 (Закон распределения Рэлея). Найдем функцию распределения и математическое ожидание случайной величины X, плотность распределения которой имеет вид

$$f(x)=\begin{cases}0, & x<0,\\ \dfrac{x}{\sigma^2}\mathrm{e}^{-\frac{x}{2\sigma^2}}, & x>0.\end{cases}$$

Очевидно, $F(x)=0$ при $x<0$. При $x>0$

$$F(x)=\int\limits_{-\infty}^{x} f(r)dr=\int\limits_{-\infty}^{0} 0dr+\int\limits_{0}^{x}\frac{r}{\sigma^2}\mathrm{e}^{-\frac{r^2}{2\sigma^2}}dr=\int\limits_{0}^{x}\mathrm{e}^{-\left(\frac{r}{\sigma\sqrt{2}}\right)^2}\frac{2r}{2\sigma^2}dr=$$

$$-\int\limits_{0}^{x}\mathrm{e}^{-\left(\frac{r}{\sigma\sqrt{2}}\right)^2}d\left[-\left(\frac{r}{\sigma\sqrt{2}}\right)^2\right]=-\mathrm{e}^{-\left(\frac{r}{\sigma\sqrt{2}}\right)^2}\Big|_0^x=-\mathrm{e}^{-\frac{x^2}{2\sigma^2}}+e^0=1-\mathrm{e}^{-\frac{x^2}{2\sigma^2}};$$

$$\mathbf{E}(X)=\int\limits_{-\infty}^{+\infty} rf(r)dr=\int\limits_{0}^{+\infty} r\frac{r}{\sigma^2}\mathrm{e}^{-\frac{r^2}{2\sigma^2}}dr=\sigma\sqrt{2}\int\limits_{0}^{+\infty}2\left(\frac{r}{\sigma\sqrt{2}}\right)^2\mathrm{e}^{-\left(\frac{r}{\sigma\sqrt{2}}\right)^2}d\left(\frac{r}{\sigma\sqrt{2}}\right)=$$

$$=\left\{\text{замена } t=\frac{r}{\sigma\sqrt{2}}\right\}=\sigma\sqrt{2}\int\limits_{0}^{+\infty}2t^2\mathrm{e}^{-t^2}dt=\sigma\sqrt{2}\int\limits_{0}^{+\infty}t\mathrm{e}^{-t^2}2tdt=-\sigma\sqrt{2}\int\limits_{0}^{+\infty}t\mathrm{e}^{-t^2}d(-t^2)=$$

$$-\sigma\sqrt{2}\int\limits_{0}^{+\infty}td\,\mathrm{e}^{-t^2}=\{\text{по частям}\}=-\sigma\sqrt{2}\left(t\mathrm{e}^{-t^2}\Big|_0^{+\infty}-\int\limits_{0}^{+\infty}\mathrm{e}^{-t^2}dt\right)=\sigma\sqrt{2}\int\limits_{0}^{+\infty}\mathrm{e}^{-t^2}dt=$$

$$=\left\{\text{замена } z=\sqrt{2}t\right\}=\sigma\sqrt{2}\int\limits_{0}^{+\infty}\mathrm{e}^{-\frac{z^2}{2}}\frac{1}{\sqrt{2}}dz=\sigma\int\limits_{0}^{+\infty}\mathrm{e}^{-\frac{z^2}{2}}dz=$$

$$=\left\{\text{заметим, что }\int\limits_{0}^{+\infty}\mathrm{e}^{-\frac{z^2}{2}}dz=\frac{1}{2}\int\limits_{-\infty}^{+\infty}\mathrm{e}^{-\frac{z^2}{2}}dz=\frac{\sqrt{2\pi}}{2}\right\}=\frac{1}{2}\sigma\sqrt{2\pi}=\sigma\sqrt{\frac{\pi}{2}}.\ \square$$

Задачи

218. В условиях примера 3.5.5 найдите долю людей, у которых коэффициент интеллекта окажется: *а*) меньше 60; *б*) меньше 75; *в*) не больше 95; *г*) меньше 100; *д*) меньше 120; *е*) в пределах от 80 до 120; *ж*) больше 120 среди тех, у кого он больше 110.

219. В условиях примера 3.5.5 найдите долю людей, у которых коэффициент интеллекта отклонится от 100 менее, чем на 48.

220. В условиях примера 3.5.5 найдите вероятность того, что из шести независимо отобранных человек у двоих коэффициент интеллекта будет выше 92.

221. Случайная величина $X=\mathcal{N}(a=0;\sigma=1)$. Постройте кривую распределения этой случайной величины, график ее функции распределения. Найдите координаты плотности распределения $f(x)$ и функции распределения $F(x)$ при $x=1;\ \ -1;\ \ 2{,}25$. Вычислите и укажите на обоих графиках следующие вероятности: $\mathbf{P}\{X<1\}$, $\mathbf{P}\{X>-1\}$, $\mathbf{P}\{|X|<1\}$, $\mathbf{P}\{|X|<3\}$ и $\mathbf{P}\{0<X<3\}$.

222. Случайная величина $X=\mathcal{N}(a=1;\sigma=1)$. Найдите вероятности $\mathbf{P}\{X>2\}$, $\mathbf{P}\{X<2\}$, $\mathbf{P}\{0<X<2\}$ и $\mathbf{P}\{X<2\,|\,X>0\}$.

223. Случайная величина $X = \mathcal{N}(a = 2; \sigma = 3)$. Найдите вероятности $\mathbf{P}\{X > 1\}$, $\mathbf{P}\{-2 < X \leqslant 2\}$, $\mathbf{P}\{X < 2\}$ и $\mathbf{P}\{X < 2 \mid X > 0\}$. Запишите правило трех сигм для этой случайной величины.

224. Текущая цена акции может быть приближена нормальным распределением с математическим ожиданием 15,28 руб. и средним квадратическим отклонением 0,12 руб. Рассчитайте вероятности того, что цена акции окажется: *а*) не ниже 15,50 руб.; *б*) не выше 15,00 руб.; *в*) между 15,10 руб. и 15,40 руб.; *г*) между 15,05 руб. и 15,10 руб.

225. Цена некоторой акции распределена по нормальному закону. В течение последнего года в 20% рабочих дней цена была меньше 20 руб., а в 75% рабочих дней она была больше 25 руб. Найдите математическое ожидание и среднее квадратическое отклонение цены этой акции.

226. Инвестор покупает ценные бумаги за счет кредита, взятого под i процентов годовых под залог своей недвижимости. Годовая доходность ценных бумаг R представляет собой случайную величину с математическим ожиданием $r > i$ и средним квадратическим отклонением σ. Предполагая случайную величину R распределенной по нормальному закону, оцените вероятность того, что инвестор не сможет вернуть кредит.

227. При расследовании причин аварии было установлено, что она могла произойти из-за установки на автомобиль детали, размеры которой выходят за пределы допустимого интервала (0,95 мм; 1,05 мм). Известно, что размер деталей, поступающих на конвейер автозавода, представляет собой случайную величину, распределенную по нормальному закону с математическим ожиданием, равным 1,00 мм, и средним квадратическим отклонением, равным 0,05 мм. Оцените вероятность того, что причиной аварии послужила установка на автомобиль детали нестандартного размера.

228. Докажите, что если случайная величина X распределена по нормальному закону с параметрами a и σ, то случайная величина $Y = kX + b$ (где $k < 0$) распределена по нормальному закону с параметрами $a_Y = ka + b$ и $\sigma_Y = k\sigma$.

229. Правило шести сигм. При статистическом контроле качества сложной электроники в настоящее время используется правило шести сигм. Напомним, что правило трех сигм означает, что некачественная продукция [не попадающая в интервал $(\mathbf{E}(X) - 3\sigma_X; \mathbf{E}(X) + 3\sigma_X)$] составляет $100\% - 99{,}73\% = 0{,}27\%$, т. е. на каждые 10 000 единиц продукции допустимо изготовление не более, чем 27 некачественных. Поясните, в чем заключается правило шести сигм: укажите максимальную допустимую долю некачественной продукции при использовании этого правила.

3.5.4. Логарифмически нормальный закон распределения и ценообразование финансовых инструментов

Как мы заметили в предыдущем пункте, по нормальному закону распределены многие случайные величины, встречающиеся на практике, а именно (говоря нестрого) такие случайные величины, которые представляют собой с у м м у большого числа независимых и приблизительно одинаковых случайных слагаемых. Но, конечно, встречаются и другие случайные величины. Например, рассмотрим акцию, стоимость которой в конце периода времени t составляет S_t. С п е к у л я т и в н а я о п е р а ц и я, состоящая в том, что в мо-

мент времени $(n-1)$ инвестор покупает акцию по цене S_{n-1}, а в момент n продает ее по цене S_n, обеспечивает *доходность*

$$R_n = \frac{S_n - S_{n-1}}{S_{n-1}}, \tag{3.5.15}$$

которая может быть гораздо выше процентной ставки банковского счета (так, в России за 1996 г. цены на акции выросли в среднем в четыре раза, т. е. средняя доходность была выше 300% годовых, что было намного выше доходности по вложениям в банки).

Конечно, предсказать заранее доходность по вложениям в акцию невозможно, поэтому предположим, что доходности R_n (3.5.19) в различные моменты времени $n = 1, 2, 3, \ldots$ представляют собой независимые одинаково распределенные случайные величины. Пусть в начальный момент акция стоила $S_0 = a$ ден. ед., тогда в момент n ее стоимость составит

$$S_n = S_0(1 + R_1)(1 + R_2) \cdots (1 + R_n).$$

Преобразуем эту формулу:

$$S_n = S_0(1+R_1)(1+R_2)\cdots(1+R_n) =$$
$$= S_0\, \mathrm{e}^{\ln[(1+R_1)(1+R_2)\cdots(1+R_n)]} = S_0\, \mathrm{e}^{\ln(1+R_1)+\ln(1+R_2)+\cdots+\ln(1+R_n)} = S_0\, \mathrm{e}^{Y_1+Y_2+\cdots+Y_n},$$

где $Y_i = \ln(1 + R_i)$, $i = 1, 2, \ldots, n$.

Таким образом,

$$\ln S_n = \ln\left(S_0\, \mathrm{e}^{Y_1+Y_2+\cdots+Y_n}\right) = \ln S_0 + Y_1 + Y_2 + \cdots + Y_n.$$

или

$$\ln S_n - \ln S_0 = Y_1 + Y_2 + \cdots + Y_n.$$

Но случайные величины $Y_1, Y_2, \ldots, Y_n$ имеют один и тот же (вообще говоря, неизвестный) закон распределения и независимы (предоставляем читателю доказать, что из независимости $R_1, R_2, \ldots, R_n$ следует независимость $Y_1, Y_2, \ldots, Y_n$), поэтому их сумма в правой части распределена приблизительно по нормальному закону, значит и левая часть $[\ln S_n - \ln S_0]$ также распределена по нормальному закону.

Приведенный пример показывает, что часто встречаются случайные величины, которые представляют собой п р о и з в е д е н и е большого числа независимых и приблизительно одинаковых случайных слагаемых, тогда логарифм произведения будет равен сумме логарифмов множителей и, следовательно, будет иметь нормальный закон распределения.

Случайная величина X называется распределенной *по логарифмически нормальному закону* с параметрами a и σ, если ее логарифм $\ln X$ распределен по нормальному закону с параметрами a и σ.

Найдем функцию распределения такой случайной величины X. Имеем при $x \geqslant 0$:

$$F(x) = \mathbf{P}\{X < x\} = \mathbf{P}\{\ln X < \ln x\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\ln x} \mathrm{e}^{-\frac{(t-a)^2}{2\sigma^2}}\, dt = F_{\mathcal{N}(a;\sigma)}(\ln x).$$

При $x \leqslant 0$, очевидно, $F(x) = \mathbf{P}\{X \leqslant x\} = 0$, так как логарифм от неположительных чисел не определен. Итак,

$$F(x)=\begin{cases}0, & x<0,\\ \dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\ln x} e^{-\frac{(t-a)^2}{2\sigma^2}}\,dt, & x\geqslant 0\end{cases}=\begin{cases}0, & x<0,\\ F_{\mathcal{N}(a;\sigma)}(\ln x), & x\geqslant 0.\end{cases}\qquad(3.5.16)$$

Плотность распределения получается из (3.5.20) дифференцированием:

$$f(x)=F'(x)=\begin{cases}0, & x<0,\\ \dfrac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}, & x>0\end{cases}=\begin{cases}0, & x<0,\\ \dfrac{1}{x}f_{\mathcal{N}(a;\sigma)}(\ln x), & x>0.\end{cases}\qquad(3.5.17)$$

Тот факт, что случайная величина X распределена по логарифмически нормальному закону с параметрами a и σ, кратко обозначается так: $X=\mathcal{LN}(a;\sigma)$.

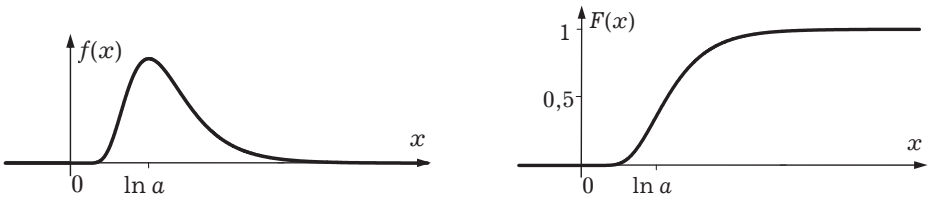
Для вычисления плотности и функции логарифмически нормального распределения можно воспользоваться пакетом Microsoft Excel, в котором при $x\geqslant 0$

$$f(x)=\text{ЛОГНОРМ.РАСП(<}x\text{>; <}a\text{>; <}\sigma\text{>; ЛОЖЬ)},$$

а

$$F(x)=\text{ЛОГНОРМ.РАСП(<}x\text{>; <}a\text{>; <}\sigma\text{>; ИСТИНА).}$$

Графики функции распределения (3.5.16) и плотности распределения (3.5.17) случайной величины $X=\mathcal{LN}(a;\sigma)$ представлены на рис. 3.5.6.



а) *б*)

Рис. 3.5.6. *График функции* (*а*) *и плотности* (*б*) *распределения случайной величины* $X=\mathcal{LN}(a;\sigma)$

Математическое ожидание и дисперсия случайной величины, распределенной по логарифмически нормальному закону с параметрами a и σ, равны соответственно

$$\mathbf{E}(X)=e^{a+\frac{\sigma^2}{2}},\qquad(3.5.18)$$

$$\mathbf{Var}(X)=e^{2a+\sigma^2}(e^{\sigma^2}-1).\qquad(3.5.19)$$

Доказательство. Поскольку плотность распределения

$$f(x)=\begin{cases}0, & x<0,\\ \dfrac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}, & x>0,\end{cases}$$

математическое ожидание данной случайной величины равно

$$\mathbf{E}(X)=\int_{-\infty}^{+\infty}xf(x)dx=\int_{-\infty}^{0}x\cdot 0dx+\int_{0}^{+\infty}x\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{0}^{+\infty}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}dx=$$

$$=\left\{\begin{array}{c}\text{замена } u=\dfrac{\ln x-a}{\sigma},\ \text{при этом } x=e^{a+\sigma u},\ dx=\sigma e^{a+\sigma u}du,\\ \text{если } x=0,\ \text{то } u=-\infty,\ \text{если } x=+\infty,\ \text{то } u=+\infty\end{array}\right\}=$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} \sigma e^{a+\sigma u}\, du = e^a \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{u^2}{2}+\sigma u}\, du = e^a \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{u^2-2\sigma u+\sigma^2-\sigma^2}{2}}\, du =$$

$$= e^a \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{(u-\sigma)^2}{2}} e^{\frac{\sigma^2}{2}}\, du = e^a e^{\frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{(u-\sigma)^2}{2}}\, d(u-\sigma) = e^{a+\frac{\sigma^2}{2}},$$

так как $\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{(u-\sigma)^2}{2}}\, d(u-\sigma) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\, dt = 1$ как интеграл от плотности нормального закона распределения. Далее:

$$\mathbf{E}(X^2) = \int\limits_{-\infty}^{+\infty} x^2 f(x)dx = \int\limits_{-\infty}^{0} x^2 \cdot 0 dx + \int\limits_{0}^{+\infty} x^2 \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x-a)^2}{2\sigma^2}}\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{0}^{+\infty} x e^{-\frac{(\ln x-a)^2}{2\sigma^2}}\, dx =$$

$$= \{\text{та же замена}\} = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{0}^{+\infty} e^{a+\sigma u} e^{-\frac{u^2}{2}} \sigma e^{a+\sigma u}\, du = e^{2a} \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{+\infty} e^{-\frac{u^2}{2}+2\sigma u}\, du =$$

$$= e^{2a} \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{+\infty} e^{-\frac{u^2-4\sigma u+4\sigma^2-4\sigma^2}{2}}\, du = e^{2a} e^{2\sigma^2} \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{+\infty} e^{-\frac{(u-2\sigma)^2}{2}}\, d(u-2\sigma) = e^{2a+2\sigma^2},$$

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 = e^{2a+2\sigma^2} - \left(e^{a+\frac{\sigma^2}{2}}\right)^2 = e^{2a+2\sigma^2} - e^{2a+\sigma^2} = e^{2a+\sigma^2}(e^{\sigma^2}-1),$$

что и требовалось доказать. ❑

По логарифмически нормальному закону распределены обычно такие величины, как стоимости ценных бумаг; размеры заработной платы работника; доходы семей; долговечность изделий, работающих в режиме износа, и многие другие.

Задачи

230. Месячный доход (в ден. ед.) случайно выбранной семьи из некоторой социальной группы описывается логарифмически нормальным законом распределения с параметрами $a = \ln 100\,000$, $\sigma = \ln 150$. Найдите долю семей, имеющих доход менее 150 000 ден. ед.

231. Статистика по вкладам населения в некоторый банк говорит о том, что размер вклада случайно выбранного клиента (в ден. ед.) распределен по логарифмически нормальному закону с параметрами $a = \ln 120\,000$, $\sigma = \ln 144$. Определите: *а*) средний размер вклада; *б*) долю клиентов, размер вклада которых составляет не менее 100 000 ден. ед.

3.5.5. Законы распределения, важные в математической статистике

Закон распределения Пирсона. Пусть $Z_1, Z_2, \ldots, Z_k$ — независимые случайные величины, распределенные по нормальному закону с параметрами $a = 0$, $\sigma = 1$ [$Z_i = \mathcal{N}(0; 1)$, $i = 1, 2, \ldots, k$]. Закон распределения случайной величины

$$\chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \tag{3.5.20}$$

по определению называется *законом распределения Пирсона* с k степенями свободы или *законом распределения «Хи-квадрат»* с k степенями свободы.

Математическое ожидание и дисперсия случайной величины χ_k^2 равны соответственно

$$\mathbf{E}\left(\chi_k^2\right) = k, \tag{3.5.21}$$

$$\mathbf{Var}\left(\chi_k^2\right) = 2k. \tag{3.5.22}$$

Доказательство. $\mathbf{E}(Z_i^2)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}x^2\,\mathrm{e}^{-\frac{(x-0)^2}{2\cdot1^2}}\,dx=\frac{1}{1\cdot\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}(x-0)^2\,\mathrm{e}^{-\frac{(x-0)^2}{2\cdot1^2}}\,dx=\mathbf{Var}(Z_i)=1,\mathbf{E}\left(\chi_k^2\right)=$

$=\mathbf{E}(Z_1^2)+\mathbf{E}(Z_2^2)+\cdots+\mathbf{E}(Z_k^2)=k.$

Доказательство формулы (3.5.22) для дисперсии оставляем читателю в качестве задачи **235**. ❑

Очевидно, что если случайные величины $\chi_{k_1}^2$ и $\chi_{k_2}^2$ независимы, то

$$\chi_{k_1}^2+\chi_{k_2}^2=\chi_{k_1+k_2}^2.$$

Закон распределения Стьюдента. Пусть $Z=\mathcal{N}(0;1)$ — случайная величина, распределенная по нормальному закону с параметрами $a=0$, $\sigma=1$, а χ_k^2 — случайная величина, распределенная по закону «Хи-квадрат» с k степенями свободы, причем случайные величины Z и χ_k^2 независимы. Закон распределения случайной величины

$$T_k=\frac{Z}{\sqrt{\frac{\chi_k^2}{k}}} \tag{3.5.23}$$

по определению называется *законом распределения Стьюдента* с k степенями свободы.

Закон распределения Фишера — Снедекора. Пусть теперь случайные величины $\chi_{k_1}^2$ и $\chi_{k_2}^2$ независимы. Закон распределения случайной величины

$$F_{k_1;k_2}=\frac{\chi_{k_1}^2/k_1}{\chi_{k_2}^2/k_2} \tag{3.5.24}$$

по определению называется *законом распределения Фишера — Снедекора* с k_1 и k_2 степенями свободы.

Формулы для функций и плотностей распределения случайных величин, введенных в данном пункте, довольно сложны, и мы не будем их здесь приводить. В пакете Microsoft Excel функции распределения случайных величин χ_k^2, T_k и $F_{k_1;k_2}$ вычисляются как

$$F(x)=\text{ХИ2.РАСП}(\langle x\rangle;\langle k\rangle;\text{ИСТИНА});$$

$$F(x)=\text{СТЬЮДЕНТ.РАСП}(\langle x\rangle;\langle k\rangle;\text{ИСТИНА});$$

$$F(x)=\text{F.РАСП}(\langle x\rangle;\langle k_1\rangle;\langle k_2\rangle;\text{ИСТИНА})$$

соответственно.

На самом деле на практике не приходится пользоваться непосредственно формулами для вычисления функций и плотностей распределения данных случайных величин; нас больше будет интересовать в некотором смысле обратная задача: не поиск вероятности того, что случайная величина попадет в некоторое множество, а поиск такого множества, в которое случайная величина попадет с заданной вероятностью (см. **главу 5**).

Задачи

232. Вычислите с помощью статистических функций Microsoft Excel: *а)* $\mathbf{P}\{\chi_{20}^2>10{,}9\}$; *б)* $\mathbf{P}\{\chi_{20}^2<28{,}4\}$; *в)* $\mathbf{P}\{8{,}26\leqslant\chi_{20}^2<31{,}4\}$; *г)* $\mathbf{P}\{\chi_{40}^2>10{,}9\}$; *д)* $\mathbf{P}\{\chi_{40}^2<28{,}4\}$; *е)* $\mathbf{P}\{8{,}26\leqslant\chi_{40}^2<31{,}4\}$.

233. С помощью статистических функций Microsoft Excel найдите: *а*) $\mathbf{P}\{T_{10} < 2{,}23\}$; *б*) $\mathbf{P}\{T_{10} > -2{,}23\}$; *в*) $\mathbf{P}\{|T_{10}| < 2{,}23\}$; *г*) $\mathbf{P}\{-1{,}81 < T_{10} < 3{,}17\}$; *д*) $\mathbf{P}\{-1{,}81 < T_{40} < 3{,}17\}$.

234. С помощью статистических функций Microsoft Excel найдите $\mathbf{P}\{|F_{3;16}| < 3{,}24\}$.

235. Докажите формулу (3.5.22) для дисперсии случайной величины χ^2_k.

236. Докажите, что $T_k^2 = F_{1;k}$.

237. Докажите, что $\dfrac{\chi^2_k}{k} = F_{k;+\infty}$.

3.5.6. Смеси распределений

Кроме дискретных случайных величин, которые принимают конечное или счетное число значений и имеют кусочно-постоянную функцию распределения с соответственно конечным или счетным числом точек разрыва, и абсолютно непрерывных случайных величин, которые имеют непрерывную функцию распределения вида $F(x) = \int\limits_{-\infty}^{x} f(z)dz$, существуют и другие типы случайных величин.

Рассмотрим случайную величину X_1 с функцией распределения $F_{X_1}(x)$ и случайную величину X_2 с функцией распределения $F_{X_2}(x)$.

Функции распределения случайных величин X_1 и X_2 обладают свойствами монотонности (3.1.4):

$$\text{если } x < y, \text{ то } F_{X_1}(x) \leqslant F_{X_1}(y), \quad F_{X_1}(x) \leqslant F_{X_2}(y), \tag{3.5.25}$$

ограниченности (3.1.5)—(3.1.6):

$$0 \leqslant F_{X_1}(x) \leqslant 1, \quad 0 \leqslant F_{X_2}(x) \leqslant 1, \tag{3.5.26}$$

$$F_{X_1}(-\infty) = F_{X_2}(-\infty) = 0, \quad F_{X_1}(+\infty) = F_{X_2}(+\infty) = 1 \tag{3.5.27}$$

и непрерывности (3.1.7):

$$F_{X_1}(x) = F_{X_1}(x+0) = \lim_{\substack{z \to x \\ z < x}} F_{X_1}(z), \quad F_{X_2}(x) = F_{X_2}(x+0) = \lim_{\substack{z \to x \\ z < x}} F_{X_2}(z). \tag{3.5.28}$$

Пусть теперь $p_1, p_2 \geqslant 0$, $p_1 + p_2 = 1$,

$$F_X(x) = p_1 F_{X_1}(x) + p_2 F_{X_2}(x). \tag{3.5.29}$$

Функция $F(x)$ также обладает свойствами монотонности (3.1.4), ограниченности (3.1.5)—(3.1.6), и непрерывности (3.1.7), а потому является функцией распределения некоторой случайной величины X.

Доказательство. Если $x < y$, то согласно (3.5.3) $F_{X_1}(x) \leqslant F_{X_1}(y)$, $F_{X_2}(x) \leqslant F_{X_2}(y)$. Умножив первое из этих неравенств на p_1, а второе на p_2, и сложив полученные неравенства, получим, что $p_1 F_{X_1}(x) + p_2 F_{X_2}(x) \leqslant p_1 F_{X_1}(y) + p_2 F_{X_2}(y)$ или $F(x) \leqslant F(y)$, т. е. функция $F(x)$ обладает свойством (2.1.4). Если первое из неравенств (3.5.1) умножить на (неотрицательное число) p_1, а второе из этих неравенств умножить на (неотрицательное число) p_2, получим: $0 \leqslant p_1 F_{X_1}(x) \leqslant p_1$, $0 \leqslant p_2 F_{X_2}(x) \leqslant p_2$.

Складывая эти неравенства, получаем, что $0 \leqslant p_1 F_{X_1}(x) + p_2 F_{X_2}(x) \leqslant p_1 + p_2$, и поскольку $p_1 + p_2 = 1$, убеждаемся, что функция $F(x)$ обладают свойством (3.1.5). При этом $F(-\infty) = p_1 F_{X_1}(-\infty) + p_2 F_{X_2}(-\infty) = 0, F(+\infty) = p_1 F_{X_1}(+\infty) + p_2 F_{X_2}(+\infty) = p_1 + p_2 = 1$, т. е. функция $F(x)$ обладает свойством (2.1.6).

Из свойств предела функции следует, что

$$F(x+0) = \lim_{\substack{z \to x \\ z > x}} F(z) = \lim_{\substack{z \to x \\ z > x}} \left[p_1 F_{X_1}(x) + p_2 F_{X_2}(x) \right] = p_1 \lim_{\substack{z \to x \\ z > x}} F_{X_1}(x) + p_2 \lim_{\substack{z \to x \\ z > x}} F_{X_2}(x) =$$

$$= p_1 F_{X_1}(x+0) + p_2 F_{X_2}(x+0) = p_1 F_{X_1}(x) + p_2 F_{X_2}(x) = p_1 F_{X_1}(x) + p_2 F_{X_2}(x) = F(x),$$

значит, функция $F(x)$ обладает свойством (2.1.7).

Таким образом, функция $F(x)$ определяет некоторую случайную величину X, что и требовалось доказать. ❑

Такая случайная величина X, функция распределения которой определяется формулой (3.5.5), называется *смесью* случайных величин X_1 и X_2.

Читателю предлагается убедиться в том, что смесь двух дискретных случайных величин будет дискретной случайной величиной, а смесь двух абсолютно непрерывных случайных величин — абсолютно непрерывной случайной величиной (задача **243**).

Пусть теперь X_1 — дискретная случайная величина, а X_2 — абсолютно непрерывная случайная величина. Смесь этих случайных величин, очевидно, уже не будет ни дискретной случайной величиной, ни абсолютно непрерывной.

Функция распределения смеси дискретной и абсолютно непрерывной случайной величин на некоторых участках непрерывна, а в отдельных точках терпит разрывы (рис. 3.5.7). При этом вероятность того, что смесь X примет значение из какого-либо участка непрерывности $F(x)$, равна нулю, а вероятность того, что смесь X примет значение, равное какой-либо точке разрыва функции $F(x)$, равна величине скачка $F(x)$ в этой точке.

Замечание. К сожалению, простым представлением (3.5.5), где X_1 — дискретная случайная величина, а X_2 — абсолютно непрерывная случайная величина, в с е случайные величины не исчерпываются. Существуют также *сингулярные* случайные величины (такие, что для любого $x \in \mathrm{R}$ вероятность $\mathbf{P}\{X = x\} = 0$, но при этом существует такое конечное или счетное множество точек $B = \{x_1, x_2, \ldots, x_l, \ldots \in \mathrm{R}\}$, что $\mathbf{P}\{X \in B\} = 1$). Легко видеть, что дискретные случайные величины не удовлетворяют первому из этих условий, а для абсолютно непрерывных случайных величин нарушается второе условие, поэтому сингулярные случайные величины не могут быть ни дискретными, ни абсолютно непрерывными. Однако сингулярные случайные величины являются экзотиическими объектами и в практических задачах не встречаются.

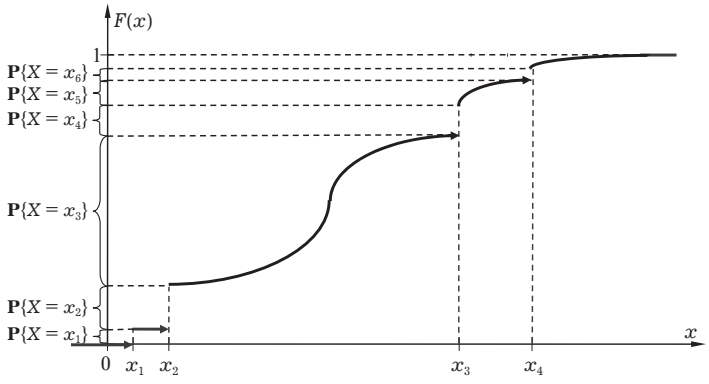


Рис. 3.5.7. *График функции распределения смеси дискретной и абсолютно непрерывной случайной величин*

Пример 3.5.7. Объем продаж X дилера некоторой компании является случайной величиной, распределенной по нормальному закону с параметрами $a = 10\,000$ ден. ед. и $\sigma = 2500$ ден. ед. Комиссионное вознаграждение дилера составляет 10% от объема продаж, но не может быть меньше 500 ден. ед. и больше 1500 ден. ед. Найдем функцию распределения комиссионного вознаграждения Y.

Представим случайную величину Y в виде

$$Y = \begin{cases} 500, & 0{,}1X \leqslant 500, \\ 0{,}1X, & 500 < 0{,}1X \leqslant 1500, \\ 1500, & 0{,}1X > 1500 \end{cases} \quad \text{или} \quad Y = \begin{cases} 500, & X \leqslant 5000, \\ 0{,}1X, & 5000 < X \leqslant 15\,000, \\ 1500, & X > 15\,000. \end{cases}$$

При этом функция распределения случайной величины Y

$$F_Y(y) = \begin{cases} 0, & y < 500, \\ \mathbf{P}\{0{,}1X < y\}, & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500 \end{cases} = \begin{cases} 0, & y < 500, \\ \mathbf{P}\{X < y / 0{,}1\}, & 500 \leqslant y < 1500, = \\ 1, & y \geqslant 1500 \end{cases}$$

$$= \begin{cases} 0, & y < 500, \\ \mathbf{P}\{X < 10y\}, & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500 \end{cases} = \begin{cases} 0, & y < 500, \\ F_X(10y), & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500 \end{cases}$$

$$= \begin{cases} 0, & y < 500, \\ F_{\mathcal{N}(a=10\,000;\,\sigma=2500)}(10y), & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500 \end{cases} = \begin{cases} 0, & y < 500, \\ F_{\mathcal{N}(a=1000;\,\sigma=250)}(y), & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500. \end{cases}$$

Эта функция $F_Y(y)$ может быть представлена в виде

$$F_Y(y) = p_1 F_{Y_1}(y) + p_2 F_{Y_2}(y),$$

где

$$p_2 = F_{\mathcal{N}(a=1000;\,\sigma=250)}(1500) - F_{\mathcal{N}(a=1000;\,\sigma=250)}(500) = F_{\mathcal{N}(a=0;\,\sigma=1)}(2) - F_{\mathcal{N}(a=0;\,\sigma=1)}(-2),$$

$$p_1 = 1 - F_{\mathcal{N}(a=0;\,\sigma=1)}(2) + F_{\mathcal{N}(a=0;\,\sigma=1)}(-2),$$

$$F_{Y_1}(y) = \begin{cases} 0, & y < 500, \\ \dfrac{1}{2}, & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500, \end{cases}$$

$$F_{Y_2}(y) = \begin{cases} 0, & y < 500, \\ \dfrac{F_{\mathcal{N}(a=1000;\,\sigma=250)}(y)}{F_{\mathcal{N}(a=1000;\,\sigma=250)}(1500) - F_{\mathcal{N}(a=1000;\,\sigma=250)}(500)}, & 500 \leqslant y < 1500, \\ 1, & y \geqslant 1500. \end{cases}$$

Значения 500 и 1500 эта случайная величина Y принимает с пренебрежимо небольшой вероятностью $\dfrac{1 - F_{\mathcal{N}(a=0;\,\sigma=1)}(2) + F_{\mathcal{N}(a=0;\,\sigma=1)}(-2)}{2} \approx 0{,}02275$. ❑

Задачи

238. Объем продаж X дилера является случайной величиной, распределенной по нормальному закону с параметрами $a = 10\,000$ ден. ед. и $= 2500$ ден. ед.

Комиссионное вознаграждение Y дилера составляет 20% от объема продаж и не может быть меньше 1000 ден. ед. Найдите функцию распределения комиссионного вознаграждения Y.

239. Поезд должен прийти на станцию по расписанию, но иногда, в силу случайных причин, может опаздывать (прибытие раньше назначенного времени исключено). Время опоздания поезда (при условии, что поезд опаздывает) представляет собой случайную величину, распределенную по показательному закону с параметром $\mu = 1/5$ (мин$^{-1}$). Вероятность того, что поезд не опоздает, равна 0,7. Найдите функцию распределения времени опоздания Y.

240. Случайная величина X с вероятностью 0,2 распределена по нормальному закону с параметрами $a = 0$, $\sigma = 2$, а с вероятностью 0,8 — по нормальному закону с параметрами $a = 2$, $\sigma = 1$. Найдите плотность распределения случайной величины X.

241. Случайная величина X с вероятностью p имеет функцию распределения $F_1(x)$, а с вероятностью $(1 - p)$ — функцию распределения $F_2(x)$. Найдите функцию распределения случайной величины X.

242. Абсолютно непрерывная случайная величина X с вероятностью p имеет плотность распределения $f_1(x)$, а с вероятностью $(1 - p)$ — плотность распределения $f_2(x)$. Найдите плотность распределения случайной величины X.

243. Докажите, что смесь двух дискретных случайных величин будет дискретной случайной величиной, а смесь двух абсолютно непрерывных случайных величин — абсолютно непрерывной случайной величиной.

§ 3.6. Моменты и критические границы случайной величины

3.6.1. Моменты случайной величины

Начальным моментом k-го порядка случайной величины X называется математическое ожидание k-й степени случайной величины X:

$$\nu_k(X) = \mathbf{E}(X^k), \quad k = 0, 1, 2, \ldots \tag{3.6.1}$$

Центральным моментом k-го порядка случайной величины X называется математическое ожидание k-й степени отклонения случайной величины X от своего математического ожидания:

$$\mu_k(X) = \mathbf{E}\{[X - \mathbf{E}(X)]^k\}, \quad k = 0, 1, 2, \ldots \tag{3.6.2}$$

Начальный момент первого порядка случайной величины X — это просто ее математическое ожидание: $\nu_1(X) = \mathbf{E}(X)$, а центральный момент второго порядка — это ее дисперсия: $\mu_2(X) = \mathbf{Var}(X)$.

Справедливы следующие выражения для центральных моментов:

$$\mu_0(X) = \nu_0(X) = 1; \tag{3.6.3}$$

$$\mu_1(X) = 0; \tag{3.6.4}$$

$$\mu_2(X) = \nu_2(X) - \nu_1^2(X); \tag{3.6.5}$$

$$\mu_3(X) = \nu_3(X) - 3\nu_1(X)\nu_2(X) + 2\nu_1^3(X); \tag{3.6.6}$$

$$\mu_4(X) = \nu_4(X) - 4\nu_1(X)\nu_3(X) + 6\nu_1^2(X)\nu_2(X) - 3\nu_1^4(X). \tag{3.6.7}$$

Доказательство. Действительно, $\mu_0(X) = \mathbf{E}[(X - \mathbf{M}X)^0] = \mathbf{E}(1) = 1$, $\nu_0(X) = \mathbf{E}(X^0) = \mathbf{E}(1) = 1$, $\mu_1(X) = \mathbf{E}[X - \mathbf{E}(X)] = \mathbf{E}(X) - \mathbf{E}(X) = 0$, формула (3.6.5) следует непосредственно из формулы (3.2.14), $\mu_3(X) = \mathbf{E}\{[X - \mathbf{E}(X)]^3\} = \mathbf{E}\{[X - \nu_1(X)]^3\} = \mathbf{E}(X^3) - 3\nu_1(X)\mathbf{E}(X^2) + 3\{\nu_1(X)\}^2\mathbf{E}(X) - \{\nu_1(X)\}^3 = \nu_3(X) - 3\nu_1(X)\nu_2(X) + 2\{\nu_1(X)\}^3$. Доказательство формулы (3.6.7) оставляем читателю в задаче **246**. ❑

Замечание. Как показал в 20-х гг. XX в. Т. Карлеман, если

$$\sum_{i=1}^{\infty}[\nu_{2i}(X)]^{-\frac{1}{2i}}=+\infty,$$

то по всем моментам случайной величины X можно однозначно восстановить ее функцию распределения. Например, если известно математическое ожидание $a=\nu_1(X)$ и дисперсия $\sigma^2=\mu_2(X)$ случайной величины, распределенной по нормальному закону, то известна и ее функция распределения

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^3}}dt=\text{НОРМ.РАСП}(\langle x\rangle;\langle a\rangle;\langle\sigma\rangle;\text{ИСТИНА}),$$

поэтому случайная величина, распределенная по нормальному закону, полностью определяется первыми д в у м я начальными моментами [так как $\mu_2(X)=\nu_2(X)-\nu_1^2(X)$ по формуле (3.6.5)].

Задачи

244. Найдите начальные и центральные моменты первых четырех порядков для случайной величины $X=\text{Unif}(a;b)$.

245. Докажите, что величина $\mathbf{E}(X-c)^2$ достигает своего наименьшего значения при $c=\mathbf{E}(X)$.

246. Докажите формулу (3.6.7).

3.6.2. Асимметрия и эксцесс случайной величины

Коэффициент асимметрии $\mathbf{A}_X$ случайной величины X характеризует с к о ш е н н о с т ь кривой распределения этой случайной величины относительно ее математического ожидания и вычисляется по формуле

$$\mathbf{A}_X=\frac{\mu_3(X)}{\sigma_X^3}. \tag{3.6.8}$$

Обычно $|\mathbf{A}_X|<2$. Для симметричных распределений $\mathbf{A}_X=0$, если левая ветвь кривой распределения длиннее правой, то $\mathbf{A}_X<0$, если же левая ветвь кривой распределения короче правой, то $\mathbf{A}_X>0$.

Эксцесс $\mathbf{E}_X$ случайной величины X характеризует т я ж е с т ь х в о с т о в распределения этой случайной величины по сравнению с нормальным законом распределения и вычисляется по формуле

$$\mathbf{E}_X=\frac{\mu_4(X)}{\sigma_X^4}-3. \tag{3.6.9}$$

Обычно $-1\leqslant\mathbf{E}_X\leqslant 6$. Для нормального распределения $\mathbf{E}_{\mathcal{N}(a;\sigma)}=0$; если кривая распределения случайной величины X имеет м е н е е т я ж е л ы е х в о с т ы, чем кривая нормального распределения, то $\mathbf{E}_X<0$; если кривая распределения случайной величины X имеет б о л е е т я ж е л ы е х в о с т ы, чем кривая нормального распределения, то $\mathbf{E}_X>0$.

Задачи

247. Найдите коэффициент асимметрии и эксцесс случайной величины $X=\text{Unif}(a;b)$.

248. Найдите коэффициент асимметрии и эксцесс случайной величины $X=\text{Exp}(\mu)$.

249. Найдите коэффициент асимметрии и эксцесс случайной величины $X=\mathcal{N}(a;\sigma)$.

250. **Закон распределения Лапласа.** Найдите коэффициент асимметрии и эксцесс случайной величины X, заданной с помощью плотности распределения $f(x)=\frac{\mu}{2}e^{-\mu|x|}$, $x\in\mathbb{R}$.

251. Пусть X — некоторая случайная величина, $Y = aX + b$ (a и b — неслучайные постоянные, $a \neq 0$). Докажите, что $\mathbf{A}_Y = \begin{cases} \mathbf{A}_X, & a > 0, \\ -\mathbf{A}_X, & a < 0, \end{cases}$ $\mathbf{E}_Y = \mathbf{E}_X$.

252. Пусть X — некоторая случайная величина, $\overset{\circ}{X} = \dfrac{X - \mathbf{E}(X)}{\sigma_X}$. Докажите, что $\mathbf{A}_{\overset{\circ}{X}} = \mathbf{A}_X$, $\mathbf{E}_{\overset{\circ}{X}} = \mathbf{E}_X$.

3.6.3. Квантили и процентные точки случайной величины

Квантилью (или *левосторонней критической границей*) *уровня* α случайной величины X называется такое число x_α, что

$$F_X(x_\alpha) = \alpha, \tag{3.6.10}$$

(т. е. $\mathbf{P}\{X \leqslant x_\alpha\} = \alpha$), а $100\alpha\%$*-ной точкой* (или *правосторонней критической границей уровня* α) случайной величины X называется такое число ω_α, что

$$F_X(\omega_\alpha) = 1 - \alpha \tag{3.6.11}$$

(т. е. $\mathbf{P}\{X > \omega_\alpha\} = \alpha$).

Левосторонняя и правосторонняя критические границы одного и того же уровня связаны между собой очевидным соотношением

$$x_\alpha = \omega_{1-\alpha}. \tag{3.6.12}$$

Замечание. Определения (3.6.10) и (3.6.1) корректны только для случая абсолютно непрерывной случайной величины. Если же случайная величина X является дискретной или смешанной, то может оказаться так, что либо числа, определяемого формулой (3.6.10) [или (3.6.11)] не существует, либо таких чисел бесконечно много. Но поскольку для любого $\alpha \in (0; 1)$ можно найти два таких числа x'_α и x''_α, что $F_X(x'_\alpha) \leqslant \alpha$, а $F_X(x''_\alpha) > \alpha$, причем $F_X(x) \geqslant \alpha$ для всех $x > x'_\alpha$ и $F_X(x) \leqslant \alpha$ для всех $x < x''_\alpha$, то в этом случае *квантилью уровня* α случайной величины X называется любое число $x_\alpha \in [x'_\alpha; x''_\alpha]$, а $100\alpha\%$*-ной точкой* этой случайной величины — любая ее квантиль уровня $(1 - \alpha)$.

Двусторонними критическими границами уровня α случайной величины X называются такие числа $\underline{x}_\alpha$, $\overline{x}_\alpha$, что одновременно

$$F_X(\underline{x}_\alpha) = \frac{\alpha}{2}, \qquad F_X(\overline{x}_\alpha) = 1 - \frac{\alpha}{2}, \tag{3.6.13}$$

т. е. $\mathbf{P}\{X \leqslant \underline{x}_\alpha\} = \mathbf{P}\{X > \overline{x}_\alpha\} = \dfrac{\alpha}{2}$.

Односторонние и двусторонние критические границы случайной величины X связаны следующими соотношениями:

$$\underline{x}_\alpha = x_{\alpha/2} = \omega_{1-\alpha/2}; \quad \overline{x}_\alpha = x_{1-\alpha/2} = \omega_{\alpha/2}. \tag{3.6.14}$$

Если плотность распределения абсолютно непрерывной случайной величины X с и м м е т р и ч н а относительно оси ординат и $\omega_{\alpha/2} = x_{1-\alpha/2}$ — $50\alpha\%$-ная точка случайной величины X (т. е. $\mathbf{P}\{X > \omega_{\alpha/2}\} = \alpha/2$), то, очевидно, квантиль уровня $\alpha/2$ этой случайной величины будет равна $x_{\alpha/2} = -\omega_{\alpha/2}$, причем $\mathbf{P}\{X \leqslant -\omega_{\alpha/2}\} = \mathbf{P}\{X > \omega_{\alpha/2}\} = \alpha/2$ или

$$\mathbf{P}\{|X| > \omega_{\alpha/2}\} = \alpha. \tag{3.6.15}$$

В этом случае двусторонние критические границы уровня α случайной величины X будут также симметричны относительно оси ординат и равны соответственно $\underline{x}_\alpha = -\omega_{\alpha/2} = -x_{1-\alpha/2}$ и $\overline{x}_\alpha = \omega_{\alpha/2} = x_{1-\alpha/2}$.

На рис. 3.6.1 отмечены квантиль уровня α, 100α%-ная критическая точка (светлой штриховкой) и двусторонние критические границы уровня α (темной штриховкой) случайной величины $X = \mathcal{N}(0; 1)$.

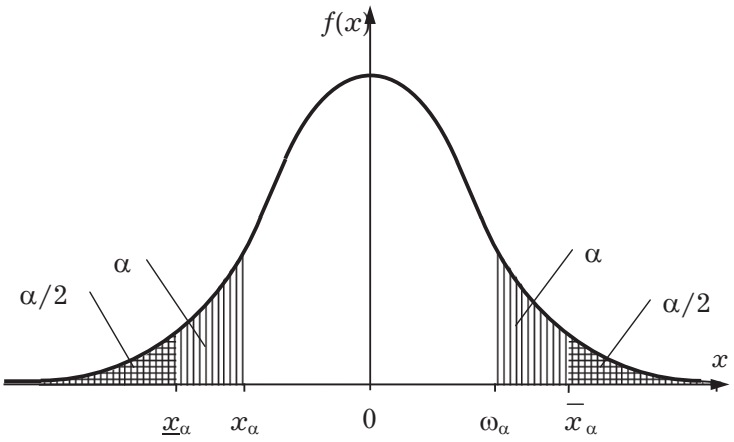


Рис. 3.6.1. *Критические границы случайной величины* $X = \mathcal{N}(0; 1)$

В пакете Microsoft Excel существуют статистические функции для вычисления критических границ наиболее распространенных случайных величин:

u_α = НОРМ.ОБР(<α>; <a>; <σ>) —
квантиль уровня α случайной величины $X = \mathcal{N}(a; \sigma)$;

z_α = НОРМ.СТ.ОБР(<α>) —
квантиль уровня α случайной величины $Z = \mathcal{N}(0; 1)$;

u_α = ЛОГНОРМ.ОБР(<α>; <a>; <σ>) —
квантиль уровня α случайной величины $X = \mathcal{LN}(a; \sigma)$;

$t_{k;\alpha}$ = СТЬЮДЕНТ.ОБР(<α>; <k>) —
квантиль уровня α случайной величины T_k;

$\overline{t}_{k;\alpha} = t_{k;1-\alpha/2}$ = СТЬЮДЕНТ.ОБР.2Х(<α>; <k>) —
абсолютное значение двусторонних критических границ уровня α случайной величины T_k (при этом $\underline{t}_{k;\alpha} - \overline{t}_{k;\alpha} = -t_{k;1-\alpha/2}$, так как плотность закона распределения Стьюдента симметрична относительно оси ординат);

$\chi^2_{k;\alpha}$ = ХИ2.ОБР(<α>; <k>) —
квантиль уровня α случайной величины χ^2_k;

$f_{k_1;k_2;\alpha}$ = F.ОБР(<α>; <k1>; <k2>) —

квантиль уровня случайной величины $F_{k_1;k_2}$;

x_α = БИНОМ.ОБР(<n>; <p>; <α>) —

наименьшая из квантилей уровня α случайной величины $X = \mathrm{Bin}(n; p)$.

Пример 3.6.1. В условиях примера 3.4.1 определим размер годового дохода $_{0,5}$, не ниже которого с вероятностью 0,5 окажется годовой доход случайно выбранного налогоплательщика.

Функция распределения годового дохода случайно выбранного налогоплательщика нам известна:

$$F(x) = \begin{cases} 0, & x < 1, \\ 1 - \dfrac{1}{x^{2,5}}, & x \geqslant 1. \end{cases}$$

По условию $\mathbf{P}\{X > \omega_{0,5}\} = 1 - \mathbf{P}\{X \leqslant \omega_{0,5}\} = 1 - F(\omega_{0,5}) = 0{,}5$, откуда $F(\omega_{0,5}) = 0{,}5$, т. е. $1 - 1/\omega_{0,5}^{2,5} = 0{,}5$ или $\omega_{0,5}^{2,5} = 2$, поэтому $\omega_{0,5} = 2^{1/2,5} = 2^{0,4} \approx 1{,}32$. Эта величина $\omega_{0,5} \approx 1{,}32$ является квантилью уровня 0,5 и 50%-ной точкой распределения годового дохода налогоплательщика. ❑

Задачи

253. В условиях задачи **201** определите размер годового дохода $\omega_{0,1}$, которым обладают 10% самых богатых налогоплательщиков.

254. Найдите для случайной величины $Z = \mathcal{N}(0; 1)$: *а*) квантили уровня 0,025 и уровня 0,975; *б*) 5%-ную точку.

255. Вычислите с помощью статистических функций Microsoft Excel квантили уровня 0,05 и 0,95 случайных величин: *а*) T_{10}; *б*) χ^2_{20}.

256. Вычислите с помощью статистических функций Microsoft Excel квантили уровня 0,05 и 0,95 случайной величины $F_{2;3}$.

3.6.4. Сумма под риском

Пусть случайная величина X вновь имеет смысл дохода от инвестиционной операции, которая продолжается в течение некоторого периода T, и мы хотим оценить риск этой операции.

В практической деятельности для этого часто рассчитывают показатель суммы под риском (*Value at Risk*, **VaR**): *100γ%-ной суммой под риском* данной инвестиционной операции за период T называется такое число $\mathbf{VaR}_\gamma$, что с вероятностью γ убыток $(-X)$ от проведения операции в течение данного периода будет не больше, чем $\mathbf{VaR}_\gamma$:

$$\mathbf{P}\{-X \leqslant \mathbf{VaR}_\gamma\} = \gamma \tag{3.6.16}$$

или

$$\mathbf{P}\{X \geqslant -\mathbf{VaR}_\gamma\} = \gamma.$$

Но тогда, согласно формуле (3.6.2),

$$-\mathbf{VaR}_\gamma = \omega_\gamma$$

или, окончательно,

$$\mathbf{VaR}_\gamma = -\omega_\gamma = -u_{1-\gamma}, \tag{3.6.17}$$

т. е. 100γ%-ная сумма под риском — это число, противоположное 100γ%-ной точке случайной величины X, что иллюстрируется рис. 3.6.2.

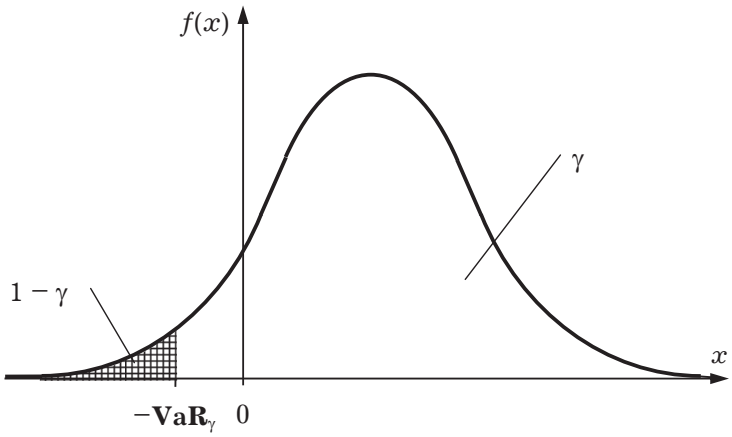


Рис. 3.6.2. *Сумма под риском*

Несмотря на ряд недостатков этого показателя, активно обсуждаемых в финансовых кругах, на сегодняшний день **VaR** является стандартом де-факто для оценки рыночных рисков в банках и инвестиционных компаниях. Базельский комитет по банковскому надзору рекомендует банкам резервировать (для покрытия возможных убытков) средства в размере, в 3—5 раз превышающем **VaR** инвестиционного портфеля банка (т. е. всех его операций).

Пример 3.6.2. Текущая цена акции равна 15 руб. Стоимость акции через семь дней описывается случайной величиной $Y = \mathcal{N}(16;\ 1)$. Рассчитаем недельную 95%-ную сумму под риском операции по покупке пакета 1000 таких акций и его последующей продаже через семь дней.

Доход от данной операции равен, очевидно, $X = 1000(Y - 15)$. По результатам **п. 3.5.3** доход X будет при этом распределен по нормальному закону с параметрами $a = 1000$, $\sigma = 1000$. Найдем 95%-ную точку случайной величины X. Воспользуемся пакетом Microsoft Excel: $\omega_{0,95} = $ НОРМ.ОБР(1 – 0,95; 1000; 1000) $= -644{,}85$, поэтому $\mathbf{VaR}_{0,95} = -\omega_{0,95} = 644{,}85$ ден. ед. ❑

Задачи

257. Текущая цена акции равна 100 руб. Стоимость акции через семь дней описывается случайной величиной $Y = \mathcal{N}(110;\ 7)$. Рассчитайте недельную 95%-ную сумму под риском операции по покупке пакета из 1000 таких акций и его последующей продаже.

258. Текущая цена акции равна 100 руб. Стоимость акции через пять лет описывается случайной величиной $Y =$ (300; 700). Вычислите пятилетнюю

95%-ную сумму под риском операции по покупке пакета из 1000 таких акций и его последующей продаже.

3.6.5. Медиана и мода случайной величины

В качестве показателя центра группирования значений случайной величины, наряду с математическим ожиданием, используются также медиана и мода.

Медианой x_{med} случайной величины X называется ее квантиль уровня 0,5 (или, что то же самое, ее 50%-ная точка):

$$x_{\text{med}} = x_{0,5} = \omega_{0,5}. \tag{3.6.18}$$

В качестве медианы дискретной случайной величины (поскольку квантиль дискретной случайной величины может быть определена не однозначно) обычно берут значение x_{med}, получаемое линейной аппроксимацией:

$$x_{\text{med}} = x_l + \frac{x_{l+1} - x_l}{p_{l+1}}\left(\frac{1}{2} - \sum_{i=1}^{l} p_i\right), \tag{3.6.19}$$

где *медианный отрезок* $[x_l; x_{l+1}]$ определяется из условий

$$\sum_{i=1}^{l} p_i \leqslant 0{,}5; \qquad \sum_{i=1}^{l+1} p_i > 0{,}5. \tag{3.6.20}$$

Модой абсолютно непрерывной случайной величины X называется точка локального максимума плотности распределения:

$$f_X(x_{\text{mod}}) = \max_{x \in \hat{Y}} f_X(x). \tag{3.6.21}$$

Модой дискретной случайной величины X называется значение этой случайной величины, соответствующее наибольшей вероятности:

$$x_{\text{mod}} = x_i, \text{ такое что } p_i = \max_j p_j. \tag{3.6.22}$$

Случайные величины, имеющие одну моду, называются *унимодальными*.

Для непрерывных случайных величин, плотность распределения которых симметрична относительно оси ординат, математическое ожидание, мода и медиана совпадают, что предлагаем доказать читателю в задаче **263**.

В случае же, когда распределение случайной величины не является симметричным, мода и медиана являются достаточно важными характеристиками.

Задачи

259. Найдите моду и медиану случайной величины $X = \text{Geom}(p)$.

260. Найдите моду и медиану случайной величины $X = \mathcal{N}(a; b)$.

261. Найдите моду и медиану случайной величины $X = \text{Unif}(a; b)$.

262. Найдите моду и медиану случайной величины $X = \text{Exp}(\mu)$.

263. Докажите, что для непрерывных случайных величин, плотность распределения которых симметрична относительно оси ординат, математическое ожидание, мода и медиана совпадают.

§ 3.7. Меры связи случайных величин

3.7.1. Случайные векторы и условные законы распределения

Случайный вектор (или *многомерная случайная величина*)

$$\mathbf{X} = \left(X^{(1)}, X^{(1)}, \ldots, X^{(k)}\right) —$$

это совокупность k случайных величин $X^{(i)}$ ($i = 1, 2, \ldots, k$), заданных на одном и том же вероятностном пространстве $\{\Omega, \mathcal{S}, \mathbf{P}\}$.

Закон распределения вероятностей случайного вектора задается его *функцией распределения*

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X^{(1)}, X^{(2)}, \ldots, X^{(k)}}\left(x^{(1)}, x^{(2)}, \ldots, x^{(k)}\right) = \mathbf{P}\left\{\left(X^{(1)} \leqslant x^{(1)}\right) \cap \left(X^{(2)} \leqslant x^{(2)}\right) \cap \cdots \cap \left(X^{(k)} \leqslant x^{(k)}\right)\right\},$$

которая является числовой функцией k переменных.

Многомерные случайные величины, так же, как и одномерные, могут быть дискретными (когда наборы возможных значений образуют конечное или счетное множество), абсолютно непрерывными (когда функция распределения представима в виде интеграла

$$F_{X^{(1)}, X^{(2)}, \ldots, X^{(k)}}\left(x^{(1)}, x^{(2)}, \ldots, x^{(k)}\right) =$$

$$= \int_{-\infty}^{x^{(1)}} \int_{-\infty}^{x^{(2)}} \cdots \int_{-\infty}^{x^{(k)}} f_{X^{(1)}, X^{(2)}, \ldots, X^{(k)}}\left(z^{(1)}, z^{(2)}, \ldots, z^{(k)}\right) dz^{(1)} dz^{(2)} \cdots dz^{(k)}$$

от своей частной производной по всем аргументам) или смешанными.

Двумерный дискретный случайный вектор удобно задавать с помощью *таблицы распределения*

X \ Y	y_1	y_2	$\cdots$	y_n	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}	$\cdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
x_m	p_{m1}	p_{m2}	$\cdots$	p_{mn}	$\cdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	

(3.7.1)

В такой таблице названия строк x_j соответствуют всем возможным значениям первой компоненты X, а заголовки столбцов y_i — всем возможным значениям второй компоненты Y. При этом в клетку, находящуюся в i-й строке и в j-м столбце, записывается значение вероятности $p_{ij} = \mathbf{P}\{(X = x_i) \cap (Y = y_j)\}$. Естественно, все $p_{ij} \geqslant 0$ и

$$\sum_i \sum_j p_{ij} = 1.$$

Законы распределения каждой из компонент такого двумерного дискретного случайного вектора восстанавливаются по таблице распределения (3.7.1) при помощи формул

$$\mathbf{P}\{X=x_i\}=\sum_j p_{ij},\quad \mathbf{P}\{Y=y_j\}=\sum_i p_{ij}. \tag{3.7.2}$$

Пусть с помощью таблицы распределения (3.1.1) задан дискретный двумерный случайный вектор $(X;Y)$. Условная вероятность события $\{Y=y_j\}$ при условии $\{X=x_i\}$ вычисляется, согласно определению условной вероятности (2.3.1), в соответствии с формулой

$$\mathbf{P}\{Y=y_j\,|\,X=x_i\}=\frac{\mathbf{P}\{(X=x_i)\cap(Y=y_j)\}}{\mathbf{P}\{X=x_i\}}=\frac{p_{ij}}{\sum\limits_k p_{ik}}.$$

Таким образом, можно получить *условный закон распределения* дискретной случайной величины Y при условии $\{X=x_i\}$, он будет задаваться рядом распределения

$Y\,\|\,X=x_i$	y_1	y_2	$\cdots$	y_n	$\cdots$
p	$\dfrac{p_{i1}}{\sum\limits_j p_{ij}}$	$\dfrac{p_{i2}}{\sum\limits_j p_{ij}}$	$\ldots$	$\dfrac{p_{in}}{\sum\limits_j p_{ij}}$	$\ldots$

(3.7.3)

Математическое ожидание случайной величины (3.7.3) называется *условным математическим ожиданием* $\mathbf{E}(Y|X=x_i)$ дискретной случайной величины Y при условии $\{X=x_i\}$:

$$\mathbf{E}(Y\,|\,X=x_i)=\sum_j y_j\mathbf{P}\{Y=y_j\,|\,X=x_i\}=\frac{\sum\limits_j y_j p_{ij}}{\sum\limits_j p_{ij}}$$

Для дискретных случайных величин X и Y условная вероятность $\mathbf{P}\{Y=y_j|X\}$ события $\{Y=y_j\}$ относительно случайной величины X определяется как случайная величина, заданная рядом распределения

$\mathbf{P}\{Y=y_j\,\|\,X\}$	$\dfrac{p_{1j}}{\sum\limits_k p_{1k}}$	$\dfrac{p_{2j}}{\sum\limits_k p_{2k}}$	$\cdots$	$\dfrac{p_{mj}}{\sum\limits_k p_{mk}}$	$\cdots$
p	$\sum\limits_k p_{1k}$	$\sum\limits_k p_{2k}$	$\cdots$	$\sum\limits_k p_{mk}$	$\cdots$

а *условное математическое ожидание* $\mathbf{E}(Y|X)$ случайной величины Y относительно случайной величины X определяется как случайная величина с рядом распределения

$\mathbf{E}(Y\,\|\,X)$	$\dfrac{\sum\limits_j y_j p_{1j}}{\sum\limits_j p_{1j}}$	$\dfrac{\sum\limits_j y_j p_{2j}}{\sum\limits_j p_{2j}}$	$\ldots$	$\dfrac{\sum\limits_j y_j p_{mj}}{\sum\limits_j p_{mj}}$	$\cdots$
p	$\sum\limits_j p_{1j}$	$\sum\limits_j p_{2j}$	$\cdots$	$\sum\limits_j p_{mj}$	$\cdots$

Здесь мы учли, что $\mathbf{E}(Y|X)$ является функцией случайной величины X, которая принимает значения

$$\mathbf{E}(Y \mid X = x_i) = \frac{\sum\limits_j y_j p_{ij}}{\sum\limits_j p_{ij}}$$

при $X = x_i$ ($i = 1, 2, \ldots, m, \ldots$).

При этом

$$\mathbf{E}\big(\mathbf{P}\{Y = y_j \mid X\}\big) = \sum_i \left(\frac{p_{ij}}{\sum\limits_k p_{ik}} \sum_k p_{ik} \right) = \sum_i p_{ij} = \mathbf{P}\{Y = y_j\},$$

$$\mathbf{E}\big(\mathbf{E}(Y \mid X)\big) = \sum_i \left(\frac{\sum\limits_j y_j p_{ij}}{\sum\limits_j p_{ij}} \sum_j p_{ij} \right) = \sum_i \sum_j y_j p_{ij} = \sum_j y_j \sum_i p_{ij} = \tag{3.7.4}$$

$$= \sum_j y_j \mathbf{P}\{Y = y_j\} = \mathbf{E}(Y).$$

Формула (3.7.4) называется *формулой полного математического ожидания.*

Замечание. В случае непрерывных случайных величин условные законы распределения и условные математические ожидания определяются аналогично (с заменой сумм интегралами).

Условная дисперсия определяется формулой

$$\mathbf{Var}(Y|X) = \mathbf{E}\{[Y - \mathbf{E}(Y|X)]^2 | X\},$$

при этом справедливо

ДИСПЕРСИОННОЕ ТОЖДЕСТВО:

$$\mathbf{Var}(Y) = \mathbf{E}[\mathbf{Var}(Y|X)] + \mathbf{Var}[\mathbf{E}(Y|X)]. \tag{3.7.5}$$

ПРИМЕР 3.7.1. Рассмотрим дискретный случайный вектор $(X; Y)$, заданный следующей таблицей распределения вероятностей:

X \ Y	0	1
-1	0	0,2
0	0,1	0,2
1	0,4	0,1

Составим ряды распределения его компонент X и Y.

Случайная величина X принимает значения -1; 0 и 1 с вероятностями $0{,}2 = 0 + 0{,}2$; $0{,}3 = 0{,}1 + 0{,}2$ и $0{,}5 = 0{,}4 + 0{,}1$ соответственно. Таким образом, эта случайная величина имеет ряд распределения

X	-1	0	1
p	0,2	0,3	0,5

Вычислим ее математическое ожидание:

$$\mathbf{E}(X) = -1 \cdot 0{,}2 + 0 \cdot 0{,}3 + 1 \cdot 0{,}5 = -0{,}2 + 0 + 0{,}5 = 0{,}3.$$

Аналогично получаем ряд распределения случайной величины Y:

Y	0	1
p	0,5	0,5

Математическое ожидание $\mathbf{E}(Y) = 0{\cdot}0{,}5 + 1{\cdot}0{,}5 = 0 + 0{,}5 = 0{,}5$.

Составим ряды распределения случайной величины XY. Вначале выпишем вероятности для всех вариантов произведений X и Y:

XY	$-1\cdot 0$	$0\cdot 0$	$1\cdot 0$	$-1\cdot 1$	$0\cdot 1$	$1\cdot 1$
p	0	0,1	0,4	0,2	0,2	0,1

Теперь сгруппируем одинаковые значения и упорядочим по возрастанию:

XY	-1	0	1
p	0,2	0,7	0,1

Математическое ожидание

$$\mathbf{E}(XY) = -1{\cdot}0{,}2 + 0{\cdot}0{,}7 + 1{\cdot}0{,}1 = -0{,}2 + 0 + 0{,}1 = -0{,}1.$$

Теперь найдем, например, вероятность $\mathbf{P}\{X < Y\}$:

$$\mathbf{P}\{X < Y\} = \mathbf{P}\{(X = -1) \cap (Y = 0)\} + \mathbf{P}\{(X = -1) \cap (Y = 1)\} + \mathbf{P}\{(X = 0) \cap (Y = 1)\} = \\ = 0 + 0{,}2 + 0{,}2 = 0{,}4. \ \square$$

Найдем условный закон распределения случайной величины X при условии $Y = 0$, для этого воспользуемся формулой условной вероятности:

$$\mathbf{P}\{X = x_i \mid Y = 0\} = \frac{\mathbf{P}\{(X = x_i) \cap (Y = 0)\}}{\mathbf{P}\{Y = 0\}},$$

т. е.

$$\mathbf{P}\{X = -1 \mid Y = 0\} = \frac{\mathbf{P}\{(X = -1) \cap (Y = 0)\}}{\mathbf{P}\{Y = 0\}} = \frac{0}{0{,}5} = 0,$$

$$\mathbf{P}\{X = 0 \mid Y = 0\} = \frac{\mathbf{P}\{(X = 0) \cap (Y = 0)\}}{\mathbf{P}\{Y = 0\}} = \frac{0{,}1}{0{,}5} = \frac{1}{5} = 0{,}2,$$

$$\mathbf{P}\{X = 1 \mid Y = 0\} = \frac{\mathbf{P}\{(X = 1) \cap (Y = 0)\}}{\mathbf{P}\{Y = 0\}} = \frac{0{,}4}{0{,}5} = \frac{4}{5} = 0{,}8.$$

Таким образом, условный закон распределения случайной величины X при условии $Y = 0$ таков:

$X \mid Y = 0$	0	1
p	0,2	0,8

$\square$

Пример 3.7.2 (Тождество Вальда). Пусть X_1, X_2, ... — независимые случайные величины с одинаковыми математическими ожиданиями: $\mathbf{E}(X_i) = a$, $i = 1, 2, \ldots$ Пусть N — случайная величина, независимая с X_1, X_2, ..., а случайная величина Y представляет собой сумму со случайным числом N слагаемых X_i:

$$Y = X_1 + X_2 + \cdots + X_N.$$

Найдем $\mathbf{E}(Y)$. При конкретном k $\quad \mathbf{E}(Y \mid N = k) = \mathbf{E}\left(\sum_{i=1}^{k} X_i\right) = ka$, $\quad$ поэтому

$$\mathbf{E}(Y) = \mathbf{E}\big(\mathbf{E}(Y \mid N = k)\big) = \sum_{i=1}^{k} ka\mathbf{P}\{N = k\} = a\sum_{i=1}^{k} k\mathbf{P}\{N = k\} = a\mathbf{E}(N).$$

Случайные величины X_1, X_2, ... в тождестве Вальда могут иметь смысл объемов выданных кредитов, а случайная величина N — числа кредитов, выданных, например, за неделю. Тогда $Y = X_1 + X_2 + \cdots + X_n$ будет суммарным объемом кредитов, выданных за неделю. $\square$

Замечание. Интересно, что тождество Вальда справедливо и в случае зависимых слагаемых!

Задачи

264. Задано распределение вероятностей дискретного двумерного случайного вектора $(X; Y)$:

X \ Y	1	2
-1	0,15	0,1
0	0,3	0,05
1	0,3	0,1

Здесь случайная величина X описывает доход инвестиционной компании на рынке акций, а случайная величина Y — доход на рынке облигаций. Составить ряды распределения компонент X и Y, а также условный закон распределения компоненты X при условии $Y = 2$. Выясните, зависимы ли маргинальные случайные величины X и Y. Найдите закон распределения суммарного дохода компании $X + Y$.

265. В условиях задачи **264** найдите условное математическое ожидание случайной величины X относительно случайной величины Y.

266. Докажите справедливость дисперсионного тождества (3.7.5).

267. Докажите, что если случайные величины X и Y независимы, то $\mathbf{E}(Y|X) = \mathrm{E}(Y)$, $\mathbf{E}(X|Y) = \mathbf{E}(X)$.

3.7.2. Ковариация и коэффициент корреляции

Напомним [см. определение (3.1.8)], что две случайные величины X и Y называются *независимыми*, если для всех $x,\ y \in \mathbb{R}$

$$\mathbf{P}\{(X \leqslant x) \cap (Y \leqslant y)\} = \mathbf{P}\{X \leqslant x\}\mathbf{P}\{Y \leqslant y\},$$

т. е. если для всех $x,\ y \in \mathbb{R}$ события $\{X \leqslant x\}$ и $\{Y \leqslant y\}$ независимы.

Для измерения «степени зависимости» случайных величин вводится *ковариация* случайных величин X и Y:

$$\operatorname{cov}(X, Y) = \mathbf{E}\{[X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\}. \tag{3.7.6}$$

Последняя формула легко преобразуется к виду

$$\operatorname{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y). \tag{3.7.7}$$

Доказательство. По определению ковариации и свойствам математического ожидания $\operatorname{cov}(X, Y) = \mathbf{E}\{[X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\} = \mathbf{E}[XY - Y\mathbf{E}(X) - X\mathbf{E}(Y) - \mathbf{E}(X)\mathbf{E}(Y)] = \mathbf{E}(XY) - \mathbf{E}(Y)\mathbf{E}(X) - \mathbf{E}(X)\mathbf{E}(Y) + \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$. ❑

Пример 3.7.3. Вычислим ковариацию случайных величин X и Y из примера3.7.1.

Чтобы найти $\mathbf{E}(XY)$, перемножим все возможные значения x_j, y_i и соответствующих вероятностей p_{ij} из таблицы распределения данной случайной величины и произведения сложим: $\mathbf{E}(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_j y_i \mathbf{P}\{(X = x_j) \cap (Y = y_i)\} =$

$= \sum_{i=1}^{m}\sum_{j=1}^{n} x_j y_i p_{ij} = 0\cdot(-1)\cdot 0 + 0\cdot 0\cdot 0{,}1 + 0\cdot 1\cdot 0{,}4 + 1\cdot(-1)\cdot 0{,}2 + 1\cdot 0\cdot 0{,}2 + 1\cdot 1\cdot 0{,}1 = 0 + 0 + 0 - 0{,}2 + 0 + 0{,}1 = -0{,}1$. Значения $\mathbf{E}(X)$ и $\mathbf{E}(Y)$ найдем по рядам распределения случайных величин X и Y, полученным в примере 3.7.1: $\mathbf{E}(X) = -1\cdot 0{,}2 +$

$+ 0{\cdot}0{,}3 + 1{\cdot}0{,}5 = 0{,}3$, $\mathbf{E}(Y) = 0{\cdot}0{,}5 + 1{\cdot}0{,}5 = 0{,}5$. Поэтому $\operatorname{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = -0{,}1 - 0{,}3{\cdot}0{,}5 = -0{,}25$. ❑

Ковариация случайных величин обладает следующими **свойствами.**

Коммутативность ковариации. *Для любых случайных величин X и Y*

$$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X). \tag{3.7.8}$$

Доказательство. Имеем: $\operatorname{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(YX) - \mathbf{E}(Y)\mathbf{E}(X) = \operatorname{cov}(Y, X)$, что и требовалось доказать. ❑

Связь ковариации и дисперсии. *Для любой случайной величины X*

$$\operatorname{cov}(X, X) = \mathbf{Var}(X). \tag{3.7.9}$$

Доказательство. Действительно, $\operatorname{cov}(X, X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 = \mathbf{Var}(X)$, что и требовалось доказать. ❑

Линейность ковариации по каждому аргументу. *Для любых случайных величин X и Y и любых чисел $a,\ b \in \mathbb{R}$*

$$\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y). \tag{3.7.10}$$

Доказательство. Действительно, для любых случайных величин X и Y и любых чисел $a,\ b \in \mathbb{R}$

$$\operatorname{cov}(aX, bY) = \mathbf{E}\{[aX - \mathbf{E}(aX)][bY - \mathbf{E}(bY)]\} = \mathbf{E}\{[aX - a\mathbf{E}(X)][bY - b\mathbf{E}(Y)]\} =$$
$$= \mathbf{E}\{a[X - \mathbf{E}(X)]b[Y - \mathbf{E}(Y)]\} = \mathbf{E}\{ab[X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\} = ab\mathbf{E}\{[X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\} =$$
$$= ab \operatorname{cov}(X, Y),$$

что и требовалось доказать. ❑

Аддитивность ковариации по каждому агрументу. *Для любых случайных величин X, Y и Z*

$$\operatorname{cov}(X + Y, Z) = \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z), \tag{3.7.11}$$
$$\operatorname{cov}(X,\ Y + Z) = \operatorname{cov}(X, Y) + \operatorname{cov}(X, Z).$$

Доказательство. Имеем:

$$\operatorname{cov}(X + Y, Z) = \mathbf{E}\{[(X + Y) - \mathbf{E}(X + Y)][Z - \mathbf{E}(Z)]\} = \mathbf{E}\{[X + Y - \mathbf{E}(X) - \mathbf{E}(Y)][Z - \mathbf{E}(Z)]\} =$$

$$= \mathbf{E}\{[X - \mathbf{E}(X) + Y - \mathbf{E}(Y)][Z - \mathbf{E}(Z)]\} = \mathbf{E}\{[X - \mathbf{E}(X)])][Z - \mathbf{E}(Z)] + [Y - \mathbf{E}(Y)][Z - \mathbf{E}(Z)]\} =$$
$$= \mathbf{E}\{[X - \mathbf{E}(X)])][Z - \mathbf{E}(Z)]\} + \mathbf{E}\{[Y - \mathbf{E}(Y)][Z - \mathbf{E}(Z)]\} = \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z),$$

что и требовалось доказать. Справедливость второй формулы следует из коммутативности ковариации. ❑

Формула для ковариации независимых случайных величин. *Для независимых случайных величин X и Y*

$$\operatorname{cov}(X, Y) = 0. \tag{3.7.12}$$

Доказательство. Так как по формуле (3.2.11) для независимых случайных величин X и Y $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$, то для независимых случайных величин X и Y $\operatorname{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(X)\mathbf{E}(Y) - \mathbf{E}(X)\mathbf{E}(Y) = 0$, что и требовалось доказать. ❑

Для случайных величин X и Y, имеющих тенденцию изменяться одновременно в одну и ту же сторону, $\operatorname{cov}(X, Y) > 0$; для случайных величин X и Y, имеющих тенденцию изменяться одновременно в разные стороны, $\operatorname{cov}(X, Y) < 0$.

Формула для дисперсии суммы произвольных случайных величин. *Дисперсия суммы произвольных (зависимых или независимых) случайных величин X и Y рассчитывается по формуле*

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\operatorname{cov}(X, Y). \tag{3.7.13}$$

Доказательство. По определению дисперсии (3.2.13) и свойствам (3.2.8), (3.2.10) математического ожидания имеем: $\mathbf{Var}(X+Y)=\mathbf{E}\{[(X+Y)-\mathbf{E}(X+Y)]^2\}=\mathbf{E}(\{[X-\mathbf{E}(X)]+(Y-\mathbf{E}(Y)]\}^2)=$ $=\mathbf{E}\{[X-\mathbf{E}(X)]^2\}+2\mathbf{E}\{[X-\mathbf{E}(X)][Y-\mathbf{E}(Y)]\}+\mathbf{E}\{[Y-\mathbf{E}(Y)]^2\}=\mathbf{Var}(X)+\mathbf{Var}(Y)+2\mathrm{cov}(X,Y)$, что и требовалось доказать. ❑

В случае, когда суммируется n случайных величин,

$$\mathbf{Var}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathbf{Var}(X_i)+\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathrm{cov}(X_i,X_j)$$

или, учитывая, что $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$ для всех $i = 1, 2, ..., n$, $j = 1, 2, ..., n$,

$$\mathbf{Var}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathbf{Var}(X_i)+2\sum_{i=1}^{n}\sum_{j=1}^{i-1}\mathrm{cov}(X_i,X_j)\,. \qquad (3.7.14)$$

Эту формулу легко доказать методом математической индукции, что оставляем читателю в задаче **273**.

Теперь мы можем доказать формулу (3.3.20) для дисперсии случайной величины X, распределенной по гипергеометрическому закону.

Доказательство формулы (3.3.20). Как и при доказательстве формулы (3.3.19) для математического ожидания гипергеометрического распределения, рассмотрим последовательность из l (з а в и с и м ы х) испытаний. В каждом испытании из партии (состоящей из L изделий, K из которых первого сорта, а остальные ($L - K$) — второго сорта), извлекается по одному изделию. Пусть события A_i состоят в том, что в i-м из l таких испытаний произошел успех ($i = 0, 1, 2, ..., l$). Вероятность каждого из этих событий. Рассмотрим l случайных величин $X_i = I_{A_i}$ — индикаторов событий A_i. При этом по формуле (3.2.19) $\mathbf{Var}(X_i)=\mathbf{Var}(I_{A_i})=$ $=\mathbf{P}(A_i)[1-\mathbf{P}(A_i)]=\dfrac{K}{L}\left(1-\dfrac{K}{L}\right)$, а случайная величина X, очевидно, равна $X=\sum_{i=1}^{l}X_i$.

Теперь достаточно заметить, что при $i \neq j$ случайная величина $X_iX_j = I_{A_i}I_{A_j} = I_{A_i\cap A_j}$ (так как эта случайная величина принимает значение 1 в том и только в том случае, когда одновременно наступили и событие A_i, и событие A_j). При этом

$$\mathrm{cov}(X_i,X_j)=\mathbf{E}(X_iX_j)-\mathbf{E}(X_i)\mathbf{E}(X_j)=\mathbf{E}\left(I_{A_i}I_{A_j}\right)-\mathbf{E}\left(I_{A_i}\right)\mathbf{E}\left(I_{A_j}\right)=\mathbf{P}\{A_i\cap A_j\}-\mathbf{P}\{A_i\}\mathbf{P}\{A_j\}=$$

$$=\frac{K}{L}\frac{K-1}{L-1}-\left(\frac{K}{L}\right)^2=\frac{K}{L}\left(\frac{K-1}{L-1}-\frac{K}{L}\right)=\frac{K}{L}\frac{L(K-1)-K(L-1)}{L(L-1)}=\frac{K}{L}\frac{LK-L-KL+K}{L(L-1)}=\frac{K}{L}\frac{K-L}{L(L-1)},$$

значит,

$$\mathbf{Var}(X)=\mathbf{Var}\left(\sum_{i=1}^{l}X_i\right)=\sum_{i=1}^{l}\mathbf{Var}(X_i)+\sum_{i=1}^{l}\sum_{\substack{j=1\\ j\neq i}}^{l}\mathrm{cov}(X_i,X_j)=\sum_{i=1}^{l}\left[\frac{K}{L}\left(1-\frac{K}{L}\right)\right]+\sum_{i=1}^{l}\sum_{\substack{j=1\\ j\neq i}}^{l}\left[\frac{K}{L}\frac{K-L}{L(L-1)}\right]=$$

$$=l\frac{K}{L}\frac{L-K}{L}+l(l-1)\frac{K}{L}\frac{K-L}{L(L-1)}=l\frac{K}{L}\frac{(L-K)(L-1)+(l-1)(K-L)}{L(L-1)}=$$

$$=l\frac{K}{L}\frac{(L-K)[L-1-(l-1)]}{L(L-1)}=l\frac{K}{L}\frac{L-K}{L}\frac{L-l}{L-1}=l\frac{K}{L}\left(1-\frac{K}{L}\right)\frac{1-\dfrac{l}{L}}{1-\dfrac{1}{L}},$$

что и доказывает формулу (3.3.20) для дисперсии случайной величины, распределенной по гипергеометрическому закону. ❑

Ковариация может принимать произвольные вещественные значения, поэтому не вполне пригодна к использованию в качестве меры связи случайных величин. Для этого лучше подходит *коэффициент корреляции* случайных величин X и Y

$$R(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\mathbf{E}\{[X-\mathbf{E}(X)][Y-\mathbf{E}(Y)]\}}{\sigma_X\sigma_Y}=\frac{\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)}{\sigma_X\sigma_Y}. \qquad (3.7.15)$$

Пример 3.7.4. Вычислим коэффициент корреляции случайных величин X и Y из примеров 3.7.1, 3.7.3.

Зная из примера 3.7.3 значения математических ожиданий $\mathbf{E}(X) = 0{,}3$ и $\mathbf{E}(Y) = 0{,}5$, найдем дисперсии:

$$\begin{gathered}\mathbf{Var}(X) = (-1 - 0{,}3)^2 \cdot 0{,}2 + (0 - 0{,}3)^2 \cdot 0{,}3 + (1 - 0{,}3)^2 \cdot 0{,}5 = \\ = (-1{,}3)^2 \cdot 0{,}2 + (-0{,}3)^2 \cdot 0{,}3 + 0{,}7^2 \cdot 0{,}5 = \\ = 1{,}69 \cdot 0{,}2 + 0{,}09 \cdot 0{,}3 + 0{,}49 \cdot 0{,}5 = 0{,}338 + 0{,}027 + 0{,}245 = 0{,}61;\end{gathered}$$

$$\mathbf{Var}(Y) = (0 - 0{,}5)^2 \cdot 0{,}5 + (1 - 0{,}5)^2 \cdot 0{,}5 = 0{,}5^3 + 0{,}5^3 = 2 \cdot 0{,}5^3 = 0{,}5^2 = 0{,}25.$$

Отсюда

$$\sigma_X = \sqrt{\mathbf{Var}(X)} = \sqrt{0{,}61} \approx 0{,}781, \ \sigma_Y = \sqrt{\mathbf{Var}(Y)} = \sqrt{0{,}25} = 0{,}5.$$

Ковариация $\operatorname{cov}(X, Y) = -0{,}25$ нам также известна из примера 3.7.3, поэтому окончательно получаем:

$$R(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \approx \frac{-0{,}25}{0{,}781 \cdot 0{,}5} \approx -0{,}64. \ \square$$

Коэффициент корреляции обладает следующими **свойствами.**

Инвариантность коэффициента корреляции относительно линейных преобразований аргументов. *Для любых случайных величин X и Y и любых чисел a, b, c, $d \in \mathbb{R}$*

$$R(aX + b, cY + d) = R(X, Y). \tag{3.7.16}$$

Доказательство. Действительно, по свойствам ковариации и дисперсии случайных величин получаем: $R(aX+b, cY+d) = \dfrac{\operatorname{cov}(aX+b, cY+d)}{\sigma_{aX+b}\sigma_{cY+d}} = \dfrac{ac\operatorname{cov}(X,Y)}{ac\sigma_X\sigma_Y} = R(X,Y)$, что и требовалось доказать. $\square$

Из формулы (3.7.16) следует, в частности, что если

$$\overset{\circ}{X} = \frac{X - \mathbf{E}(X)}{\sigma_X}, \qquad \overset{\circ}{Y} = \frac{Y - \mathbf{E}(Y)}{\sigma_Y}$$

(где X и Y — некоторые случайные величины), то

$$R\left(\overset{\circ}{X}, \overset{\circ}{Y}\right) = R(X, Y). \tag{3.7.17}$$

При этом $\mathbf{E}\left(\overset{\circ}{X}\right) = \mathbf{E}\left(\overset{\circ}{Y}\right) = 0$, $\sigma_{\overset{\circ}{X}} = \sigma_{\overset{\circ}{Y}} = 1$, поэтому

$$\operatorname{cov}\left(\overset{\circ}{X}, \overset{\circ}{Y}\right) = \sigma_{\overset{\circ}{X}} \sigma_{\overset{\circ}{Y}} \rho\left(\overset{\circ}{X}, \overset{\circ}{Y}\right) = R\left(\overset{\circ}{X}, \overset{\circ}{Y}\right) = R(X, Y). \tag{3.7.18}$$

Нормированность коэффициента корреляции. *Для любых случайных величин X и Y*

$$-1 \leqslant R(X, Y) \leqslant 1. \tag{3.7.19}$$

Доказательство. Рассмотрим две случайные величины $Z^{\pm} = \sigma_Y X \pm \sigma_X Y$. По формуле 3.7.13

$$\begin{gathered}\mathbf{Var}(Z^{\pm}) = \mathbf{Var}(\sigma_Y X \pm \sigma_X Y) = \mathbf{Var}(\sigma_Y X) + \mathbf{Var}(\pm\sigma_X Y) + 2\operatorname{cov}(\sigma_Y X, \pm\sigma_X Y) = \\ = \sigma_Y^2 \mathbf{Var}(X) + \sigma_X^2 \mathbf{Var}(Y) \pm 2\sigma_Y \sigma_X \operatorname{cov}(X, Y) = \sigma_Y^2 \sigma_X^2 + \sigma_X^2 \sigma_Y^2 \pm 2\sigma_X^2 \sigma_Y^2 \frac{\operatorname{cov}(X,Y)}{\sigma_Y \sigma_X} = \\ = 2\sigma_X^2 \sigma_Y^2 (1 \pm R(X, Y)).\end{gathered}$$

Из неотрицательности дисперсий случайных величин Z^+ и Z^- следует, что

$$1+R(X,Y)=\frac{\mathbf{Var}(Z^+)}{2\sigma_X^2\sigma_Y^2}\geqslant 0;\quad 1-R(X,Y)=\frac{\mathbf{Var}(Z^-)}{2\sigma_X^2\sigma_Y^2}\geqslant 0, \tag{3.7.20}$$

откуда

$$R(X,Y)\geqslant -1;\quad R(X,Y)\leqslant 1,$$

что и требовалось доказать. ❑

Формула для коэффициента корреляции независимых случайных величин. *Для независимых случайных величин X и Y*

$$R(X,\,Y)=0; \tag{3.7.21}$$

Доказательство. По формуле (3.7.9) $\operatorname{cov}(X,\,Y)=0$, значит, $R(X,Y)=\dfrac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=0$, что и требовалось доказать. ❑

Формула для коэффициента корреляции линейно связанных случайных величин. *Для линейно связанных случайных величин X и $Y=aX+b$ ($a,\ b\in\mathbb{R},\ a\ \ 0$)*

$$|R(X,\,Y)|=1, \tag{3.7.22}$$

если для каких-либо случайных величин X и Y выполняется (3.7.22), *то существуют такие числа $a,\ b\in\mathbb{R},\ a\neq 0$, что $\mathbf{P}\{Y=aX+b\}=1$.*

Доказательство. По свойствам математического ожидания и дисперсии

$$\rho(X,Y)=\frac{\mathbf{E}[X(aX+b)]-\mathbf{E}(X)\mathbf{E}(aX+b)}{\sigma_X\sigma_{aX+b}}=\frac{a[\mathbf{E}(X^2)]+b\mathbf{E}(X)-a[\mathbf{E}(X)]^2-b\mathbf{E}(X)}{\sqrt{\mathbf{D}X}\ \sqrt{\mathbf{M}[(aX+b)^2]-(a\mathbf{M}X+b)^2}}=$$

$$=\frac{a\mathbf{Var}(X)+b[\mathbf{E}(X)-\mathbf{E}(X)]}{\sqrt{\mathbf{Var}(X)}\ \sqrt{a^2\mathbf{E}(X^2)+2ab\mathbf{E}(X)+b^2-a^2[\mathbf{E}(X)]^2-2ab\mathbf{E}(X)-b^2}}=\frac{a\mathbf{Var}(X)}{|a|\cdot|\mathbf{Var}(X)|}=\begin{cases}-1, & a<0,\\ 1 & a>0,\end{cases}$$

поэтому $|R(X,\,Y)|=1$, что доказывает первую часть утверждения.

Докажем вторую часть. Пусть $R(X,\,Y)=+1$, Это возможно только если второе неравенство в формуле (3.7.20) превращается в равенство:

$$1-R(X,Y)=\frac{\mathbf{Var}(Z^-)}{2\sigma_X^2\sigma_Y^2}=0, \tag{3.7.23}$$

откуда следует, что $\mathbf{Var}(Z^-)=0$. Но по свойству (3.7.21) если для случайной величины Z^- выполняется равенство $\mathbf{Var}(Z^-)=0$, то существует такое число $c=\mathbf{E}(Z^-)$, что $\mathbf{P}\{Z^-=c\}=1$.

Таким образом, $\mathbf{P}\{\sigma_Y X-\sigma_X Y=c\}=1$ или $\mathbf{P}\{\sigma_X Y=\sigma_Y X-c\}=1$ или $\mathbf{P}\{Y=aX+b\}=1$, где $a=\dfrac{\sigma_Y}{\sigma_X},\ b=-\dfrac{c}{\sigma_X}$, что и требовалось доказать. ❑

Пример 3.7.5. Рассмотрим две случайные величины X и Y — число появлений событий простейшего потока в интервалах времени $(0;\,t)$ и $(0;\,t+\tau)$ соответственно. Найдем $R(X,\,Y)$.

Пусть случайная величина Z — число появлений событий простейшего потока в интервале $(t;\,t+\tau)$. Тогда $X=\mathrm{Pois}(\lambda t)$, $Y=\mathrm{Pois}(\lambda(t+\tau))$, $Z=\mathrm{Pois}(\lambda\tau)$, $\mathbf{E}(X)=\mathbf{Var}(X)=\lambda t$, $\mathbf{E}(Y)=\mathbf{Var}(Y)=\lambda(t+\tau)$, $\mathbf{E}(Z)=\mathbf{Var}(Z)=\lambda\tau$. Поэтому $\mathbf{E}(X^2)=$ $=\mathbf{Var}(X)+[\mathbf{E}(X)]^2=\lambda t+(\lambda t)^2$.

Учтем, что в силу свойства отсутствия последействия простейшего потока случайные величины X, Y и Z независимы, поэтому $\mathbf{E}(XZ)=\mathbf{E}(X)\mathbf{E}(Z)=\lambda^2 t\tau$. Окончательно получаем:

$$R(X,Y)=\frac{\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)}{\sigma_X\sigma_Y}=\frac{\mathbf{E}[X(X+Z)]-\mathbf{E}(X)\mathbf{E}(Y)}{\sigma_X\sigma_Y}=$$

$$=\frac{\mathbf{E}(X^2)+\mathbf{E}(XZ)-\mathbf{E}(X)\mathbf{E}(Y)}{\sqrt{\mathbf{Var}(X)}\sqrt{\mathbf{Var}(Y)}}=\frac{\mathbf{E}(X^2)+\mathbf{E}(X)\mathbf{E}(Z)-\mathbf{E}(X)\mathbf{E}(Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}=$$

$$=\frac{\lambda t+(\lambda t)^2+\lambda t\lambda\tau-\lambda t\lambda(t+\tau)}{\sqrt{\lambda t\lambda(t+\tau)}}=\frac{\lambda t}{\sqrt{\lambda^2 t(t+\tau)}}=\sqrt{\frac{t}{t+\tau}}.$$

Таким образом, при $t=\ \tau\ \ R(X, Y)$ близок к нулю (что мы понимаем как свидетельство того, что связь между числом появлений событий простейшего потока в интервалах времени $(0; t)$ и $(0; t+\ )$ если и есть, то далека от линейной), а при t ? τ — к единице (что понимается как свидетельство того, что число появлений событий простейшего потока в интервалах времени $(0; t)$ и $(0; t+\tau)$ связаны линейной функциональной зависимостью, что, впрочем, очевидно). ❑

Если коэффициент корреляции $\rho(X, Y) = 0$, то это не обязательно означает независимость случайных величин X и Y. В этом случае говорят, что данные случайные величины *некоррелированы*. **Из независимости следует некоррелированность, но наоборот — не всегда!**

Пример 3.7.6. Рассмотрим случайную величину X, заданную рядом распределения

X	-1	0	1
p	$1/3$	$1/3$	$1/3$

и случайную величину $Y = X^2$.

Составим ряды распределения случайных величин Y:

Y	0	1
p	$1/3$	$2/3$

и $XY = X^3$:

XY	-1	0	1
p	$1/3$	$1/3$	$1/3$

Найдем математические ожидания случайных величин X, Y и XY:

$$\mathbf{E}(X)=-1\cdot\frac{1}{3}+0\cdot\frac{1}{3}+1\cdot\frac{1}{3}=0;\ \mathbf{E}(Y)=0\cdot\frac{1}{3}+1\cdot\frac{2}{3}=\frac{2}{3};\ \mathbf{E}(XY)=-1\cdot\frac{1}{3}+0\cdot\frac{1}{3}+1\cdot\frac{1}{3}=0.$$

Отсюда

$$\operatorname{cov}(X,Y)=\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)=0-0\cdot\frac{2}{3}=0,\ R(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_y}=\frac{0}{\sigma_X\sigma_y}=0,$$

т. е. данные случайные величины некоррелированы. Однако они зависимы, так как, например,

$$\mathbf{P}\{(X=1)\cap(Y=0)\}=\mathbf{P}\{(X=1)\cap(X^2=0)\}=0\neq\frac{1}{3}\cdot\frac{1}{3}=\mathbf{P}\{X=1\}\mathbf{P}\{Y=0\}.\ ❑$$

Задачи

268. Задан закон распределения вероятностей двумерного дискретного случайного вектора $(X; Y)$:

X \ Y	2,3	2,7
26	0,05	0,09
30	0,12	0,30
41	0,08	0,11
50	0,04	0,21

Найдите ковариацию и коэффициент корреляции его компонент X и Y.

269. В условиях задачи **264** убедитесь в том, что $\mathbf{E}(X+Y)=\mathbf{E}(X)+\mathbf{E}(Y)$, $\mathbf{Var}(X+Y)=\mathbf{Var}(X)+\mathbf{Var}(Y)+2\mathrm{cov}(X,Y)$.

270. Петя вычислил ковариацию роста X спортсменов из университетской баскетбольной команды, измеренного в см, и скорости бега Y (тех же спортсменов), измеренной в м/с. Маша для той же совокупности баскетболистов вычислила ковариацию роста X, измеренного в м, и скорости бега Y, измеренной в м/с. В каком отношении находятся эти ковариации?

271. В условиях предыдущей задачи сравните коэффициенты корреляции, полученные Петей и Машей.

272. Средние квадратические отклонения случайных величин X и Y равны соответственно 5 и 4. Определите наибольшее возможное значение $\mathrm{cov}(X,Y)$.

273. Докажите формулу (3.7.14).

274. Пусть X, Y, Z — независимые случайные величины с конечными положительными дисперсиями. Могут ли случайные величины $X+Z$ и $Y+Z$ быть: *а*) зависимыми; *б*) независимыми?

275. Монета с вероятностью выпадения орла, равной 1/2, бросается 10 раз. Найдите коэффициент корреляции между количеством выпадений орла при первых семи бросаниях (X) и количеством выпадений орла при последних шести бросаниях (Y).

276. Студенты Иванов и Петров одновременно и независимо друг от друга сдают экзамен по теории вероятностей доценту Честнову и профессору Доброву (один студент сдает экзамен одному преподавателю, а другой — другому, причем заранее неизвестно, кто кому будет сдавать). И Иванов, и Петров знают каждый одну треть билетов на «3», одну треть на «4» и одну треть на «5». Доцент Честнов всегда ставит студенту ту оценку, которую тот заслуживает, а профессор Добров — на балл выше (если студент отвечает на «5», то Добров ставит «5»). Рассматриваются следующие случайные величины: $X_{\text{и}}$ — оценка Иванова, $X_{\text{п}}$ — оценка Петрова, $Y_{\text{ч}}$ — оценка, которую поставил Честнов, $Y_{\text{д}}$ — оценка, которую поставил Добров. Найдите коэффициенты корреляции между всеми возможными парами случайных величин.

3.7.3. Портфель финансовых инструментов

Рассмотрим ситуацию, когда в некоторый момент времени t инвестор может часть своих средств оставить на банковском счете, а другую часть потратить на приобретение ценных бумаг. *Портфелем финансовых инструментов* на (B, S)-рынке назовем вектор

$$\boldsymbol{\pi}=(x_0,x_1,x_2,\ldots,x_n),$$

где $x_0\in\mathbb{R}$ — доля капитала инвестора, вложенная в безрисковый актив (банковский счет), $x_k\in\mathbb{R}$ — доля капитала инвестора, вложенная в акцию с номером k ($k=1, 2, \ldots, n$), при этом, очевидно,

$$x_0+\sum_{k=1}^{n}x_k=1.$$

Числа x_k могут быть как положительными, так и отрицательными, в последнем случае инвестор берет средства в долг с банковского счета либо совершает короткую продажу[1] акций (конечно, эти числа могут быть и нулями).

Доходность портфеля финансовых инструментов вычисляется, очевидно, как

$$R_\pi = x_0 i + \sum_{k=1}^{n} x_k R_k ,$$

где i — доходность банковского счета (процентная ставка), R_k — доходность k-й акции.

Ожидаемой доходностью портфеля финансовых инструментов называется математическое ожидание его доходности, а *риском* портфеля — среднее квадратическое отклонение доходности.

Возникает **задача оптимизации портфеля** с целью получения максимальной доходности при минимальном риске. К сожалению, одновременно этого достичь невозможно, поэтому инвестор, оптимизирующий свой портфель, должен выбрать один из критериев:

- либо минимизировать риск $\sigma_\pi = \sqrt{\mathbf{Var}(R_\pi)}$ при заданном уровне доходности портфеля,
- либо максимизировать ожидаемую доходность $\mathbf{E}(R_\pi)$ при заданном уровне риска портфеля.

В общем случае математическая формулировка **задачи минимизации риска портфеля ценных бумаг** при заданной допустимой ожидаемой доходности $\tilde{r}_\pi$ выглядит так:

$$\sqrt{\mathbf{Var}\left(x_0 i + \sum_{k=1}^{n} x_k R_k\right)} \to \min , \qquad (3.7.24)$$

$$\mathbf{E}\left(x_0 i + \sum_{k=1}^{n} x_k R_k\right) = \tilde{r}_\pi ,$$

$$x_0 + \sum_{k=1}^{n} x_k = 1 ,$$

$$x_k \in \mathbb{R} \quad (k = 0, 1, 2, \ldots, n).$$

Воспользовавшись тем, что квадратный корень — возрастающая функция, и применив свойства математического ожидания и дисперсии, эту модель можно переписать в виде

$$\sum_{k=1}^{n}\sum_{l=1}^{n} \operatorname{cov}(R_i, R_j) x_k x_l \to \min ,$$

$$i x_0 + \sum_{k=1}^{n} \mathbf{E}(R_k) x_k = \tilde{r}_\pi , \quad x_0 + \sum_{k=1}^{n} x_k = 1 , \quad x_k \in \mathbb{R} \quad (k = 0, 1, 2, \ldots, n).$$

[1] При *короткой продаже* инвестор вначале продает некоторый актив, которого у него в наличии нет (например, в надежде, что цена на этот актив в будущем понизится), а затем (через некоторое время) приобретает этот актив на рынке и поставляет его покупателю (если при этом надежды сбылись и актив действительно упал в цене, то инвестор получает прибыль, иначе остается в убытке).

Можно поставить и другую задачу — **задачу максимизации ожидаемой доходности портфеля** при заданном максимальном допустимом риске $\tilde{\sigma}_{\pi}$:

$$\mathbf{E}\left(x_0 i + \sum_{k=1}^{n} x_k R_k\right) \to \min, \tag{3.7.25}$$

$$\sqrt{\mathbf{Var}\left(x_0 i + \sum_{k=1}^{n} x_k R_k\right)} = \tilde{\sigma}_{\pi},$$

$$x_0 + \sum_{k=1}^{n} x_k = 1,$$

$$x_k \in \mathbb{R} \quad (k = 0, 1, 2, \ldots, n).$$

Задачи (3.7.1) и (3.7.2) были поставлены и успешно решены Г. Марковицем в 1952 г. для частного случая отсутствия безрискового актива ($x_0 = 0$) и обобщены Дж. Тобином, У. Шарпом, Дж. Линтнером и др. Дж. Тобин стал лауреатом Нобелевской премии в области экономики в 1981 г., Г. Марковиц и У. Шарп стали Нобелевскими лауреатами в 1990 г.

Подробно эти задачи рассматриваются в курсах финансовой математики.[1] Мы же рассмотрим задачи управления портфелями ценных бумаг на конкретных примерах.

Пример 3.7.7. Ожидаемая доходность первого актива равна 8% со средним квадратичным отклонением 7%, ожидаемая доходность второго актива равна 11% со средним квадратичным отклонением 10%. Коэффициент корреляции между доходностями этих активов составляет 0,7.

Найдем ожидаемую доходность и риск портфеля, состоящего на 35% из первого актива и на 65% из второго.

Так как $\mathbf{E}(R_1) = 0{,}08$, $\sigma_1 = 0{,}07$, $\mathbf{E}(R_2) = 0{,}11$, $\sigma_2 = 0{,}10$, $R(R_1, R_2) = 0{,}7$, ожидаемая доходность портфеля

$$\mathbf{E}(R_{\pi}) = \mathbf{E}(x_1 R_1 + x_2 R_2) = x_1 \mathbf{E}(R_1) + x_2 \mathbf{E}(R_2) = 0{,}35 \cdot 0{,}08 + 0{,}65 \cdot 0{,}11 = 0{,}0995;$$

а риск портфеля

$$\sigma_{\pi} = \sqrt{\mathbf{Var}(x_1 R_1 + x_2 R_2)} = \sqrt{x_1^2 \mathbf{Var}(R_1) + x_2^2 \mathbf{Var}(R_1) + x_1 x_2 \operatorname{cov}(R_2, R_2)} =$$

$$= \sqrt{0{,}35^2 0{,}07^2 + 0{,}65^2 0{,}10^2 + 0{,}35 \cdot 0{,}65 \cdot 0{,}7} =$$

$$= \sqrt{0{,}35^2 0{,}07^2 + 0{,}65^2 0{,}10^2 + 0{,}35 \cdot 0{,}65 \cdot 0{,}7} = \sqrt{0{,}16407525} \approx 0{,}405.$$

Видим, что риск такого портфеля существенно выше, чем риск его составляющих! ❑

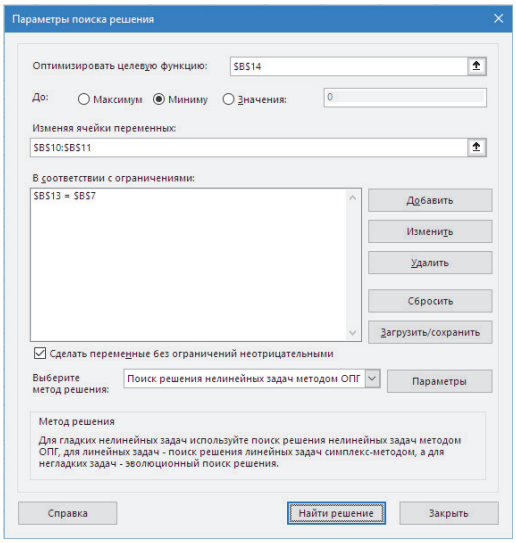
Пример 3.7.8. Выясним, с каким наименьшим риском можно достичь 20%-ной доходности инвестиций, если есть возможность банковских вложений и заимствований по ставке $i = 10\%$ годовых, а на рынке ценных бумаг обращаются две акции, их ожидаемые доходности равны соответственно $r_1 = 16\%$ и $r_2 = 23\%$, риски $\sigma_1 = 5\%$, $\sigma_2 = 14\%$, а коэффициент корреляции доходностей данных акций равен $\rho_{12} = 0{,}36$.

Введем данные в рабочий лист Microsoft Excel, как показано на рис. 3.7.1, *а*.

[1] См., например, *Соловьев В.И.* Финансовая математика. М.: КНОРУС, 20167.

	A	B
1	i =	0,10
2	r_1 =	0,16
3	r_2 =	0,23
4	sigma_1 =	0,05
5	sigma_2 =	0,14
6	R_12 =	0,36
7	r_портфеля =	0,20
8		
9	x_0 =	= 1 – B10 – B11
10	x_1 =	
11	x_2 =	
12		
13	E(r_портфеля)=	= B9*B1 + B10*B2 + B11*B3
14	Var(r_портфеля)=	= B10^2*B4^2 + B11^2*B5^2 + 2*B10*B11*B6*B4*B5

а) ввод исходных данных



б) окно «Поиск решения»

	A	B
1	i =	0,10
2	r_1 =	0,16
3	r_2 =	0,23
4	sigma_1 =	0,05
5	sigma_2 =	0,14
6	R_12 =	0,36
7	r_портфеля =	0,20
8		
9	x_0 =	–0,3908
10	x_1 =	1,1543
11	x_2 =	0,2365
12		
13	E(r_портфеля)=	0,2000
14	Var(r_портфеля)=	0,0058

в) результаты расчетов

Рис. 3.7.1. *Расчет оптимального портфеля в примере 3.7.7*

Пусть ячейки B9 и B10 соответствуют долям рисковых вложений x_1 и x_2, в ячейку B8, соответствующую доле безрисковых вложений x_0, введем формулу, соответствующую разности всех вложений (единицы) и вложений в акции x_1 и x_2, в ячейку B12 введем формулу для ожидаемой доходности портфеля $\mathbf{E}(R_\pi)$, а в ячейку B13 введем формулу для дисперсии доходности портфеля $\mathbf{Var}(R_\pi)$; учтем здесь, что $\sigma_{12} = \rho_{12}\sigma_1\sigma_2$ (эти формулы приводятся справа от соответствующих ячеек).

Воспользуемся инструментом «Поиск решения». Для этого нажмем на вкладке **«Данные»** кнопку **«Поиск решения...»**,[1] и в появившемся окне (рис. 3.7.1, *б*) укажем, что мы хотим **оптимизировать целевую функцию** в ячейке **$B$14** (в которой рассчитывается дисперсия портфеля) **до минимума, изменяя ячейки переменных $B$10:$B$11** (в которых находятся доли рисковых составляющих портфеля) **в соответствии с ограничением $B$13 = $B$7**.

После нажатия на кнопку «Найти решение» в рабочем листе произойдут изменения: в ячейках B9, B10 и B11 появятся значения $x_0^* = -0{,}3908$, $x_1^* = 1{,}1543$, $x_2^* = 0{,}2365$, в ячейке B13 будет рассчитана ожидаемая доходность портфеля (она равна требуемой доходности $r_\pi = 0{,}20$, а в ячейке B14 появится рассчитанное значение дисперсии доходности портфеля $\mathbf{Var}(R_\pi) = 0{,}0058$. При этом риск портфеля равен $\sigma_\pi = \sqrt{0{,}0058} \approx 0{,}076 = 7{,}6\%$. Интерпретация оптимального решения

$$x_0^* = -0{,}3908,\ x_1^* = 1{,}1543,\ x_2^* = 0{,}2365$$

такова: необходимо 23,65% потратить на приобретение акций второго вида, взять банковский кредит в размере 39,08% от общей суммы собственных средств, после чего все оставшиеся после покупки акций второго вида собственные средства вместе со средствами, полученными в кредит, вложить в покупку акций первого вида. Результаты работы программы представлены на рис. 3.7.1, *в*. ❑

Задачи

277. На рынке обращаются две акции, их ожидаемые доходности равны соответственно r_1 и r_2, средние квадратичные отклонения доходностей равны σ_1, σ_2, а коэффициент корреляции доходностей данных акций равен ρ_{12}. Составьте из этих акций портфель, среднее квадратическое отклонение доходности которого было бы равно нулю, и найдите математическое ожидание доходности такого портфеля в случае, если: *а*) $r_1 = 15\%$, $r_2 = 25\%$, $\sigma_1 = 8\%$, $\sigma_2 = 16\%$, $\rho_{12} = -1$; *б*) $r_1 = 10\%$, $r_2 = 20\%$, $\sigma_1 = 14\%$, $\sigma_2 = 12\%$, $\rho_{12} = 1$.

278. С каким наименьшим риском можно достичь 30%-ной доходности инвестиций, если есть возможность банковских вложений и заимствований по ставке $i = 12\%$ годовых, а на рынке ценных бумаг обращаются две акции, их ожидаемые доходности равны соответственно $r_1 = 22\%$ и $r_2 = 43\%$, риски $\sigma_1 = 15\%$, $\sigma_2 = 25\%$, а коэффициент корреляции доходностей данных акций равен $\rho_{12} = -0{,}68$?

---

[1] Если кнопка «Поиск решения» на вкладке **«Данные»** отсутствует, то это означает, что надстройка «Поиск решения» не установлена. Чтобы ее установить, необходимо выбрать в окне управления параметрами Excel (меню **«Файл»** — **«Параметры»**) пункт **«Надстройки»** (рис. 1.2.2, *а*), далее в нижней строке **«Управление»** выбрать **«Надстройки Excel»**, нажать кнопку **«Перейти...»**, в появившемся окне (рис. 1.2.2, *б*) отметить флажок **«Поиск решения»** и нажать **«ОК»**.

§ 3.8. Функции случайных величин

3.8.1. Функции одной случайной величины

Часто оказывается, что случайная величина является функцией некоторых других случайных величин: число качественных изделий в партии равно разности между размером партии и числом некачественных изделий в этой партии; время, затраченное наладчиком на ремонт двух станков, равно сумме двух случайных величин — времени, затраченного на ремонт первого и второго станка; дивиденды на акции в портфеле инвестора равны сумме случайных величин — дивидендов, выплаченных на каждую из акций портфеля.

Многие функции $\varphi(X^{(1)}, X^{(2)}, \ldots, X^{(n)})$ от случайных величин $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ также оказываются случайными величинами.

Например, можно показать, что если $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ — случайные величины, заданные на вероятностном пространстве $\{\Omega, \mathcal{S}, \mathbf{P}\}$, а $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ — непрерывная функция, то $\varphi(X^{(1)}, X^{(2)}, \ldots, X^{(n)})$ — случайная величина.

С функциями случайных величин мы уже имели дело, например, когда рассчитывали дисперсию случайной величины X как математическое ожидание непрерывной функции $\varphi(X) = [X - \mathbf{E}(X)]^2$.

Рассмотрим теперь о б щ и й с л у ч а й.

Пусть дана дискретная случайная величина X, заданная рядом распределения

$$\begin{array}{c|ccccc} X & x_1 & x_2 & \cdots & x_n & \cdots \\ \hline p & p_1 & p_2 & \cdots & p_n & \cdots \end{array} \qquad (3.8.1)$$

и с т р о г о м о н о т о н н а я функция $y = \varphi(x)$. Тогда различным значениям X соответствуют различные значения $Y = \varphi(X)$, причем вероятности соответствующих значений $X = x_i$ и $Y = \varphi(x_i)$ одинаковы.

Поэтому ряд распределения случайной величины $Y = \varphi(X)$ будет таким:

$$\begin{array}{c|ccccc} Y & \varphi(x_1) & \varphi(x_2) & \cdots & \varphi(x_n) & \cdots \\ \hline p & p_1 & p_2 & \cdots & p_n & \cdots \end{array} \qquad (3.8.2)$$

В случае же, когда функция $\varphi(X)$ немонотонна, различным значениям X могут, вообще говоря, соответствовать одинаковые значения $Y = \varphi(X)$, при этом для отыскания вероятностей возможных значений случайной величины $Y = \varphi(X)$ нужно сложить соответствующие вероятности тех возможных значений X, при которых $Y = \varphi(X)$ принимает одинаковые значения:

$$\begin{array}{c|cccc} Y & y_1 = \varphi(x_1) & y_2 = \varphi(x_2) & \cdots & y_n = \varphi(x_n) \quad \cdots \\ \hline p & \displaystyle\sum_{i:\,\varphi(x_i)=y_1} p_i & \displaystyle\sum_{i:\,\varphi(x_i)=y_2} p_i & \cdots & \displaystyle\sum_{i:\,\varphi(x_i)=y_n} p_i \quad \cdots \end{array} \qquad (3.8.3)$$

Пусть теперь X — абсолютно непрерывная случайная величина, заданная плотностью распределения $f_X(x)$. Если функция $y = \varphi(x)$ является с т р о г о м о н о т о н н о й и дифференцируемой [при этом существует обратная функция $x = \psi(y)$], то плотность распределения случайной величины Y вычисляется по формуле

$$f_Y(y) = f_X\big(\psi(y)\big)\,\big|\psi'(y)\big|. \qquad (3.8.4)$$

Доказательство. Пусть функция $y = \varphi(x)$ является строго возрастающей. Тогда

$$\int_{-\infty}^{y} f_Y(t)dt = F_Y(y) = \mathbf{P}\{Y < y\} = \mathbf{P}\{\varphi(X) < \varphi(y)\} = \mathbf{P}\{X < \varphi^{-1}(y)\} = \mathbf{P}\{X < \psi(y)\} =$$

$$= F_X(\psi(y)) = \int_{-\infty}^{\psi(y)} f_X(x)dx = \{\text{замена } z = \varphi(x), \text{ тогда } x = \varphi^{-1}(z) = \psi(z) > 0,$$

$$dx = \psi'(z)dz = |\psi'(z)|\,dz, \text{ при } x = -\infty \;\; z = -\infty, \text{ при } x = \psi(y) \;\; z = y\} = \int_{-\infty}^{y} f_X(\psi(z))\,|\psi'(z)|\,dz.$$

Если же функция $y = \varphi(x)$ является строго убывающей, то

$$\int_{-\infty}^{y} f_Y(t)dt = F_Y(y) = \mathbf{P}\{Y < y\} = \mathbf{P}\{\varphi(X) < \varphi(y)\} = \mathbf{P}\{X > \varphi^{-1}(y)\} = 1 - \mathbf{P}\{X < \psi(y)\} = 1 - F_X(\psi(y)) =$$

$$= 1 - \int_{-\infty}^{\psi(y)} f_X(x)dx = \int_{\psi(y)}^{+\infty} f_X(x)dx = \{\text{замена } z = \varphi(x), \;\; \text{тогда } x = \varphi^{-1}(z) = \psi(z) < 0,$$

$$dx = \psi'(z)dz = -|\psi'(z)|\,dz, \;\; \text{при } x = \psi(y) \;\; z = y, \;\; \text{при } x = +\infty \;\; z = -\infty\} =$$

$$= -\int_{y}^{-\infty} f_X(\psi(z))\,|\psi'(z)|\,dz = \int_{-\infty}^{y} f_X(\psi(z))\,|\psi'(z)|\,dz.$$

Итак, и в случае строгого возрастания функции $y = \varphi(x)$, и в случае ее строгого убывания $\int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} f_X(\psi(z))\,|\psi'(z)|\,dz$, поэтому $f_Y(y) = f_X(\psi(y))\,|\psi'(y)|$, что и требовалось доказать. ❑

Если же функция $y = \varphi(x)$ немонотонна, то область возможных значений случайной величины X разбивается на участки монотонности функции $\varphi(x)$, в каждом интервале по формуле (3.8.4) рассчитывается соответствующая функция $f_Y^{(i)}(y) = f_X(\psi_i(y))\,|\psi_i'(y)|$, а затем все они суммируются:

$$f_Y(y) = \sum_i f_Y^{(i)}(y). \tag{3.8.5}$$

Пример 3.8.1. Случайная величина X имеет плотность распределения $f_X(x)$. Найдем плотность распределения случайной величины $Y = |1 - X|$.

Вычисление плотности распределения $f_Y(y)$ случайной величины Y при $y > 0$ сведем в табл. 3.8.1.

Таблица 3.8.1

Вычисление плотности распределения случайной величины $Y = |1 - X|$

$f_X(x)$	$f_X(x)$
$y = \varphi(x)$	$y = \lvert 1 - x \rvert$
$x = \begin{cases} \psi_1(y), \\ \psi_2(y) \end{cases}$	$x_1 = 1 - y$ $x_2 = 1 + y$
$\lvert\psi_1'(y)\rvert = \lvert\psi_2'(y)\rvert$	1
$f_Y(y) = \sum_i f_X(\psi_i(y))\,\lvert\psi_i'(y)\rvert$	$f_Y(y) = f_X(1-y) + f_X(1+y)$

При $y < 0$ $\;\; f_Y(y) = 0$, так как $Y = |1 - X| \geqslant 0$. Окончательно получаем:

$$f_Y(y) = \begin{cases} f_X(1-y) + f_X(1+y), & y > 0, \\ 0, & y < 0. \end{cases} \;\; ❑$$

Пример 3.8.2. Случайная величина $X = \mathcal{N}(0; 1)$. Найдем плотность распределения случайной величины $Y = X^2$ (т. е. $Y = \chi_1^2$).

По определению функции распределения

$$F_Y(y) = \mathbf{P}\{Y \leqslant y\} = \mathbf{P}\{X^2 \leqslant y\} = \mathbf{P}\left\{|X| \leqslant \sqrt{y}\right\} = \mathbf{P}\left\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\right\} =$$

$$= F_{\mathcal{N}}\left(\sqrt{y}\right) - F_{\mathcal{N}}\left(-\sqrt{y}\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{-\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{0} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{-\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{-\sqrt{y}}^{0} \mathrm{e}^{-\frac{t^2}{2}}\, dt = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - \frac{1}{2} == \frac{2}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - 1.$$

При этом плотность распределения случайной величины Y по свойству (3.4.2) равна

$$f_Y(y) = F_Y'(y) = \left(\frac{2}{\sqrt{2\pi}} \int\limits_{-\infty}^{\sqrt{y}} \mathrm{e}^{-\frac{t^2}{2}}\, dt - 1 \right)' = \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y}{2}} \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} \mathrm{e}^{-\frac{y}{2}}. \ \square$$

Пример 3.8.3. Случайная величина $X = \mathrm{Exp}(\mu)$. Найдем функцию распределения случайной величины $Y = \mathrm{e}^{-\mu X}$.

Решение. По определению функции распределения

$$F_Y(y) = \mathbf{P}\{Y \leqslant y\} = \mathbf{P}\{\mathrm{e}^{-\mu X} \leqslant y\} = \begin{cases} 0, & y < 0, \\ \mathbf{P}\{-\mu X \leqslant \ln y\}, & 0 \leqslant y < 1, \\ 1, & y \geqslant 1. \end{cases}$$

Учитывая, что

$$\mathbf{P}\{-\mu X \leqslant \ln y\} = \mathbf{P}\left\{X \geqslant -\frac{1}{\mu} \ln y\right\} = 1 - \mathbf{P}\left\{X < -\frac{1}{\mu} \ln y\right\} =$$

$$= 1 - F_X\left(-\frac{1}{\mu} \ln y\right) = 1 - \left[1 - \mathrm{e}^{-\mu\left(-\frac{1}{\mu} \ln y\right)}\right] = y,$$

получаем окончательно:

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leqslant y < 1, \\ 1, & y \geqslant 1, \end{cases}$$

т. е. $Y = \mathrm{Unif}(0; 1)$. $\square$

Задачи

279. Случайная величина $X = \mathrm{Unif}(-\pi/2; \pi/2)$. Найдите плотность распределения случайной величины $Y = \cos X$.

280. Случайная величина X имеет плотность распределения $f_X(x)$. Найдите плотность распределения случайной величины $Y = aX + b$ (где a и b —числа).

281. Случайная величина X имеет строго возрастающую функцию распределения $F(x)$. Найдите функцию распределения случайной величины $Y = F(X)$.

282. При контроле качества на предприятии ежедневно проверяются десять изделий. Случайная величина Y — число дефектных изделий из десяти проверенных. Вероятность P того, что изделие окажется дефектным, является случайной величиной, распределенной по равномерному закону на отрезке $[0; 1/4]$. Найдите математическое ожидание и дисперсию случайной величины Y.

3.8.2. Функции нескольких случайных величин

Во многих простых случаях бывает удобно вычислять функцию распределения функции нескольких случайных величин непосредственным образом. В более сложных случаях приходится искать специальные приемы для решения таких задач.

Например, для функции нескольких абсолютно непрерывных случайных величин обычно удобнее искать не плотность распределения, а функцию распределения. Так, в случае функции двух аргументов: $Z = \varphi(X, Y)$ функция распределения случайной величины Z вычисляется по формуле

$$F_Z(z) = \iint\limits_{\varphi(x,y)<z} f_{XY}(x,y)dxdy\,. \tag{3.8.6}$$

В частности, функция распределения суммы двух абсолютно непрерывных случайных величин $Z = X + Y$ равна

$$F_Z(z) = \iint\limits_{x+y<z} f_{XY}(x,y)dxdy = \{\text{замена } u = x+y, y = u-x\} = \int\limits_{-\infty}^{z}\left(\int\limits_{-\infty}^{+\infty} f_{XY}(x,u-x)dx\right)du,$$

поэтому

$$f_Z(z) = \int\limits_{-\infty}^{+\infty} f_{XY}(x,z-x)dx\,. \tag{3.8.7}$$

Формула (3.8.7) называется *формулой композиции* или *формулой свертки*.

Пример 3.8.4. $X_1, X_2, \ldots, X_n$ — независимые случайные величины, равномерно распределенные на отрезке $[0; 1]$: $X_i = \mathrm{Unif}(0; 1)$, $i = 1, 2, \ldots, n$. Найдем функцию распределения случайной величины $X = \max\{X_1, X_2, \ldots, X_n\}$.

Функция распределения случайной величины $X = \max\{X_1, X_2, \ldots, X_n\}$ равна

$$F_X(x) = \mathbf{P}\{X \leqslant x\} = \mathbf{P}\left\{\max\{X_1, X_2, \ldots, X_n\} \leqslant x\right\} =$$
$$= \mathbf{P}\{(X_1 \leqslant x) \cap (X_2 \leqslant x) \cap \cdots \cap (X_n \leqslant x)\} =$$
$$= \{\text{независимость}\} = \prod_{i=1}^{n} \mathbf{P}\{X_i \leqslant x\} = \prod_{i=1}^{n} F_{X_i}(x) = \begin{cases} 0, & x < 0, \\ x^n, & 0 \leqslant x < 1, \\ 1, & x \geqslant 1. \end{cases} \quad \square$$

Пример 3.8.5. Пусть $X_1, X_2, \ldots, X_n$ — независимые случайные величины, каждая из которых имеет плотность распределения

$$f_{X_i}(x)=\begin{cases}ax, & x\in[0;1],\\ 0, & x\notin[0;1].\end{cases}$$

Найдем плотность распределения случайной величины $X=\min\{X_1, X_2, \ldots, X_n\}$.

Из условия нормированности плотности распределения: $\int\limits_{-\infty}^{+\infty} f(x)dx=1$ получаем, что $a\int\limits_0^1 xdx=1$ или $a\dfrac{x^2}{2}\Big|_{x=0}^{1}=1$, откуда $a=\dfrac{1}{1/2-0/2}=2$.

Поэтому плотность распределения случайной величины X_i равна $f_{X_i}(x)=\begin{cases}2x, & x\in[0;1],\\ 0, & x\notin[0;1],\end{cases}$ а функция распределения случайной величины X_i

$$F_{X_i}(x)=\int\limits_{-\infty}^{x} f(z)dz=\begin{cases}0, & x<0,\\ \int\limits_0^x 2zdz, & 0\leqslant x<1,\\ 1, & x\geqslant 1\end{cases}=\begin{cases}0, & x<0,\\ 2\dfrac{z^2}{2}\Big|_{z=0}^{x}, & 0\leqslant x<1,\\ 1, & x\geqslant 1\end{cases}=\begin{cases}0, & x<0,\\ x^2, & 0\leqslant x<1,\\ 1, & x\geqslant 1.\end{cases}$$

При этом функция распределения случайной величины $X=\min\{X_1, X_2, \ldots, X_n\}$ равна

$$F_X(x)=\mathbf{P}\{X\leqslant x\}=1-\mathbf{P}\{X>x\}=1-\mathbf{P}\{\min(X_1, X_2,\ldots, X_n)>x\}=$$

$$=1-\mathbf{P}\{(X_1>x)\cap(X_2>x)\cap\cdots\cap(X_n>x)\}=\{\text{независимость}\}=1-\prod_{i=1}^{n}\mathbf{P}\{X_i>x\}=$$

$$=1-\prod_{i=1}^{n}\bigl(1-\mathbf{P}\{X_i\leqslant x\}\bigr)=1-\prod_{i=1}^{n}\bigl(1-F_{X_i}(x)\bigr)=\begin{cases}0, & x<0,\\ 1-\bigl(1-x^2\bigr)^n, & 0\leqslant x<1,\\ 1, & x\geqslant 1,\end{cases}$$

а плотность распределения $f_X(x)=F_X'(x)=\begin{cases}2nx\bigl(1-x^2\bigr)^{n-1}, & x\in[0;1],\\ 0, & x\notin[0;1].\end{cases}$ ❑

Пример 3.8.6. Случайные величины $X_1=\mathrm{Unif}(1;3)$ и $X_2=\mathrm{Unif}(2;6)$. Найдем плотность распределения случайной величины $X=X_1+X_2$. По формуле (3.8.7)

$$f_X(x)=\int\limits_{-\infty}^{+\infty} f_{X_1X_2}(x_1, x-x_1)dx_1=\{\text{учтем независимость } X_1 \text{ и } X_2\}=$$

$$=\int\limits_{-\infty}^{+\infty} f_{X_1}(x_1)f_{X_2}(x-x_1)dx_1.$$

Учитывая, что

$$f_{X_1}(x_1)=\begin{cases}1/2, & x_1\in[1;3],\\ 0, & x_1\notin[1;3],\end{cases}$$

можно записать:

$$f_X(x)=\int\limits_{-\infty}^{+\infty} f_{X_1}(x_1)f_{X_2}(x-x_1)dx_1=\int\limits_1^3 \frac{1}{2}f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1.$$

Теперь учтем, что

$$f_{X_2}(x-x_1)=\begin{cases}1/4 & x-x_1\in[2;6],\\ 0, & x-x_1\notin[2;6],\end{cases}$$

и рассмотрим пять различных областей, в которых может находиться z.

Если $x<3$ и $x_1\in[1;3]$, то $x-x_1<2$, поэтому $f_{X_2}(x-x_1)=0$ при всех $x_1\in[1;3]$ и

$$f_X(x)=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_1^3 0dx_1=0.$$

Если $x\in[3;5]$ и $x_1\in[1;3]$, то $x-x_1\in[2;6]$ при $x_1\in[1;x-2]$, поэтому

$$f_{X_2}(x-x_1)=\begin{cases}1/4 & x_1\in[1;x-2],\\ 0, & x_1\in(x-2;3]\end{cases}$$

и

$$f_X(x)=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_1^{x-2}\frac{1}{4}dx_1=\frac{1}{8}(x-2-1)=\frac{x-3}{8}.$$

Если $x\in[5;7]$ и $x_1\in[1;3]$, то $x-x_1\in[2;6]$ поэтому $f_{X_2}(x-x_1)=\frac{1}{4}$ при всех $x_1\in[1;3]$ и

$$f_X(x)=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_1^3\frac{1}{4}dx_1=\frac{1}{8}(3-1)=\frac{2}{8}=\frac{1}{4}.$$

Если $x\in[7;9]$ и $x_1\in[1;3]$, то $x-x_1\in[2;6]$ при $x_1\in[x-6;3]$, поэтому

$$f_{X_2}(x-x_1)=\begin{cases}1/4 & x_1\in[x-6;3],\\ 0, & x_1\in[1;x-6)\end{cases}$$

и

$$f_X(x)=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_{x-6}^3\frac{1}{4}dx_1=\frac{1}{8}(3-x+6)=\frac{9-x}{8}.$$

Наконец, если $x>9$ и $x_1\in[1;3]$, то $x-x_1<2$, поэтому $f_{X_2}(x-x_1)=0$ и

$$f_X(x)=\frac{1}{2}\int\limits_1^3 f_{X_2}(x-x_1)dx_1=\frac{1}{2}\int\limits_1^3 0dx_1=0.$$

Собирая эти варианты, получаем:

$$f_X(x)=\begin{cases}0, & x<3,\\ (z-3)/8, & 3\leqslant x<5,\\ 1/4, & 5\leqslant x<7,\\ (9-z)/8, & 7\leqslant x<9,\\ 0, & x\geqslant 9.\end{cases}$$

График плотности распределения представлен на рис. 3.8.1. Если мы сложим три, четыре, пять равномерных случайных величин, то график плотности суммы будет становиться все больше похожим на «колокол» нормального закона распределения. ❑

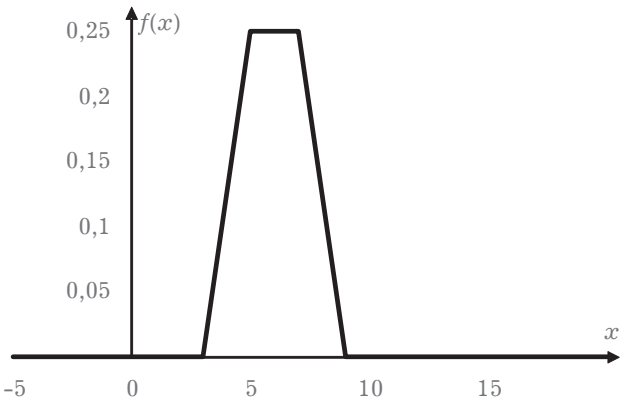


Рис. 3.8.1. *График плотности распределения суммы двух равномерно распределенных случайных величин в примере 3.8.6*

Задачи

283. Троллейбусы движутся с интервалом 8 мин, поезда метро — с интервалом 2 мин. Определите закон суммарного времени ожидания транспорта случайно выбранным пассажиром, пользующимся, чтобы добраться на работу, троллейбусом и метро (без пересадок в метро).

284. Случайные величины $X_1 = \mathcal{N}(a_1; \sigma_1)$ и $X_2 = \mathcal{N}(a_2; \sigma_2)$ независимы. Найдите закон распределения случайной величины $X = X_1 + X_2$.

285. Независимые случайные величины X_i распределены по нормальному закону: $X_i = \mathcal{N}(a_i; \sigma_i)$, c_i — неслучайные постоянные ($i = 1, 2, \ldots, n$). Найдите закон распределения случайной величины $X = \sum_{i=1}^{n} c_i X_i$.

286. Автомат заполняет банки кофе. Масса кофе и масса банки распределены по нормальному закону с математическими ожиданиями 500 г, 50 г и средними квадратичными отклонениями 8 г, 6 г соответственно. Найдите вероятность того, что масса готовой к продаже банки будет меньше 540 г.

287. Найдите закон распределения суммы двух независимых случайных величин, распределенных по закону Пуассона с параметрами λ_1 и λ_2 соответственно.

288. Докажите, что если случайные величины $\chi^2_{k_1}$ и $\chi^2_{k_2}$ независимы, то $\chi^2_{k_1} + \chi^2_{k_2} = \chi^2_{k_1+k_2}$.

Глава 4. Предельные теоремы теории вероятностей

§ 4.1. Закон больших чисел

4.1.1. Неравенство Чебышёва

Закон больших чисел утверждает, что среднее арифметическое значение некоррелированных случайных величин при возрастании их числа в определенном смысле стабилизируется, стремится к неслучайному числу — среднему арифметическому математических ожиданий этих случайных величин.

Доказательство этого факта основывается на неравенстве Чебышёва, смысл которого состоит в том, что любая случайная величина принимает значения, близкие к математическому ожиданию, и большие отклонения значений случайной величины от математического ожидания маловероятны.

Неравенство Чебышёва.[1] *Если случайная величина X имеет конечные математическое ожидание $\mathbf{E}(X)$ и дисперсию $\mathbf{Var}(X)$, то для любого $\varepsilon > 0$ справедливо неравенство*

$$\mathbf{P}\{|X-\mathbf{E}(X)| \geqslant \varepsilon\} \leqslant \frac{\mathbf{Var}(X)}{\varepsilon^2}. \tag{4.1.1}$$

Доказательство. Проведем доказательство для дискретных случайных величин. Если в выражении для дисперсии $\mathbf{Var}(X)=\sum_i (x_i-\mathbf{E}(X))^2 p_i$ отбросить из суммы в правой части те слагаемые, для которых $|x_i-\mathbf{E}(X)|<\varepsilon$, то сумма может только уменьшиться:

$$\mathbf{Var}(X)=\sum_i (x_i-\mathbf{E}(X))^2 p_i \geqslant \sum_{i:|x_i-\mathbf{E}(X)|\geqslant\varepsilon} (x_i-\mathbf{E}(X))^2 p_i.$$

$|x_i-\mathbf{E}(X)| \geqslant \varepsilon$, т. е.

$$\mathbf{Var}(X)=\sum_i (x_i-\mathbf{E}(X))^2 p_i \geqslant \sum_{i:|x_i-\mathbf{E}(X)|\geqslant\varepsilon} (x_i-\mathbf{E}(X))^2 p_i \geqslant \varepsilon^2 \sum_{i:|x_i-\mathbf{E}(X)|\geqslant\varepsilon} p_i.$$

Но $\sum_{i:|x_i-\mathbf{E}(X)|\geqslant\varepsilon} p_i = \mathbf{P}\{|X-\mathbf{E}(X)| \geqslant \varepsilon\}$, поэтому

$$\mathbf{Var}(X)=\sum_i (x_i-\mathbf{E}(X))^2 p_i \geqslant \sum_{i:|x_i-\mathbf{E}(X)|\geqslant\varepsilon} (x_i-\mathbf{E}(X))^2 p_i \geqslant \varepsilon^2 \mathbf{P}\{|X-\mathbf{E}(X)| \geqslant \varepsilon\}.$$

Отсюда получаем:

$$\mathbf{P}\{|X-\mathbf{E}(X)| \geqslant \varepsilon\} \leqslant \frac{\mathbf{Var}(X)}{\varepsilon^2},$$

что доказывает формулу (4.1.1) для дискретных случайных величин. Доказательство неравенства Чебышёва для абсолютно непрерывных случайных величин проводится по аналогичной схеме с заменой сумм интегралами.[2] ❑

---

[1] Пафнутий Львович Чебышёв очень обижался, когда его фамилию произносили с ударением на первый слог и с «е» вместо «ё».

[2] Справедливость неравенства Чебышёва для произвольных случайных величин следует из того, что любая случайная величина может быть приближена монотонно неубывающей последовательностью дискретных случайных величин, и математическое ожидание случайной величины определяется как предел последовательности соответствующих математических ожиданий.

Очевидно, что неравенство Чебышёва (4.1.1) можно записать в эквивалентной форме:

$$\mathbf{P}\{|X-\mathbf{E}(X)|<\varepsilon\}\geqslant 1-\frac{\mathbf{Var}(X)}{\varepsilon^2}. \qquad (4.1.2)$$

Неравенство (4.1.1), названное в честь П.Л. Чебышёва, впервые было сформулировано без доказательства в 1853 г. И.-Ж. Бьенеме, другом и коллегой П.Л. Чебышёва, а сам П.Л. Чебышёв в 1867 г. впервые доказал данное неравенство. Другое доказательство неравенства Чебышёва предложил в 1884 г. в своей кандидатской диссертации А.А. Марков, ученик П.Л. Чебышёва.

Пример 4.1.1. Средний размер вклада клиента некоторого банка равен 300 тыс. руб. при среднем квадратическом отклонении, равном 100 тыс. руб. Оценим с помощью неравенства Чебышёва долю вкладов, которые имеют размер от 100 до 500 тыс. руб.:

$$\mathbf{P}\{100<X<500\}=\mathbf{P}\{|X-300|<200\}=\mathbf{P}\{|X-\mathbf{E}(X)|<200\}\geqslant$$

$$\geqslant 1-\frac{\mathbf{Var}(X)}{200^2}=1-\frac{\sigma_X^2}{200^2}=1-\frac{100^2}{200^2}=1-\frac{1}{4}=\frac{3}{4}=0{,}75,$$

т. е. 75% всех вкладов имеют размер от 100 до 500 тыс. руб. ❑

С помощью неравенства Чебышёва (4.1.2) можно доказать

Правило трех сигм для произвольной случайной величины. *Вероятность* $\mathbf{P}\{|X-\mathbf{E}(X)|<3\sigma_X\}$ *для произвольной случайной величины с конечным математическим ожиданием и конечной дисперсией составляет не менее* $8/9=0{,}(8)\approx 0{,}89$:

$$\mathbf{P}\{|X-\mathbf{E}(X)|<3\sigma_X\}\geqslant\frac{8}{9}. \qquad (4.1.3)$$

Доказательство. Согласно неравенству Чебышёва (4.1.2) для любого $\varepsilon>0$ $\mathbf{P}\{|X-\mathbf{E}(X)|<\varepsilon\}\geqslant 1-\frac{\mathbf{Var}(X)}{\varepsilon^2}$. Пусть $\varepsilon=3\sigma_X$, тогда $\mathbf{P}\{|X-\mathbf{E}(X)|<3\sigma_X\}\geqslant 1-\frac{\sigma_X^2}{(3\sigma_X)^2}=1-\frac{\sigma_X^2}{9\sigma_X^2}=$

$=1-\frac{1}{9}=\frac{8}{9}$, что и требовалось доказать. ❑

Неравенство Чебышёва позволяет легко провести доказательство, обещанное в **п. 3.2.2**.

Доказательство второй части утверждения о формуле для дисперсии константы (3.2.21). Покажем, что если для некоторой случайной величины X выполняется равенство $\mathbf{Var}(X)=0$, то существует такое число $c=\mathbf{E}(X)$, что $\mathbf{P}\{X=c\}=1$, т. е. с вероятностью, равной единице, эта случайная величина равна константе. Согласно неравенству Чебышёва (4.1.2) для любого $\varepsilon>0$ $\mathbf{P}\{|X-\mathbf{E}(X)|<\varepsilon\}\geqslant 1-\mathbf{Var}(X)/\varepsilon^2$. Но в данном случае $\mathbf{Var}(X)=0$, т. е. для любого $\varepsilon>0$ $\mathbf{P}\{|X-\mathbf{E}(X)|<\varepsilon\}\geqslant 1$. Учитывая, что по теореме об ограниченности вероятности (2.2.28) вероятность любого события, в том числе, и события $\{|X-\mathbf{E}(X)|<\varepsilon\}$, не превосходит единицы, заключаем, что для любого $\varepsilon>0$ $\mathbf{P}\{|X-\mathbf{E}(X)|<\varepsilon\}=1$, откуда $\mathbf{P}\{|X-\mathbf{E}(X)|=0\}=1$, значит, $\mathbf{P}\{X=\mathbf{E}(X)\}=1$, т. е. существует такое число $c=\mathbf{E}(X)$, что $\mathbf{P}\{X=c\}=1$, что и требовалось доказать. ❑

Задачи

289. Для новогоднего праздника Петя должен сделать гирлянду из 400 электрических лампочек. Лампочки оказались очень низкого качества —

вероятность того, что какая-либо из них погаснет во время праздника, составляет 0,5. Поэтому Петя решил включить их параллельно. С помощью неравенства Чебышёва оценить вероятность того, что число горящих лампочек будет заключено между 100 и 300.

290. Средние ежедневные расходы на покупку канцелярских принадлежностей для офиса банка составляют 1000 руб., а среднее квадратическое отклонение этой случайной величины не превышает 200 руб. С помощью неравенства Чебышёва оцените вероятность того, что расходы на канцелярские принадлежности в любой наугад выбранный день не превысят 2000 руб.

4.1.2. Теорема Чебышёва и теорема Бернулли

Под законом больших чисел понимается обобщенное название группы теорем, утверждающих, что при неограниченном увеличении числа испытаний средние величины сходятся (в определенном смысле) к некоторым постоянным. Наиболее общей из этих теорем является теорема Чебышёва.

Теорема Чебышёва. *Если дисперсии некоррелированных случайных величин* $X_1, X_2, \ldots, X_n$ *ограничены сверху числом* B, *то для произвольного сколь угодно малого* $\varepsilon > 0$ *справедливо неравенство*

$$\mathbf{P}\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} \mathbf{E}(X_i)}{n}\right| < \varepsilon\right\} \geqslant 1 - \frac{B}{n\varepsilon^2} \tag{4.1.4}$$

и предельное равенство

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} \mathbf{E}(X_i)}{n}\right| < \varepsilon\right\} = 1. \tag{4.1.5}$$

Доказательство. Очевидно, если $X_1, X_2, \ldots, X_n$ — случайные величины, то величина $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ также является случайной, причем по свойствам математического ожидания и дисперсии $\mathbf{E}(\overline{X}) = \dfrac{\sum_{i=1}^{n} \mathbf{E}(X_i)}{n}$, и поскольку случайные величины $X_1, X_2, \ldots, X_n$ некоррелированы, то $\mathbf{E}(\overline{X}) = \dfrac{\sum_{i=1}^{n} \mathbf{Var}(X_i)}{n^2}$.

Применим к случайной величине $\overline{X}$ неравенство Чебышёва: $\mathbf{P}\left\{\left|\dfrac{\sum_{i=1}^{n} X_i}{n} - \dfrac{\sum_{i=1}^{n} \mathbf{E}(X_i)}{n}\right| < \varepsilon\right\} = \mathbf{P}\{|\overline{X} - \mathbf{E}(\overline{X})| < \varepsilon\} \geqslant 1 - \dfrac{\mathbf{Var}(\overline{X})}{\varepsilon^2} = 1 - \dfrac{\sum_{i=1}^{n} \mathbf{Var}(X_i)}{n^2\varepsilon^2}$.

Учитывая, что все $\mathbf{Var}(X_i) \leqslant B$, получим, что $1 - \dfrac{\sum_{i=1}^{n} \mathbf{Var}(X_i)}{n^2\varepsilon^2} \geqslant 1 - \dfrac{\sum_{i=1}^{n} B}{n^2\varepsilon^2} = 1 - \dfrac{nB}{n^2\varepsilon^2} = 1 - \dfrac{B}{n\varepsilon^2}$, т. е. доказана справедливость неравенства (4.1.4).

Переходя в этом неравенстве к пределу при $n \to \infty$, получаем равенство (4.1.5). ❑

Закон больших чисел в форме Чебышёва утверждает, что среднее арифметическое случайных величин при возрастании их числа обладает свойством *статистической устойчивости*, т. е. сходится по вероятности к неслучайной величине — среднему арифметическому математических ожиданий этих случайных величин. Практическое применение закона больших чисел состоит в том, что среднее арифметическое, вычисленное по достаточно большому числу результатов измерений какой-либо величины, будет сколь угодно близко к измеряемой величине.

А в следующей теореме доказывается статистическая устойчивость относительной частоты появления успеха в серии независимых испытаний.

Теорема Бернулли. *Если вероятность успеха в каждом из n независимых испытаний постоянна и равна p, то для произвольного сколь угодно малого $\varepsilon > 0$ справедливо предельное равенство*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{m}{n} - p\right| < \varepsilon\right\} = 1, \tag{4.1.6}$$

где m — число успехов в серии из n испытаний.

Доказательство. Рассмотрим случайные величины $I_{A_1}, I_{A_2}, \text{K}, I_{A_n}$ — индикаторы событий A_1, A_2, K, A_n:

$$I_{A_i} = \begin{cases} 1, & \text{если произошел успех в } i\text{-м испытании [с вероятностью } p\text{]}, \\ 0, & \text{если не произошел успех в } i\text{-м испытании [с вероятностью } (1-p)\text{]}. \end{cases}$$

Тогда $\mathbf{E}(I_{A_i}) = p$, $\mathbf{Var}(I_{A_i}) = p(1-p)$, $m = \sum_{i=1}^{n} I_{A_i}$. Поскольку $0 \leqslant p \leqslant 1$, дисперсии случайных величин X_i ограничены сверху единицей (так как $\mathbf{Var}(I_{A_i}) = p(1-p) \leqslant 1$), и можно воспользоваться теоремой Чебышёва (4.1.5), согласно которой

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{m}{n} - p\right| < \varepsilon\right\} = \lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{m}{n} - \frac{np}{n}\right| < \varepsilon\right\} = \lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{\sum_{i=1}^{n} I_{A_i}}{n} - \frac{\sum_{i=1}^{n} \mathbf{E}(I_{A_i})}{n}\right| < \varepsilon\right\} = 1,$$

что и требовалось доказать. ❑

Теорему Бернулли, очевидно, можно записать и в форме, аналогичной (4.1.4):

$$\mathbf{P}\left\{\left|\frac{m}{n} - p\right| < \varepsilon\right\} \geqslant 1 - \frac{p(1-p)}{n\varepsilon^2}. \tag{4.1.7}$$

Пример 4.1.2. Для определения среднего дохода налогоплательщиков города налоговой инспекцией была проведена проверка 250 жителей этого города, отобранных случайным образом. Оценим вероятность того, что средний годовой доход жителей города отклонится от среднего арифметического $\overline{X} = \sum_{i=1}^{250} X_i / 250$ годовых доходов выбранных 250 жителей не более, чем на 1000 руб., если известно, что среднее квадратическое отклонение годового дохода не превышает 2500 руб.

Согласно неравенству (4.1.4), которым можно пользоваться, поскольку все $\mathbf{Var}(X_i) \leqslant (2500)^2$,

$$\mathbf{P}\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} \mathbf{E}(X_i)}{n}\right| < 1000\right\} \geqslant 1 - \frac{2500 \cdot 2500}{250 \cdot 1000 \cdot 1000} = 1 - \frac{25}{1000} = 0{,}975. \ \square$$

Задачи

291. По статистическим данным в среднем по 87% договорам страхования от несчастных случаев страховых событий не происходит (т. е. вероятность отсутствия страховых событий по случайно выбранному договору равна 0,87). Оцените вероятность того, что из 1000 заключенных договоров страхования от несчастных случаев доля (относительная частота) договоров, по которым не произойдет страховых событий, будет отличаться от вероятности не более, чем на 0,04 (по модулю).

292. Математическое ожидание ущерба по договору страхования автогражданской ответственности (для определенной группы страхователей одинакового возраста, с одинаковым стажем вождения и автомобилями с одинаковым объемом двигателя) составляет 10 000 руб., а среднее квадратическое отклонение ущерба равно 2100 руб. Оцените, в каких границах может находиться реальное значение ущерба по конкретному договору с вероятностью, не меньшей чем: а) 0,75; б) 8/9.

§ 4.2. Центральная предельная теорема

4.2.1. Теорема Леви

Законы больших чисел устанавливают факт приближения среднего значения большого числа случайных величин к некоторым постоянным в виде сходимости последовательностей случайных величин по вероятности и почти наверное. Но этим не ограничиваются закономерности, возникающие в результате суммарного действия случайных величин. Центральная предельная теорема — это общее название группы теорем, утверждающих, что достаточно большая сумма сравнительно малых случайных величин распределена приближенно по нормальному закону.

В ряде задач приходится сталкиваться с ситуацией, когда исследуемая случайная величина является суммой большого числа независимых слагаемых, влияние каждого из которых на сумму очень мало. Такими случайными величинами являются, например, капиталы банков и страховых компаний (доля каждого отдельно взятого вкладчика не зависит от доли других вкладчиков и относительно мала, но в сумме все эти доли весьма весомы), выручка торговых предприятий (покупатели действуют независимо друг от друга и покупают товары на относительно небольшие суммы) и др. — мы уже говорили об этом в **п. 3.5.3**.

На основании центральной предельной теоремы часто можно до наблюдения того или иного явления сказать, что соответствующая случайная величина должна иметь нормальное распределение или близкое к нему.

Практическое значение центральной предельной теоремы велико — она составляет теоретико-вероятностную основу методов математической статистики.

В частном случае суммы большого числа биномиальных случайных величин центральная предельная теорема была впервые опубликована в 1733 г.

А. де Муавром. Этот результат опередил свое время и был забыт до 1812 г., когда П.-С. Лаплас включил его в свою знаменитую книгу «Аналитическая теория вероятностей».

В 1887 г. П.Л. Чебышёв опубликовал статью «О двух теоремах относительно вероятностей», в которой он установил, что при некоторых весьма общих условиях сумма большого числа независимых случайных величин с нулевыми математическими ожиданиями распределена приблизительно по нормальному закону. Доказывая свой вариант центральной предельной теоремы, П.Л. Чебышёв допустил логический пробел: он не учел необходимость выполнения условия существования предела у среднего арифметического дисперсий. Этот недостаток был вскоре исправлен А. А. Марковым, учеником П.Л. Чебышёва.

В общей форме центральная предельная теорема была впервые сформулирована и доказана в 1901 г. другим учеником П.Л. Чебышёва — А.М. Ляпуновым. Назвать теорему Ляпунова «центральной предельной» предложил в 1920 г. Д. Пойа. В том же 1920 г. Я.В. Линдеберг показал, что центральная предельная теорема справедлива в более слабых условиях, чем было доказано А.М. Ляпуновым.

Суммы независимых случайных величин легли в основу теории случайных процессов, разработанной А.А. Марковым, А. Н. Колмогоровым, Н. Винером, Е.Е. Слуцким, П.П. Леви, Б.В. Гнеденко, А.Я. Хинчиным, В. Феллером и другими исследователями.

Наиболее простая из формулировок центральной предельной теоремы — предложена П.П. Леви. В теореме Леви рассматривается сумма случайных величин $X_1, X_2, \ldots, X_n, \ldots$, распределенных по одному и тому же закону. В более сложных вариантах (например, в теореме Ляпунова), складывающиеся случайные ведичины $X_1, X_2, \ldots, X_n, \ldots$ могут не обязательно быть распределенными по одному и тому же закону — главное, чтобы они были независимы, и вклад каждой из них в общую дисперсию был относительно мал.

Теорема Леви. *Если независимые случайные величины* $X_1, X_2, \ldots, X_n, \ldots$ *распределены по одному и тому же закону с математическим ожиданием* a *и средним квадратическим отклонением* σ, *то при* $n \to \infty$ *функция распределения случайной величины*

$$Z_n = \frac{\sum_{i=1}^{n} (X_i - a)}{\sigma\sqrt{n}} \tag{4.2.1}$$

сходится (в определенном смысле) к функции распределения стандартной нормальной случайной величины $\mathcal{N}(0;1)$. Это записывается так:

$$Z_n \Rightarrow \mathcal{N}(0;1).$$

Доказательство теоремы Леви выходит за рамки данной книги.

Пусть независимые случайные величины $X_1, X_2, \ldots, X_n, \ldots$ распределены по одному и тому же закону, $\mathbf{E}(X_i) = a, \mathbf{Var}(X_i) = \sigma^2$.

Из теоремы Леви, очевидно, следует, что сумма случайных величин $X_1, X_2, \ldots, X_n, \ldots$ *при* $n \to \infty$ *сходится к нормальной случайной величине* $\mathcal{N}\left(na; \sigma\sqrt{n}\right)$.

Рассмотрим теперь среднее арифметическое этих величин $\overline{X} = \sum_{i=1}^{n} X_i \Big/ n$ и перепишем выражение (4.2.1) в виде

$$Z = \frac{\sum_{i=1}^{n}(X_i - a)}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n}(X_i - a)\Big/ n}{\sigma\sqrt{n}\big/ n} = \frac{\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} a}{n}}{\sigma\big/\sqrt{n}} = \frac{\overline{X} - \frac{na}{n}}{\sigma\big/\sqrt{n}} =$$

$$= \frac{\overline{X} - a}{\sigma\big/\sqrt{n}} = \frac{(\overline{X} - a)\sqrt{n}}{\sigma},$$

откуда следует, что выборочное среднее $\overline{X}$ при $n \to \infty$ сходится по распределению к случайной величине, распределенной по нормальному закону с параметрами $a_{\overline{X}} = a, \sigma_{\overline{X}} = \sigma / \sqrt{n}$.

Пример 4.2.1. В шляпе лежат четыре карточки, на которых написаны числа 1, 3, 5, 7. Какая карточка будет извлечена, если их предварительно перемешать и не подглядывать, не известно. Это случайная величина X, заданная рядом распределения

X	1	3	5	7
p	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

.

Математическое ожидание этой случайной величины равно $\mathbf{E}(X) = 1\cdot\frac{1}{4} + 3\cdot\frac{1}{4} + 5\cdot\frac{1}{4} + 7\cdot\frac{1}{4} = \frac{1+3+5+7}{4} = \frac{16}{4} = 4$, дисперсия $\mathbf{Var}(X) = (1-4)^2\cdot\frac{1}{4} + (3-4)^2\cdot\frac{1}{4} + (5-4)^2\cdot\frac{1}{4} + (7-4)^2\cdot\frac{1}{4} = \frac{(-3)^2 + (-1)^2 + 1^2 + 3^2}{4} = \frac{9+1+1+9}{4} = \frac{20}{4} = 5$, среднее квадратическое отклонение $\sigma_X = \sqrt{\mathbf{Var}(X)} = \sqrt{5} \approx 2{,}236$.

Будем не глядя доставать из шляпы карточку, запоминать написанное на ней число, возвращать карточку обратно в шляпу, перемешивать и не глядя доставать вторую карточку. Все возможные варианты извлечений двух карточек с возвращением представлены в табл. 4.2.1 (справа от чисел, выписанных на карточках, вычислены средние значения).

Таблица 4.2.1

Варианты извлечений двух карточек с возвращением

Извлеченные карточки X_1, X_2	Среднее $\overline{X} = (X_1 + X_2)/2$	Извлеченные карточки X_1, X_2	Среднее $\overline{X} = (X_1 + X_2)/2$
1; 1	1	5; 1	3
1; 3	2	5; 3	4
1; 5	3	5; 5	5
1; 7	4	5; 7	6
3; 1	2	7; 1	4
3; 3	3	7; 3	5
3; 5	4	7; 5	6
3; 7	5	7; 7	7

В столбце средних единица встречается один раз из 16, двойка два раза, тройка три раза, четверка четыре раза, пятерка три раза, шестерка два раза и семерка — один раз. Составим ряд распределения случайной величины $\bar{X}$:

$\bar{X}$	1	2	3	4	5	6	7
p	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

.

Математическое ожидание случайной величины $\bar{X}$ равно

$$\mathbf{E}(\bar{X}) = 1\cdot\frac{1}{16} + 2\cdot\frac{2}{16} + 3\cdot\frac{3}{16} + 4\cdot\frac{4}{16} + 5\cdot\frac{3}{16} + 6\cdot\frac{2}{16} + 7\cdot\frac{1}{16} = \\ = \frac{1+4+9+16+15+12+7}{16} = \frac{64}{16} = 4,$$

,

дисперсия

$$\mathbf{Var}(\bar{X}) = (1-4)^2\cdot\frac{1}{16} + (2-4)^2\cdot\frac{2}{16} + (3-4)^2\cdot\frac{3}{16} + \\ + (4-4)^2\cdot\frac{4}{16} + (5-4)^2\cdot\frac{3}{16} + (6-4)^2\cdot\frac{2}{16} + (7-4)^2\cdot\frac{1}{16} = \\ = \frac{(-3)^2\cdot 1 + (-2)^2\cdot 2 + (-1)^2\cdot 3 + 0^2\cdot 4 + 1^2\cdot 3 + 2^2\cdot 2 + 3^2\cdot 1}{16} = \frac{9+8+3+0+3+8+9}{16} = \\ = \frac{40}{16} = \frac{5}{2},$$

среднее квадратическое отклонение $\sigma_{\bar{X}} = \sqrt{\mathbf{Var}(\bar{X})} = \sqrt{\frac{5}{2}} = \frac{\sqrt{10}}{2} \approx 1{,}581.$

Можно заметить, что $\mathbf{E}(\bar{X}) = \mathbf{E}(X), \sigma_{\bar{X}} = \sigma_X / \sqrt{n}$. ❑

Пример 4.2.2. Случайная величина X — размер месячного счета абонента сотового оператора за пользование услугами телефонной связи.

Математическое ожидание случайной величины X составляет 630 руб., при этом ее среднее квадратическое отклонение равно 110 руб. График плотности распределения данной случайной величины представлен на рис. 4.2.1, *а*, из которого видно, что закон распределения случайной величины является достаточно сложным и нестандартным — асимметричным, с несколькими модами.

Неравенство Чебышёва позволяет гарантировать, что по крайней мере у 75% абонентов размер счета отклоняется от математического ожидания менее чем на два средних квадратических отклонения:

$$\mathbf{P}\{|X - \mathbf{E}(X)| < 2\sigma_X\} \geqslant 1 - \frac{\sigma_X^2}{(2\sigma_X)^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0{,}75$$

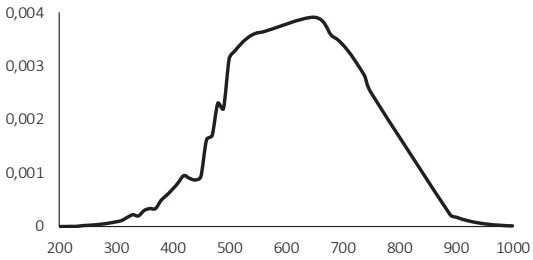
или

$$\mathbf{P}\{|X - 630| < 2\cdot 110\} = \mathbf{P}\{630 - 220 < X < 630 + 220\} = \mathbf{P}\{410 < X < 850\} \geqslant 0{,}75,$$

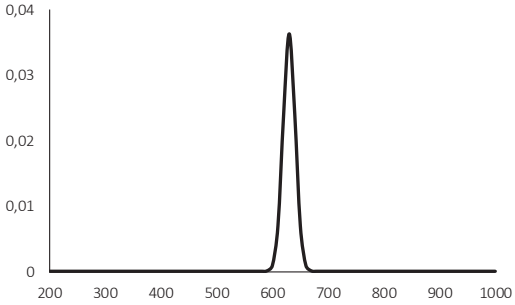
т. е. не менее 75% всех абонентов получают счета на суммы от 410 до 850 руб.

Пусть из всех абонентов случайным образом выбраны 100 человек, и $\bar{X}$ — средний размер счета у выбранных $n = 100$ абонентов. Так как по теореме Чебышёва $\mathbf{P}\{|\bar{X} - \mathbf{E}(X)| < \varepsilon\} \geqslant 1 - \frac{\mathbf{Var}(X)}{n\varepsilon^2}$, можно записать: $\mathbf{P}\{|\bar{X} - \mathbf{E}(X)| < 2\sigma_X\} \geqslant$

$\geqslant 1-\dfrac{\mathbf{Var}(X)}{n\left(2\sigma_X\right)^2}=1-\dfrac{\sigma_X^2}{4n\sigma_X^2}=1-\dfrac{1}{4n}=1-\dfrac{1}{400}=\dfrac{399}{400}=0{,}9975$, т. е. средний счет у 100 случайно выбранных абонентов не менее чем в 99,75% случаев будет находиться в границах от 410 до 850 руб.



а)



б)

Рис. 4.2.1. *Графики функций плотности распределения случайных величин X (а) и $\overline{X}$ (б) из примера 4.2.2*

Согласно центральной предельной теореме случайная величина $\overline{X}$ подчиняется нормальному закону распределения с параметрами $a=\mathbf{E}(\overline{X})=\mathbf{E}(X)=$ $=630$ руб., $\sigma_{\overline{X}}=\dfrac{\sigma_X}{\sqrt{n}}=\dfrac{110}{\sqrt{100}}=\dfrac{110}{10}=11$ руб.

График плотности распределения случайной величины $\overline{X}$ представлен на рис. 4.2.1, *б*.

Вычислим вероятность

$$\mathbf{P}\left\{410<\overline{X}<850\right\}=\text{НОРМ.РАСП(850; 630; 11; ИСТИНА) –}$$
$$\text{– НОРМ.РАСП(410; 630; 11; ИСТИНА) = 1.}$$

Теорема Чебышёва поволяет утверждать, что средний счет у 100 случайно выбранных абонентов не менее чем в 99,75% случаев будет находиться в границах от 410 до 850 руб., тогда как с помощью центральной предельной теоремы мы можем заключить, что вероятность выхода среднего размера счета 100 абонентов за эти границы равна нулю.

Рассчитаем вероятность

$$\mathbf{P}\left\{597<\overline{X}<663\right\}=\text{НОРМ.РАСП(663; 630; 11; ИСТИНА) –}$$
$$\text{– НОРМ.РАСП(597; 630; 11; ИСТИНА) = 0,997300. ❑}$$

С практической точки зрения теорема Леви позволяет гораздо точнее, чем теорема Чебышёва, оценивать границы доверительных интервалов для случайных величин, т. е. интервалов, в которые с определенной (близкой к единице) вероятностью попадают значения случайных величин.

Пример 4.2.3. Математическое ожидание остатка по вкладу случайно выбранного клиента банка составляет 450 000 руб., при этом среднее квадратическое отклонение равно 85 000 руб.

Вычислим вероятность, с которой остаток на вкладе у одного случайно выбранного вкладчика окажется более 500 000 руб.:

$$\mathbf{P}\{X > 500\,000\} = \mathbf{P}\{\mathcal{N}(450\,000; 85\,000) > 500\,000\} =$$

$$= 1 - \text{НОРМ.РАСП}(500\,000; 450\,000; 85\,000; \text{ИСТИНА}) = 0{,}278187.$$

Рассчитаем теперь вероятность, с которой средний остаток на вкладе у 100 случайно выбранных вкладчиков окажется более 500 000 руб.:

$$\mathbf{P}\{\bar{X} > 500\,000\} = \mathbf{P}\left\{\mathcal{N}\left(450\,000; \frac{85\,000}{\sqrt{100}}\right) > 500\,000\right\} =$$

$$= \mathbf{P}\{\mathcal{N}(450\,000; 8500) > 500\,000\} =$$

$$= 1 - \text{НОРМ.РАСП}(500\,000; 450\,000; 8500; \text{ИСТИНА}) = 0{,}00000000202237.$$

Видим, что хотя почти 30% клиентов имеют остатки, большие 500 000 руб., очень маловероятно, чтобы средний остаток по вкладу у достаточно большого числа вкладчиков оказался больше 500 000 руб. ❑

Пример 4.2.4. Ежедневный суммарный объем выдач банкомата равен в среднем 100 000 руб. и в 95% случаев не превосходит от 110 000 руб. Найдем вероятность того, что в очередной день суммарный объем выдач банкомата окажется в пределах от 80 000 до 120 000 руб.

Пусть X — суммарный объем выдач банкомата. Клиенты банка, снимающие в банкомате наличные деньги, действуют независимо друг от друга и снимают относительно небольшие суммы $X_i \ll X$, но всего клиентов у банка достаточно много, так что можно считать, что их количество $n \to \infty$. Поэтому суммарный объем выдач банкомата будет иметь нормальное распределение с некоторыми параметрами a и σ. Поскольку для нормального закона распределения $a = \mathbf{E}(X)$, то по условию $a = \mathbf{E}(X) = 100\,000$. Также в условии сказано, что $\mathbf{P}\{X \leqslant 110\,000\} = 0{,}95$.

Рассмотрим случайную величину $Z = \dfrac{X - a}{\sigma}$, она будет иметь стандартный нормальный закон распределения $\mathcal{N}(0;1)$, при этом $\mathbf{P}\{X \leqslant x\} = \mathbf{P}\left\{\dfrac{X-a}{\sigma} \leqslant \dfrac{x-a}{\sigma}\right\} = \mathbf{P}\left\{Z \leqslant \dfrac{x-a}{\sigma}\right\}$. Из того, что $\mathbf{P}\{X \leqslant 110\,000\} = 0{,}95$, следует: $\mathbf{P}\{X \leqslant x\} = \mathbf{P}\left\{\dfrac{X-a}{\sigma} \leqslant \dfrac{x-a}{\sigma}\right\} = \mathbf{P}\left\{Z \leqslant \dfrac{x-a}{\sigma}\right\}$, откуда $\mathbf{P}\{X \leqslant 110\,000\} =$ $= \mathbf{P}\left\{Z \leqslant \dfrac{110\,000 - 100\,000}{\sigma}\right\} = \mathbf{P}\left\{Z \leqslant \dfrac{10\,000}{\sigma}\right\}$. Но $\mathbf{P}\{X \leqslant 110\,000\} = 0{,}95$, поэтому $\mathbf{P}\left\{Z \leqslant \dfrac{10\,000}{\sigma}\right\} = 0{,}95$.

Теперь можно воспользоваться функцией НОРМ.СТ.ОБР и вычислить $10\,000 / \sigma = \text{НОРМ.СТ.ОБР}(0{,}95) = 1{,}644854$.

Отсюда $\sigma = 1{,}644854 / 10\,000 = 6079{,}568$, и мы теперь можем вычислить вероятность

$$\mathbf{P}\{80\,000 < X < 120\,000\} = \text{НОРМ.РАСП}(120\,000;100\,000;\ 6079{,}568;\ \text{ИСТИНА}) -$$
$$- \text{НОРМ.РАСП}(100\,000;100\,000;\ 6079{,}568;\ \text{ИСТИНА}) = 0{,}998997. \ \square$$

Задачи

293. В условиях примера 4.2.1 выпишите все выборки объема $n = 3$ (все возможные варианты извлечений $n = 3$ карточек из шляпыс возвращением. Составьте ряд распределения случайной величины $\bar{X}$, вычислите $\mathbf{E}(\bar{X})$ и $\sigma_{\bar{X}}$ и сравните с результатами, получен в примере 4.2.1.

294. В условиях примера 4.2.2 оцените с помощью теоремы Чебышёва и теоремы Леви, с какой вероятностью средний счет у $n = 64$ абонентов отклонится от своего математического ожидания менее чем на два средних квадратических отклонения.

295. Производитель автомобильных покрышек утверждает , что математическое ожидание срока службы покрышки равно 100 000 км при среднем квадратическом отклонении 1600 км. Для тестов было случайным образом выбрано 100 покрышек, и оказалось, чтот их средний срок службы составил 98 721 км. Вычислите: *а*) с какой вероятностью срок службы одной случайно выбранной покрышки будет не больше 98 721 км; *б*) с какой вероятностью средний срок службы у 100 случайно выбранных покрышек будет не больше 98 721 км. На основании сделанных выводов дайте оценку, насколько соответстует действительности утверждение производителя.

296. Математическое ожидание возраста работника некоторого крупного банка составляет 47 лет при среднем квадпратическом отклонении, равном 3,6 г. Вычислите математическое ожидание и среднее квадратическое отклонение для среднего возраста n случайно выбранных работников банка при n, равном: *а*) 36; *б*) 64; *в*) 100.

297. Банк, признанный лучшим работодателем года, утверждает, что средняя заработная плата его сотрудников составляет 160 тыс. руб. при среднем квадратическом отклонении 7200 руб. С какой вероятностью средняя заработная плата у 30 случайно выбранных работников окажется менее 120 тыс. руб.?

298. Банкомат выдает стандартные суммы в 5000, 1000 и 500 руб., причем первые составляют 10%, а последние — 20% всех выдач. В среднем банкомат производит 3000 выдач в сутки. Определите размер денежной суммы, которую необходимо заложить в банкомат утром, чтобы этой суммы с вероятностью 0,9 хватило для выдачи наличности вкладчикам до следующего утра.

299. Торговец чашками-непроливайками ходит по вагонам электропоездов. В каждом из вагонов он может продать чашку с вероятностью 1/3. Случайная величина X — число вагонов, в которые заходил торговец прежде, чем продал первые 100 чашек. Найти распределение случайной величины X.

Указание. Пусть Y_i — число вагонов, которые обошел торговец за время от продажи $(i-1)$-й чашки до продажи i-й. Тогда все Y_i $(i = 1,\ 2,\ \ldots,\ n)$ имеют одинаковый (геометрический) закон распределения, $X = \sum_{i=1}^{n} Y_i$. Согласно центральной предельной теореме, при большом n случайная величина X имеет нормальный закон распределения. Предоставляем читателю показать, что параметры этого распределения равны $a = 300, \sigma = 30$. $\square$

4.2.2. Интегральная теорема Муавра — Лапласа

Приведем теперь следствие из центральной предельной теоремы, относящееся к независимым испытаниям.

Интегральная теорема Муавра — Лапласа. *Если вероятность p успеха в каждом испытании отлична от нуля и единицы, а число испытаний n достаточно велико, то число успехов в последовательности из n испытаний распределено примерно по нормальному закону с параметрами $a = np$ и* $\sigma = \sqrt{np(1-p)}$.

Доказательство. Пусть в серии из n испытаний Бернулли произошло X успехов. Тогда случайную величину $X = \mathrm{Bin}(n; p)$ можно представить в виде суммы n независимых одинаково распределенных случайных величин — индикаторов событий

$$I_{A_i} = \begin{cases} 1, & \text{если произошел успех в } i\text{-м испытании (с вероятностью } p), \\ 0, & \text{если не произошел успех в } i\text{-м испытании [с вероятностью } (1-p)]: \end{cases}$$

$$X = \sum_{i=1}^{n} I_{A_i}.$$

При этом $\mathbf{E}\left(I_{A_i}\right) = p, \mathbf{Var}\left(I_{A_i}\right) = p(1-p)$, и по теореме Леви случайная величина

$$Z_n = \frac{\sum_{i=1}^{n}(I_{A_i} - p)}{\sqrt{p(1-p)}\sqrt{n}} = \frac{X - np}{\sqrt{np(1-p)}}$$ при $n \to \infty$ $X = \sum_{i=1}^{n} I_{A_i} \Rightarrow \mathcal{N}\left(np; \sqrt{np(1-p)}\right)$, что и требовалось доказать. ❑

На практике приближение Муавра — Лапласа достаточно хорошо работает, когда $n > 60$, а $np > 10$.

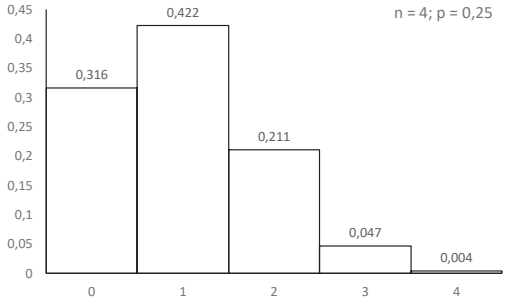
Пример 4.2.5. Рассмотрим несколько биномиальных случайных с одной и той же вероятностью успеха, например, $p = 0{,}25$, и увеличивающимся числом испытаний, например, $n = 4, 10, 25, 50, 100, 300$.

Построим ряд распределения случайной величины $\mathrm{Bin}(n = 4;\ p = 0{,}25)$. Вычисления в Microsoft Excel иллюстрирует рис. 4.2.2 (справа от ячеек, в которых есть формулы, приведены эти формулы). Теперь визуализируем ряд распределения, построив соответствующую гистограмму (рис. 4.2.3, *а*).

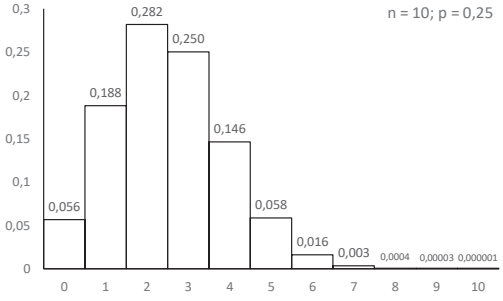
Аналогичным образом на рис. 4.2.3, *б—е* построим гистограммы, визуализирующие ряды распределения случайных величин $\mathrm{Bin}(n = 10;\ p = 0{,}25)$, $\mathrm{Bin}(n = 25;\ p = 0{,}25)$, $\mathrm{Bin}(n = 50;\ p = 0{,}25)$, $\mathrm{Bin}(n = 100;\ p = 0{,}25)$, $\mathrm{Bin}(n = 300;\ p = 0{,}25)$. На рис. 4.2.3, *ж* для $n = 300;\ p = 0{,}25$ построим график плотности случайной величины $\mathcal{N}\left(a = np = 75; \sigma = \sqrt{np(1-p)} = 7{,}5\right)$. Видно, что с увеличением числа испытаний n биномиальная случайная величина становится все более похожей на нормальную с параметрами $a = np; \sigma = \sqrt{np(1-p)}$. ❑

	A	B	C
1	*n*	4	
2	*p*	0,25	
3	*k*	P{*X* = *k*}	
4	0	0,316406	=БИНОМ.РАСП($A4;B$1;B$2;ЛОЖЬ)
5	1	0,421875	=БИНОМ.РАСП($A5;B$1;B$2;ЛОЖЬ)
6	2	0,210938	=БИНОМ.РАСП($A6;B$1;B$2;ЛОЖЬ)
7	3	0,046875	=БИНОМ.РАСП($A7;B$1;B$2;ЛОЖЬ)
8	4	0,003906	=БИНОМ.РАСП($A8;B$1;B$2;ЛОЖЬ)

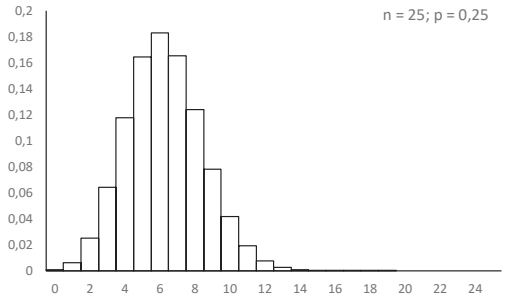
Рис. 4.2.2. *Ряд распределения случайной величины* $\mathrm{Bin}(n = 4;\ p = 0{,}25)$



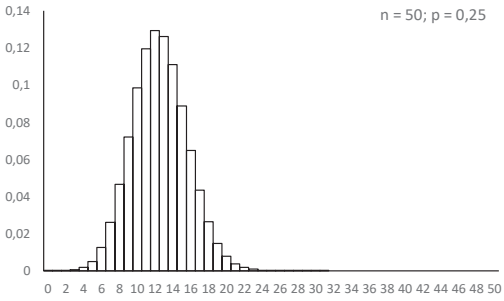
а) Bin($n = 4$; $p = 0{,}25$)



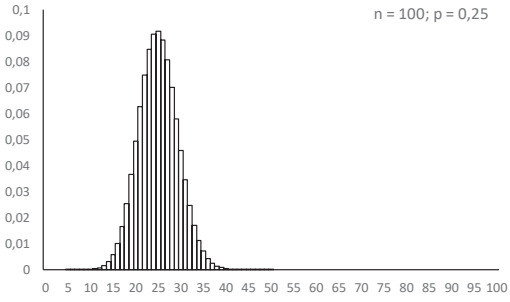
б) Bin($n = 10$; $p = 0{,}25$)



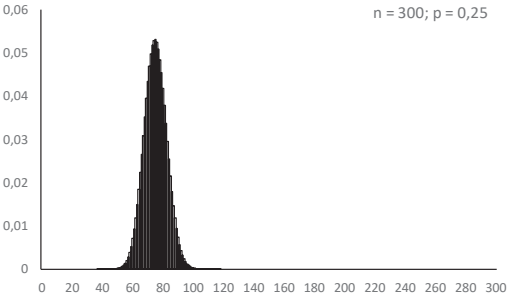
в) Bin($n = 25$; $p = 0{,}25$)



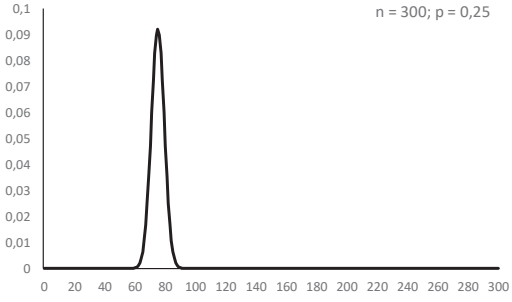
г) Bin($n = 50$; $p = 0{,}25$)



д) Bin($n = 100$; $p = 0{,}25$)



е) Bin($n = 300$; $p = 0{,}25$)



ж) *нормальное приближение* $\mathcal{N}\left(a = np; \sigma = \sqrt{np(1-p)}\right)$
биномиального распределения Bin($n = 300$; $p = 0{,}25$)

Рис. 4.2.3. *Сходимость биномиальных случайных величин к нормальному закону*

Пример 4.2.6. Строительная фирма для привлечения инвестиций в строительство нового дома собирается воспользоваться банковским кредитом. Вероятность того, что какой-либо банк в ответ на поступление бизнес-плана примет положительное решение о кредитовании фирмы, равна 0,3. Строительная фирма обратилась в 100 банков. Найдем вероятности того, что решение о предоставлении кредита этой фирме: *а*) не примет ни один банк; *б*) примет хотя бы один банк; *в*) примут больше 10 банков; *г*) примут больше 40 банков.

Данную ситуацию можно рассматривать как серию из $n=100$ испытаний Бернулли, в которых успехом считается принятие банком решения о кредитовании. Вероятность успеха в единичном испытании равна по условию $p=0{,}3$. Поскольку число испытаний n велико, а произведение $np=30>10$, можно воспользоваться интегральной теоремой Муавра — Лапласа:

$$а)\ \mathbf{P}(A_0)= \text{НОРМ.РАСП(0; 100*0,3; КОРЕНЬ(100*0,3*(1 – 0,3)); ИСТИНА)} =$$
$$= 0{,}000000000029443;$$
$$б)\ \mathbf{P}(A_{\geqslant 1})=1-\mathbf{P}(A_0)=$$
$$= 1 - \text{НОРМ.РАСП(0; 100*0,3; КОРЕНЬ(100*0,3*(1 – 0,3)); ИСТИНА)} =$$
$$= 0{,}999999999970557;$$
$$в)\ \mathbf{P}(A_{>10})=1-\mathbf{P}(A_{\leqslant 10})=$$
$$= 1 - \text{НОРМ.РАСП(9; 100*0,3; КОРЕНЬ(100*0,3*(1 – 0,3)); ИСТИНА)} =$$
$$= 0{,}999993625162539;$$
$$г)\ \mathbf{P}(A_{>40})=1-\mathbf{P}(A_{\leqslant 40})=$$
$$= 1 - \text{НОРМ.РАСП(9; 100*0,3; КОРЕНЬ(100*0,3*(1 – 0,3)); ИСТИНА)} =$$
$$= 0{,}014548165870626.$$

Задачи

300. Вероятность появления успеха в каждом из независимых испытаний равна 0,25. Найти вероятность того, что в серии из 300 испытаний успех наступит ровно 75 раз.

301. Вероятность появления успеха в каждом из независимых испытаний равна 0,25. Найти вероятность того, что в серии из 300 испытаний успех наступит от 70 до 100 раз.

302. Менеджер ресторана по своему опыту знает, что в среднем около 70% клиентов, заказавших в ресторане столик на вечер, приходят вечером в ресторан. В ресторане 30 столиков, но сегодня менеджер принял заказы у 35 клиентов. Определить, с какой вероятностью вечером в ресторан придут более чем 30 посетителей, заказавших столики.

303. Во время каникул Петя работал в предвыборном штабе кандидата в депутаты, который проводил выборочный опрос избирателей. Примерное распределение голосов было известно: по 40% избирателей «*за*» и «*против*» кандидата, остальные воздержались. Сколько нужно опросить людей, чтобы с вероятностью, не меньшей 0,9, гарантировать отклонение процента голосов, отданных за кандидата при выборочном опросе, от истинного мнения избирателей не более, чем на 2% от всего электората?

304. В дачном поселке 2500 жителей, каждый из которых примерно шесть раз в месяц ездит на поезде в город, выбирая дни поездок случайным образом и независимо от других жителей. Какой наименьшей вместимостью должен обладать поезд, чтобы он переполнялся в среднем не чаще одного раза в 100 дней (поезд ходит раз в сутки).

4.2.3. Математические основы теории страхования

Эффективным способом уменьшения потерь от неопределенностей является объединение отдельных людей и организаций в страховые сообщества, поскольку трудно предсказать время, место и характер событий, способных повлиять на экономическое состояние индивидуумов, вместе с тем, по закону больших чисел, средние (или суммарные) потери большой группы индивидуумов предсказать можно. В страховых сообществах каждый индивидуум вносит сумму, намного меньшую его возможного ущерба, и в случае наступления ущерба убытки возмещаются из средств, собранных всеми членами сообщества, в случае же, когда для кого-то из членов страхового сообщества ущерб не наступает, первоначально выплаченная этим индивидуумом сумма распределяется между теми членами сообщества, которые понесли убытки.

Первая математическая модель страхования была построена Т. Барруа в 1834 г., современные *актуарные* (т. е. страховые) модели восходят к Ф. Лундбергу, который в 1903 г. заложил основы актуарной теории риска.

По договору страхования одна сторона (страхователь) платит другой стороне (страховщику) определенную денежную сумму (страховую премию), и за это страховщик гарантирует возмещение возможных убытков страхователя (в случае их возникновения). Смысл договора страхования состоит в том, что страхователь подвержен определенному риску (который заключается в возможном наступлении некоторого страхового случая) и стремится от этого риска защититься, а задачей страховщика является предоставление такой защиты. В качестве страхового случая может выступать болезнь, смерть, автомобильная авария, потеря имущества при пожаре, потеря финансовых средств при неблагоприятно складывающейся рыночной ситуации, а также отмеченные выше переломы ног у балерин, пальцев у пианистов, зубов у фотомоделей и т. п. В договоре страхования указываются *срок* его действия, условия и способ возмещения ущерба. Например, в договоре страхования гражданской ответственности водителя транспортного средства обычно указывается, что если в момент наступления страхового случая (при аварии) водитель находился в состоянии алкогольного опьянения, то страховщик ответственности по полису не несет. Если в указанный в договоре страхования срок страховой случай не наступил, страхователь теряет уплаченную премию.

Далее мы более подробно рассмотрим математические модели **страхования жизни**, которое получило наибольшее развитие.

Договор страхования жизни может быть *обязательным* (в силу действия определенного закона) или *добровольным* (по взаимному волеизъявлению страховщика и страхователя), *краткосрочным* (как правило, на один год) или *долгосрочным*.

Основным источником случайности в страховании жизни является неопределенность момента смерти отдельного человека. Однако в случае, когда одновременно у одного и того же страховщика страхуется большая однородная (по возрасту, полу, типу профессии, месту проживания и т. п.) группа страхователей, в силу закона больших чисел можно говорить об устойчивости относительных частот и рассматривать *продолжительность жизни* как неотрицательную случайную величину X с функцией распределения $F(x) = \mathbf{P}\{X < x\}$. *Функция выживания*

$$s(x) = \mathbf{P}\{X \geqslant x\} = 1 - F(x), \tag{4.2.2}$$

равна вероятности того, что человек из данной однородной группы проживет не менее x лет. Функция выживания (4.2.2) предполагается монотонно убывающей (иначе в определенных интервалах времени смерть будет невозможна) и непрерывной (иначе возможны моменты, в которые смерть наступает с положительной вероятностью). Кроме того, функция выживания (4.2.2) должна удовлетворять всем свойствам, которые следуют из того, что $F(x) = 1 - s(x)$ является функцией распределения случайной величины X.

Пусть

$$T(x) = X - x \text{ —}$$

остаточное время жизни человека в возрасте x лет. Через

$${}_tp_x = \mathbf{P}\{T(x) \geqslant t\}$$

обозначается вероятность того, что человек в возрасте x лет проживет еще не менее t лет.

По определению условной вероятности

$${}_tp_x = \mathbf{P}\{T(x) \geqslant t\} = \mathbf{P}\{X \geqslant x + t | X \geqslant x\} = \frac{\mathbf{P}\{X \geqslant x + t\}}{\mathbf{P}\{X \geqslant x\}} = \frac{{}_{x+t}p_0}{{}_xp_0} = \frac{s(x+t)}{s(x)}.$$

В таблицах продолжительности жизни рассматривается группа новорожденных одного пола, проживающих в одинаковой местности, в количестве l_0 чел. Пусть X_k — продолжительность жизни k-го человека из данной группы ($k = 1, 2, \ldots, l_0$). Количество доживших до возраста x обозначим $L(x)$, и в таблицах продолжительности жизни приводится математическое ожидание случайной величины $L(x)$:

$$l_x = \mathbf{E}(L(x)) = \sum_{k=1}^{l_0} \mathbf{P}\{X_k \geqslant x\} = \sum_{k=1}^{l_0} s(x) = l_0 s(x).$$

Фрагмент такой таблицы продолжительности жизни для городского населения Российской Федерации в 1993 г. приведен в табл. 4.2.2.

Таблица 4.2.2

Фрагмент таблицы продолжительности жизни городского населения Российской Федерации в 1993 г.

x	Женщины		Мужчины		x	Женщины		Мужчины	
	l_x	$s(x)$	l_x	$s(x)$		l_x	$s(x)$	l_x	$s(x)$
0	100 000	1,00000	100 000	1,00000	55	87 007	0,87007	64 338	0,64338
1	98 324	0,98324	97 822	0,97822	60	82 469	0,82469	54 864	0,54864
5	97 922	0,97922	97 416	0,97416	65	76 558	0,76558	44 222	0,44222
10	97 790	0,97790	97 080	0,97080	70	67 118	0,67118	32 706	0,32706
15	97 623	0,97623	96 764	0,96764	75	53 628	0,53628	21 417	0,21417
20	97 278	0,97278	95 804	0,95804	80	36 986	0,36986	11 814	0,11814
25	96 832	0,96832	94 194	0,94194	85	20 192	0,20192	5113	0,05113
30	96 296	0,96296	92 009	0,92009	90	7607	0,07607	1571	0,01571
35	95 572	0,95572	89 008	0,89008	95	1591	0,01591	297	0,00297
40	94 474	0,94474	85 003	0,85003	99	237	0,00237	48	0,00048
45	92 831	0,92831	79 644	0,79644	100	130	0,00130	28	0,00028
50	90 335	0,90335	72 722	0,72722	110	0	0,00000	0	0,00000

Простейший вид краткосрочного страхования жизни заключается в следующем. Страхователь (некоторый человек) платит страховщику (страховой компании) страховую премию в сумме c ден. ед., а страховщик соглашается выплатить наследникам страхователя страховую выплату (или страховое пособие) в сумме b ден. ед. в случае его смерти в течение года (и не платить ничего в противном случае).

Величина страховой выплаты, конечно, должна быть много больше страховой премии, и одной из важнейших задач актуарной математики является вычисление соотношений между страховой выплатой b и страховой премией c.

Пример 4.2.7. Вероятность смерти тридцатилетнего мужчины составляет 0,006. Страховая компания заключила 10 000 договоров страхования с мужчинами в возрасте тридцати лет, согласно которым в случае смерти застрахованного лица в течение ближайшего года его наследникам в конце этого года выплачивается 105 000 руб. Стоимость одного договора равна 1200 руб, а годовая ставка по банковским депозитам равна 5%.

Найдем вероятность того, что к концу года страховая компания окажется в убытке.

Пусть за год наступило K страховых случаев, тогда доход страховой компании составит

$$U = 10\,000 \cdot 1200 - \frac{105\,000K}{1+0{,}05} = 100\,000(120 - K)$$

(здесь мы привели все выплаты к настоящему моменту времени: выплата 105 000 руб. через год имеет сегодня ценность $\frac{105\,000}{1+0{,}05}$).

Поэтому компания окажется в убытке ($U < 0$), если за год наступит более 120 страховых случаев. Вероятность наступления страхового случая $p_x = 0{,}006$. Всего проводится $n = 10\,000$ испытаний. Поскольку число испытаний n велико, можно воспользоваться интегральной теоремой Муавра — Лапласа:

$$\mathbf{P}\{K \geqslant 120\} = 1 - \mathbf{P}\{K \leqslant 119\} =$$
$$= 1 - \text{НОРМ.РАСП}(119;60;\text{КОРЕНЬ}(10000*0{,}006*(1-0{,}006));\text{ИСТИНА}) =$$
$$= 0{,}000000000000011,$$

т. е. страховая компания может оказаться в убытке практически с нулевой вероятностью.

Теперь вычислим вероятность того, что доход страховой компании превысит 4 000 000 руб. (очевидно, это произвойдет в том и только в том случае, если за год наступит менее 80 страховых случаев):

$$\mathbf{P}\{K \leqslant 79\} = 1 - \text{НОРМ.РАСП}(79;60;\text{КОРЕНЬ}(10000*0{,}006*(1-0{,}006));\text{ИСТИНА}) =$$
$$= 0{,}9930586,$$

значит, доход страховой компании превысит 4 000 000 руб. с вероятностью 0,995, очень близкой к единице, т. е. почти наверное. ❑

Естественно, в реальных страховых компаниях стоимость договора страхования складывается из теоретической оценки страховой премии и оценки средних транзакционных издержек на один договор. Первое из этих слагаемых одинаково для всех страховых компаний, действующих на одном рынке, и компания может обеспечить конкурентоспособность своих страховых продуктов только за счет снижения транзакционных издержек.

Задачи

305. В страховой компании 10 000 клиентов (страхователей), каждый из которых уплатил страховой компании сумму 1000 руб. (страховую премию). Вероятность наступления страхового случая равна (по оценкам экспертов компании) 0,005, а страховая выплата при наступлении страхового случая составляет 100 000 руб. Определите, на какую прибыль может рассчитывать страховая компания с вероятностью 0,99.

306. В условиях предыдущей задачи определите минимальный размер страховой премии, при котором страховая компания получит прибыль, не меньшую 1 000 000 руб., с вероятностью 0,999.

4.2.4. Моделирование случайных величин

Во многих практических задачах бывает полезно вместо проведения дорогостоящих (или принципиально невозможных) экспериментов с реальным объектом «проиграть», имитировать его поведение на компьютере.

Метод исследования поведения систем со случайными воздействиями, состоящий в компьютерной имитации изучаемых процессов при помощи моделирования случайных величин, называется *методом Монте-Карло*[1] или *методом статистических испытаний*.

Идея метода Монте-Карло состоит в том, что строится вероятностная модель реального процесса, затем с помощью компьютерно реализованных датчиков случайных чисел проводят большое число испытаний, в которых моделируется поведение исследуемой системы. По результатам большого числа испытаний на основании закона больших чисел делают выводы о поведении реального процесса.

Процесс получения значений случайной величины с заданным законом распределения с помощью датчиков случайных чисел называется *моделированием* (или *разыгрыванием*) этой случайной величины.

Для реализации метода Монте-Карло достаточно иметь датчик случайных чисел, являющихся реализациями случайной величины, распределенной по равномерному закону на отрезке [0; 1].

Моделирование произвольной случайной величины X с помощью датчика случайных чисел, распределенных по равномерному закону на отрезке [0; 1], основано на том факте, что случайная величина $Y = F_X(X)$, где $F_X(x)$ — функция распределения вероятностей произвольной случайной величины X, имеет равномерный закон распределения на отрезке [0; 1].

Доказательство. Действительно, при $y < 0$ значение функции распределения случайной величины $Y = F_X(X)$ равно $F_Y(y) = \mathbf{P}\{Y \leqslant y\} = 0$, при $y > 1$ значение $F_Y(y) = \mathbf{P}\{Y \leqslant y\} = 1$, поскольку $F_X(X) \in [0; 1]$. При $0 \leqslant y \leqslant 1$ значение $F_Y(y) = \mathbf{P}\{Y \leqslant y\} = \mathbf{P}\{F_X(X) \leqslant y\} = \mathbf{P}\left\{X \leqslant F_X^{-1}(y)\right\} = F_X\left(F_X^{-1}(y)\right) = y$ [здесь $F_X^{-1}(y)$ обозначает функцию, обратную к функции $F_X(x)$].[2] ❑

[1] Название метода связано с городом Монте-Карло, известного своими казино, и в частности, игрой в рулетку, которая представляет собой один из простейших примеров датчиков случайных чисел.

[2] Вообще говоря, обратная функция $F_X^{-1}(y)$ существует не для всякой функции распределения. Достаточным условием существования обратной функции $F_X^{-1}(y)$ является, например, монотонное возрастание функции распределения.

Отсюда следует, что если y представляет собой значение случайной величины, распределенной по равномерному закону на отрезке [0; 1], полученное с помощью датчика случайных чисел, то для получения значения x случайной величины X достаточно решить уравнение $F_X(x) = y$, т. е. найти квантиль уровня y данной случайной величины (учитывая определение квантили уровня α случайной величины X как числа x_α, являющегося решением уравнения $F_x(x_\alpha) = \alpha$ — см. **п. 3.6.3**).

Пример 4.2.8. Пусть требуется смоделировать систему массового обслуживания, состоящую из продавца и покупателей в магазине. Предполагая, что покупатели образуют простейший *поток событий*, можно считать, что количество покупателей, пришедших за время τ, имеет закон распределения Пуассона с параметром $\lambda = \mu_1\tau$ (где интенсивность поступления заказов μ_1 — число покупателей в единицу времени), а время между двумя последовательно пришедшими покупателями распределено по показательному закону с параметром μ_1.

Время обслуживания покупателя продавцом можно считать показательной случайной величиной с параметром μ_2, равным интенсивности обслуживания (числу покупателей, обслуженных в единицу времени).

Теперь можно смоделировать рабочий день магазина и выяснить значения важных параметров: среднего времени дневного простоя продавцов, средней длины очереди и т. п. Для этого нужно провести расчеты по следующей схеме.

В начале рабочего дня продавец свободен. Он будет свободен в течение времени, пока не появится первый покупатель, которое получим, как реализацию случайной величины, распределенной по показательному закону с параметром μ_1.

Для моделирования времени прихода первого покупателя, которое распределено по показательному закону с параметром μ_1, достаточно взять реализацию y случайной величины Y, распределенной по равномерному закону на отрезке [0; 1], и определить x как решение уравнения

$$1 - e^{-\mu_1 x} = y \Leftrightarrow x = -\frac{\ln(1-y)}{\mu_1}.$$

Вычисления можно сократить, если заметить, что если случайная величина Y распределена по равномерному закону на отрезке [0; 1], то точно такой же закон распределения будет и у случайной величины $1 - Y$, поэтому вместо формулы

$$x = -\frac{\ln(1-y)}{\mu_1}$$

для моделирования показательной случайной величины используют соотношение

$$x = -\frac{\ln y}{\mu_1}.$$

После прихода покупателя продавец будет занят его обслуживанием в течение времени, которое представляет собой случайную величину, распределенную по показательному закону с параметром μ_2. Реализацию этой случайной величины также получим с помощью датчика случайных чисел.

Время прихода следующего покупателя вновь определим как реализацию экспоненциальной случайной величины с параметром μ_1. Если в момент прихода нового покупателя продавец еще занят обслуживанием предыдущего, новый покупатель встает в очередь, и ее длина увеличивается на единицу.

В момент окончания обслуживания очередного покупателя продавец начинает обслуживание нового покупателя (если очередь к этому моменту была непуста) либо начинает простаивать.

Проведя достаточно много испытаний, можно быть уверенным, что в силу закона больших чисел средние значения параметров реального процесса будут близки к средним значениям, полученным в ходе моделирования. ❑

Аналогичным образом моделируют и другие случайные величины.

В пакете Microsoft Excel для получения случайных чисел, равномерно распределенных на отрезке [0; 1] служит функция СЛЧИС(). Для вычисления квантилей (т. е. функций, обратных к функциям распределения) биномиальных, нормальных и логарифмически нормальных случайных величин в Microsoft Excel удобно использовать функции БИНОМ.ОБР(<n>; <p>; <y>); НОРМ.ОБР(<y>; <a>; <σ>); ЛОГНОРМ.ОБР(<y>; <a>; <σ>).

Пример 4.2.9. Пусть в среднем каждый день в магазин приходит 600 посетителей, при этом стандартное отклонение числа посетителей равно 30. Конечно, заранее неизвестно, сколько посетителей придет в магазин в данный конкретный день, поэтому число посетителей за день представляет собой некоторую случайную величину X.

При этом, с одной стороны, если посетители образуют простейший поток, то их число в день X распределено по закону Пуассона с параметром λ, равным интенсивности поступления посетителей, т. е. $\lambda = 600$ (посетителей в день).

С другой стороны, поскольку число X посетителей в день можно представить в виде

$$X = I_{A_1} + I_{A_2} + \cdots + I_{A_n} + \cdots,$$

где A_i — событие, состоящее в том, что i-й потенциальный посетитель пришел в магазин, I_{A_n} — индикатор события A_i ($i = 1, 2, ..., n, ...$), потенциальных посетителей достаточно много, они примерно одинаковы и действуют независимо друг от друга, то можно считать, что $X = \mathcal{N}(a = 600; \ = 30)$.

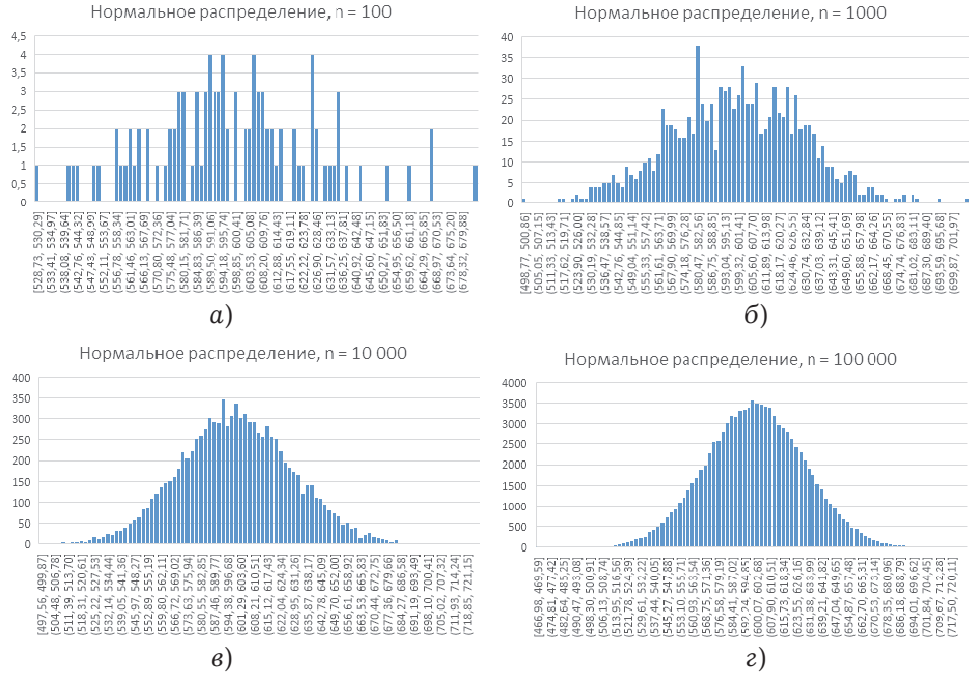
Смоделируем число покупателей X как случайную величину, распределенную по нормальному закону с математическим ожиданием $a = 600$ и средним квадратическим отклонением $\sigma = 30$.

Вначале проведем 100 испытаний. Создадим новую книгу Microsoft Excel и введем в ячейку A1 название имитационного эксперимента: «Нормальное распределение, n = 100». В ячейку A2 введем формулу «= НОРМ.ОБР(СЛЧИС(); 600; 30)», скопируем ее и вставим в ячейки A3:A101. Теперь в этих ячейках находятся 100 реализаций случайной величины $\mathcal{N}(a = 600; \sigma = 30)$. Построим по этим данным гистограмму распределения (нажав на вкладке «Данные» кнопку «Вставить статистическую диаграмму» и выбрав тип статистической диаграммы — «Гистограмма»). Результат представлен на рис. 4.2.4, *а*.

Теперь проведем 1000 испытаний и сохраним результаты в столбце B. Для этого введем в ячейку B1 название имитационного эксперимента: «Нормальное распределение, n = 1000», скопируем формулу «= НОРМ.ОБР(СЛЧИС(); 600; 30)» и вставим ее в ячейки B2:B1001.[1] Соответствующая гистограмма представлена на рис. 4.2.4, *б*. Аналогичным образом проведем 10 000 экспериментов в ячейках

[1] Проще всего для этого перейти в ячейку B1001, введя «B1001» в поле выбора текущей ячейки, находящееся слева от поля ввода формул, а затем последовательно нажать клавиши «Shift», «End» (не отпуская «Shift»), «↑» (отпустив «End», но не отпуская «Shift»). Таким образом мы выделили диапазон B1:B101, и чтобы снять выделение с ячейки B1, в которой находится название столбца, достаточно нажать клавишу «↓» (все еще не отпуская «Shift»). Теперь можно вставить формулу в выделенный диапазон B2:B101.

C2:C10001 (гистограмма представлена на рис. 4.2.4, *в*) и 100 000 экспериментов в ячейках D2:D100001 (гистограмма представлена на рис. 4.2.4, *г*). Замечаем, что с ростом числа экспериментов гистограммы становятся все больше похожими на «колокол» нормального закона распределения.



а) *б*) *в*) *г*)

Рис. 4.2.4. *Гистограммы результатов моделирования случайной величины* $\mathcal{N}(600; 30)$ *при* $n = 100$ (*а*); *при* $n = 1000$ (*б*); *при* $n = 10\,000$ (*в*); *при* $n = 100\,000$ (*г*)

Результаты сохранены на рабочем листе «Нормальная и пуассоновская» в файле «Моделирование случайных величин.xlsx». ❑

Для многих случайных величин, тем не менее, в пакете Microsoft Excel нет функций, вычисляющих их квантили, однако в надстройку «Анализ данных» пакета Microsoft Excel входит программа «Генерация случайных чисел», позволяющая получить последовательности значений различных случайных величин:

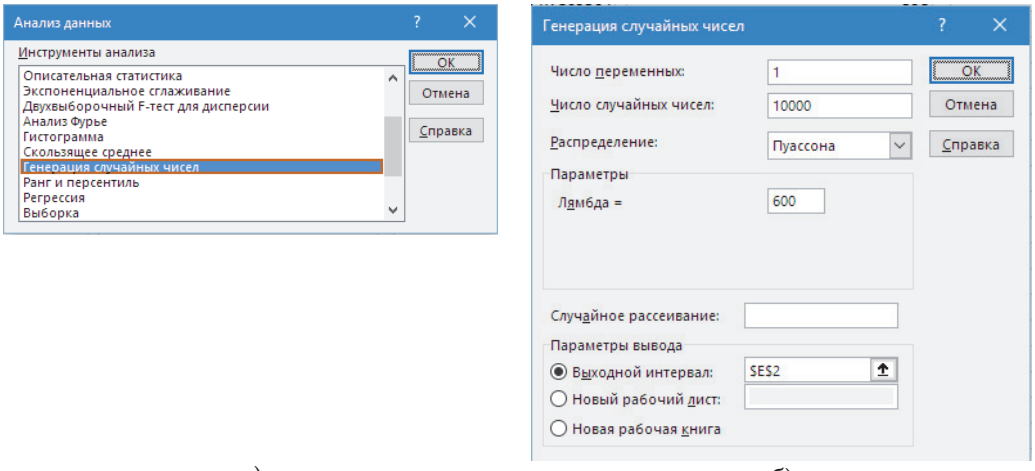
- дискретной случайной величины, заданной своим рядом распределения;
- индикатора события (т. е. случайной величины, принимающей значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$; в программе «Генерация случайных чисел» данный закон распределения назван законом Бернулли);
- биномиальной случайной величины с заданным числом испытаний n и вероятностью p успеха в единичном испытании;
- случайной величины, распределенной по закону Пуассона с параметром λ;
- случайной величины, распределенной по равномерному закону на отрезке $[a; b]$;
- случайной величины, распределенной по нормальному закону с заданным математическим ожиданием a и средним квадратическим отклонением σ.

Пример 4.2.10. Теперь смоделируем число покупателей из предыдущего примера как случайную величину, распределенную по закону Пуассона с математическим ожиданием $\lambda = 600$.

Результаты будем располагать в столбце E на рабочем листе «Нормальная и пуассоновская» в файле «Моделирование случайных величин.xlsx». Введем в ячейку E1 название имитационного эксперимента: «Распределение Пуассона, n = 10 000». Запустим программу «Генерация случайных чисел», для этого нажмем на кнопку «Анализ данных» (в правой части вкладки «Данные»), в меню выбора инструментов анализа данных выберем соответствующий инструмент «Генерация случайных чисел» (рис. 4.2.5, *а*) и нажмем «ОК».

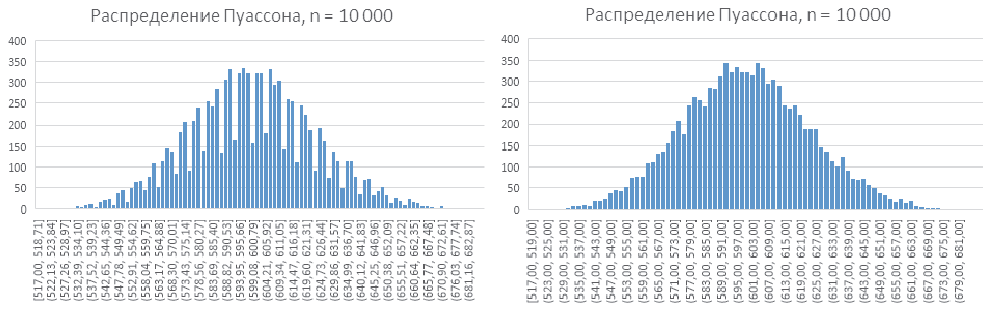
В окне ввода параметров программы «Генерация случайных чисел» (рис. 4.2.5, *б*) укажем число переменных, равное 1 (мы моделируем только одну случайную величину), и число случайных чисел, равное 10000; выберем распределение Пуассона и укажем значение параметра «Лямбда», равное 600. Укажем, что результаты мы хотим получить в выходном интервале, начинающемся с ячейки E2, и нажмем кнопку «ОК». После довольно продолжительных (в зависимости от мощности компьютера) расчетов в диапазоне E2:E10001 появится 10 000 реализаций соответствующей пуассоновской случайной величины.

Построим по этим данным гистограмму (рис. 4.2.6, *а*).



а) *б*)

Рис. 4.2.5. *Моделирование пуассоновской случайной величины: выбор инструмента анализа данных* (*а*); *ввод параметров* (*б*)



а) *б*)

Рис. 4.2.6. *Гистограмма результатов моделирования случайной величины* Pois(600) *при n* = 10 000: *с шириной интервала по умолчанию* (*а*); *с заданной вручную шириной интервала, равной* 2 (*б*)

Видим, что гистограмма распределения Пуассона в целом похожа на «колокол» нормального закона распределения.

Однако высота соседних столбцов гистограммы значительно различается. Это связано с тем, что случайная величина, распределенная по закону Пуассона, является дискретной и принимает целочисленные значения, а автоматически выбранная длина интервала оказалась равной 1,71, т. е. в какие-то из интервалов, например, в интервал (578,56; 580,27), попадает по два целых числа (579 и 580), а в какие-то [например, (599,08; 600,79)] — по одному (600).

Если вручную выбрать длину интервала, равную целому числу, то несбалансированность столбцов исчезнет (на рис. 4.2.6, *б* приведена гистограмма, при построении которой вручную выбрана длина интервала, равная 2). В итоге можно сделать вывод, что при большом числе прибывающих клиентов результаты моделирования числа прибывающих посетителей с помощью нормального распределения и с помощью распределения Пуассона различаются несущественно. ❑

Замечание. Каждый раз при запуске программы «Генерация случайных чисел» получаются различные наборы чисел. Это совершенно естественно, но в некоторых ситуациях может быть не очень удобно: например, при поиске ошибок в вычислениях или при проведении занятий со студентами. Чтобы при каждом запуске получался один и тот же псевдослучайный результат, можно задать определенное значение параметра «Случайное рассеивание» (рис. 4.2.5, *б*). Если значение этого параметра не задано, то оно выбирается автоматически в соответствии с текущим временем, это гарантирует, что каждый раз результаты будут разными.

Кроме того, программа «Генерация случайных чисел» имеет ограничение: за один раз возможно сгенерировать не более, чем 32 767 случайных чисел.

Поэтому программой «Генерация случайных чисел» целесообразно пользоваться в двух случаях: когда требуется обеспечить воспроизводимость экспериментов (с помощью поля «Случайное рассеивание») или когда использование функции, обратной к функции распределения, затруднительно (например, функцию, обратную к функции распределения Пуассона, в Microsoft Excel вычислить довольно трудоемко).

Случайные величины, с которыми приходится иметь дело в реальности, часто имеют сложные законы распределения, чем нормальный, показательный или пуассоновский.

Рассмотрим

Пример 4.2.11. Смоделируем выручку киоска «Руский квас», работающего на пляже в летние месяцы. Ежедневно попить кваса приходит в среднем 600 человек, при этом стандартное отклонение числа любителей кваса равно 30.

Стакан кваса обычно стоит 40 руб. В 30% летних дней стоит очень жаркая погода, и в таких случаях цену повышают до 50 руб. Еще в 30% летних дней на пляже сильный ветер, и в такие дни цену снижают до 35 руб.

Половина покупателей не оставляет чаевых, 20% покупателей дают на чай по 5 руб., еще 20% — по 10 руб., а 10% клиентов оставляют по 25 руб. (квас продает очень красивая девушка).

Количество покупателей в день будем моделировать как случайную величину $X = \mathcal{N}(a = 600; \sigma = 30)$ — это уже было сделано в примере 4.2.9 (соответствующая гистограмма представлена на рис. 4.2.6, *а*).

Цена кваса представляет собой дискретную случайную величину C, заданную рядом распределения

C	35	40	50
p	0,3	0,4	0,3

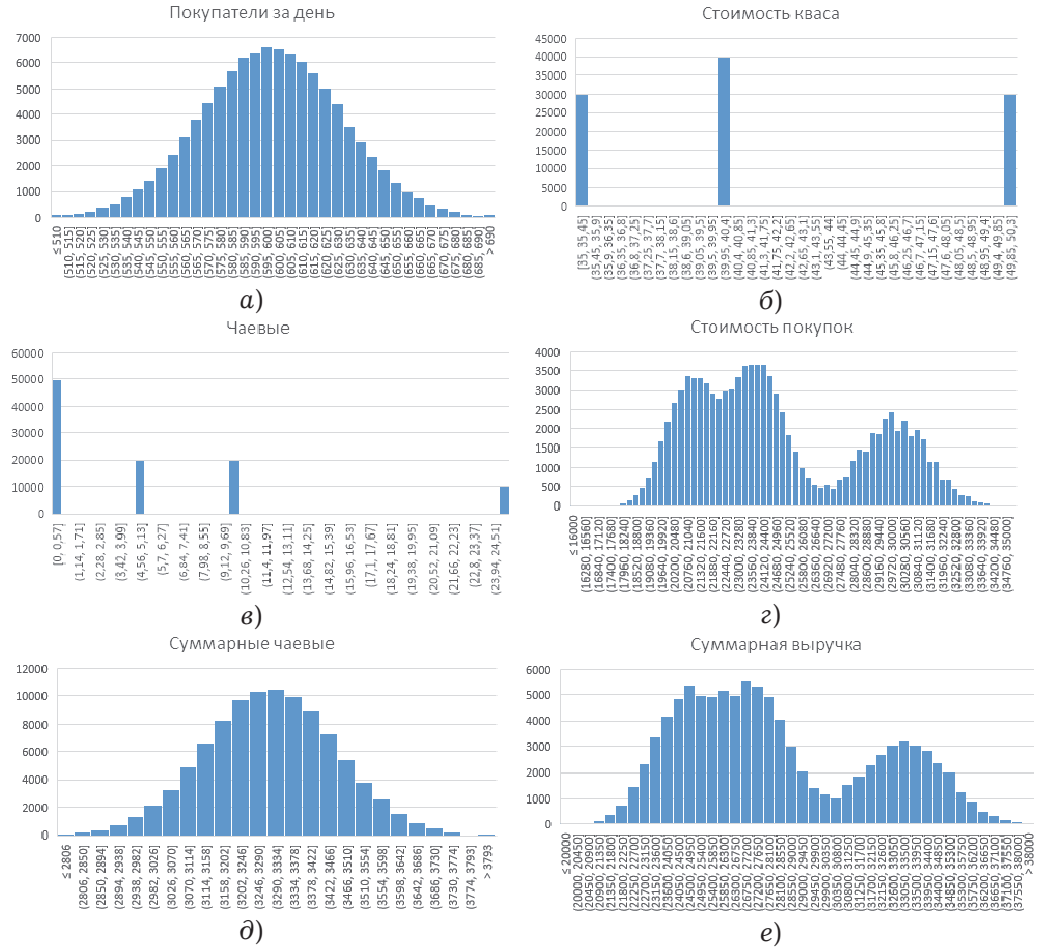


Рис. 4.2.6. *Результаты моделирования выручки от продажи кваса при 600 покупателях в день: гистограмма числа покупателей (а); гистограмма стоимости кваса (б); гистограмма чаевых на одну покупку (в); гистограмма стоимости дневных покупок (г); гистограмма чаевых за день (д); гистограмма суммарной дневной выручки (е)*

Размер чаевых на одну покупку — это случайная величина T:

T	0	5	10	25
p	0,5	0,2	0,2	0,1

Функции распределения случайных величин C и T таковы:

$$F_C(c)=\begin{cases}0, & c<35,\\ 0{,}3, & 35\leqslant c<40,\\ 0{,}7, & 40\leqslant c<50,\\ 1 & c\geqslant 50;\end{cases} \qquad F_T(t)=\begin{cases}0, & t<0,\\ 0{,}5, & 0\leqslant t<5,\\ 0{,}7, & 5\leqslant t<10,\\ 0{,}9, & 10\leqslant t<25,\\ 1 & t\geqslant 25.\end{cases}$$

Математические ожидания цены кваса и размера чаевых на одну покупку равны соответственно $\mathbf{E}(C) = 35{\cdot}0{,}3 + 40{\cdot}0{,}4 + 50{\cdot}0{,}3 = 41{,}5$; $\mathbf{E}(T) = 0{\cdot}0{,}5 + 5{\cdot}0{,}2 + 10{\cdot}0{,}2 + 25{\cdot}0{,}1 = 5{,}5$.

Чтобы смоделировать дискретную случайную величину C, поступим следующим образом. Введем в какую-либо ячейку рабочего листа заголовок «Unif-C», а в ячейку, находящуюся под заголовком, формулу «=СЛЧИС()». Теперь отформатируем эту формулу как таблицу с заголовком (кнопка «Форматировать как таблицу» находится на вкладке «Главная»), а затем на вкладке «Конструктор» нажмем кнопку «Размер таблицы» и укажем такой размер таблицы, чтобы в ней было 100 000 строк (например, если заголовок и формула «СЛЧИС()» находились в диапазоне A1:A2, то новый размер таблицы должен быть таким: «=$A$1:$A$100001».

После этого справа от заголовка «Unif-C» введем новый заголовок «C», а под ним — формулу «=ЕСЛИМН([@Unif-C] < 0,3; 35; [@Unif-C] < 0,7; 40; ИСТИНА; 50)». В новом столбце, таким образом, мы получили 100 000 реализаций случайной величины C — стоимости одной покупки. Соответствующая гистограмма представлена на рис. 4.2.6, *б*).

Аналогичным образом получим 100 000 реализаций случайной величины T — размера чаевых на одну покупку (гистограмма представлена на рис. 4.2.6, *в*).В новых столбцах рассчитаем стоимость покупок за день как произведение числа покупателей на стоимость одной покупки (рис. 4.2.6, *г*), суммарные чаевые за день как произведение числа покупателей на математическое ожидание чаевых на одну покупку[1] (рис. 4.2.6, *д*) и суммарную дневную выручку как сумму стоимости покупок за день и суммарных чаевых за день (рис. 4.2.6, *е*).

В табл. 4.2.3 представлены статистические характеристики набора данных, полученного в результате моделирования: видно, в частности, что средние значения близки к соответствующим математическим ожиданиям.

Замечаем, что стоимость покупок за день и суммарная дневная выручка устроены довольно сложно: они имеют по три пика. Проанализировать эти случайные величины другим способом, отличным от имитационного моделирования, было бы крайне сложно.

Результаты моделирования сохранены на рабочем листе «Русский квас» в файле «Моделирование случайных величин.xlsx».

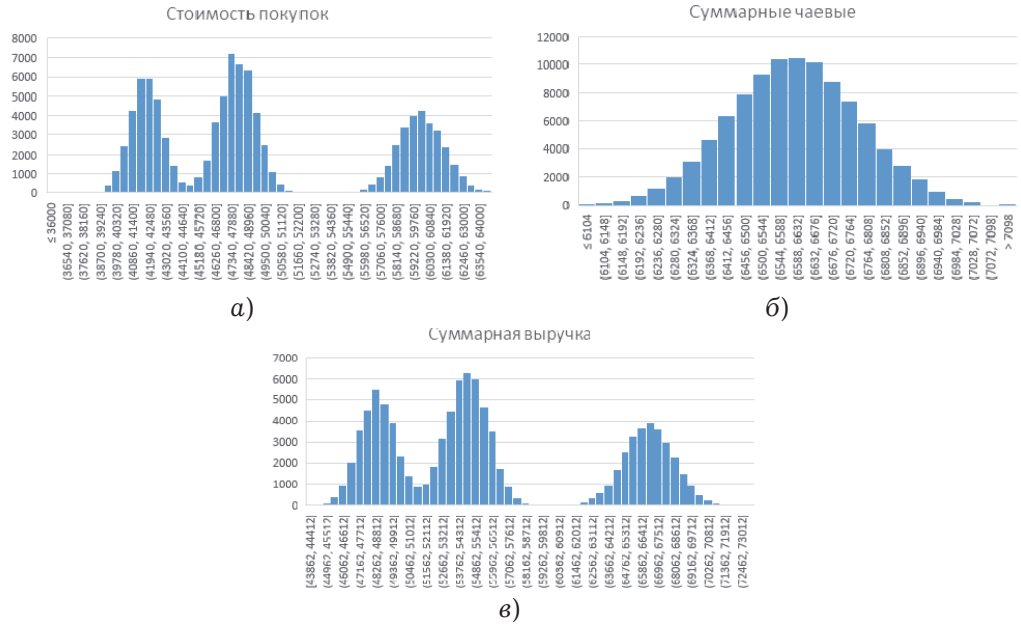
Если бы киоск нанял еще одну девушку, которая раздавала бы на пляже рекламные листовки, в результате чего число покупателей выросло бы вдвое, то распределения стоимости покупок за день, суммарных чаевых за день и суммарной дневной выручки стали бы другими (рис. 4.2.7). Оставляем анализ такой ситуации читателю в качестве задачи **307**. ❑

Таблица 4.2.3

Статистические характеристики результатов моделирования

	Число покупателей	*Цена кваса*	*Чаевые*	*Стоимость покупок*	*Суммарные чаевые*	*Суммарная выручка*
Минимум	466	35	0	17 010	2563	19 683
Максимум	746	50	25	36 650	4103	40 682
Первая квартиль	580	35	0	21 840	3190	25 162
Третья квартиль	620	50	10	28 550	3410	31 691
Медиана	600	40	5	24 000	3300	27 300
Среднее	599,9718	41,5033	5,4978	24901,03	3300,095	28201,12563

[1] Так можно делать благодаря закону больших чисел, поскольку число покупателей в день достаточно велико.



а) *б*) *в*)

Рис. 4.2.7. *Результаты моделирования выручки от продажи кваса при 1200 покупателях в день*: *гистограмма стоимости дневных покупок* (*а*); *гистограмма чаевых за день* (*б*); *гистограмма суммарной дневной выручки* (*в*)

Идеи метода Монте-Карло уходят корнями в далекое прошлое, когда классики теории вероятностей проверяли на практике теоретические результаты. В современном виде и с таким названием метод Монте-Карло был предложен в 1949 г. Н. Метрополисом и С. Уламом. Далее метод активно развивался в работах Н.П. Бусленко, А.А. Вавилова, В.С. Владимирова, И.М. Гельфанда, В.М. Глушкова, С.М. Ермакова, И.Н. Коваленко, Г.И. Марчука, Г.А. Михайлова, Н.Н. Моисеева, Дж. фон Неймана, И.М. Соболя, В.В. Чавчанидзе, Ю.А. Шрейдера и Р. Экхардта. В настоящее время метод Монте-Карло применяется в различных областях математики, экономики, физики и техники.

В качестве примера применения метода Монте-Карло, помимо имитационных моделей, можно назвать ставшие в последнее время популярными в анализе данных так называемые *генетические алгоритмы* — эвристические алгоритмы, используемые для решения задач оптимизации и моделирования путем случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, аналогичных естественному отбору в природе. Важнейшей особенностью генетических алгоритмов является процедура «скрещивания» (или «размножения»), которая производит операцию рекомбинации решений-кандидатов, роль которой аналогична роли скрещивания в живой природе («потомки» наследуют черты обоих «родителей», смешанные определенным образом).

Метод Монте-Карло может быть применен и для решения задач, на первый взгляд далеких от анализа данных. Например, для вычисления определенного интеграла

$$\int_0^a g(t)dt$$

от ограниченной функции (для определенности будем считать, что $0 \leqslant g(t) \leqslant b$) можно многократно смоделировать две случайные величины X и Y, распределенные по равномерному закону на отрезке $[0; a]$ и на отрезке $[0; b]$ соответственно, вычислить относительную частоту $\grave{p}_n$ наступления события $g(X) \leqslant Y$ и определить искомый интеграл из соотношения

$$\grave{p}_n = \frac{\int_0^a g(t)dt}{ab},$$

представляющего собой формулу геометрической вероятности.

Задачи

307. В условиях примера 4.2.10 киоск «Русский квас» нанял промоутера — девушку, раздающую на пляже листовки с рекламой кваса, в результате чего среднее число покупателей кваса в день выросло с 600 до 1200 человек (при $\sigma = 20$). Смоделируйте в этой ситуации стоимость покупок за день, суммарные чаевые за день и суммарную дневную выручку. Оцените, на сколько процентов в результате работы промоутера изменилась средняя дневная выручка, медиана, первая и третья квартиль дневной выручки.

308. Отделения банка посещают в среднем 500 человек в день. Среднее квадратическое отклонение числа посетителей отделений составляет при этом 20 человек. В банке внедрена автоматизированная система оценки работы операционистов: по окончании обслуживания клиенты на специальном устройстве указывают качество обслуживания: 50% клиентов ставят оценку «2» («средне»), 20% ставят оценку «1» («ужасно»), а 30% ставят оценку «3» («прекрасно»). Проведите имитационное моделирование суммарной дневной оценки за день, используя 100 000 реализаций соответствующих случайных величин. Оцените, в каком проценте дней суммарная оценка будет близка к 500, в каком проценте дней она будет близка к 1000 и в каком проценте дней — к 1500.

309. Смоделируйте эксперимент по 1000 подбрасываниям монеты и постройте график зависимости относительной частоты выпадения герба отт количества проведенных опытов (как на рис. 2.2.7).

Глава 5. Оценивание параметров

§ 5.1. Выборочный метод

5.1.1. Генеральная совокупность и выборка

Генеральной совокупностью называют совокупность результатов всех мысленно возможных наблюдений над какой-либо случайной величиной X (в том числе, и повторяющихся), проводимых в одинаковых условиях. Иными словами, генеральная совокупность представляет собой набор всех возможных значений данной случайной величины. Как правило, огромный объем генеральной совокупности не позволяет просто переписать все ее элементы, в таких случаях подвергают изучению ограниченное количество значений, отобранных из всей совокупности.

Выборочной совокупностью (или просто *выборкой*) называют результаты ограниченного числа наблюдений над случайной величиной X. Сущность выборочного метода состоит в том, чтобы по выборке как некоторой части генеральной совокупности делать выводы о всей генеральной совокупности в целом.

В табл. 5.1.1 представлено несколько примеров генеральных совокупностей и выборок из них.

Таблица 5.1.1

Примеры генеральных и выборочных совокупностей

Генеральная совокупность	*Выборочная совокупность*	*Изучаемый параметр*
Все население Российской Федерации	Участники очередной переписи населения	Медиана среднемесячного дохода жителей России
Все взрослое население Российской Федерации	Избиратели, пришедшие на Президентские выборы	Доля взрослых жителей России, которые поддерживают конкретного кандидата
Все жители планеты, владеющие методами интеллектуального анализа данных	Россияне — обладатели сертификата *Microsoft Professional Program for Data Science*	Средний возраст специалистов по наукам о данных
Все клиенты Сбербанка России	Все клиенты Сбербанка России, пришедшие в определенное отделение в конкретный день	Математическое ождидангие и среднее квадратическое отклонение размера вклада клиента Сбербанка
Все процессоры «Байкал-Т1», произведенные компанией «Т-Платформы» в текущем месяце	1000 процессоров «Байкал-Т1», отобранных для контроля качества	Доля бракованных процессоров «Байкал-Т1» (среди всех произведенных процессоров)
Все студенты Финансового университета	Члены Научного студенческого общества Финуниверситета	Математическое ожидание оценки по дисциплине «Анализ данных»
Все игроки Континентальной хоккейной лиги	Случайным образом отобранные 20 игроков КХЛ	Математическое ожидание возраста игрока КХЛ

Выборку называют *репрезентативной*, если она адекватно отражает исследуемые свойства генеральной совокупности. Чтобы выборка была репрезентативной, можно организовать ее следующим образом. Из генеральной совокупности случайным образом отбирается элемент и обследуется, после чего возвращается в общую совокупность и может быть отобран и обследован повторно.

Такая выборка называется *повторной случайной*. В повторной случайной выборке наблюдения $X_1, X_2, \ldots, X_n$ независимы и проводятся в одинаковых (с вероятностной точки зрения) условиях, т. е. распределены по одному и тому же закону: $F_{X_i}(x) = F_X(x), \quad i = 1, 2, \ldots, n$.

Бесповторную случаную выборку организовать, как правило, проще: из генеральной совокупности случайным образом отбирается элемент и обследуется, после чего этот элемент не может быть отобран повторно. В бесповторной случайной выборке наблюдения $X_1, X_2, \ldots, X_n$ являются з а в и с и м ы м и, что может привести к усложнению статистических вычислений. Однако при большом объеме генеральной совокупности N и относительно небольшом (по сравнению с объемом генеральной совокупности) объеме бесповторной случайной выборки $n \ll N$ различиями между бесповторной и повторной случайными выборками можно пренебречь. Например, при исследовании вероятности успеха можно воспользоваться тем, что (как было показано в **п. 2.4.1**) при $N \to \infty, K/N \to p$ гипергеометрический закон распределения, описывающий число успехов в бесповторной случайной выборке, переходит в биномиальный закон распределения, соответствующий повторной случайной выборке:

$$\lim_{\substack{N\to\infty \\ K/N\to p}} \frac{C_K^k C_{N-K}^{n-k}}{C_N^n} = C_n^k p^k (1-p)^{n-k}.$$

Пример 5.1.1. Предположим, что у банка $N = 10\,000$ клиентов, из которых нужно отобрать для детального анализа $n = 1000$. Если организовать случайную повторную выборку, то для каждого клиента вероятность попасть в выборку равна

$$1 - \left(1 - \frac{1}{N}\right)^n = 1 - \left(1 - \frac{1}{10\,000}\right)^{1000} = 0{,}095167106 \approx 9{,}5\%.$$

В случае организации случайной бесповторной выборки аналогичная вероятность составит

$$1 - \frac{N-1}{N}\frac{N-2}{N-1}\frac{N-3}{N-2}\cdots\frac{N-n}{N-(n-1)} = 1 - \frac{N-n}{N} = \frac{n}{N} = \frac{1000}{10\,000} = 0{,}10 = 10\%. \ \square$$

Пример 5.1.2. Очевидно, не все выборки, приведенные в табл. 5.1.1, являются репрезентативными. Например, скорее всего, средняя оценка по анализу данных у членов Научного студенческого общества будет выше, чем средняя оценка по анализу данных по всему университету.

Сбербанк — большой банк с представительствами во всех регионах России. Если речь идет о клиентах, пришедших в одно из московских отделений, то скорее всего, они отличаются от клиентов, пришедших в одно из астраханских отделений, а те, в свою очередь, отличаются от клиентов, пришедших в одно из отделений в Архангельской области.

Чтобы обеспечить репрезентативность, гораздо лучше выбрать случайным образом по одному студенту из каждой группы Финуниверситета или по одному клиенту из каждого отделения Сбербанка. ❑

Пример 5.1.3. В 1936 г., когда претендентами на пост Президента США были Ф. Рузвельт и А. Лэндон, в американском журнале «Литературное обозрение» был опубликован прогноз результатов выборов. Из генеральной совокупности — всех дееспособных американцев — была извлечена выборка, для этого журналисты выбрали 10 млн. адресов, по которым редакция журнала разослала открытки с просьбой высказать свое отношение к кандидатам на пост

президента. Журнал получил почти 2,5 млн. ответов, в 57% из которых предпочтение отдавалось А. Лэндону, а в 40% — Ф. Рузвельту. На основании этих результатов журнал опубликовал прогноз о том, что на предстоящих выборах с большим перевесом победит А. Лэндон. Но прогноз оказался ошибочным: победу одержал Ф. Рузвельт.

Это — типичный пример нерепрезентативной выборки. Дело в том, что в 1930-х гг. в США телефоны были только у зажиточных американцев, поддерживавших, в основном, взгляды А. Лэндона. Гораздо больше американцев не имели телефонов и поддерживали Ф. Рузвельта. ❑

Конкретной выборкой называется конкретный набор чисел $x_1, x_2, \ldots, x_n$, полученный в результате наблюдений над случайной величиной X, т. е. набор, состоящий из n р е а л и з а ц и й случайной величины X.

Число элементов в выборке называется ее *объемом.*

Пример 5.1.4. В файле **«КХЛ.xlsx»** (рис. 5.1.1, *а*) приведен список игроков Континентальной хоккейной лиги.[1]

	A	B	C	D	E	F	G
1	Номер	Фамилия, имя	Амплуа	Дата рождения	Возраст	Гражданство	СЛЧИС
2	1	Аалтонен Миро	нападающий	7 Июня 1993	24	Финляндия	0,31367088
3	2	Аалтонен Юхаматти	нападающий	4 Июня 1985	32	Финляндия	0,43261546
4	3	Абдрахманов Руслан	нападающий	25 Декабря 1984	33	Россия	0,68182273
5	4	Абдуллин Денис	нападающий	1 Января 1985	33	Россия	0,07200816
…	…	…	…	…	…	…	…

а)

	A	B	C	D	E	F	G
1	Номер	Фамилия, имя	Амплуа	Дата рождения	Возраст	Гражданство	СЛЧИС
2	1	Улмер Джефф	нападающий	27 Апреля 1977	40	Канада	0,998993948
3	2	Карцев Денис	нападающий	25 Апреля 1976	41	Россия	0,998611737
4	3	Зедник Рихард	нападающий	6 Января 1976	42	Словакия	0,998448576
5	4	Шуленин Дмитрий	защитник	17 Декабря 1995	22	Россия	0,998063951
6	5	Шёгрен Маттиас	нападающий	27 Ноября 1987	30	Швеция	0,997997704
7	6	Куинт Дерон	защитник	12 Марта 1976	41	США	0,997901511
8	7	Махановский Павел	нападающий	2 Сентября 1993	24	Россия	0,997860876
9	8	Марков Даниил	защитник	30 Июля 1976	41	Россия	0,997025259
10	9	Умарк Линус	нападающий	5 Февраля 1987	30	Швеция	0,996665514
11	10	Шефер Нолан	вратарь	15 Января 1980	38	Канада	0,996646557
12	11	Брукбэнк Шелдон	защитник	3 Октября 1980	37	Канада	0,996192011
13	12	Риссанен Расмус	защитник	13 Июля 1991	26	Финляндия	0,995888703
14	13	Леневё Давид	вратарь	23 Мая 1983	34	Канада	0,995808722
15	14	Мирасти Джон	нападающий	4 Июня 1982	35	Канада	0,995629238
16	15	Барзман Кирилл	вратарь	14 Июля 1993	24	Россия	0,995488777
17	16	Коусал Роберт	нападающий	7 Октября 1990	27	Чехия	0,994022063
18	17	Риддервалль Калле	нападающий	28 Мая 1988	29	Швеция	0,993563333
19	18	Педан Руслан	защитник	16 Ноября 1994	23	Россия	0,993132201
20	19	Делорье Джефф	вратарь	15 Мая 1984	33	Канада	0,992895696
21	20	Наместников Егор	защитник	5 Апреля 1992	25	Россия	0,992832718
22	21	Порядин Павел	нападающий	21 Июля 1996	21	Россия	0,992578559
23	22	Тортунов Борис	вратарь	14 Августа 1973	44	Россия	0,992355649
24	23	Гусманов Равиль	нападающий	25 Июля 1972	45	Россия	0,99131509
25	24	Троттер Брок	нападающий	16 Января 1987	31	Канада	0,99056004
26	25	Раутанен Юхо	защитник	25 Мая 1997	20	Финляндия	0,99046841
27	26	Толпеко Денис	нападающий	29 Января 1985	32	Россия	0,989938679
28	27	Толузаков Филипп	нападающий	10 Января 1989	29	Россия	0,989456211
29	28	Вей Линден	нападающий	17 Июля 1991	26	Канада	0,988948136
30	29	Виклунд Тобиас	защитник	8 Мая 1986	31	Швеция	0,988662558
31	30	Шэн Павел	нападающий	14 Августа 1999	18	Россия	0,988636627

б)

Рис. 5.1.1. *Генеральная совокупность игроков КХЛ по состоянию на январь 2018 г.* (*а*) *и случайная бесповторная выборка из 30 игроков КХЛ* (*б*)

[1] Источник: https://www.khl.ru/players/ (по состоянию на январь 2018 г.).

Всего этот список насчитывает 2936 человек, для каждого из которых приводится фамилия и имя, амплуа (нападающий, защитник или вратарь), дата рождения, возраст и гражданство. Добавим справа столбец «СЛЧИС» из случайных чисел, равномерно распределенных на отрезке [0; 1], отсортируем таблицу по убыванию столбца «СЛЧИС» и выберем первые 30 строк — получим конкретную бесповторную выборку объема 30 из генеральной совокупности объема 2936 (рис. 5.1.1, *б*[2]).

Теперь повторим 30 раз процедуру из двух шагов: отсортируем генеральную совокупность по убыванию столбца «СЛЧИС» и выберем один верхний элемент. Отобранные 30 элементов образуют конкретную повторную выборку объема 30 из той же генеральной совокупности. ❑

Выборочным средним называется величина

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}. \qquad (5.1.1)$$

Эта величина является выборочным аналогом математического ожидания $\mathbf{E}(X)$.

Для вычисления выборочного среднего по конкретной выборке в Microsoft Excel используется функция

$$\overline{x} = \text{СРЗНАЧ}(<\text{выборка } x_1, x_2, \ldots, x_n>). \text{ ❑}$$

Пример 5.1.5. Оценим математическое ожидание возраста игроков Континентальной хоккейной лиги, вычислив выборочное среднее для возраста игроков по выборке, построенной в предыдущем примере (рис. 5.1.1, *б*).

Вначале проведем вычисления вручную по формуле (5.1.1):

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{30}}{30} = (40+41+42+22+34+41+24+41+30+38+ \\ +37+26+34+35+24+27+29+23+33+25+ \\ +21+44+45+31+20+32+29+26+31+18)/30 = 31{,}3.$$

Теперь для получения того же результата воспользуемся функцией СРЗНАЧ: $\overline{x}$ = СРЗНАЧ(E2:E31) = 31,3. ❑

Выборочным аналогом дисперсии $\mathbf{Var}(X)$ является, например, величина

$$\hat{\sigma}_X^2 = \overline{(X-\overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}, \qquad (5.1.2)$$

называемая *выборочной дисперсией.*

По результатам наблюдений двумерной случайной величины $(X_1; Y_1), (X_2; Y_2), \ldots, (X_n; Y_n)$ можно вычислить *выборочную ковариацию*

$$\widehat{\operatorname{cov}}(X,Y) = \overline{(X-\overline{X})(Y-\overline{Y})} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n} \qquad (5.1.3)$$

[2] Очевидно, если читатель проведет такую же процедуру, как в примере, получившаяся выборка будет отличаться от приведенной на рис. 5.1.1, *б*, так как процесс построения выборки основан на вычислении случайных чисел.

и выборочный коэффициент корреляции

$$R(X,Y)=\frac{\widehat{\operatorname{cov}}(X,Y)}{\hat{\sigma}_X\hat{\sigma}_Y}. \tag{5.1.4}$$

Для выборочной дисперсии и выборочной ковариации несложно доказать формулы (предлагаем это сделать читателю в задачах **313** и **314**)

$$\hat{\sigma}_X^2=\overline{X^2}-(\overline{X})^2=\frac{\sum\limits_{i=1}^{n}X_i^2}{n}-\left(\frac{\sum\limits_{i=1}^{n}X_i}{n}\right)^2 \tag{5.1.5}$$

и

$$\widehat{\operatorname{cov}}(X,Y)=\overline{XY}-\overline{X}\cdot\overline{Y}=\frac{\sum\limits_{i=1}^{n}X_iY_i}{n}-\overline{X}\cdot\overline{Y}, \tag{5.1.6}$$

но пользоваться этими формулами можно только в теоретических выкладках, так как при практических расчетах эти формулы обладают существенно большей вычислительной погрешностью, чем их аналоги (5.1.2) и (5.1.3).

Несложно также показать, что выборочный коэффициент корреляции заключен в пределах от -1 до 1 и характеризует близость зависимости между выборками X и Y к линейной (рис. 5.1.2).

Чем ближе точки $(x_i; y_i)$ расположены к некоторой прямой, тем ближе значение модуля выборочного коэффициента корреляции $|\hat{R}(X,Y)|$ к единице, и наоборот, чем ближе $|\hat{R}(X,Y)|$ к единице, тем ближе точки $(x_i; y_i)$ расположены к некоторой прямой. При этом выборочный коэффициент корреляции $\hat{R}(X,Y)$ положителен [отрицателен] тогда и только тогда, когда при увеличении одной из величин X, Y значения другой имеют тенденцию к увеличению [соответственно, к уменьшению], т. е. прямая, около которой расположены точки $(x_i; y_i)$, имеет положительный [соответственно, отрицательный] наклон.

Значение выборочного коэффициента корреляции, близкое к нулю, означает отсутствие линейной связи между переменными, но при этом может наблюдаться сильная нелинейная корреляционная зависимость (как на рис. 5.1.1, *е*, где точки расположены близко к некоторой п а р а б о л е) или статистическая связь, которую в виде функциональной зависимости переменных представить невозможно, так, на рис. 5.1.1, *ж* приведен пример статистической зависимости, когда с увеличением x растет разброс точек вдоль оси y, однако никакой функцией $y=f(x)$ (линейной или нелинейной) такую зависимость выразить нельзя.

Пример 5.1.6. Вычислим выборочный коэффициент корреляции по результатам наблюдений двумерной случайной величины

x_i	3	4	6	7
y_i	5	10	9	12

Для этого последовательно вычислим $\overline{x}, \overline{y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\operatorname{cov}}(X,Y)$, заполняя столбцы табл. 5.1.2.

Получим следующие значения: $\overline{x}=5$; $\overline{y}=6{,}5$; $\sigma_X^2=\overline{(x-\overline{x})^2}=2{,}5$; $\hat{\sigma}_Y^2=\overline{(y-\overline{y})^2}=6{,}5$; $\widehat{\operatorname{cov}}(X,Y)=\overline{(x-\overline{x})(y-\overline{y})}=3{,}25$.

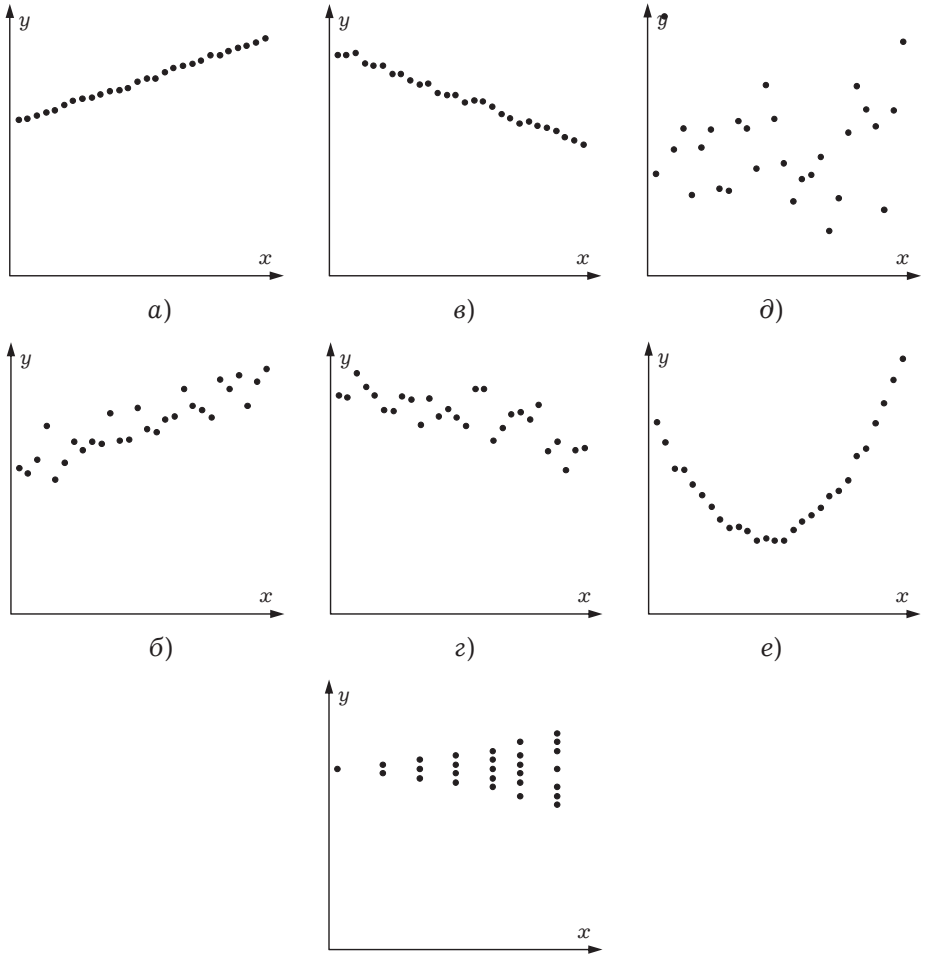


Рис. 5.1.2. *Виды зависимости между выборками:*
сильная линейная прямая зависимость, $\hat{R}(X,Y) \approx 1$ *(а);*
слабая линейная прямая зависимость, $\hat{R}(X,Y) \approx -0{,}5$ *(б);*
сильная линейная обратная зависимость, $\hat{R}(X,Y) \approx -1$ *(в);*
слабая линейная обратная зависимость, $\hat{R}(X,Y) \approx -0{,}5$ *(г);*
отсутствие зависимости, $\hat{R}(X,Y) \approx 0$ *(д);*
сильная нелинейная зависимость, $\hat{R}(X,Y) \approx 0$ *(е);*
зависимость, не являющаяся корреляционной, $\hat{R}(X,Y) \approx 0$ *(ж)*

Таблица 5.1.2

Расчет выборочных характеристик в примере 5.1.6

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
3	5	−2	-4	4	16	8
4	10	−1	1	1	1	−1
6	9	1	0	1	0	0
7	12	2	3	4	9	6
$\overline{x}=5$	$\overline{y}=9$			$\overline{(x-\overline{x})^2}=2{,}5$	$\overline{(y-\overline{y})^2}=6{,}5$	$\overline{(x-\overline{x})(y-\overline{y})}=3{,}25$

Теперь находим $\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2} \approx 1{,}58$; $\hat{\sigma}_Y = \sqrt{\hat{\sigma}_Y^2} = 2{,}55$ и подставляем рассчитанные значения в формулу (5.1.4):

$$R(X, Y) = \frac{\widehat{\mathrm{cov}}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{3{,}25}{1{,}58 \cdot 2{,}55} \approx 0{,}81.$$

Значение выборочного коэффициента корреляции, равное 0,81, позволяет считать, что между X и Y имеется достаточно сильная линейная зависимость. ❑

Кроме повторной случайной выборки и бесповторной случайной выборки бывают и другие способы организации выборок.

Например, *систематическая случайная выборка* объемом в n элементов из генеральной совокупности объема N предполагает, что сначала случайным образом выбирается первый элемент в списке всех объектов из генеральной совокупности, а далее выбираются объекты, отстоящие от выбранного первого элемента на $\pm[n/N]$, $\pm 2[n/N]$, $\pm 3[n/N]$ и т. д.

Вообще говоря, использование систематических выборок может приводить к серьезным ошибкам.

Пример 5.1.7. Предположим, что каждая десятая изготавливаемая заводом деталь является бракованной (т. е. брак составляет 10% всей продукции). Если мы неслучайным образом будем выбирать для контроля качества каждую десятую деталь, то либо каждый раз нам будет попавдаться бракованная деталь, и мы сделаем заключение о том, что брак составляет 100% производимых деталей являются бракованными, либо каждый раз нам будет попадаться деталь надлежащего качества, и мы сделаем вывод, что продукции вообще не содержит брака. ❑

Все же, как правило, систематические случайные выборки репрезентативны, поскольку очень маловероятно (хотя и возможно), чтобы какие-то характеристики повторялись только на каждом $[n/N]$-м объекте.

Стратифицированная выборка позволяет улучшить точность результатов за счет разделения всей генеральной совокупности на несколько областей — так называемых *страт* — и независимой работе с этими стратами.

Пример 5.1.8. Предположим, что в стране 3 города: в городе А живет 1 млн. заводских рабочих, в городе Б живут 2 млн. офисных работников, а в городе В живут 3 млн. пенсионеров, и нам нужно оценить среднее число голосов, поданных за кандидата на выборах. Можно организовать повторную или бесповторную случайную выборку объемом в 60 голосов со всей генеральной совокупности, но есть некоторая вероятность, что такая выборка окажется плохо сбалансированной по городам, следовательно, будет необъективной и повлечет значительную погрешность оценки.

Если же мы построим отдельно три повторных или бесповторных случайных выборки в 10, 20 и 30 голосов из городов А, Б и В соответственно, а затем объединим эти три выборки, мы можем получить меньшую погрешность оценки при том же объеме выборки. ❑

Пример 5.1.9. Предположим, что мы изучаем потенциальный спрос на новую кофеварку, зная, что 40% всех потенциальных покупателей — это женщины, а 60% — мужчины, причем женщины гораздо серьезнее, чем мужчины, относятся к выбору цвета кофеварки. В этом случае в выборку объемом в 100 клиентов может быть целесообразным включить 70 случайным образом выбранных женщин и 30 мужчин. ❑

Пример 5.1.10. Предположим, что в задаче кредитного скоринга мы выясняем, какие факторы влияют на вероятность невозврата кредита, при этом 80% клиентов банка полностью исполнили свои обязательства по кредитным договорам, а 20% — нет. В такой ситуации может быть целесооб-

разным отобрать случайным образом половину из тех клиентов, которые не вернули кредит, и восьмую часть из тех, которые вернули. ❑

Процедура организации *«удобной» выборки* состоит в установлении контактов с «удобными» единицами выборки — со студенческой группой или, например, с друзьями и соседями. Если необходимо получить информацию о реакции людей на новую концепцию или продукт, такая выборка вполне обоснована, однако никакой гарантии репрезентативности в такой выборке быть не может.

Задачи

310. Вычислите выборочные средние $\overline{x}, \overline{y}$ и выборочный коэффициент корреляции $\hat{R}(X,Y)$ по результатам наблюдений двумерного случайного вектора:

x_i	−2	−1	0	1	2
y_i	4	1	0	1	4

311. В условиях примера 1.1.1 организуйте случайную бесповторную выборку 1000 выданных кредитов и вычислите выборочные средние размера выданного кредита и готового дохода заемщика.

312. В условиях примера 5.1.4 постройте десять случайных повторных выборок и вычислите для каждой из них выборочное среднее возраста игрока.

313. Докажите формулу (5.1.5).

314. Докажите формулу (5.1.6).

5.1.2. Допустимый объем выборки для обеспечения ее репрезентативности

Из теоремы Чебышёва (4.1.4) следует, что если $\mathbf{Var}(X) < B$, то

$$\mathbf{P}\{|\overline{X} - \mathbf{E}(X)| < \varepsilon\} \geqslant 1 - \frac{B}{n\varepsilon^2}, \tag{5.1.7}$$

поэтому при объеме выборки

$$n > \frac{B}{(1-\gamma)\varepsilon^2} \tag{5.1.8}$$

с вероятностью, большей γ, выполняется неравенство $|\overline{X} - \mathbf{E}(X)| < \varepsilon$, т. е. гарантируется меньшая, чем ε, ошибка репрезентативности при замене математического ожидания $\mathbf{E}(X)$ выборочным средним $\overline{X}$.

Пусть выборка состоит из n испытаний Бернулли, в которых произошло m успехов. Тогда выборочным аналогом вероятности успеха является *относительная частота*

$$\hat{p} = \frac{m}{n}.$$

Из теоремы Бернулли (4.1.6) следует, что

$$\mathbf{P}\{|\hat{p} - p| < \varepsilon\} \geqslant 1 - \frac{p(1-p)}{n\varepsilon^2},$$

поэтому при объеме выборки

$$n > \frac{p(1-p)}{(1-\gamma)\varepsilon^2} \tag{5.1.9}$$

с вероятностью, большей γ, выполняется неравенство $|\hat{p} - p| < \varepsilon$, т. е. гарантируется меньшая, чем ε, ошибка репрезентативности при замене вероятности

p относительной частотой $\hat{p}$. Поскольку для любых $p \in [0,1]$ $p(1-p) < 1/4$, то при неизвестной p неравенство для n можно заменить на

$$n > \frac{1}{4(1-\gamma)\varepsilon^2}. \tag{5.1.10}$$

Если математическое ожидание случайной величины X равно a, ее среднее квадратичное отклонение равно σ, а объем выборки велик, то согласно следствию из теоремы Леви (см. **п. 4.2.1**) случайная величина

$$Z = \frac{(\overline{X} - a)\sqrt{n}}{\sigma}$$

имеет стандартный нормальный закон распределения $\mathcal{N}(0; 1)$, поэтому

$$\mathbf{P}\left\{\left|\overline{X} - \mathbf{E}(X)\right| < \varepsilon\right\} = \mathbf{P}\left\{-\frac{\varepsilon}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mathbf{E}(X)}{\sigma/\sqrt{n}} < \frac{\varepsilon}{\sigma/\sqrt{n}}\right\} = F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - F_{\mathcal{N}(0;1)}\left(-\frac{\varepsilon}{\sigma/\sqrt{n}}\right).$$

Обсудим смысл выражения «объем выборки достаточно велик». Как правило, для подавляющего большинства генеральных совокупностей распределение выборочных средних становится приближенно нормальным при $n = 30$. Однако если распределение генеральной совокупности обладает сильной асимметрией или имеет несколько мод, объем выборок следует увеличить.

Практику применения центральной предельной теоремы к различным генеральным совокупностям иллюстрирует рис. 5.1.3, на котором представлены графики плотностей распределения выборочных средних для различных генеральных совокупностей при разных объемах выборки. Видим, что если генеральная совокупность распределена по равномерному закону, то при $n = 5$ закон распределения выборочных средних становится приближенно нормальным. Для показательного распределения, демонстрирующего ярко выраженную положительную асимметрию, при $n = 5$ распределение выборочных средних остается асимметричным, но при $n = 30$ также становится очень близким к нормальному.

Учтем, что «колокол» распределения стандартной нормальной случайной величины $\mathcal{N}(0;1)$ симметричен относительно оси ординат, поэтому $F_{\mathcal{N}(0;1)}(-x) = 1 - F_{\mathcal{N}(0;1)}(x)$, и

$$\mathbf{P}\left\{\left|\overline{X} - \mathbf{E}(X)\right| < \varepsilon\right\} = F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - F_{\mathcal{N}(0;1)}\left(-\frac{\varepsilon}{\sigma/\sqrt{n}}\right) = \tag{5.1.11}$$

$$= F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - \left(1 - F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right)\right) = 2F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - 1.$$

Пусть объем выборки

$$n > \frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2},$$

где $z_{(1+\gamma)/2}$ — квантиль уровня $(1+\gamma)/2$ случайной величины $Z = \mathcal{N}(0; 1)$, т.е. такое число, что $F_{\mathcal{N}(0;1)}\left(z_{(1+\gamma)/2}\right) = (1+\gamma)/2$. Тогда

$$\frac{\varepsilon}{\sigma/\sqrt{n}} > \frac{\varepsilon}{\sigma}\sqrt{\frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2}} = z_{(1+\gamma)/2},$$

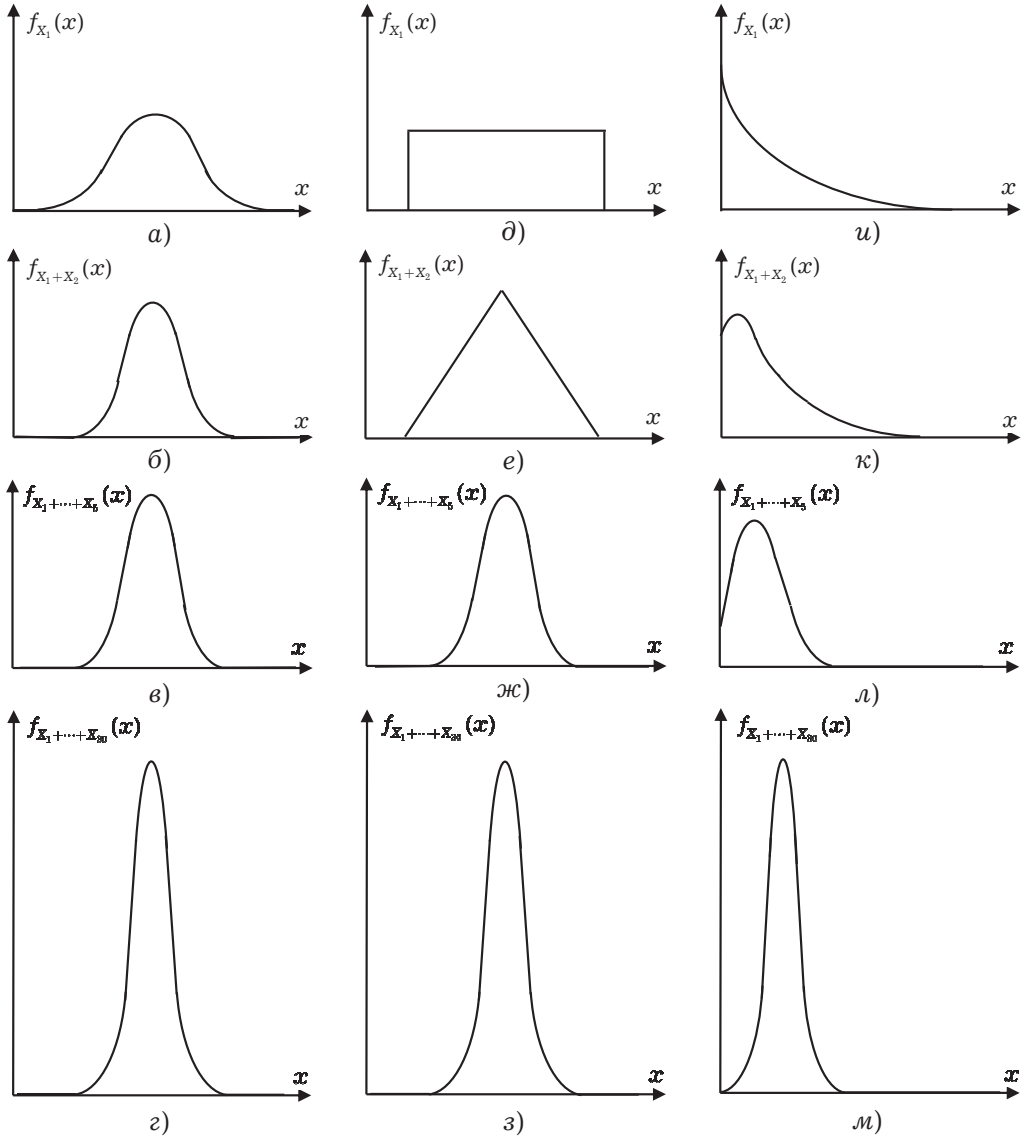


Рис. 5.1.3. *Распределения выборочных средних для различных генеральных совокупностей при различных объемах выборки: нормальное распределение при $n = 1$ (а), $n = 2$ (б), $n = 5$ (в), $n = 30$ (г); равномерное распределение при $n = 1$ (д), $n = 2$ (е), $n = 5$ (ж), $n = 30$ (з); показательное распределение при $n = 1$ (и), $n = 2$ (к), $n = 5$ (л), $n = 30$ (м)*

и учитывая, что функция распределения является возрастающей, можно записать, что

$$\frac{\varepsilon}{\sigma/\sqrt{n}} > \frac{\varepsilon}{\sigma}\sqrt{\frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2}} = z_{(1+\gamma)/2}, \qquad F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) > F_{\mathcal{N}(0;1)}\left(z_{(1+\gamma)/2}\right) = \frac{1+\gamma}{2}.$$

Пусть объем выборки

$$n > \frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2},$$

где $z_{(1+\gamma)/2}$ — квантиль уровня $(1+\gamma)/2$ случайной величины $Z=\mathcal{N}(0;1)$, т.е. такое число, что $F_{\mathcal{N}(0;1)}\left(z_{(1+\gamma)/2}\right)=(1+\gamma)/2$. Тогда

$$\frac{\varepsilon}{\sigma/\sqrt{n}}>\frac{\varepsilon}{\sigma}\sqrt{\frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2}}=z_{(1+\gamma)/2},$$

и учитывая, что функция распределения является возрастающей, можно записать, что

$$\frac{\varepsilon}{\sigma/\sqrt{n}}>\frac{\varepsilon}{\sigma}\sqrt{\frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2}}=z_{(1+\gamma)/2},\qquad F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right)>F_{\mathcal{N}(0;1)}\left(z_{(1+\gamma)/2}\right)=\frac{1+\gamma}{2}.$$

Поэтому

$$\mathbf{P}\left\{\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon\right\}=2F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right)-1>2\cdot\frac{1+\gamma}{2}-1=\gamma.$$

Значит, при объеме выборки

$$n>\frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2} \tag{5.1.12}$$

с вероятностью, большей γ, выполняется неравенство $\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon$, т. е. гарантируется меньшая, чем ε, ошибка репрезентативности при замене математического ожидания $\mathbf{E}(X)$ выборочным средним $\overline{X}$.

Пример 5.1.11. Предположим, что нам известно, что дисперсия случайной величины X не превышает 10.

Оценим вероятность того, что отклонение выборочного среднего $\overline{X}=\sum_{i=1}^{16\,000}X_i\Big/16\,000$, рассчитанного по 16 000 результатов наблюдений случайной величины (независимых, проведенных в одинаковых с вероятностной точки зрения условиях) от математического ожидания $\mathbf{E}(X)$ не превысит 0,25.

По формуле (5.1.7), в которой $n=16\,000$, $B=10$, $\varepsilon=0{,}25$, получаем:

$$\mathbf{P}\left\{\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon\right\}\geqslant 1-\frac{B}{n\varepsilon^2}=1-\frac{10}{16\,000\cdot 0{,}25^2}=0{,}99.$$

Если обратить внимание на то, что n велико, то согласно (5.1.11)

$$\mathbf{P}\left\{\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon\right\}=2F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right)-1\geqslant 2F_{\mathcal{N}(0;1)}\left(\frac{\varepsilon}{\sqrt{B}/\sqrt{n}}\right)-1.$$

Вычислим это значение в Microsoft Excel:

$$\mathbf{P}\left\{\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon\right\}=\text{2*НОРМ.СТ.РАСП(0,25/КОРЕНЬ(10)*КОРЕНЬ(16000)) – 1 =}$$
$$=1{,}0000000000000000000000000000000.$$

Таким образом, вероятность того, что выборочное среднее отличается от математического ожидания случайной величины X меньше, чем на 0,25, с точностью до 30 знаков равно единице. Это значит, что можно быть практически уверенным в том, что ошибка репрезентативности при замене математического ожидания $\mathbf{E}(X)$ выборочным средним $\overline{X}$ будет меньше 0,25.

Выясним теперь, сколько следует провести наблюдений, чтобы с вероятностью, не меньшей 0,99, можно было утверждать, что ошибка репрезентативности при замене математического ожидания $\mathbf{E}(X)$ выборочным средним $\overline{X}$ не превысит 0,2.

По формуле (5.1.8), в которой $B=10, \gamma=0{,}99, \varepsilon=0{,}2$, получим:

$$n > \frac{B}{(1-\gamma)\varepsilon^2} = \frac{10}{(1-0{,}99)0{,}2^2} = 25\,000.$$

Предположив же нормальность распределения случайной величины X и воспользовавшись формулой (5.1.12), в которой

$$z_{(1+\gamma)/2} = z_{(1+0{,}99)/2} = z_{0{,}995} = \text{НОРМ.СТ.ОБР(0,995)} = 2{,}575829304,$$

получим:

$$n > \frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2} = \frac{10 \cdot 2{,}575829304^2}{0{,}2^2} = 1625{,}6 \text{ —}$$

значительно меньше 25 000. ❑

Видим, что формулы (5.1.7) и (5.1.8) «более осторожные», чем (5.1.11) и (5.1.12). Если известно, что изучаемая генеральная совокупность имеет нормальный закон распределения либо объем выборки достаточно велик, чтобы центральная предельная теорема давала основания предполагать нормальный закон распределения выборочных средних, то, конечно, следует для определения минимального допустимого объема выборки использзовать формулы (5.1.11) и (5.1.12).

При большом числе испытаний Бернулли n и не очень малой вероятности p (такой, что $np > 10$) в силу интегральной теоремы Муавра — Лапласа (см. **п. 4.2.2**) число успехов X в последовательности из n испытаний распределено примерно по нормальному закону с параметрами $a = np$ и $\sigma = \sqrt{np(1-p)}$, относительная частота $\hat{p} = X/n$ имеет нормальный закон распределения с параметрами $a = p, \sigma = \sqrt{p(1-p)/n}$, поэтому при объеме выборки

$$n > \frac{z^2_{(1+\gamma)/2} p(1-p)}{\varepsilon^2} \tag{5.1.13}$$

с вероятностью, большей γ, выполняется неравенство $|\hat{p} - p| < \varepsilon$, т. е. гарантируется меньшая, чем ε, ошибка репрезентативности при замене вероятности p относительной частотой $\hat{p}$.

Пример 5.1.12. Известно, что в среднем из каждой тысячи кредитов, выданных на развитие малого предпринимательства, 30 не возвращаются. Определим, сколько нужно отобрать предприятий, чтобы с вероятностью, не меньшей 0,9, можно было бы ожидать, что доля предприятий в выборке, не возвращающих кредиты, будет отличаться от доли аналогичных предприятий в генеральной совокупности меньше, чем на 0,01.

По формуле (5.1.9), в которой $p = 30/1000 = 0{,}03$, $\gamma = 0{,}9$, получаем:

$$n > \frac{p(1-p)}{(1-\gamma)\varepsilon^2} = \frac{0{,}03 \cdot 0{,}97}{(1-0{,}9)0{,}01^2} = 2910.$$

По формуле (5.1.13), учитывая, что

$$z_{(1+\gamma)/2} = z_{(1+0{,}9)/2} = z_{0{,}95} = \text{НОРМ.СТ.ОБР(0,95)} = 1{,}644853627,$$

получим менее осторожный результат:

$$n > \frac{z^2_{(1+\gamma)/2} p(1-p)}{\varepsilon^2} = \frac{1{,}644853627^2 \cdot 0{,}03 \cdot 0{,}97}{0{,}01^2} = 787{,}3131451.$$

При этом, поскольку $n > 787{,}3131451$, $np > 787{,}3131451 \cdot 0{,}03 = 23{,}61939435 > 10$, есть основания пользоваться интегральной теоремой Муавра — Лапласа. Таким образом, необходимо отобрать 788 предприятий. ❑

Задачи

315. Определите количество респондентов, которых необходимо опросить, чтобы рейтинг Президента (доля граждан, поддерживающих Президента), вычисленный по выборке, с вероятностью, не меньшей 0,99, отличался от истинного рейтинга президента для всех жителей страны не более чем на 5% по абсолютной величине.

316. Компания, управляющая зданиями, желает по выборке оценить среднюю стоимость эксплуатации квартир определенного типа с надежностью 99% и ошибкой репрезентативности ±10 ден. ед. Определите объем выборки, необходимой для такой оценки, если из подобного же исследования, проведенного ранее, известно, что среднее квадратичное отклонение стоимости эксплуатации не превышает 50 ден. ед.

317. По данным 1970 г. на каждую тысячу работников всех специальностей приходился 31 человек с высшим или средним специальным образованием. Определите, сколько нужно отобрать работников, чтобы с вероятностью, не меньшей 0,9, можно было ожидать, что доля лиц с высшим и средним специальным образованием в выборке будет отличаться от доли таких лиц в генеральной совокупности не более, чем на 0,01.

318. Исследованиями установлено, что 20% школьников не знают правил уличного движения. Определите с гарантией 95% количество учеников из отобранных 1600, которые знают правила уличного движения.

319. При каждой пятой аварии на некотором производстве имеются случаи тяжелого травматизма. Определите вероятность того, что при 100 авариях доля несчастных случаев с тяжелым травматизмом находится: *а*) ниже 0,3; *б*) выше 0,2; *в*) между 0,15 и 0,25.

320. Оцените вероятность того, что число девочек среди 3000 новорожденных будет отличаться от ожидаемого числа не более, чем на 55 чел., и определить, сколько следует отобрать новорожденных, чтобы с вероятностью $\gamma > 0{,}99$ быть уверенным в том, что доля девочек среди них отличается от истинной вероятности рождения девочки не более чем на ± 0,02.

321. Выборочным путем требуется оценить среднее значение коэффициента интеллекта *IQ* по тесту Стенфорда — Бине для достаточно многочисленной совокупности обычных людей. Дисперсия *IQ* в каждом случае не превышает 17. Определите: *а*) у скольких людей необходимо измерить *IQ*, чтобы с вероятностью $\gamma > 0{,}95$ можно было утверждать, что среднее значение коэффициента интеллекта в отобранной группе будет отличаться от среднего значения *IQ* во всей совокупности не более, чем на ± 2; *б*) как изменится результат, если такую же ошибку репрезентативности требуется гарантировать с (меньшей) надежностью $\gamma > 0{,}9$; *в*) как изменится результат при увеличении ошибки репрезентативности до ± 4.

322. Найдите объем выборки, при котором с надежностью 0,85 погрешность оценивания математического ожидания нормальной генеральной совокупности выборочным средним не превысит 0,35, если известно, что среднее квадратическое отклонение генеральной совокупности равно 2,5.

323. Систематическая ошибка измерительного прибора равна нулю, а случайная распределена по нормальному закону со средним квадратическим отклонением $\sigma = 2$ мм. Определите количество независимых измерений, которые

необходимо произвести, чтобы определить истинный размер измеряемой детали с ошибкой, не превышающей 1,5 мм, при 90%-ной надежности.

324. По 16 отделениям банка вычислено выборочное среднее случайной величины X — ежемесячного суммарного вклада населения в отделение. Предполагая, что X — нормально распределенная случайная величина со средним квадратическим отклонением $\sigma = 200$ тыс. ден. ед., найдите максимальную погрешность оценивания математического ожидания ежемесячного суммарного вклада выборочным средним, которая гарантируется с надежностью 0,95.

5.1.3. Оценка функции распределения и плотности распределения

Расположив элементы выборки $x_1, x_2, \ldots, x_n$ в порядке неубывания, получим *вариационный ряд*. Далее будем считать, что именно в таком порядке уже расставлены выборочные наблюдения. *Выборочной случайной величиной* называется при этом дискретная случайная величина $\hat{X}$, задаваемая рядом распределения

$$\begin{array}{c|cccc} \hat{X} & x_1' & x_2' & \cdots & x_l' \\ \hline p & \hat{p}_1 & \hat{p}_2 & \cdots & \hat{p}_l \end{array} \qquad (5.1.14)$$

в котором $x_i'\,(i=1,2,\ldots,l)$ — это *варианты*[3] (расположенные по возрастанию различные элементы выборки), а $\hat{p}_i = m_i / n \;\; (i=1,2,\ldots,l)$ — отвечающие этим значениям *относительные частоты* (здесь m_i — *частота* варианты x_i', т. е. количество ее появлений в статистическом ряде распределения).

При этом, очевидно, $n = \sum_{i=1}^{l} m_i$.

Иногда ряд распределения (5.1.14) выборочной случайной величины называют *статистическим рядом распределения*.

Выборочное среднее и выборочную дисперсию можно при этом вычислить по формулам

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{l} x_i' \hat{p}_i = \frac{\sum_{i=1}^{l} x_i' m_i}{n}; \qquad (5.1.15)$$

$$\hat{\sigma}_X^2 = \sum_{i=1}^{k} (x_i' - \overline{x})^2 \hat{p}_i = \sum_{i=1}^{k} (x_i')^2 \hat{p}_i - (\overline{x})^2. \qquad (5.1.16)$$

В Microsoft Excel $\overline{x}$ =СРЗНАЧ(<массив x_i>), $\hat{\sigma}_X^2$ =ДИСП.Г(<массив x_i>).

Пример 5.1.13. Ежедневные суммарные денежные вклады населения (в тыс. руб.) в отделение банка в течение последних 20 рабочих дней были такими:

600; 200; 700; 700; 300; 200; 500; 500; 400; 600;
300; 400; 300; 500; 500; 600; 500; 600; 400; 400.

Расположив элементы выборки в порядке возрастания, получим вариационный ряд:

[3] Варианта — слово женского рода.

200; 200; 300; 300; 300; 400; 400; 400; 400; 500;
500; 500; 500; 500; 600; 600; 600; 600; 700; 700.

Построим теперь выборочную случайную величину (5.1.12):

$\hat{X}$	200	300	400	500	600	700
p	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{4}{20}$	$\frac{5}{20}$	$\frac{4}{20}$	$\frac{2}{20}$

или

$\hat{X}$	200	300	400	500	600	700
p	0,10	0,15	0,20	0,25	0,20	0,10

Средний суммарный дневной вклад оценим с помощью выборочного среднего (5.1.13):

$$\overline{x} = \sum_{i=1}^{l} x_i' \hat{p}_i = 200 \cdot \frac{2}{20} + 300 \cdot \frac{3}{20} + 400 \cdot \frac{4}{20} + 500 \cdot \frac{5}{20} + 60 \cdot \frac{4}{20} + 700 \cdot \frac{2}{20} = \frac{9200}{20} = 460,$$

а дисперсию — с помощью выборочной дисперсии (5.1.14):

$$\hat{\sigma}_X^2 = \sum_{i=1}^{k} (x_i')^2 \hat{p}_i - \left(\sum_{i=1}^{k} x_i' \hat{p}_i \right)^2 = 40\,000 \cdot \frac{2}{20} + 90\,000 \cdot \frac{3}{20} + 160\,000 \cdot \frac{4}{20} + 250\,000 \cdot \frac{5}{20} +$$

$$+ 360\,000 \cdot \frac{4}{20} + 490\,000 \cdot \frac{2}{20} - 460^2 = \frac{4\,660\,000}{20} - 211\,600 = 233\,000 - 211\,600 = 21\,400,$$

при этом $\hat{\sigma}_X = \sqrt{21\,400} = 146{,}29$. ❑

Для н е п р е р ы в н ы х случайных величин при достаточно больших объемах выборки n вместо выборочной случайной величины используют *интервальный вариационный ряд*

$\widehat{X}$	$[a_1; a_2)$	$[a_2; a_3)$	$\cdots$	$[a_\nu; a_{\nu+1})$
p	$\hat{p}_1$	$\hat{p}_2$	$\cdots$	$\hat{p}_\nu$

(5.1.17)

Для этого выбирается определенная длина интервала Δ, и границы интервалов $[a_i; a_{i+1})$ рассчитываются по правилу: $a_1 = x_{(\min)}, a_2 = a_1 + \Delta, a_3 = a_2 + \Delta, \ldots$ (здесь $x_{(\min)}$ — минимальный элемент выборки, а $x_{(\max)}$ — максимальный); формирование интервалов заканчивается, как только для конца $a_{\nu+1}$ очередного интервала выполняется условие $a_{\nu+1} > x_{(\max)}$. Значения $\hat{p}_i = m_i / n$ — это *выборочные интервальные относительные частоты* [здесь m_i — число вариант, попавших в i-й интервал ($i = 1, 2, \ldots, \nu$)].

Напомним (см. **п. 1.1.2**), что *гистограмма* — это фигура, состоящая из прямоугольников, в основании которых лежат интервалы группирования признака $[a_j; a_{j+1})$, а высота каждого прямоугольника соответствует числу вариант m_i, попавших в соответствующий интервал.

Выборочным аналогом плотности распределения $f_X(x)$ случайной величины X служит *выборочная плотность распределения*, рассчитываемая по интервальному вариационному ряду как

$$\hat{f}_X(x) = \hat{p}_i / \Delta \quad \text{при } x \in [a_i; a_{i+1}), i = 1, 2, \ldots, \nu, \tag{5.1.18}$$

при этом ломаная с вершинами в точках $\left(x_i'; \hat{p}_i / \Delta\right)$, где через $x_i' = (a_i + a_{i+1})/2$ обозначены середины интервалов, называется *полигоном частот. Кривая распределения относительных частот* — это ломаная с вершинами $\left(a_{i+1}, \hat{f}_X(a_{i+1})\right)$.

Замечание. Мы уже обсуждали в **п. 1.1.2**, что длину интервала группирования Δ целесообразнее всего подбирать опытным путем таким образом, чтобы гистограмма и полигон выглядели наиболее гладкими.

Зная выборочную плотность распределения, легко построить *выборочную функцию распределения*

$$\hat{F}_X(x)=\begin{cases}0, & x<a_1,\\ \sum\limits_{k=1}^{i}\hat{p}_k, & a_i\leqslant x<a_{i+1}\quad (i=1,2,\ldots,\nu),\\ 1, & x\geqslant a_{\nu+1},\end{cases}\tag{5.1.19}$$

при этом ломаная с вершинами в точках $\left(x_i';\sum\limits_{k=1}^{i}\hat{p}_i\right)$ называется *кумулятой*.

Пример 5.1.14. Построим интервальный вариационный ряд, полигон частот и кумуляту по данным о росте 30 студентов (в см) из примера 1.1.2.

В примере 1.1.2 мы опытным путем выяснили, что оптимальная длина интервала группирования составляет 3 см, и гистограмма, построенная на рис. 1.1.2, *г* дает возможность построить интервальный вариационный ряд (5.1.15):

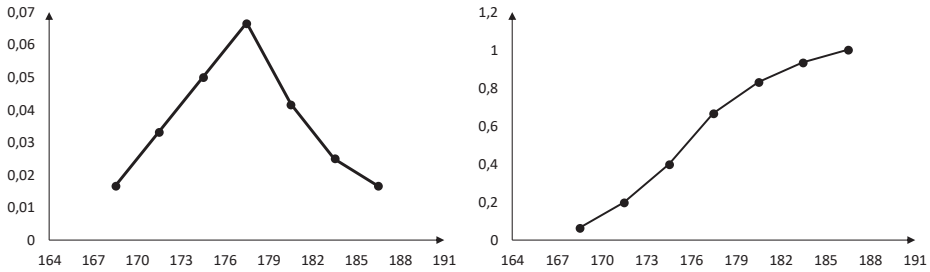
X	[167;170)	[170;173)	[173;176)	[176;179)	[179;182)	[182;185)	[185;188)
p	$\frac{2}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{8}{30}$	$\frac{5}{30}$	$\frac{3}{30}$	$\frac{2}{30}$

Теперь по интервальному вариационному ряду построим выборочную плотность распределения (5.1.17) и выборочную функцию распределения (5.1.18):

$$\hat{f}_X(x)=\begin{cases}0, & x<167,\\ \frac{2/30}{4}, & 167\leqslant x<170,\\ \frac{4/30}{4}, & 170\leqslant x<173,\\ \frac{6/30}{4}, & 173\leqslant x<176,\\ \frac{8/30}{4}, & 176\leqslant x<179,\\ \frac{5/30}{4}, & 179\leqslant x<182,\\ \frac{3/30}{4}, & 182\leqslant x<185,\\ \frac{2/30}{4}, & 185\leqslant x<188,\\ 0, & x\geqslant 188\end{cases}=\begin{cases}0, & x<167,\\ \frac{2}{120}, & 167\leqslant x<170,\\ \frac{4}{120}, & 170\leqslant x<173,\\ \frac{6}{120}, & 173\leqslant x<176,\\ \frac{8}{120}, & 176\leqslant x<179,\\ \frac{5}{120}, & 179\leqslant x<182,\\ \frac{3}{120}, & 182\leqslant x<185,\\ \frac{2}{120}, & 185\leqslant x<188,\\ 0, & x\geqslant 188;\end{cases}$$

$$\hat{F}_X(x)=\begin{cases}0, & x<167,\\ \frac{2}{30}, & 167\leqslant x<170,\\ \frac{2}{30}+\frac{4}{30}, & 170\leqslant x<173,\\ \frac{2}{30}+\frac{4}{30}+\frac{6}{30}, & 173\leqslant x<176,\\ \frac{2}{30}+\frac{4}{30}+\frac{6}{30}+\frac{8}{30}, & 176\leqslant x<179,\\ \frac{2}{30}+\frac{4}{30}+\frac{6}{30}+\frac{8}{30}+\frac{5}{30}, & 179\leqslant x<182,\\ \frac{2}{30}+\frac{4}{30}+\frac{6}{30}+\frac{8}{30}+\frac{5}{30}+\frac{3}{30}, & 182\leqslant x<185,\\ \frac{2}{30}+\frac{4}{30}+\frac{6}{30}+\frac{8}{30}+\frac{5}{30}+\frac{3}{30}+\frac{2}{30}, & 185\leqslant x<188,\\ 1, & x\geqslant 188;\end{cases}=\begin{cases}0, & x<167,\\ \frac{2}{30}, & 167\leqslant x<170,\\ \frac{6}{30}, & 170\leqslant x<173,\\ \frac{12}{30}, & 173\leqslant x<176,\\ \frac{20}{30}, & 176\leqslant x<179,\\ \frac{25}{30}, & 179\leqslant x<182,\\ \frac{28}{30}, & 182\leqslant x<185,\\ 1, & x\geqslant 185.\end{cases}$$

Полигон частот представлен на рис. 5.1.4 *а*, а кумулята — на рис. 5.1.4, *б* соответственно.



а) *б*)

Рис. 5.1.4. *Полигон частот* (*а*) *и кумулята* (*б*) *в примере 5.1.14*

Замечаем, что выборочная функция распределения напоминает функцию нормального закона распределения, но при этом полигон частот имеет слишком острую для нормального закона распределения вершину. ❑

Для того, чтобы по выборке построить интервальный вариационный ряд, оценить плотность распределения и функцию распределения, удобно пользоваться программой «Гистограмма» из надстройки «Анализ данных» пакета Microsoft Excel.

Пример 5.1.15. Служба маркетинга оценивает дилеров фирмы по объему продаж. Сведения об объеме ежедневных продаж товара (в тыс. ден. ед.) дилером за последние 100 дней приведены в табл. 5.1.3.

Таблица 5.1.3

Объем продаж в примере 5.1.15

47,0	37,2	52,4	62,8	62,0	67,3	28,2	47,7	61,0	39,1	43,1	33,1	31,5	40,2	42,3	28,8	44,3	46,0	51,3	46,3
46,7	46,3	63,4	49,1	48,1	44,9	69,7	58,7	73,8	43,5	66,6	33,9	55,4	59,0	69,2	49,2	44,8	56,8	46,2	57,6
35,6	41,5	34,8	46,4	49,7	50,3	46,8	71,9	32,6	42,6	24,2	64,5	37,2	43,5	57,6	54,7	58,7	56,0	36,3	38,8
56,9	53,2	40,6	47,6	51,3	55,6	51,4	40,9	68,8	54,9	50,7	58,3	58,6	43,6	40,8	61,1	38,0	34,4	57,1	56,4
72,1	64,4	63,0	51,1	50,0	54,5	49,7	39,5	32,3	58,3	54,4	56,2	52,1	39,7	62,4	46,9	41,6	41,8	45,7	45,5

Построим интервальный вариационный ряд. В ячейки A1:A101 рабочего листа Microsoft Excel введем данные об объеме продаж из табл. 5.1.3 (в первой строке — заголовок, как показано на рис. 5.1.5). Выберем длину интервала группирования равной 7 и рассчитаем границы интервалов (рис. 5.1.5).

Для расчета интервальных частот и построения полигона частот и гистограммы воспользуемся программой «Гистограмма» из надстройки «Анализ данных» Microsoft Excel. Для этого нажмем на кнопку «Анализ данных» (напомним, что она расположена в правой части вкладки «Данные»), в меню выбора инструментов анализа данных выберем соответствующий инструмент «Гистограмма» и нажмем «ОК».

В окне ввода исходных данных программы «Гистограмма» (рис. 5.1.6) укажем входной интервал (ссылку на ячейки A1:A101, содержащие данные об объеме продаж, включая заголовок столбца), интервал карманов (ссылку на ячейки C8:C16, содержащие правые границы интервалов), установим флажок «Метки», который означает, что в первой строке диапазона A1:A101 содержится текстовый заголовок. Укажем, что результаты работы программы необходимо вывести на новый рабочий лист. Установим флажок для генерации интегральных процентов — значений выборочной функции распределения, также установим флажок автоматического вывода графика — гистограммы и кумуляты. Результаты работы программы «Гистограмма» представлены на рис. 5.1.7. В столбце «Правые границы» на рис. 5.1.7, *а* указаны правые границы интервалов, в столбце «Частота» — интервальные частоты, а в столбце «Интегральный %» — накопленные частоты, рассчитанные программой. На рис. 5.1.7, *б* представлен график, построенный программой, — на одной диаграмме построены гистограмма и кумулята.

	A	B	C
1	Объем продаж за 100 дней	Параметры	Значения параметров
2	47,0	Объем выборки *n*	=СЧЕТ(A2:A101)
3	37,2	$x_{(min)}$	=МИН(A2:A101)
4	52,4	$x_{(max)}$	=МАКС(A2:A101)
5	62,8	Ширина интервала	7
6	62,0	Границы интервалов	
7	67,3	Левые границы	Правые границы
8	28,2	=C3–C5/2	=B8+$C$5
9	47,7	=C8	=B9+$C$5
10	61,0	=C9	=B10+$C$5
11	39,1	=C10	=B11+$C$5
12	43,1	=C11	=B12+$C$5
13	33,1	=C12	=B13+$C$5
14	31,5	=C13	=B14+$C$5
15	40,2	=C14	=B15+$C$5
16	⋮	=C15	=B16+$C$5

а) формулы Microsoft Excel

	A	B	C
1	Объем продаж за 100 дней	Параметры	Значения параметров
2	47,0	Объем выборки *n*	100
3	37,2	$x_{(min)}$	24,2
4	52,4	$x_{(max)}$	73,8
5	62,8	Ширина интервала	7
6	62,0	Границы интервалов	
7	67,3	Левые границы	Правые границы
8	28,2	24,2	30,2
9	47,7	30,2	36,2
10	61,0	36,2	42,2
11	39,1	42,2	48,2
12	43,1	48,2	54,2
13	33,1	54,2	60,2
14	31,5	60,2	66,2
15	40,2	66,2	72,2
16	⋮	72,2	78,2

б) результаты расчетов

Рис. 5.1.5. *Расчет границ интервалов в примере 5.1.15*

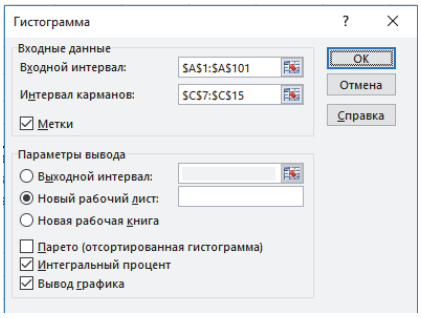


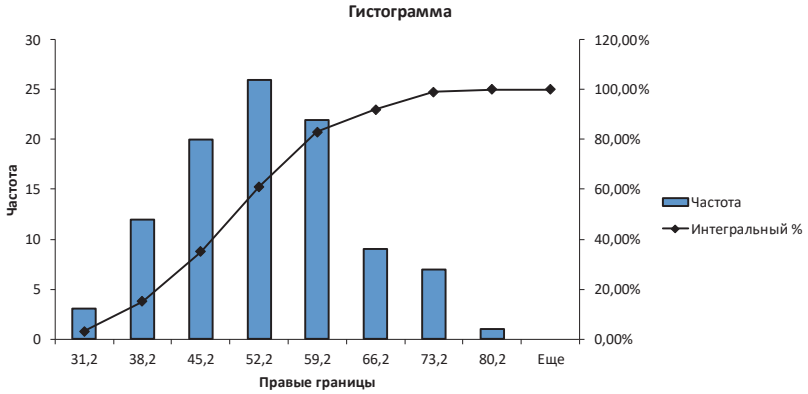
Рис. 5.1.6. *Окно ввода данных программы «Гистограмма»*

Добавим к таблице, полученной в результате работы программы **«Гистограмма»** и уже содержащей правые границы интервалов $[a_j; a_{j+1})$ и интервальные частоты m_j, столбцы, содержащие:

- левые границы интервалов;
- середины интервалов $x'_j = (a_j + a_{j+1})/2$;
- интервальные относительные частоты $\hat{p}_j = m_j / n$;
- оценки функции плотности внутри интервалов $\hat{f}_X(x'_j) = \hat{p}_j / \Delta$.

Правые границы	Частота	Интегральный %
31,2	3	3,00%
38,2	12	15,00%
45,2	20	35,00%
52,2	26	61,00%
59,2	22	83,00%
66,2	9	92,00%
73,2	7	99,00%
80,2	1	100,00%
Еще	0	100%

а) числовые результаты



б) графические результаты

Рис. 5.1.7. *Результаты работы программы «Гистограмма»: числовые (а), графические (б)*

Для получения оценок функции распределения в концах интервалов $\hat{F}_X(a_{j+1}) = \sum_{k \leqslant j+1} \hat{p}_k$ установим в столбце «Интегральный %» (в результатах работы программы «Гистограмма» — рис. 5.1.5) числовой формат значений с двумя десятичными знаками после запятой.

Теперь заполним табл. 5.1.4 с итоговыми результатами. ❑

Таблица 5.1.4

Оценка функции плотности и функции распределения в примере 5.1.15

Интервал $[a_j; a_{j+1})$	Середина интервала x'_j	Интервальная частота m_j	Интервальная относительная частота $\hat{p}_j$	Оценка функции плотности $\hat{f}_X(x'_j)$	Оценка функции распределения $\hat{F}_X(a_{j+1})$
[24,2; 31,2)	27,7	3	0,03	0,004286	0,03
[31,2; 38,2)	34,7	12	0,12	0,017143	0,15
[38,2; 45,2)	41,7	20	0,20	0,028571	0,35
[45,2; 52,2)	48,7	26	0,26	0,037143	0,61
[52,2; 59,2)	55,7	22	0,22	0,031429	0,83
[59,2; 66,2)	62,7	9	0,09	0,012857	0,92
[66,2; 73,2)	69,7	7	0,07	0,010000	0,99
[73,2; 80,2)	76,7	1	0,01	0,001429	1,00
Итого	—	100	1,00	—	—

По интервальному вариационному ряду (5.1.15) оценки математического ожидания и дисперсии вычисляются точно так же, как и по статистическому ряду

распределения, при этом $x_i' = (a_i + a_{i+1})/2$ и, например, выборочное среднее можно рассчитывать по формуле (5.1.13), а выборочную дисперсию — по формуле (5.1.14).

В качестве оценок медианы и моды генеральной совокупности проще всего, как обсуждалось в **п. 1.2.1**, вычислить выборочную медиану как серединный элемент выборки, а выборочную моду — как наиболее часто встречающийся элемент выборки. Однако более целесообразно получать оценки медианы и моды по интервальному вариационному ряду:

$$\hat{x}_{\text{med}} = a_l + \Delta \frac{0{,}5 - \hat{F}(a_l)}{\hat{p}_l}, \tag{5.1.20}$$

где a_l — начало *медианного интервала*, т. е. такого интервала группирования $(a_l; a_{l+1})$, что $\hat{F}(a_l) < 0{,}5$, а $\hat{F}(a_{l+1}) \geqslant 0{,}5$;

$$\hat{x}_{\text{mod}} = a_m + \Delta \frac{\hat{p}_m - \hat{p}_{m-1}}{2\hat{p}_m - \hat{p}_{m-1} - \hat{p}_{m+1}}, \tag{5.1.21}$$

где a_m — начало *модального интервала*, т. е. такого интервала группирования $(a_m; a_{m+1})$, что $\hat{p}_m = \max\limits_{i=1,2,\ldots,v} \hat{p}_i$.

Пример 5.1.16. Вычислим в условиях примера 5.1.15 выборочное среднее, выборочную дисперсию, выборочную медиану и выборочную моду.

Выборочное среднее СРЗНАЧ(A2:A101) = 49,596, выборочная дисперсия ДИСП.Г(A2:A101) = 116,634784. Выборочная медиана, рассчитанная по исходной выборке, МЕДИАНА(A2:A101) = 49,15, выборочная мода, рассчитанная по исходной выборке, МОДА(A2:A101) = 37,2.

Если оценивать медиану и моду по интервальному вариационному ряду, то получим следующие значения:

$$\hat{x}_{\text{med}} = a_l + \Delta \frac{0{,}5 - \hat{F}(a_l)}{\hat{p}_l} = 45{,}2 + 7 \cdot \frac{0{,}5 - 0{,}35}{0{,}26} = 49{,}23846;$$

$$\hat{x}_{\text{mod}} = a_m + \Delta \frac{\hat{p}_m - \hat{p}_{m-1}}{2\hat{p}_m - \hat{p}_{m-1} - \hat{p}_{m+1}} = 45{,}2 + 7 \cdot \frac{0{,}26 - 0{,}20}{2 \cdot 0{,}26 - 0{,}20 - 0{,}22} = 49{,}4.$$

Медиана в данном примере изменилась не сильно, а вот мода — весьма значительно. Значение, полученное по интервальному вариационному ряду, содержит в себе больше смысла, так как значение 37,2 действитиельно встречается в выборке чаще всего (два раза), а значение 49,4 относится к интервалу [45,2; 52,2), в который попало 26 элементов выборки. ❑

Задачи

325. По выборке 11; 11; 11; 9; 9; 5; 10; 10; 10; 14 постройте вариационный ряд, выборочную случайную величину, полигон частот, гистограмму, кумуляту, вычислите выборочное среднее и выборочную дисперсию.

326. Распределение 200 пострадавших от несчастных случаев по возрасту таково:

Возраст	[16;21)	[21;26)	[26;31)	[31;36)	[36;41)	[41;46)
Число пострадавших	133	45	15	4	2	1

Требуется: *а*) найти средний возраст пострадавшего и оценку дисперсии среднего возраста; *б*) предположив, что возраст пострадавшего подчиняется показательному закону, найти выражение для плотности распределения и функции распределения; рассчитать вероятности несчастных случаев в каждом из возрастных интервалов и ожидаемое распределение числа

пострадавших; *в*) построить на одном рисунке гистограмму, полигон и график плотности распределения, на другом рисунке — кумуляту и график функции распределения.

327. По данным о выданных банком кредитах из примера 1.1.1 (файл «Кредиты.xlsx») постройте интервальные вариационные ряды, гистограммы, полигоны частот и графики выборочных функций распределения размера кредита и годового дохода заемщиков.

328. По данным об акциях из задачи 5 постройте интервальные вариационные ряды, гистограммы, полигоны частот и графики выборочных функций распределения для каждого из признаков: цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов.

§ 5.2. Точечные оценки параметров

5.2.1. Свойства точечных оценок

Статистикой называется любая функция $\gamma = \gamma(X_1, X_2, \ldots, X_n)$ от элементов выборки $X_1, X_2, \ldots, X_n$ [на конкретной выборке $x_1, x_2, \ldots, x_n$ эта статистика принимает конкретное значение $\gamma = \gamma(x_1, x_2, \ldots, x_n)$]. Очевидно, любая статистика является случайной величиной, поскольку она является функцией случайных величин.

Оценкой $\hat{\theta}$ числовой характеристики или параметра θ называется любая статистика $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$, используемая в качестве приближенного значения θ.

Например, в качестве оценки параметра $a = \mathbf{E}(X)$ случайной величины X, распределенной по нормальному закону (с неизвестными параметрами), можно использовать:

- результат единичного наблюдения $\hat{a} = X_1$;
- полусумму минимального и максимального элементов выборки

$$\hat{a} = \frac{X_{(\min)} + X_{(\max)}}{2};$$

- выборочную медиану $\hat{a} = \hat{X}_{\text{med}}$;
- выборочную моду $\hat{a} = \hat{X}_{\text{mod}}$;
- выборочное среднее $\hat{a} = \overline{X}$ и др.

Как определить, какая из этих оценок лучше?

Качество оценки определяется по выполнению следующих трех **свойств**: состоятельности, несмещенности и эффективности.

Прежде всего, от оценки $\hat{\theta}_n$ хотелось бы требовать, чтобы по мере роста числа наблюдений она сходилась по вероятности к оцениваемому параметру, т. е. чтобы для любого сколь угодно малого $\varepsilon > 0$ было справедливо предельное равенство

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\hat{\theta}_n - \theta_n\right| < \varepsilon\right\} = 1. \qquad (5.2.1)$$

Оценка $\hat{\theta}_n$, обладающая свойством (5.2.1), называется *состоятельной оценкой параметра* θ.

Также от «хорошей» оценки $\hat{\theta}_n$ естественно требовать, чтобы она не содержала систематической ошибки, т. е. при любом фиксированном объеме выборки n результат осреднения оценки по всем возможным выборкам данного объема должен приводить к точному значению параметра:

$$\mathbf{E}\left(\hat{\theta}_n\right)=\theta. \tag{5.2.2}$$

Оценка $\hat{\theta}_n$, обладающая свойством (5.2.2), называется *несмещенной оценкой параметра* θ.

Наконец, от оценки $\hat{\theta}_n$ желательно требовать, чтобы ее дисперсия была минимальной в некотором классе оценок $\widehat{\Theta}$:

$$\mathbf{Var}\left(\hat{\theta}_n^*\right)=\min_{\hat{\theta}_n\in\widehat{\Theta}}\mathbf{Var}\left(\hat{\theta}_n\right). \tag{5.2.3}$$

Оценка $\hat{\theta}_n^*$, обладающая свойством (5.2.3), называется *эффективной оценкой параметра* θ в классе оценок $\widehat{\Theta}$.

Выборочное среднее $\overline{X}$ (5.1.1) является состоятельной, несмещенной и эффективной (в классе линейных несмещенных оценок) оценкой математического ожидания $\mathbf{E}(X)$.

Доказательство. Выборочное среднее $\overline{X}$ является состоятельной оценкой математического ожидания, если генеральная совокупность имеет ограниченную дисперсию, поскольку предельное равенство $\lim\limits_{n\to\infty}\mathbf{P}\left\{\left|\overline{X}-\mathbf{E}(X)\right|<\varepsilon\right\}=1$ непосредственно следует из теоремы Чебышёва (4.3.2), так как $\overline{X}=\dfrac{\sum\limits_{i=1}^{n}X_i}{n}$, а $\mathbf{E}\left(\overline{X}\right)=\mathbf{E}\left(\dfrac{\sum\limits_{i=1}^{n}X_i}{n}\right)=\dfrac{\sum\limits_{i=1}^{n}\mathbf{E}\left(X_i\right)}{n}=\mathbf{E}(X)$. Из того, что $\mathbf{E}\left(\overline{X}\right)=\mathbf{E}(X)$, следует также, что выборочное среднее $\overline{X}$ является несмещенной оценкой математического ожидания $\mathbf{E}(X)$.

Линейные оценки имеют вид $\tilde{a}=\sum\limits_{i=1}^{n}c_iX_i$; линейная несмещенная оценка должна удовлетворять условию $\mathbf{E}(\tilde{a})=a$, поэтому $a=\mathbf{E}(\tilde{a})=\mathbf{E}\left(\sum\limits_{i=1}^{n}c_iX_i\right)=\sum\limits_{i=1}^{n}c_i\mathbf{E}\left(X_i\right)=a\sum\limits_{i=1}^{n}c_i$, откуда следует, что для несмещенности необходимо, чтобы $\sum\limits_{i=1}^{n}c_i=1$. Дисперсия линейной оценки $\mathbf{Var}(\tilde{a})=\mathbf{Var}\left(\sum\limits_{i=1}^{n}c_iX_i\right)=\sum\limits_{i=1}^{n}c_i^2\mathbf{Var}\left(X_i\right)=\sigma^2\sum\limits_{i=1}^{n}c_i^2$. Для определения эффективной оценки (в классе линейных несмещенных оценок) необходимо минимизировать дисперсию $\mathbf{Var}(\tilde{a})$ при выполнении условия несмещенности $\sum\limits_{i=1}^{n}c_i=1$, и поскольку $\sigma^2=\text{const}$, получаем задачу на условный экстремум: $\sum\limits_{i=1}^{n}c_i^2\to\min,\sum\limits_{i=1}^{n}c_i=1$. Составим функцию Лагранжа для этой задачи: $L(c_1,c_2,\ldots,c_n,\lambda)=\sum\limits_{i=1}^{n}c_i^2-\lambda\left(\sum\limits_{i=1}^{n}c_i-1\right)$. Для определения точки минимума найдем производные функции Лагранжа и приравняем их нулю:

$$\begin{cases}\dfrac{\partial L}{\partial c_i}=2c_i-\lambda=0,\ i=1,2,\ldots,n,\\[2ex] \dfrac{\partial L}{\partial\lambda}=\sum\limits_{i=1}^{n}c_i-1=0.\end{cases}$$

Подставляя значения $c_i=\lambda/2$, выраженные из первых уравнений, в последнее уравнение, получаем: $\lambda/2=1/n$ или $c_i^*=\lambda/2=1/n$, поэтому эффективной (в данном классе) оценкой является $\tilde{a}^*=\sum\limits_{i=1}^{n}c_i^*X_i=\sum\limits_{i=1}^{n}\dfrac{1}{n}X_i=\sum\limits_{i=1}^{n}X_i\Big/n=\overline{X}$.

Дисперсия выборочного среднего равна

$$\mathbf{Var}(\overline{X}) = \mathbf{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathbf{Var}(X_1 + X_2 + \cdots + X_n) = \{\text{с учетом независимости}$$

$$X_1, X_2, \ldots, X_n\} = \frac{1}{n^2}(\mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n)) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}. \square$$

Пример 5.2.1. Предположим, что мы исследуем генеральную совокупность, представляющую собой случайную величину X — число очков, выпавших при бросании игрального кубика.

Вероятность выпадения каждой из шести граней равна $1/6$, поэтому $\mathbf{E}(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3{,}5$.

Вычислим дисперсию случайной величины X:

$$\mathbf{Var}(X) = \mathbf{E}\left([X - \mathbf{E}(X)]^2\right) =$$

$$= (1-3{,}5)^2 \cdot \frac{1}{6} + (2-3{,}5)^2 \cdot \frac{1}{6} + (3-3{,}5)^2 \cdot \frac{1}{6} + (4-3{,}5)^2 \cdot \frac{1}{6} + (5-3{,}5)^2 \cdot \frac{1}{6} + (6-3{,}5)^2 \cdot \frac{1}{6} =$$

$$= (-2{,}5)^2 \cdot \frac{1}{6} + (-1{,}5)^2 \cdot \frac{1}{6} + (-0{,}5)^2 \cdot \frac{1}{6} + (0{,}5)^2 \cdot \frac{1}{6} + (1{,}5)^2 \cdot \frac{1}{6} + (2{,}5)^2 \cdot \frac{1}{6} =$$

$$= 6{,}25 \cdot \frac{1}{6} + 2{,}25 \cdot \frac{1}{6} + 0{,}25 \cdot \frac{1}{6} + 0{,}25 \cdot \frac{1}{6} + 2{,}25 \cdot \frac{1}{6} + 6{,}25 \cdot \frac{1}{6} = \frac{17{,}5}{6} = 2{,}91(6).$$

Выпишем все варианты случайных повторных выборок объемом в два элемента (на рис. 5.2.1 они представлены в столбцах «X1» и «X2»), вероятность каждой комбинации из двух элементов при этом равна $1/6^2 = 1/36 = 0{,}02(7)$.

Для каждой выборки вычислим в столбце «X_средн. = (X1 + X2)/2» выборочное среднее $\overline{X}$. Теперь математическое ожидание случайной величины $\overline{X}$ можно вычислить в ячейке G2 с помощью функции СУММПРОИЗВ как сумму произведений элементов столбца «X_средн. = (X1 + X2)/2» на элементы столбца «p». Замечаем, что $\mathbf{E}(\overline{X}) = \mathbf{E}(X) = 3{,}5$.

Теперь в столбце «[X_средн – E(X_средн.)]^2» вычислим значения $\left((\overline{X})_i - \mathbf{E}(\overline{X})\right)^2$, и можно в в ячейке G5 рассчитать дисперсию случайной величины $\overline{X}$ с помощью функции СУММПРОИЗВ как сумму произведений элементов столбца «[X _средн – E(X_средн.)]^2» на элементы столбца «p». Замечаем, что $\dfrac{\mathbf{Var}(X)}{n} = \dfrac{2{,}91(6)}{2} = 1{,}458(3)$, $\mathbf{Var}(\overline{X}) = 1{,}458(3)$, т. е. $\mathbf{Var}(\overline{X}) = \dfrac{\mathbf{Var}(X)}{n}$. $\square$

Относительная частота $\hat{p}$ (5.1.8) является состоятельной, несмещенной и эффективной (в классе линейных несмещенных оценок) оценкой вероятности успеха p.

Доказательство этого утверждения оставляем читателю в задаче **338**.

Выборочная дисперсия $\hat{\sigma}_X^2$ (5.1.2) является смещенной оценкой дисперсии $\mathbf{Var}(X)$:

$$\mathbf{E}\left(\hat{\sigma}_X^2\right) = \frac{n-1}{n}\mathbf{Var}(X) \neq \mathbf{Var}(X). \tag{5.2.4}$$

Доказательство. Действительно,

$$\mathbf{E}\left(\hat{\sigma}_X^2\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}\left(X_i^2\right) - 2\mathbf{E}\left(X_i\overline{X}\right) + \mathbf{E}\left(\overline{X}^2\right)\right),$$

при этом: по определению дисперсии $\mathbf{E}(X_i^2) = \mathbf{Var}(X_i) + [\mathbf{E}(X_i)]^2 = \sigma^2 + a^2$; поскольку наблюдения X_i и X_j независимы, $\mathbf{E}(X_iX_j) = \mathbf{E}(X_i)\mathbf{E}(X_j)$ (при $i \neq j$), и

	A	B	C	D	E	F	G
1	X1	X2	p	X_средн. = (X1 + X2)/2	[X_средн – E(X_средн.)]^2		E(X_средн.)
2	1	1	0,027778	1	6,25		3,5
3	1	2	0,027778	1,5	4		
4	1	3	0,027778	2	2,25		Var(X_средн.)
5	1	4	0,027778	2,5	1		1,458333
6	1	5	0,027778	3	0,25		
7	1	6	0,027778	3,5	0		
8	2	1	0,027778	1,5	4		
9	2	2	0,027778	2	2,25		
10	2	3	0,027778	2,5	1		
11	2	4	0,027778	3	0,25		
12	2	5	0,027778	3,5	0		
13	2	6	0,027778	4	0,25		
14	3	1	0,027778	2	2,25		
15	3	2	0,027778	2,5	1		
16	3	3	0,027778	3	0,25		
17	3	4	0,027778	3,5	0		
18	3	5	0,027778	4	0,25		
19	3	6	0,027778	4,5	1		
20	4	1	0,027778	2,5	1		
21	4	2	0,027778	3	0,25		
22	4	3	0,027778	3,5	0		
23	4	4	0,027778	4	0,25		
24	4	5	0,027778	4,5	1		
25	4	6	0,027778	5	2,25		
26	5	1	0,027778	3	0,25		
27	5	2	0,027778	3,5	0		
28	5	3	0,027778	4	0,25		
29	5	4	0,027778	4,5	1		
30	5	5	0,027778	5	2,25		
31	5	6	0,027778	5,5	4		
32	6	1	0,027778	3,5	0		
33	6	2	0,027778	4	0,25		
34	6	3	0,027778	4,5	1		
35	6	4	0,027778	5	2,25		
36	6	5	0,027778	5,5	4		
37	6	6	0,027778	6	6,25		

Рис. 5.2.1. *Выборочные средние при подбрасывании двух кубиков*

$$\mathbf{E}\left(X_i\overline{X}\right)=\mathbf{E}\left(X_i\frac{\sum\limits_{j=1}^{n}X_j}{n}\right)=\frac{1}{n}\left(\mathbf{E}\left(X_i^2\right)+\sum_{\substack{j=1\\j\neq i}}^{n}\mathbf{E}\left(X_i\right)\mathbf{E}\left(X_j\right)\right)=\frac{1}{n}(\sigma^2+a^2+(n-1)a^2)=\frac{\sigma^2}{n}+a^2;$$

наконец,

$$\mathbf{E}\left(\overline{X}^2\right)=\mathbf{E}\left(\frac{\sum\limits_{i=1}^{n}X_i}{n}\frac{\sum\limits_{j=1}^{n}X_j}{n}\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbf{E}\left(X_i^2\right)+\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\mathbf{E}\left(X_i\right)\mathbf{E}\left(X_j\right)\right)=\frac{1}{n^2}(n(\sigma^2+a^2)+n(n-1)a^2)=\frac{\sigma^2}{n}+a^2.$$

Подставляя найденные выражения для математических ожиданий в формулу для $\mathbf{E}\left(\hat{\sigma}_X^2\right)$, получим:

$$\mathbf{E}\left(\hat{\sigma}_X^2\right)=\frac{1}{n}\sum_{i=1}^{n}\left[(\sigma^2+a^2)-2\left(\frac{\sigma^2}{n}+a^2\right)+\left(\frac{\sigma^2}{n}+a^2\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2-\frac{\sigma^2}{n}\right)=\frac{1}{n}n\left(\sigma^2-\frac{\sigma^2}{n}\right)=\frac{n-1}{n}\sigma^2=\frac{n-1}{n}\mathbf{Var}(X),$$

поэтому выборочная дисперсия $\hat{\sigma}_X^2$ является смещенной оценкой дисперсии. ❑

Однако смещенность выборочной дисперсии $\hat{\sigma}_X^2$ легко исправима, достаточно в качестве оценки дисперсии выбрать величину

$$s_X^2=\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}=\frac{n}{n-1}\hat{\sigma}_X^2, \qquad (5.2.5)$$

называемую *исправленной выборочной дисперсией*; исправленная выборочная дисперсия s_X^2 является состоятельной и несмещенной оценкой дисперсии.

Замечение. При доказательстве состоятельности выборочной дисперсии (и исправленной выборочной дисперсии), что мы предлагаем сделать читателю самостоятельно в задаче **337**, полезно следующее

Достаточное условие состоятельности. *Для того, чтобы несмещенная оценка* $\hat{\theta}_n$ *была состоятельной, достаточно, чтобы* $\lim\limits_{n\to\infty}\mathbf{Var}(\hat{\theta}_n)=0$.

Доказательство. Согласно неравенству Чебышёва (4.1.2), для любого $\varepsilon>0$ $\mathbf{P}\{|\hat{\theta}_n-\mathbf{E}(\hat{\theta}_n)|\leqslant\varepsilon\}>1-\mathbf{Var}(\hat{\theta}_n)/\varepsilon^2$, откуда, учитывая несмещенность оценки $\hat{\theta}_n$ (т. е. $\mathbf{E}(\hat{\theta}_n)=\theta$), получаем: $\mathbf{P}\{|\hat{\theta}_n-\theta|\leqslant\varepsilon\}>1-\mathbf{Var}(\hat{\theta}_n)/\varepsilon^2$, и поскольку $\lim\limits_{n\to\infty}\mathbf{Var}(\hat{\theta}_n)=0$, окончательно имеем, что $\lim\limits_{n\to\infty}\mathbf{P}\{|\hat{\theta}_n-\theta|\leqslant\varepsilon\}=1$, т. е. $\hat{\theta}_n$ является состоятельной оценкой θ. ❑

Замечание. Если известно точное значение $a=\mathbf{E}(X)$ математического ожидания нормальной случайной величины X, распределенной по нормальному закону, то статистика

$$s_0^2=\frac{\sum\limits_{i=1}^{n}(X_i-a)^2}{n} \tag{5.2.6}$$

является состоятельной, несмещенной и эффективной (в классе всех несмещенных оценок) оценкой дисперсии $\mathbf{Var}(X)$ случайной величины X. Однако на практике оценка (5.2.6) не используется, так как точное значение математического ожидания почти никогда не известно.

Выборочное среднее квадратическое отклонение — это $\hat{\sigma}_X=\sqrt{\hat{\sigma}_X^2}$, а *стандартное отклонение* — это $s_X=\sqrt{s_X^2}$.

В Microsoft Excel s_X^2 =ДИСП.В(<массив x_i>), s_X = СТАНДОТКЛОН(<массив x_i>).

Пример 5.2.2. В условиях примера 5.1.15 мы уже вычислили выборочное среднее СРЗНАЧ(A2:A101) = 49,596 — несмещенную, состоятельную и эффективную оценку математического ожидания объема продаж. Однако выборочная дисперсия ДИСП.Г(A2:A101) = 116,634784 является смещенной оценкой дитсперсии объема продаж. Несмещенной оценкой дисперсии является исправленная выборочная дисперсия s_X^2 =ДИСП.В(A2:A101) = 117,8129131 (она больше, чем $\hat{\sigma}_X^2$). Стандартное отклонение s_X = СТАНДОТКЛОН(A2:A101) = 10,8541657.

В примере 5.1.15 мы заполнили табл. 5.1.4, в которой вычислили оценки функции плотности распределения и функции распределения объема продаж в серединах интервалов. По данным табл. 5.1.4 построим полигон частот на рис. 5.2.2, *а* и *кумуляту* на рис. 5.2.2, *б*.

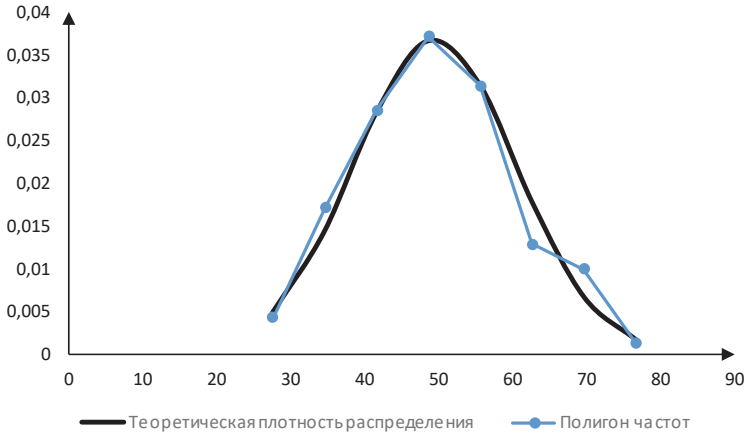
Заменим параметры нормального закона a и σ их оценками $a=\bar{x}=49{,}596$, $\sigma=s_X=10{,}85417$ и рассчитаем значения функции плотности нормального закона

$$f_{\mathcal{N}}(x_j')=\text{НОРМ.РАСП}(<x_j'>;\ <a>;\ <\sigma>;\ \text{ЛОЖЬ})$$

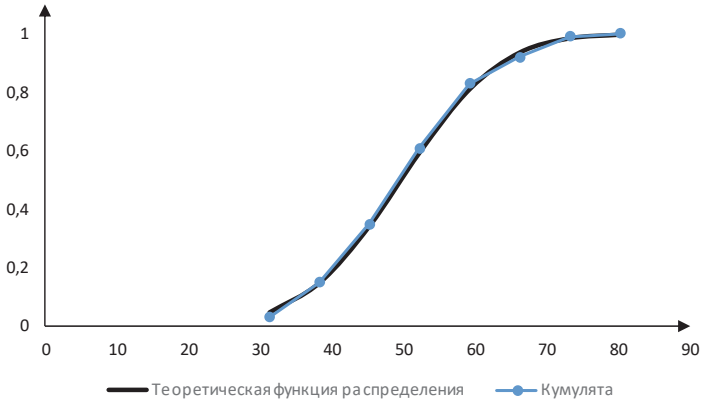
в серединах интервалов и функции распределения

$$F_{\mathcal{N}}(a_{j+1})=\text{НОРМ.РАСП}(<a_{j+1}>;\ <a>;\ <\sigma>;\ \text{ИСТИНА})$$

в правых концах интервалов. Результаты расчетов представлены в седьмом и восьмом столбцах табл. 5.2.1 (первые шесть столбцов табл. 5.2.1 взяты из табл. 5.1.4). Графики функций $f(x)$ и $F(x)$ построены на рис. 5.2.2, *а* и 5.2.2, *б* соответственно. ❑



а)



б)

Рис. 5.2.2. *Полигон частот и функция плотности нормального закона* (*а*), *кумулята и функция распределения нормального закона* (*б*) *в примере 5.2.2*

Таблица 5.2.1

Оценка функции плотности и функции распределения в примере 5.2.15

Интервал $[a_j; a_{j+1})$	Середина интервала x'_j	Интервальная частота m_j	Интервальная относительная частота $\hat{p}_j$	Оценка функции плотности $\hat{f}_X(x'_j)$	Оценка функции распределения $\hat{F}_X(a_{j+1})$	Функция плотности нормального закона $f_N(x'_j)$	Функция распределения нормального закона $F_N(a_{j+1})$
[24,2; 31,2)	27,7	3	0,03	0,004286	0,03	0,0048044	0,045054
[31,2; 38,2)	34,7	12	0,12	0,017143	0,15	0,0143329	0,146878
[38,2; 45,2)	41,7	20	0,20	0,028571	0,35	0,0282097	0,342737
[45,2; 52,2)	48,7	26	0,26	0,037143	0,61	0,0366297	0,594799
[52,2; 59,2)	55,7	22	0,22	0,031429	0,83	0,0313791	0,811874
[59,2; 66,2)	62,7	9	0,09	0,012857	0,92	0,0177344	0,936959
[66,2; 73,2)	69,7	7	0,07	0,010000	0,99	0,0066125	0,985172
[73,2; 80,2)	76,7	1	0,01	0,001429	1,00	0,0016266	0,997596

По аналогии с теоретическим начальным моментом порядка k ($k = 0$, 1, 2, ...) случайной величины X, т. е. величиной $\nu_k(X) = \mathbf{E}(X^k)$ — см. **п. 3.6.1**, и теоретическим центральным моментом порядка k ($k = 0$, 1, 2, ...) этой случайной величины, т. е. величиной $\mu_k(X) = \mathbf{E}\left[(X - MX)^k\right]$, определяются *выборочные начальный* и *центральный моменты* порядка k ($k = 0$, 1, 2, ...) — это статистики $\hat{\nu}_k = \overline{X^k}$ и $\hat{\mu}_k = \overline{(X - \bar{\bar{X}})^k}$ соответственно.

Свойства выборочных моментов аналогичны свойствам соответствующих теоретических моментов (3.6.3)—(3.6.7).

В качестве выборочных оценок асимметрии (3.6.8) и эксцесса (3.6.9) можно взять *выборочную асимметрию*

$$\hat{\mathbf{A}}_X = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$$

и *выборочный эксцесс*

$$\hat{\mathbf{E}}_X = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3$$

соответственно; эти оценки являются смещенными, однако их можно подправить и получить несмещенные оценки $\mathbf{A}_X$ и $\mathbf{E}_X$ — *исправленную выборочную асимметрию*

$$\tilde{\mathbf{A}}_X = \frac{\sqrt{n(n-1)}}{n-2}\hat{\mathbf{A}}_X$$

и *исправленный выборочный эксцесс*

$$\tilde{\mathbf{E}}_X = \frac{(n-1)[(n+1)\hat{\mathbf{E}}_X + 6]}{(n-2)(n-3)}.$$

Отношение

$$\hat{V}_X = \frac{s_X}{\bar{\bar{X}}}$$

называется *выборочным коэффициентом вариации*. Выборочный коэффициент вариации оценивает относительный разброс значений случайной величины вокруг математического ожидания.

Теория статистического оценивания параметров была разработана К. Пирсоном, Р.Фишером, Г. Крамером, К.Р. Рао, М.Р. Фреше и другими исследователями.

Задачи

329. По результатам наблюдений $x_1, x_2, ..., x_n$ случайной величины X найдены ее выборочные характеристики: $\bar{x} = 25$, $\hat{\sigma}_X^2 = 0{,}6$, $\hat{\mathbf{A}}_X = -0{,}5$, $\hat{\mathbf{E}}_X = 1{,}2$. Найдите соответствующие характеристики величины $Y = -4X - 15$.

330. По результатам наблюдений $x_1, x_2, ..., x_n$ случайной величины X найдены ее выборочные характеристики: $\bar{x} = 10$, $\hat{\sigma}_X^2 = 4$. Найдите соответствующие характеристики для ряда чисел $y_i = -3x_i + 5$, $i = 1, 2, \ldots, n$.

331. Средняя масса жителя некоторого большого города составляет 75 кг при среднем квадратичном отклонении, равном 10 кг. Городской автобус рассчитан на то, чтобы за один раз везти 64 пассажира. Оцените: *а*) среднюю массу пассажиров в полностью загруженном автобусе и среднее квадратическое отклонение этой массы; *б*) общую массу всех пассажиров и среднее квадратическое отклонение этой массы.

332. По данным об акциях из задачи 5 в предположении нормального закона распределения цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов замените для каждого из этих признаков параметры нормального закона a и σ их несмещенными состоятельными оценками, рассчитайте значения функции плотности нормального закона в серединах интервалов и функции распределения в правых концах интервалов. Для каждого из анализируемых признаков постройте на одном рисунке полигон частот и график теоретической функции плотности распределения, а на другом — кумуляту и график теоретической функции распределения.

333. По данным о выданных банком кредитах из примера 1.1.1 (файл «Кредиты.xlsx») в предположении нормального закона распределения размера кредита, кредитногто рейтинга и годового дохода замените для каждого из этих признаков параметры нормального закона a и σ их несмещенными состоятельными оценками, рассчитайте значения функции плотности нормального закона в серединах интервалов и функции распределения в правых концах интервалов. Для каждого из анализируемых признаков постройте на одном рисунке полигон частот и график теоретической функции плотности распределения, а на другом — кумуляту и график теоретической функции распределения. Предварительно предложите и реализуйте варианты обработки пропущенных значений и выбросов.

334. Генеральная совокупность содержит четыре числа: 3; 6; 9; 10. Требуется: *а*) вычислить $\mathbf{E}(X)$ и $\mathbf{Var}(X)$; *б*) предположив, что из этой генеральной совокупности составляются случайные выборки с возвращением по $n=2$ элемента, указать все возможные пары чисел, которые могут попасть в выборку, и вероятности появления каждой пары; *в*) для каждой такой пары вычислить выборочное среднее $\overline{x}$ и выборочную дисперсию $\hat{\sigma}_X^2$; *г*) составить ряды распределения случайных величин $\overline{X}$ и $\hat{\sigma}_X^2$; *д*) используя ряд распределения случайной величины $\overline{X}$, вычислить $\mathbf{E}(\overline{X})$ и $\mathbf{Var}(\overline{X})$ и убедиться в справедливости соотношений $\mathbf{E}(\overline{X})=\mathbf{E}(X)$ и $\mathbf{Var}(\overline{X})=\mathbf{Var}(X)/n$; *е*) используя ряд распределения случайной величины $\hat{\sigma}_X^2$, вычислить $\mathbf{Var}(\hat{\sigma}_X^2)$ и убедиться, что $\mathbf{Var}(\hat{\sigma}_X^2)=\dfrac{n-1}{n}\mathbf{Var}(X)$.

335. Докажите, что выборочный центральный момент первого порядка равен нулю.

336. Докажите, что выборочное среднее $\overline{X}$ (5.1.1) является эффективной (в классе линейных несмещенных оценок) оценкой математического ожидания $\mathbf{E}(X)$.

337. Докажите, что исправленная выборочная дисперсия s_X^2 является состоятельной и несмещенной оценкой дисперсии.

338. Докажите, что относительная частота является состоятельной, несмещенной и эффективной (в классе всех несмещенных оценок) оценкой вероятности наступления события.

339. Средняя производительность труда a в одном отделе оказалась равной $x_1=10$ (при числе сотрудников $n_1=9$), в другом отделе — $x_2=12$ (при числе сотрудников $n_2=16$). Определите, при каком $\alpha\in(0,1)$ объединенная оценка $\hat{a}=\alpha\overline{x}_1+(1-\alpha)\overline{x}_2$ будет наиболее точной (с точки зрения всех мыслимых выборок указанных объемов), если известно, что $\sigma_1=0{,}1$, $\sigma_2=0{,}5$.

340. Докажите, что нижняя граница дисперсии различных несмещенных оценок дисперсии нормального распределения равна $2\sigma^4/n$.

341. Докажите, что дисперсия оценки s_X^2 равна $2(n-1)\sigma^2/n^2$, т. е. s_X^2 не является эффективной оценкой дисперсии.

342. Докажите, что если известно точное значение $a=\mathbf{E}(X)$ математического ожидания нормальной случайной величины $X=\mathcal{N}(a;\sigma)$, то статистика $s_0^2=\dfrac{\sum_{i=1}^{n}(X_i-a)^2}{n}$ является состоятельной, несмещенной и эффективной (в классе всех возможных несмещенных оценок) оценкой дисперсии $\mathbf{Var}(X)$ случайной величины X.

343. Определите значение k, при котором статистика $\delta = k\sum_{i=1}^{n-1}(X_{i+1}-X_i)^2$ будет несмещенной оценкой дисперсии случайной величины X.

344. Докажите, что $\operatorname{cov}\left(\overline{X}, s_X^2\right) = \mu_3(X)$.

345. Докажите, что случайные величины $\overline{X}$ и s_X^2, рассчитанные по выборке из нормальной генеральной совокупности, являются независимыми.

5.2.2. Методы построения точечных оценок

Будем считать, что известен общий вид $p(x;\theta)$ закона распределения случайной величины X, соответствующей генеральной совокупности, из которой извлечена выборка $x_1, x_2, \ldots, x_n$ (функция вероятности $p(x_i;\theta) = \mathbf{P}\{X = x_i\}$ для дискретной случайной величины или плотность распределения $p(x;\theta) = f_X(x)$ для непрерывной случайной величины), где θ — некоторый неизвестный параметр закона распределения.

Метод моментов (разработанный П.Л. Чебышёвым в 1887 г. в процессе доказательства центральной предельной теоремы и далее обобщенный и развитый К. Пирсоном в 1894 г.) заключается в следующем.

Определяется зависимость

$$\theta = g(\nu_1, \nu_2, \ldots, \nu_k, \mu_1, \mu_2, \ldots, \mu_l) \tag{5.2.7}$$

параметра θ от начальных моментов с первого по k-й и центральных моментов с первого по l-й, после чего для вычисления *оценки* $\hat{\theta}_{\text{мм}}$ *параметра* θ *по методу моментов* в эту зависимость (5.2.7) вместо неизвестных теоретических моментов подставляют их выборочные аналоги $\hat{\nu}_i$ и $\hat{\mu}_j$:

$$\hat{\theta}_{\text{мм}} = g(\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_k, \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_l). \tag{5.2.8}$$

Очевидным достоинством метода моментов является его простота, однако качество оценок, полученных с помощью этого метода, не всегда бывает высоким, особенно при небольших объемах выборки.

Изложим теперь алгоритм **метода максимального правдоподобия** (разработанного Р. Фишером в 1912—1922 гг.[4]). По выборке $x_1, x_2, \ldots, x_n$ составляется *функция правдоподобия*

$$L(\theta) = p(x_1;\theta)p(x_2;\theta)\cdots p(x_n;\theta), \tag{5.2.9}$$

равная вероятности получения именно набора $x_1, x_2, \ldots, x_n$ при извлечении выборки объемом n из генеральной совокупности в случае дискретной случайной величины (и плотности распределения этой вероятности — в случае непрерывной случайной величины). Чем больше $L(\theta)$, тем вероятнее (или правдоподобнее) получить при наблюдениях именно *данную конкретную* выборку $x_1, x_2, \ldots, x_n$.

Оценкой максимального правдоподобия параметра θ называют при этом такое значение $\hat{\theta}_{\text{ммп}}$, которое максимизирует функцию правдоподобия $L(\theta)$:

---

[4] В частных случаях метод максимального правдоподобия применялся еще в начале XIX в. К.Ф. Гауссом и П.-С. Лапласом.

$$L\left(\hat{\theta}_{\text{ммп}}\right)=\max_{\theta} L(\theta). \tag{5.2.10}$$

Для многих распределений вместо оптимизационной задачи (5.2.11) удобнее решать эквивалентную задачу

$$\ln L\left(\hat{\theta}_{\text{ммп}}\right)=\max_{\theta} \ln L(\theta), \tag{5.2.11}$$

так как натуральный логарифм является монотонно возрастающей функцией.

Если существует состоятельная и эффективная оценка $\hat{\theta}$ параметра θ (в классе всех оценок — несмещенных и смещенных), то при весьма общих условиях такой оценкой является $\hat{\theta}_{\text{ммп}}$, если при этом оценка $\hat{\theta}_{\text{ммп}}$ оказывается смещенной, то ее подправляют аналогично тому, как это было сделано с выборочной дисперсией [см. формулу (5.2.5)].

Пример 5.2.3. По выборке $x_1, x_2, \ldots, x_n$ найдем оценку $\hat{\lambda}$ параметра λ случайной величины, распределенной по закону Пуассона методом моментов и методом максимального правдоподобия.

Известно, что начальный момент первого порядка распределения Пуассона равен $\nu_1=\mathbf{E}(X)=\lambda$. Поэтому оценкой метода моментов для параметра λ будет $\hat{\lambda}_{\text{мм}}=\hat{\nu}_1=\overline{x}$.

Для распределения Пуассона

$$p(x;\lambda)=\mathbf{P}\{X=x\}=\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}.$$

Составим функцию правдоподобия:

$$L(\lambda)=p(x_1;\lambda)p(x_2;\lambda)\cdots p(x_n;\lambda)=\frac{\lambda^{x_1}}{x_1!}\mathrm{e}^{-\lambda}\frac{\lambda^{x_2}}{x_2!}e^{-\lambda}\cdots\frac{\lambda^{x_n}}{x_n!}\mathrm{e}^{-\lambda}=\frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\cdots x_n!}\mathrm{e}^{-n\lambda}.$$

Ее логарифм равен

$$l(\lambda)=\ln L(\lambda)=\ln\left(\frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!x_2!\cdots x_n!}\mathrm{e}^{-n\lambda}\right)=\ln\lambda\sum_{i=1}^{n}x_i-n\lambda-\ln(x_1!x_2!\cdots x_n!).$$

Чтобы функция $l(\lambda)$ достигала максимального значения, необходимо, чтобы ее производная $l'(\lambda)=0$, а вторая производная $l''(\lambda)<0$. В нашем случае

$$l'(\lambda)=\frac{d}{d\lambda}\left[\ln\lambda\sum_{i=1}^{n}x_i-n\lambda-\ln(x_1!x_2!\cdots x_n!)\right]=\frac{\sum_{i=1}^{n}x_i}{\lambda}-n,$$

$$l''(\lambda)=\frac{d}{d\lambda}\left[\frac{\sum_{i=1}^{n}x_i}{\lambda}-n\right]=-\frac{\sum_{i=1}^{n}x_i}{\lambda^2},$$

поэтому оценку максимального правдоподобия найдем из условия

$$\frac{\sum_{i=1}^{n}x_i}{\hat{\lambda}_{\text{ммп}}}-n=0:\ \hat{\lambda}_{\text{ммп}}=\frac{\sum_{i=1}^{n}x_i}{n}=\overline{x}.$$

При этом вторая производная отрицательна, так как в числителе дроби стоит положительное число (поскольку генеральная совокупность распределена по закону Пуассона и ее элементы не могут принимать отрицательных значений), в знаменателе дроби также стоит положительное число $\hat{\lambda}^2_{\text{ммп}}$, а перед дробью стоит знак «минус». ❑

Пример 5.2.4. Рассмотрим случайную величину $X = \text{Unif}(a;b)$. По выборке $x_1, x_2, \ldots, x_n$ найдем оценки параметров a и b методом моментов и методом максимального правдоподобия.

Найдем вначале оценки $\hat{a}_{\text{мм}}$ и $\hat{b}_{\text{мм}}$ с помощью метода моментов. Известно, что $\nu_1(X) = \mathbf{E}(X) = \dfrac{a+b}{2}$, $\mu_2(X) = \mathbf{Var}(X) = \dfrac{(b-a)^2}{12}$. Решив систему

$$\begin{cases} \nu_1 = \dfrac{a+b}{2}, \\ \mu_2 = \dfrac{(b-a)^2}{12} \end{cases}$$

относительно a и b (учитывая, что $a < b$), получим: $a = \nu_1 - \sqrt{3\mu_2}$, $b = \nu_1 + \sqrt{3\mu_2}$, а потому $\hat{a}_{\text{мм}} = \hat{\nu}_1 - \sqrt{3\hat{\mu}_2} = \overline{x} - \sqrt{3\hat{\sigma}^2_X}$, $\hat{b}_{\text{мм}} = \hat{\nu}_1 + \sqrt{3\hat{\mu}_2} = \overline{x} + \sqrt{3\hat{\sigma}^2_X}$.

Так как

$$f_X = \begin{cases} \dfrac{1}{b-a}, & x \in [a;b], \\ 0, & x \notin [a;b], \end{cases}$$

то функция правдоподобия

$$L(a;b) = \begin{cases} \dfrac{1}{(b-a)^n}, & \text{все } x_i \in [a;b],\ i = 1,2,\ldots,n, \\ 0, & \text{хотя бы одно значение } x_i \notin [a;b]. \end{cases}$$

Поскольку $a \leqslant \min\{x_1, x_2, \ldots, x_n\}$, а $b \geqslant \max\{x_1, x_2, \ldots, x_n\}$, максимум функции L достигается при $\hat{a}_{\text{ммп}} = \min\{x_1, x_2, \ldots, x_n\}$ и $\hat{b}_{\text{ммп}} = \max\{x_1, x_2, \ldots, x_n\}$. ❑

Задачи

346. Постройте по выборке $x_1, x_2, \ldots, x_n$ оценку параметра p геометрического распределения *а*) методом моментов; *б*) методом максимального правдоподобия.

347. По выборке $x_1, x_2, \ldots, x_n$ найдите оценку параметра μ случайной величины, распределенной по показательному закону: *а*) методом моментов; *б*) методом максимального правдоподобия.

348. Случайная величина $X = \mathcal{N}(a;\sigma)$. По выборке $x_1, x_2, \ldots, x_n$ найдите оценки параметров a и σ: *а*) методом моментов; *б*) методом максимального правдоподобия.

349. По данным социологического опроса получено распределение числа групп по числу респондентов в группе, отрицательно отзывающихся о новой рекламной политике фирмы (в каждой группе по 10 респондентов):

Число респондентов, не поддерживающих новую рекламную политику	0	1	2	3	4
Число групп	132	43	20	3	2

Предполагая, что число респондентов в группе, не поддерживающих новую рекламную политику, распределено по закону Пуассона, оцените параметр λ этого закона и определите долю групп, в которых все респонденты поддерживают новую рекламную политику.

§ 5.3. Интервальные оценки параметров

5.3.1. Двусторонние доверительные интервалы для математического ожидания

Основываясь на фундаментальных трудах К. Пирсона, посвященных теории статистического оценивания параметров, Р. Фишер и Е. Нейман разработали метод доверительных интервалов.

Вычисляя по выборке оценку $\hat{\theta}$ какого-либо параметра θ, мы отдаем себе отчет, что даже если она будет обладать свойствами состоятельности, несмещенности и эффективности, она все равно остается всего лишь приближенным значением параметра θ. Насколько же может отклониться это приближенное значение от истинного? Иными словами, можно ли указать интервал $(\hat{\theta}_1;\hat{\theta}_2)$, который с заранее заданной вероятностью γ (близкой к единице) накрывал бы истинное значение параметра θ. Такой интервал называется *доверительным интервалом* или *интервальной оценкой*, а вероятность γ — *доверительной вероятностью* или *надежностью* интервальной оценки.

Предположим, что наблюдения независимы и проводятся в одинаковых условиях, т. е. элементы выборки $X_1, X_2, \ldots, X_n$ представляют собой независимые случайные величины, распределенные по одному и тому же закону с математическим ожиданием a и дисперсией σ^2. Тогда, если объем выборки n велик (на практике — при $n > 30$), то в силу центральной предельной теоремы случайная величина

$$Z = \frac{(\bar{X} - a)\sqrt{n}}{\sigma}$$

имеет стандартный нормальный закон распределения $\mathcal{N}(0; 1)$.

Пример 5.3.1. При выборочной проверке $n = 100$ банковских счетов была получена оценка среднего остатка на счете: $\bar{x} = 25$ тыс. руб. Известно, что остаток на случайно выбранном счете X — это случайная величина, распределенная по нормальному закону, причем среднее квадратическое отклонение остатка на счете известно и равно $\sigma = 8$ тыс. руб.

Вычислим вероятность $\mathbf{P}\{0{,}95\bar{X} < \mathbf{E}(X) < 1{,}15\bar{X}\}$, с которой интервал $(0{,}95\bar{x};\ 1{,}15\bar{x})$ [или $(0{,}95 \cdot 25;\ 1{,}15 \cdot 25)$, т. е. $(23{,}75;\ 26{,}25)$] накроет неизвестное математическое ожидание случайной величины X (т. е. параметр a нормального закона распределения):

$$\mathbf{P}\{0{,}95\bar{X} < \mathbf{E}(X) < 1{,}15\bar{X}\} = \mathbf{P}\{\bar{X} - 1{,}15\bar{X} < \bar{X} - \mathbf{E}(X) < \bar{X} - 0{,}95\bar{X}\} =$$

$$= \mathbf{P}\left\{\frac{(\bar{X} - 1{,}15\bar{X})\sqrt{n}}{\sigma} < \frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{\sigma} < \frac{(\bar{X} - 0{,}95\bar{X})\sqrt{n}}{\sigma}\right\} =$$

$$= \mathbf{P}\left\{-\frac{0{,}05\bar{X}\sqrt{n}}{\sigma} < \frac{(\bar{X} - a)\sqrt{n}}{\sigma} < \frac{0{,}05\bar{X}\sqrt{n}}{\sigma}\right\} = \mathbf{P}\left\{-\frac{0{,}05\bar{X}\sqrt{n}}{\sigma} < Z < \frac{0{,}05\bar{X}\sqrt{n}}{\sigma}\right\} =$$

= НОРМ.СТ.РАСП(0,05*25*КОРЕНЬ(100)/8; ИСТИНА) –
– НОРМ.СТ.РАСП(0,05*25*КОРЕНЬ(100)/8; ИСТИНА) = 0,881829754. ❑

В этом примере мы вычислили вероятность, с которой заданный интервал накрывает неизвестное значение параметра a. А для того, чтобы построить доверительный интервал, нужно, наоборот, задавшись заданным уровнем вероятности, найти границы интервала, который с заданной вероятностью накроет неизвестное значение математического ожидания.

Пусть $z_{(1+\gamma)/2}$ — квантиль уровня $(1+\gamma)/2$ случайной величины $Z = \mathcal{N}(0; 1)$, т.е. такое число, что $\mathbf{P}\{Z \leqslant z_{(1+\gamma)/2}\} = F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2}) = (1+\gamma)/2$.

Тогда

$$\mathbf{P}\{|Z| < z_{(1+\gamma)/2}\} = \mathbf{P}\{-z_{(1+\gamma)/2} < Z < z_{(1+\gamma)/2}\} = F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2}) - F_{\mathcal{N}(0;1)}(-z_{(1+\gamma)/2}),$$

и учитывая, что $F_{\mathcal{N}(0;1)}(-x) = 1 - F_{\mathcal{N}(0;1)}(x)$, получаем:

$$\mathbf{P}\{|Z| < z_{(1+\gamma)/2}\} = \mathbf{P}\{-z_{(1+\gamma)/2} < Z < z_{(1+\gamma)/2}\} = F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2}) - F_{\mathcal{N}(0;1)}(-z_{(1+\gamma)/2}) =$$

$$= F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2}) - (1 - F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2})) = 2F_{\mathcal{N}(0;1)}(z_{(1+\gamma)/2}) - 1 = 2 \cdot \frac{1+\gamma}{2} - 1 = \gamma.$$

Итак,

$$\gamma = \mathbf{P}\{|Z| < z_{(1+\gamma)/2}\} = \mathbf{P}\left\{\left|\frac{(\overline{X} - a)\sqrt{n}}{\sigma}\right| < z_{(1+\gamma)/2}\right\}.$$

Но

$$\left|\frac{(\overline{X} - a)\sqrt{n}}{\sigma}\right| < z_{(1+\gamma)/2} \Leftrightarrow |a - \overline{X}| < \frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}} \Leftrightarrow$$

$$\Leftrightarrow -\frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}} < a - \overline{X} < \frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}} \Leftrightarrow \overline{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}} < a < \overline{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}},$$

значит,

$$\gamma = \mathbf{P}\left\{\left|\frac{(\overline{X} - a)\sqrt{n}}{\sigma}\right| < z_{(1+\gamma)/2}\right\} = \mathbf{P}\left\{\overline{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}} < a < \overline{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right\}, \quad (5.3.1)$$

т. е. вероятность того, что интервал

$$\left(\overline{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right)$$

накроет истинное значение математического ожидания a, равна γ (см. рис. 5.3.1).

Таким образом, получена **интервальная оценка математического ожидания при известном значении дисперсии** (или, что то же самое, при известном значении среднего квадратического отклонения).

Для удобства рассмотренные в этом и следующих пунктах формулы доверительных интервалов собраны в **прил. 3**.

Замечание. В качестве надежности, как правило, берут числа, близкие к единице: 0,9; 0,95; 0,99; 0,999; 0,999999 и т. д.

Пример 5.3.2. Построим в условиях примера 5.3.1 90%-ный доверительный интервал для математического ожидания остатка на случайно выбранном счете (среди всех счетов).

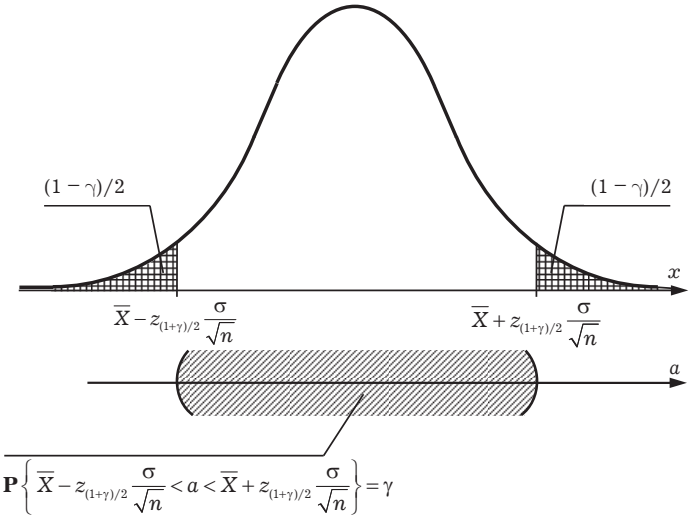


Рис. 5.3.1. *Доверительный интервал для математического ожидания при известном среднем квадратическом отклонении*

Поскольку по условию примера 5.3.1 остаток на случайно выбранном счете X — это случайная величина, распределенная по нормальному закону, подставим $\gamma = 0{,}9$, $u_{\gamma/2} = 1{,}65$, $\overline{x} = 25$, $\sigma = 8$, $n = 100$ в формулу доверительного интервала для математического ожидания нормальной случайной величины с известным средним квадратичным отклонением (5.3.1):

$$\mathbf{P}\left\{\overline{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}} < \mathbf{E}(X) < \overline{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right\} = \gamma$$

и получим искомый 90%-ный доверительный интервал для математического ожидания остатка на счете:

$$\left(25 - \frac{1{,}64485627 \cdot 8}{\sqrt{100}};\ 25 + \frac{1{,}64485627 \cdot 8}{\sqrt{100}}\right)$$

или, окончательно, (23,51963174; 26,48036826).

Квантиль $z_{(1+\gamma)/2} = z_{(1+0{,}9)/2} = z_{0{,}95} = 1{,}644853627$ стандартной нормальной случайной величины мы нашли с помощью статистической функции НОРМ.СТ.ОБР пакета Microsoft Excel: $z_{(1+\gamma)/2} =$ НОРМ.СТ.ОБР$((1 + \gamma)/2) =$ $=$ НОРМ.СТ.ОБР$((1 + 0{,}9)/2) = 1{,}644853627$. ❑

На практике точное значение среднего квадратического отклонения σ неизвестно, но по выборке можно получить его состоятельную, несмещенную и (для нормальной случайной величины) эффективную оценку

$$s = \sqrt{s_X^2} = \sqrt{\frac{n}{n-1}\hat{\sigma}_X^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}.$$

При большом объеме выборки n оправдано использование s в качестве приближенного значения σ в доверительном интервале (5.3.2) для математического ожидания a. Если же объем выборки не очень велик ($n < 30$), так делать нельзя.

Рассмотрим статистику

$$T_{n-1}=\frac{(\overline{X}-a)\sqrt{n}}{s}. \tag{5.3.2}$$

Оказывается, распределение величины T_{n-1} не зависит ни от $\overline{X}$, ни от s, а зависит только от числа $n-1$, называемого *числом степеней свободы*. Это распределение называется р а с п р е д е л е н и е м С т ь ю д е н т а (см. **п. 3.5.5**).

В Microsoft Excel есть функция $\overline{t}_{k;\alpha}=t_{k;1-\alpha/2}=$ СТЬЮДЕНТ.ОБР.2Х(<α>; <k>), вычисляющая абсолютное значение двусторонних критических границ уровня α случайной величины T_k, распределенной по закону Стьюдента с k степенями свободы.

Вспомним (см. **п. 3.6.3**), что двусторонние критические границы уровня случайной величины T_k — это такие числа $\underline{t}_{k;\alpha}=-\overline{t}_{k;\alpha}=-t_{k;1-\alpha/2}$, $\overline{t}_{k;\alpha}=-t_{k;1-\alpha/2}$, что $\mathbf{P}\{T_k\leqslant\underline{t}_{k;\alpha}\}=\mathbf{P}\{T_k>\overline{t}_{k;\alpha}\}=\alpha/2$. Значит, мы можем записать, что

$$\mathbf{P}\left\{T_{n-1}\leqslant-\overline{t}_{n-1;1-\gamma}\right\}=\frac{1-\gamma}{2};\quad \mathbf{P}\left\{T_{n-1}>\overline{t}_{n-1;1-\gamma}\right\}=\frac{1-\gamma}{2}.$$

Отсюда

$$\mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma}<T_{n-1}\leqslant\overline{t}_{n-1;1-\gamma}\right\}= \tag{5.3.3}$$

$$=1-\mathbf{P}\left\{T_{n-1}\leqslant-\overline{t}_{n-1;1-\gamma}\right\}-\mathbf{P}\left\{T_{n-1}>\overline{t}_{n-1;1-\gamma}\right\}=1-\frac{1-\gamma}{2}-\frac{1-\gamma}{2}=\gamma.$$

Подставим теперь в формулу (5.3.3) выражение статистики T_{n-1} из формулы (5.3.2) и учтем, что случайная величина T_k является абсолютно непрерывной (откуда $\mathbf{P}\{T_{n-1}\leqslant t\}=\mathbf{P}\{T_{n-1}<t\}$):

$$\mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma}<\frac{(\overline{X}-a)\sqrt{n}}{s}<\overline{t}_{n-1;1-\gamma}\right\}=\gamma\Leftrightarrow\mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma}<\frac{(a-\overline{X})\sqrt{n}}{s}<\overline{t}_{n-1;1-\gamma}\right\}=\gamma\Leftrightarrow$$

$$\Leftrightarrow\mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}<a-\overline{X}<\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\}=\gamma\Leftrightarrow$$

$$\Leftrightarrow\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}<a<\overline{X}+\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\}=\gamma.$$

Итак,

$$\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}<a<\overline{X}+\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\}=\gamma, \tag{5.3.4}$$

т. е. **интервальная оценка математического ожидания при неизвестном значении дисперсии** имеет вид

$$\left(\overline{X}-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}};\overline{X}+\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right),$$

где $\overline{X}$ — выборочное среднее, s — стандартное отклонение, а $\overline{t}_{n-1;1-\gamma}=$ = СТЬЮДЕНТ.ОБР.2Х($1-\gamma$; $n-1$).

Пример 5.3.3. Из некоторой отрасли было случайным образом отобрано 20 компаний. По выборочным данным оказалось, что выборочное среднее для срока, прошедшего с момента основания компании, составило $\overline{x} = 6$ лет при стандартном отклонении $s = 2$ г. Построим 99%-ный доверительный интервал для среднего времени, прошедшего с момента основания компании, для всех компаний отрасли.

Интервальная оценка генерального среднего (т. е. математического ожидания) при неизвестной дисперсии определяется формулой (5.3.4):

$$\mathbf{P}\left\{\overline{X} - \overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}} < \mathbf{E}(X) < \overline{X} + \overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma.$$

Подставляя в эту формулу значения $n = 20$, $\overline{x} = 6$ лет, $s = 2$ г., $\gamma = 0{,}99$ и

$$\overline{t}_{n-1;1-\gamma} = \overline{t}_{20-1;1-0{,}99} = \overline{t}_{19;0{,}01} = \text{СТЬЮДЕНТ.ОБР.2Х}(0{,}01;\ 19) = 2{,}860934606,$$

получим искомый 99%-ный доверительный интервал для $\mathbf{E}(X)$: (4,720551148; 7,279448852). ❑

Важно отметить, что речь идет о том, что доверительный интервал накроет истинное значение математического ожидания, но никак не о том, что математическое ожидание попадет в доверительный интервал. Ведь математическое ожидание, хоть оно нам и не известно, представляет собой совершенно конкретное число, а не случайную величину. А границы доверительного интервала зависят от элементов случайной выборки, поэтому являются случайными величинами. Это хорошо иллюстрирует следующий

Пример 5.3.4. Смоделируем выборку объемом 36 элементов из генеральной совокупности значений теста *IQ* (коэффициента интеллекта), распределенных по нормальному закону с математическим ожиданием $a = 100$ и средним квадратическим отклонением $\sigma = 16$.

Для этого введем в диапазон H7:AQ7 заголовки элементов выборки: «IQ 1», «IQ 2», ..., «IQ 36». В ячейку H8 введем формулу «=НОРМ.ОБР(СЛЧИС();100;16)» (для разыгрывания значения случайной величины $\mathcal{N}(a = 100; \sigma = 16)$. Распространим формулу из ячейки H8 на диапазон H8:AQ8 — теперь в этом диапазоне смоделирована выборка из 36 значений случайной величины $\mathcal{N}(a = 100; \sigma = 16)$.

В ячейки D7 и E7 введем заголовки «X_среднее» и «s», а в ячейках вычислим соответственно выборочное среднее для нашей выборки («=СРЗНАЧ(H8:AQ8)») и стандартное отклонение («=СТАНДОТКЛОН.В(H8:AQ8)»).

Теперь по данной выборке вычислим границы доверительного интервала для математического ожидания коэффициента интеллекта. Для этого нам потребуются двусторонние критические точки.

Введем в ячейку A1 заголовок «t_верх_35; 0,05», и в ячейке A2 вычислим абсолютное значение двусторонних критических точек $\overline{t}_{36-1;1-0{,}95} = \overline{t}_{35;0{,}05} =$ = СТЬЮДЕНТ.ОБР.2Х(0,05; 35) = 2,030107928.

Введем заголовки «Нижн. граница» и «Верхн. граница» в ячейки B7 и C7, а в ячейках B7 и C7 вычислим нижнюю границу («=D8-($A$2*E8/КОРЕНЬ(36))») доверительного интервала для математического ожидания коэффициента интеллекта и верхнюю границу («=D8+($A$2*E8/КОРЕНЬ(36))») этого доверительного интервала. В этих формулах зафиксируем с помощью знака «$» ссылки на ячейку A2, содержащую $\overline{t}_{35;0{,}05}$.

Видим (см. рис. 5.3.2), что данный доверительный интервал (95,040041; 107,394608) действительно накрывает истинное значение математического ожидания коэффициента интеллекта (число 100).

	A	B	C	D	E	F	G	H	I	J	...	AQ
1	t_верх_35; 0,05										...	
2	2,030107928										...	
3											...	
4	Число доверительных интервалов, накрывших истинное значение математического ожидания (равное 100)										...	
5	93										...	
6											...	
7	Включает 100	Нижн. граница	Верхн. граница	X_среднее	s	Выборка		IQ 1	IQ 2	IQ 3	...	IQ 36
8	Да	95,040041	107,394608	101,217325	18,257009	1		108,399046	90,502323	99,440488	...	123,938125
9	Да	96,456032	108,384680	102,420356	17,627607	2		113,416995	99,375340	103,508761	...	111,030132
10	Да	93,616995	107,440625	100,528810	20,427923	3		120,947248	88,419535	80,879546	...	110,332770
11	Да	91,522305	101,954886	96,738595	15,416787	4		120,074231	93,327581	111,498393	...	108,679442
12	Да	97,789717	108,017226	102,903472	15,113742	5		107,298900	115,125925	106,032218	...	104,183382
13	Да	94,088077	103,625515	98,856796	14,093987	6		93,487543	94,399719	104,613888	...	73,411087
14	**Нет**	100,883224	110,855735	105,869480	14,736920	7		98,541239	129,19559	106,20553	...	114,831898
15	Да	94,506054	108,336213	101,421134	20,437573	8		132,792126	82,700691	102,396642	...	143,656881
...	...	...	...	...	...	...	...	...	...	...	...	...
107	Да	92,786515	101,879249	97,332882	13,74759	100		125,7889	81,967643	83,480403	...	98,568532

Рис. 5.3.2. *Расчет 100 доверительных интервалов для генерального среднего коэффициента интеллекта*

Если нажать клавишу F9 (или кнопку «Пересчет» на вкладке «Формулы»), то все формулы в ячейках будут пересчитаны, и мы получим другую выборку, которой соответствует другой доверительный интервал.

Автоматизируем процесс идентификации ситуации, в которой доверительный интервал накрывает истинное значение математического ожидания. Для этого введем в ячейку A7 заголовок «Включает 100», а в ячейку A8 — формулу «=ЕСЛИ(И(B8<100; C8>100); "Да"; "Нет")». Теперь ячейка A8 принимает значение «Да», если доверительный интервал, построенный по выборке из диапазона H8:AQ8, накрывает число 100 — истинное значение математического ожидания коэффициента интеллекта (и значение «Нет» в противном случае).

Построим еще 99 выборок, по которым также рассчитаем границы доверительных интервалов для генерального среднего коэффициента интеллекта. Для этого выделим диапазон A8:AQ8 и «протянем» его (за правый угол) на 99 строк вниз. Введем в ячейку F8 заголовок «Выборка», и в диапазоне F8:F107 пронумеруем выборки от 1 до 100.

Видим, что в большинстве ячеек из диапазона A8:A107 содержится значение «Да», но есть и такие строки, где в первом столбце написано «Нет».

Вычислим, сколько из 100 построенных доверительных интервалов действительно накрывают истинное значение математического ожидания (т. е. число 100).

Для этого в ячейку A1 введем заголовок «Число доверительных интервалов, накрывших истинное значение математического ожидания (равное 100)», а в ячейку A2 — формулу «=СЧЁТЕСЛИ(A8:A107; "Да")».

Результаты сохранены в файле «100 доверительных интервалов.xlsx» и проиллюстрированы рис. 5.3.2. Замечаем, что 93 из 100 построенных доверительных интервалов (т. е. 93% построенных доверительных интервалов) накрыли истинное значение генерального среднего коэффициента интеллекта.

Нажмем клавишу «F9», тогда все формулы будут пересчитаны, мы получим 100 новых выборок и 100 новых доверительных интервалов. Из них накроют истинное значение математического ожидания коэффициента интеллекта уже не 93, а, например, 97 или 89. Но поскольку доверительный интервал с 95%-ной надежностью накрывает истинное значение математического ожидания, количество доверительных интервалов, накрывших число 100, будет около 95. ❑

Пример 5.3.5. Построим 95%-ные доверительные интервалы для математических ожиданий месячных доходностей акций российских компаний из примеров 1.1.3, 1.2.10, 1.2.13, 1.4.10.

В примере 1.2.4 были рассчитаны выборочные средние и стандартные отклонения месячных доходностей акций:

$$\overline{x}_{\text{ВТБ}} = 0{,}006526318;\ s_{\text{ВТБ}} = 0{,}094495296;$$
$$\overline{x}_{\text{Газпром}} = -0{,}001153938;\ s_{\text{Газпром}} = 0{,}067803139;$$
$$\overline{x}_{\text{Сбербанк}} = 0{,}015489170;\ s_{\text{Сбербанк}} = 0{,}093977345$$

(см. вкладку «Описательная статистика» в файле «Акции.xlsx» и табл. 1.2.4).

Учитывая, что каждая из трех выборок доходностей акций состоит из $n = 56$ наблюдений, найдем абсолютное значение двусторонних критических точек $\overline{t}_{56-1;1-0{,}95} = \overline{t}_{55;0{,}05}$ = СТЬЮДЕНТ.ОБР.2Х(0,05; 35) = 2,004044783.

Теперь рассчитаем по формуле (5.3.4) границы 95%-ного доверительного интервала для математического ожидания месячной доходности акций ВТБ:

$$\left(\overline{x}_{\text{ВТБ}} - \overline{t}_{55;0{,}05}\frac{s_{\text{ВТБ}}}{\sqrt{n}};\ \overline{x}_{\text{ВТБ}} + \overline{t}_{55;0{,}05}\frac{s_{\text{ВТБ}}}{\sqrt{n}}\right)$$

или

$$\left(0{,}006526318 - \frac{2{,}004044783 \cdot 0{,}094495296}{\sqrt{56}};\ 0{,}006526318 + \frac{2{,}004044783 \cdot 0{,}094495296}{\sqrt{56}}\right)$$

или, окончательно,

$$(-0{,}018779687;\ 0{,}031832324).$$

Аналогично вычислим границы доверительных интервалов для генеральных средних месячных доходностей акций Газпрома:

$$\left(-0{,}001153938 - \frac{2{,}004044783 \cdot 0{,}067803139}{\sqrt{56}};\ -0{,}001153938 + \frac{2{,}004044783 \cdot 0{,}067803139}{\sqrt{56}}\right)$$

или

$$(-0{,}019311737;\ 0{,}017003861)$$

и Сбербанка:

$$\left(0{,}015489170 - \frac{2{,}004044783 \cdot 0{,}093977345}{\sqrt{56}};\ 0{,}015489170 + \frac{2{,}004044783 \cdot 0{,}093977345}{\sqrt{56}}\right)$$

или

$$(-0{,}009678128;\ 0{,}040656467).$$

Визуализируем три доверительных интервала на рис. 5.3.3.

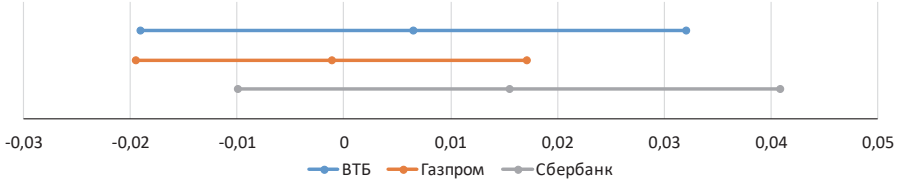


Рис. 5.3.3. *Доверительные интервалы для генеральных средних доходностей акций российских компаний*

Сравнивая доверительные интервалы для генеральных средних месячных доходностей акций ВТБ и Газпрома, видим, что их нижние границы практически совпадают, а верхняя граница доверительного интервала для генеральной средней месячной доходности акций ВТБ существенно больше верхней границы доверительного интервала для генеральной средней доходности акций Газпрома. Это дает основания заключить, что инвестиции в акции ВТБ в среднем могут в лучшем случае принести большую доходность, чем вложения в акции Газпрома, а в худшем случае убытки по вложениям в акции обеих компаний будут приблизительно одинаковы.

Если же мы сравним доверительные интервалы для генеральных средних месячных доходностей акций ВТБ и Сбербанка, то заметим, что длины обоих интервалов приблизительно одинаковы, но при этом нижняя и верхняя границы доверительного интервала для генеральной средней месячной доходности акций ВТБ меньше, чем соответственно нижняя и верхняя границы доверительного интервала для генеральной средней месячной доходности акций Сбербанка.

Из этого можно сделать вывод, что с 95%-ной надежностью на инвестициях в акции Сбербанка в лучшем случае можно в среднем больше заработать, чем на инвестициях в акции ВТБ, а в худшем случае — меньше потерять. ❑

Замечание. Важно отметить, что увеличение надежности при том же объеме выборке приводит к более широкому, т. е. менее точному, доверительному интервалу, а увеличение объема выборки при той же надежности — к менее широкому, т. е. более точному доверительному интервалу.

Для удобства рассмотренные в этом и следующих пунктах формулы для вычисления интервальных оценок основных числовых характеристик и параметров случайных величин сведены в табл. П.3, приведенную **в приложении**.

Задачи

350. На контрольных испытаниях 16 осветительных ламп были определены оценки математического ожидания и среднего квадратичного отклонения срока их службы: $\overline{x} = 3000$ ч и $s = 200$ ч. Считая, что срок службы лампы является нормально распределенной случайной величиной, определите: *а*) вероятность, с которой можно утверждать, что ошибка оценивания математического ожидания выборочным средним $\overline{x}$ не превзойдет 100 ч; *б*) 90%-ный доверительный интервал для математического ожидания срока службы лампы.

351. По данным о сумме пополнения вклада за месяц для 225 клиентов вычислены выборочное среднее $\overline{x} = 48$ тыс. руб. и стандартное отклонение $s = 12$ тыс. руб. Предполагая, что суммарное пополнение вклада случайно выбранным клиентом — случайная величина X, распределенная по нормальному закону, определите: *а*) вероятность того, что интервал $(0{,}9\overline{x};\ 1{,}1\overline{x})$ накроет математическое ожидание случайной величины X; *б*) 95%-ный доверительный интервал для математического ожидания случайной величины X.

352. По данным о цене закрытия для некоторой акции за последние 20 дней были получены выборочное среднее $\overline{x} = 100$ ден. ед. и стандартное отклонение $s = 1{,}6$ ден. ед. В предположении о нормальном распределении цены закрытия акции найдите вероятность того, что ее математическое ожидание накрывается интервалом $(0{,}9\overline{x};\ 1{,}1\overline{x})$.

353. Постройте 95%-ный доверительный интервал для математического ожидания роста студентов из задачи 4.

354. В условиях примера 5.1.15 постройте 90%-ный, 95%-ный и 99%-ный доверительные интервалы для математического ожидания объема ежедневных продаж.

355. Постройте 95%-ные доверительные интервалы для математических ожиданий цен, объемов торгов, логарифмов объемов торгов и доходностей акций из задачи 5.

5.3.2. Двусторонний доверительный интервал для вероятности

Предположим, что выборка $X_1, X_2, \ldots, X_n$ взята из генеральной совокупности, соответствующей индикатору некоторого события (успеха), тогда смысл элементов выборки таков:

$$X_i = \begin{cases} 1, & \text{если произошел успех в } i\text{-м испытании,} \\ 0, & \text{если не произошел успех в } i\text{-м испытании.} \end{cases}$$

При этом $\mathbf{E}(X_i) = p, \mathbf{Var}(X_i) = p(1-p)$, величина $m = \sum_{i=1}^{n} X_i$ представляет собой количество успехов в серии из n испытаний Бернулли, а выборочное среднее $\overline{X}$ совпадает с относительной частотой $\hat{p}$:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{m}{n} = \hat{p},$$

поэтому

$$Z = \frac{(\overline{X} - a)\sqrt{n}}{\sigma} = \frac{\frac{m}{n} - p}{\sqrt{p(1-p)/n}},$$

и по формуле 5.3.1

$$\gamma = \mathbf{P}\left\{\left|\frac{(\overline{X} - a)\sqrt{n}}{\sigma}\right| < z_{(1+\gamma)/2}\right\} = \mathbf{P}\left\{\left|\frac{\frac{m}{n} - p}{\sqrt{p(1-p)/n}}\right| < z_{(1+\gamma)/2}\right\} =$$

$$= \mathbf{P}\left\{\frac{m}{n} - z_{(1+\gamma)/2}\sqrt{\frac{p(1-p)}{n}} < p < \frac{m}{n} + z_{(1+\gamma)/2}\sqrt{\frac{p(1-p)}{n}}\right\}.$$

Истинное значение вероятности успеха p в единичном испытании нам неизвестно, но при большом объеме выборки n можно в качестве p взять приближенное значение — относительную частоту $\hat{p} = m/n$, являющуюся состоятельной, несмещенной и эффективной оценкой вероятности (см. **п. 5.2.1**). В результате получим **приближенный доверительный интервал**

$$\left(\frac{m}{n} - z_{(1+\gamma)/2}\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n}}; \frac{m}{n} + z_{(1+\gamma)/2}\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n}}\right), \tag{5.3.5}$$

который с вероятностью γ накроет истинное значение вероятности успеха p в единичном испытании.

Пример 5.3.6. Банк хочет установить степень известности некоторого кредитного продукта в данном городе. Для этого было опрошено $n = 400$ человек, $m = 80$ из которых сказали, что знакомы с продуктом (остальные — что незнакомы). Построим 95%-ный доверительный интервал для степени известности продукта среди всех жителей города.

Степень известности продукта — это неизвестная вероятность успеха в серии испытаний Бернулли (если успехом считать ситуацию, в которой случайно выбранный житель города окажется знакомым с данным продуктом).

Относительная частота успехов $\hat{p} = m / n = 80 / 400 = 0{,}2$, и поскольку объем выборки n достаточно велик, справедлива формула (5.3.4), подставив в которую $n = 400$, $\hat{p} = 0{,}2$, $\gamma = 0{,}95$, $z_{(1+\gamma)/2} = z_{0{,}975} =$ НОРМ.СТ.ОБР(0,975) $= 1{,}959963985$, получим 95%-ный доверительный интервал для степени известности продукта: (0,16080072; 0,23919928).

Таким образом, с 95%-ной надежностью можно утверждать, что с продуктом знакомы от 16 до 24% жителей города. ❑

В случае, когда число успехов или число неудач в выборке близко к нулю, доверительный интервал, рассчитываемый по формуле (5.3.4), вырождается в точку и перестает работать.

В таких ситуациях, когда $m = 0$, $m = 1$, $m = n - 1$ или $m = n$, можно пользоваться **формулами Блита:**

$$\text{при } m = 0 \qquad \mathbf{P}\left\{0 < p < 1 - (1-\gamma)^{1/n}\right\} = \gamma;$$

$$\text{при } m = 1 \qquad \mathbf{P}\left\{1 - \left(1 - \frac{1-\gamma}{2}\right)^{1/n} < p < 1 - \left(\frac{1-\gamma}{2}\right)^{1/n}\right\} = \gamma;$$

$$\text{при } m = n-1 \quad \mathbf{P}\left\{\left(\frac{1-\gamma}{2}\right)^{1/n} < p < \left(1 - \frac{1-\gamma}{2}\right)^{1/n}\right\} = \gamma;$$

$$\text{при } m = n \qquad \mathbf{P}\left\{(1-\gamma)^{1/n} < p < 1\right\} = \gamma.$$

Пример 5.3.7. Предположим, что банк выдал $n = 500$ кредитов, и все заемщики полностью исполнили условия кредитных договоров (т. е. $m = n = 500$, $\hat{p} = 500 / 500 = 1$). Зададим надежность $\gamma = 0{,}99$ и построим 99%-ный доверительный интервал для вероятности невозврата кредита среди всех потенциальных заемщиков данного банка.

Подставим в формулу (5.3.5) $n = 500$, $\hat{p} = 1$ $\gamma = 0{,}99$, $z_{(1+\gamma)/2} = z_{0{,}995} = 2{,}575829$, получим:

$$\mathbf{P}\left\{1 - 2{,}575829\sqrt{\frac{1 \cdot 0}{n}} < p < 1 + 2{,}575829\sqrt{\frac{1 \cdot 0}{n}}\right\} = 0{,}99 \ —$$

доверительный интервал превратился в единственную точку, равную единице. Конечно, это ошибка, в такой ситуации нельзя вместо вероятности подставлять относительную частоту!

Воспользовавшись формулой Блита, получим более адекватный доверительный интервал:

$$\mathbf{P}\left\{(1-0{,}99)^{1/500} < p < 1\right\} = 0{,}99 \qquad \text{или} \qquad \mathbf{P}\left\{0{,}9908319 < p < 1\right\} = 0{,}99\,. \ ❑$$

Задачи

356. На $n = 250$ разосланных рекламных писем пришло $m = 32$ ответа с заказами продукции фирмы. Постройте доверительный интервал, накрывающий с надежностью $\gamma = 0{,}99$ неизвестную вероятность p воздействия рекламного письма на клиента.

357. Из 2000 выданных некоторым банком кредитов 375 не были возвращены в срок. Постройте 95%-ный и 99%-ный доверительные интервалы для вероятности невозврата кредита.

358. Из n выданных банком кредитов все, кроме одного, были возвращены вовремя. Постройте 95%-ные доверительные интервалы для вероятности невозврата кредита: *а*) при $n = 6$; *а*) при $n = 600$.

5.3.3. Двусторонний доверительный интервал для дисперсии

Рассмотрим статистику

$$\frac{(n-1)s^2}{\sigma^2} = \chi^2_{n-1}. \tag{5.3.6}$$

Распределение статистики (5.3.6) зависит только от числа степеней свободы $n-1$ (где n — объем выборки). Это распределение называется распределением «Хи-квадрат» (см. **п. 3.5.5**).

В Microsoft Excel есть функция $\chi^2_{k;\alpha}$ = ХИ2.ОБР(<α>; <*k*>), вычисляющая квантиль уровня случайной величины χ^2_k, т. е. такое число, что $\mathbf{P}\{\chi^2_k \leqslant \chi^2_{k;\alpha}\} = \alpha$.

Поэтому можно записать, что

$$\mathbf{P}\{\chi^2_{n-1} \leqslant \chi^2_{n-1;(1+\gamma)/2}\} = (1+\gamma)/2,\ \mathbf{P}\{\chi^2_{n-1} \leqslant \chi^2_{n-1;(1-\gamma)/2}\} = (1-\gamma)/2,$$

откуда

$$\mathbf{P}\left\{\frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{n-1;(1+\gamma)/2}\right\} = (1+\gamma)/2,\ \mathbf{P}\left\{\frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{n-1;(1-\gamma)/2}\right\} = (1-\gamma)/2$$

или

$$\mathbf{P}\left\{\sigma^2 \geqslant \frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}}\right\} = (1+\gamma)/2,\ \mathbf{P}\left\{\sigma^2 \geqslant \frac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right\} = (1-\gamma)/2.$$

Теперь можно вычислить вероятность

$$\mathbf{P}\left\{\frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right\} = \mathbf{P}\left\{\sigma^2 > \frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}}\right\} - \mathbf{P}\left\{\sigma^2 \geqslant \frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}}\right\} =$$

$$= (1+\gamma)/2 - (1-\gamma)/2 = \gamma.$$

Итак,

$$\mathbf{P}\left\{\frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right\} = \gamma \tag{5.3.7}$$

т. е. **интервальная оценка дисперсии** имеет вид

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}};\ \frac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right),$$

где s^2 — исправленная выборочная дисперсия,

$\chi^2_{n-1;(1+\gamma)/2}$ = ХИ2.ОБР((1 + γ)/2; $n-1$), $\chi^2_{n-1;(1-\gamma)/2}$ = ХИ2.ОБР((1 – γ)/2; $n-1$).

Пример 5.3.8. По данным за последние $n = 20$ дней были получены выборочное среднее $\bar{x} = 100{,}00$ руб. и стандартное отклонение $s = 1{,}60$ руб. случайной величины X — цены некоторой акции. В предположении о нормальном распределении случайной величины X найдем вероятность того, что ее среднее квадратическое отклонение σ будет накрываться интервалом $(s - 0{,}08; s + 0{,}08)$, а математическое ожидание $\mathbf{E}(X)$ — интервалом $(0{,}9\bar{x};\ 1{,}1\bar{x})$.

Поскольку случайная величина X распределена по нормальному закону, статистика $\chi^2_{n-1} = (n-1)s^2/\sigma^2$ распределена по закону «Хи-квадрат» с $n-1$ степенью свободы, поэтому

$$\mathbf{P}\{s - 0{,}08 < \sigma < s + 0{,}08\} =$$

$$= \mathbf{P}\left\{\frac{1}{s+0{,}08} < \frac{1}{\sigma} < \frac{1}{s-0{,}08}\right\} = \mathbf{P}\left\{\frac{1}{(s+0{,}08)^2} < \frac{1}{\sigma^2} < \frac{1}{(s-0{,}08)^2}\right\} =$$

$$= \mathbf{P}\left\{\frac{(n-1)s^2}{(s+0{,}08)^2} < \frac{(n-1)s^2}{\sigma^2} < \frac{(n-1)s^2}{(s-0{,}08)^2}\right\} = \mathbf{P}\left\{\frac{(n-1)s^2}{(s+0{,}08)^2} < \chi^2_{n-1} < \frac{(n-1)s^2}{(s-0{,}08)^2}\right\} =$$

$$= \mathbf{P}\left\{\frac{19 \cdot 1{,}60^2}{1{,}68^2} < \chi^2_{19} < \frac{19 \cdot 1{,}60^2}{1{,}52^2}\right\} = \mathbf{P}\{17{,}23356009 < \chi^2_{19} < 21{,}05263158\} =$$

$$= \mathbf{P}\{\chi^2_{19} \leqslant 21{,}05263158\} - \mathbf{P}\{\chi^2_{19} \leqslant 17{,}23356009\} =$$

= ХИ2.РАСП(21,05263158;19;ИСТИНА) – ХИ2.РАСП(17,23356009;19;ИСТИНА) =

$$= 0{,}66609997 - 0{,}42595187 = 0{,}240148098.$$

Далее, статистика $T_{n-1} = (\bar{X} - \mathbf{E}(X))\sqrt{n}/s$ имеет распределение Стьюдента с $n-1$ степенью свободы, поэтому

$$\mathbf{P}\{0{,}9\bar{X} < \mathbf{E}(X) < 1{,}1\bar{X}\} = \mathbf{P}\left\{\frac{(\bar{X} - 1{,}1\bar{X})\sqrt{n}}{s} < \frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} < \frac{(\bar{X} - 0{,}9\bar{X})\sqrt{n}}{s}\right\} =$$

$$= \mathbf{P}\left\{-\frac{0{,}1\bar{X}\sqrt{n}}{s} < T_{n-1} < \frac{0{,}1\bar{X}\sqrt{n}}{s}\right\} = \mathbf{P}\left\{-\frac{0{,}1 \cdot 100{,}00 \cdot \sqrt{20}}{1{,}6} < T_{n-1} < \frac{0{,}1 \cdot 100{,}00 \cdot \sqrt{20}}{1{,}6}\right\} =$$

$$= \mathbf{P}\{-27{,}95084972 < T_{n-1} < 27{,}95084972\} =$$

= СТЬЮДЕНТ.РАСП(27,95084972;19;ИСТИНА) –

– СТЬЮДЕНТ.РАСП(–27,95084972;19;ИСТИНА) ≈ 1. ❑

Пример 5.3.9. По данным о случайной величине X — ежемесячной сумме пополнения вклада для $n = 100$ вкладчиков банка месяцев вычислены выборочное среднее $\bar{x} = 90$ тыс. руб. и стандартное отклонение $s = 8$ тыс. руб.

Предполагая, что случайная величина X распределена по нормальному закону, определим вероятность того, что интервал $(0{,}9\bar{x}; 1{,}1\bar{x})$ накроет математическое ожидание и генеральное среднее квадратическое отклонение случайной величины X, а интервал $(0{,}8s; 1{,}1s)$ — генеральное среднее квадратическое отклонение случайной величины X.

Поскольку случайная величина X распределена по нормальному закону, запишем:

$$\mathbf{P}\left\{0{,}9\bar{X} < a < 1{,}1\bar{X}\right\} = \mathbf{P}\left\{-1{,}1\bar{X} < -a < -0{,}9\bar{X}\right\} =$$

$$= \mathbf{P}\left\{\frac{(\bar{X} - 1{,}1\bar{X})\sqrt{n}}{s} < \frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} < \frac{(\bar{X} - 0{,}9\bar{X})\sqrt{n}}{s}\right\} =$$

$$= \mathbf{P}\left\{-\frac{0{,}1\bar{X}\sqrt{n}}{s} < \frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} < \frac{0{,}1\bar{X}\sqrt{n}}{s}\right\} = \mathbf{P}\left\{-\frac{0{,}1 \cdot 90\sqrt{100}}{8} < T_{99} < \frac{0{,}1 \cdot 90\sqrt{100}}{8}\right\} =$$

$$= \mathbf{P}\left\{-\frac{90}{8} < T_{99} < \frac{90}{8}\right\} = \mathbf{P}\left\{-11{,}25 < T_{99} < 11{,}25\right\} =$$

$$= \text{СТЬЮДЕНТ.РАСП}(11{,}25;\ 99;\ \text{ИСТИНА}) -$$

$$- \text{СТЬЮДЕНТ.РАСП}(-11{,}25;\ 99;\ \text{ИСТИНА}) \approx 1;$$

$$\mathbf{P}\left\{0{,}8s < \sigma < 1{,}1s\right\} = \mathbf{P}\left\{0{,}64s^2 < \sigma^2 < 1{,}21s^2\right\} = \mathbf{P}\left\{\frac{1}{1{,}21s^2} < \frac{1}{\sigma^2} < \frac{1}{0{,}64s^2}\right\} =$$

$$= \mathbf{P}\left\{\frac{n-1}{1{,}21} < \frac{(n-1)s^2}{\sigma^2} < \frac{n-1}{0{,}64}\right\} = \mathbf{P}\left\{\frac{n-1}{1{,}21} < \chi^2_{n-1} < \frac{n-1}{0{,}64}\right\} = \mathbf{P}\left\{\frac{99}{1{,}21} < \chi^2_{99} < \frac{99}{0{,}64}\right\} =$$

$$= \mathbf{P}\left\{81{,}81818182 < \chi^2_{99} < 154{,}6875\right\} = \mathbf{P}\left\{\chi^2_{99} \leqslant 154{,}6875\right\} - \mathbf{P}\left\{\chi^2_{99} \leqslant 81{,}81818182\right\} =$$

$$= \text{ХИ2.РАСП}(154{,}6875;\ 99;\ \text{ИСТИНА}) - \text{ХИ2.РАСП}(81{,}81818182;\ 99;\ \text{ИСТИНА}) =$$

$$= 0{,}89437053. \ \square$$

Задачи

359. Считая в условиях задачи 350, что срок службы лампы является нормально распределенной случайной величиной, распределенной по нормальному закону, определите: *а*) вероятность, с которой можно утверждать, что ошибка при замене генерального среднего квадратического отклонения величиной s не превзойдет 10 ч; *б*) 90%-ный доверительный интервал для генерального среднего квадратического отклонения срока службы лампы.

360. В условиях задачи 351 определите: *а*) вероятность того, что интервал $(0{,}8s;\ 1{,}1s)$ накроет генеральное среднее квадратическое отклонение случайной величины X; *б*) 95%-ный доверительный интервал для генерального среднего квадратического отклонения случайной величины X.

361. В условиях задачи 352 найдите вероятность того, что среднее квадратическое отклонение σ цены закрытия накрывается интервалом $(s - 0{,}08;\ s + 0{,}08)$.

362. Определите вероятность, с которой можно ожидать, что при объеме выборки $n = 27$ элементов ошибка от замены генерального среднего квадратического отклонения σ оценкой s не превзойдет $0{,}388s$.

363. Определите вероятность, с которой можно ожидать, что при выборке объемом $n = 11$ отношение s^2 / σ^2 будет заключено в интервале (0,256; 0,487).

364. Постройте 95%-ные доверительные интервалы для дисперсий и средних квадратических отклонений цен, объемов торгов, логарифмов объемов торгов и доходностей акций из задачи 5.

5.3.4. Поправка на конечный объем генеральной совокупности

Если объем генеральной совокупности конечен, а выборка представляет собой относительно большую долю генеральной совокупности (больше 5%), необходимо использовать вместо формул (5.3.1), (5.3.4) и (5.3.5) скорректированые формулы, учитывающие конечный объем генеральной совокупности.

При конечном объеме генеральной совокупности **доверительный интервал для математического ожидания** при известной дисперсии определяется формулой

$$\mathbf{P}\left\{\overline{X}-z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}<a<\overline{X}+z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right\}=\gamma\,, \qquad (5.3.8)$$

при неизвестной дисперсии — формулой

$$\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}<a<\overline{X}+\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right\}=\gamma. \qquad (5.3.9)$$

При конечном объеме генеральной совокупности **приближенный доверительный интервал для вероятности** принимает вид

$$\mathbf{P}\left\{\frac{m}{n}-z_{(1+\gamma)/2}\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)(N-n)}{n(N-1)}}<p<\frac{m}{n}+z_{(1+\gamma)/2}\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)(N-n)}{n(N-1)}}\right\}=\gamma. \qquad (5.3.10)$$

Замечание. Величина $fpc=\sqrt{(N-n)/(N-1)}$ называется поправочным коэффициентом для конечной генеральной совокупности. Поскольку числитель под корнем всегда меньше знаменателя, поправочный коэффициент fpc всегда меньше единицы, а это означает, что формулы (5.3.8), (5.3.9) и (5.3.10) дают для конечной генеральной совокупности более точные (т. е. более узкие) доверительные интервалы, чем формулы (5.3.1), (5.3.3) и (5.3.4).

Пример 5.3.10. В университете $N=500$ студенческих групп. Из них было случайным образом отобрано $n=25$ групп. По выборочным данным оказалось, что в среднем в группе учатся $\overline{x}=22$ девушки при стандартном отклонении $s=5$. Пользуясь 95%-ным доверительным интервалом, оценим число девушек во всем университете.

По формуле (5.3.9), в которой $\gamma=0{,}95$, $N=500$, $n=25$, $\overline{x}=22$, $s=5$ и $\overline{t}_{25-1;1-0{,}95}=\overline{t}_{24;0{,}05}=$ СТЬЮДЕНТ.ОБР.2Х(0,05; 35) $=2{,}063898562$, получим искомый доверительный интервал:

$$\mathbf{P}\left\{22-\frac{2{,}063898562\cdot 5}{\sqrt{25}}\sqrt{\frac{500-25}{500-1}}<a<22+\frac{2{,}063898562\cdot 5}{\sqrt{25}}\sqrt{\frac{500-25}{500-1}}\right\}=0{,}95\,.$$

Таким образом, интервал $(19{,}98634586;\ 24{,}01365414)$ с вероятностью 0,95 накрывает среднее количество девушек, которые учатся в одной случайно выбранной группе. Чтобы оценить общее число девушек в университете, учтем, что в институте $N=500$ групп, и получим, что с вероятностью 95% в институте учатся от $500\cdot 19{,}98634586=9993$ до $500\cdot 24{,}01365414=12\,007$ девушек. ❑

В случае конечной генеральной совокупности корректируются и **формулы для определения допустимого объема выборки**: вместо формул (5.1.12) и (5.1.13) используются соответственно формулы

$$n > \frac{N_0 N}{N_0 + N - 1}, \text{ где } N_0 = \frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2} \qquad (5.3.11)$$

и

$$n > \frac{N_0 N}{N_0 + N - 1}, \text{ где } N_0 = \frac{z^2_{(1+\gamma)/2} p(1-p)}{\varepsilon^2}. \qquad (5.3.12)$$

Пример 5.3.11. Из отрасли, включающей $N = 1200$ фирм, было случайным образом отобрано $n = 20$ фирм. По выборочным данным оказалось, что в фирме работают в среднем $\bar{x} = 77{,}5$ человек при исправленном среднем квадратичном отклонении $s = 20$ человек.

Пользуясь 95%-ными интервальными оценками, оценим среднее число, дисперсию и среднее квадратическое отклонение числа работающих в фирме по отрасли в целом, а также общее число работников отрасли.

Подставим значения $n = 20$, $\bar{x} = 77{,}5$, $s = 20$, $\gamma = 0{,}95$ и критическую точку $t_{1-\gamma;n-1} = t_{0{,}05;19}$ =СТЬЮДЕНТ.ОБР.2Х(0,05;19) = 2,093024054 в формулу (5.3.6):

$$\mathbf{P}\left\{\overline{X} - \overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < a < \overline{X} + \overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right\} = \gamma \text{ —}$$

получим 95%-ный доверительный интервал для среднего (по отрасли) числа работающих в одной фирме:

$$\mathbf{P}\left\{77{,}5 - \frac{2{,}093024054 \cdot 20}{\sqrt{20}}\sqrt{\frac{1200-20}{1200-1}} < a < 77{,}5 - \frac{2{,}093024054 \cdot 20}{\sqrt{20}}\sqrt{\frac{1200-20}{1200-1}}\right\} = 0{,}95$$

или

$$\mathbf{P}\{68{,}21417212 < a < 86{,}78582788\} = 0{,}95.$$

Далее подставим $n = 20$, $\bar{x} = 77{,}5$, $s = 20$, $\gamma = 0{,}95$, а также квантили $\chi^2_{(1-\gamma)/2;n-1} = \chi^2_{0{,}025;19}$ = ХИ2.ОБР((0,025; 19) = 8,90651648 и $\chi^2_{(1+\gamma)/2;n-1} = \chi^2_{0{,}975;19}$ = = ХИ2.ОБР((0,025; 19) = 32,85232686 в формулу

$$\mathbf{P}\left\{\frac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right\} = \gamma \text{ —}$$

получим 95%-ный доверительный интервал для генеральной дисперсии: (231,3382559; 853,3078017).

Теперь легко определить 95%-ный доверительный интервал для генерального среднего квадратического отклонения числа работающих в фирме: $\left(\sqrt{231{,}3382559};\ \sqrt{853{,}3078017}\right)$ или, окончательно, (15,20980789; 29,21143272).

Общее число работающих в отрасли в 1200 раз больше, чем общее число работающих в одной фирме, поэтому оно с 95%-ной надежностью заключено в интервале $(1200 \cdot 68{,}21417212; 1200 \cdot 86{,}78582788)$ или от 81 857 до 104 143 человек. ❑

Пример 5.3.12. Предположим, что мы хотим оценить среднюю годовую заработную плату руководителей 500 крупнейших компаний из списка *Fortune* 500. Предположим, что нам известно генеральное среднее квадратическое отклонение этих зарплат, равное 5 млн долл., а по выборке из 100 руководителей вычислена выборочная средняя годовая заработная плата $\bar{x} = 40$ млн долл.

По формуле (5.3.10), в которой $\gamma = 0{,}95$, $N = 500$, $n = 100$, $\overline{x} = 40$, $s = 5$ и $z_{(1+\gamma)/2} =$ $=$НОРМ.СТ.ОБР((1+0,95)/2) $= 1{,}959963985$, получаем:

$$\mathbf{P}\left\{40 - \frac{1{,}959963985 \cdot 5}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} < a < 40 + \frac{1{,}959963985 \cdot 5}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}}\right\} = 0{,}95$$

или (39,02001801; 40,97998199).

Отметим, что без учета поправки на конечную генеральную совокупность мы получили бы по формуле (5.3.4) чуть менее точный доверительный интервал: (39,12259962; 40,87740038).

Теперь определим объем выборки, который позволил бы нам гарантировать, что ошибка от замены математического ожидания годовой заработной платы выборочным средним не превысит $\varepsilon = 1$ млн долл.

Если не учитывать конечность генеральной совокупности, то объем выборкив соответствии с формулой (5.1.12) должен быть не меньше, чем 96 человек:

$$N_0 = \frac{\sigma^2 z^2_{(1+\gamma)/2}}{\varepsilon^2} = \frac{5^2 1{,}959963985^2}{1^2} = 96{,}03647052.$$

Коррекция на конечный объем генеральной совокупности позволяет существенно уменьшить объем выборки, гарантирующий, что погрешность генеральной средней годовой заработной платы выборочным средним окажется не более 1 млн долл.:

$$n > \frac{96{,}03647052 \cdot 500}{96{,}03647052 + 500 - 1} = 80{,}69797002,$$

т. е. достаточно отобрать 81 руководителя, что существенно меньше 96. ❑

Задачи

365. Для отрасли, включающей 800 фирм, была составлена случайная выборка из 20 фирм. По выборочным данным оказалось, что в фирме работают в среднем $\overline{x} = 77{,}5$ человек при стандартном отклонении $s = 20$ человек. Пользуясь 95%-ным доверительным интервалом, оцените: *а*) среднее число, дисперсию и среднее квадратичное отклонение работающих в фирме для отрасли в целом; *б*) общее число работающих в отрасли.

366. В случайной выборке из $n = 200$ аспирантов, специализирующихся в финтех-предпринимательстве, составленной по нескольким ведущим российским университетам, $m = 90$ человек оказались сыновьями банкиров. Оцените с 90%-ной надежностью: *а*) долю аспирантов обследованных университетов, специализирующихся в финтех-предпринимательстве, отцы которых являются банкирами; *б*) число таких аспирантов среди $N = 2000$ аспирантов.

367. Производственные рабочие $m = 20$ из $n = 50$ фирм принадлежат к национальным профсоюзам. Требуется: *а*) пользуясь 95%-ным доверительным интервалом, оценить долю тех предприятий, рабочие которых являются членами национальных профсоюзов, среди $N = 800$ фирм; *б*) с той же надежностью оценить общее число таких предприятий (среди 800 фирм).

5.3.5. Двусторонний доверительный интервал для прогноза

Предположим, что мы располагаем выборкой $X_1, X_2, \ldots, X_n$ и хотим построить интервальный прогноз будущего значения X_{n+1}. Доверительный интервал для математического ожидания для этого не подойдет, ведь он оценивает не будущее значение случайной величины X_{n+1}, а неизвестный числовой параметр — математическое ожидание этой случайной величины.

Чтобы построить доверительный интервал для прогноза X_{n+1}, воспользуемся тем, что статистика

$$T_{n-1} = \frac{\left(X_{n+1} - \overline{X}\right)\sqrt{n}}{s\sqrt{n+1}} \tag{5.3.13}$$

имеет закон распределения Стьюдента с $n-1$ степенью свободы, и подставим выражение (5.3.13) в формулу (5.3.3):

$$\gamma = \mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma} < T_{n-1} < \overline{t}_{n-1;1-\gamma}\right\} = \mathbf{P}\left\{-\overline{t}_{n-1;1-\gamma} < \frac{\left(X_{n+1} - \overline{X}\right)\sqrt{n}}{s\sqrt{n+1}} < \overline{t}_{n-1;1-\gamma}\right\} =$$

$$= \mathbf{P}\left\{-\frac{\overline{t}_{n-1;1-\gamma}s\sqrt{n+1}}{\sqrt{n}} < X_{n+1} - \overline{X} < \frac{\overline{t}_{n-1;1-\gamma}s\sqrt{n+1}}{\sqrt{n}}\right\} =$$

$$= \mathbf{P}\left\{\overline{X} - \overline{t}_{n-1;1-\gamma}s\sqrt{1+\frac{1}{n}} < X_{n+1} < \overline{X} + \overline{t}_{n-1;1-\gamma}s\sqrt{1+\frac{1}{n}}\right\}.$$

Таким образом, **доверительный интервал для прогноза будущего значения** X_{n+1} с надежностью γ определяется формулой

$$\mathbf{P}\left\{\overline{X} - \overline{t}_{n-1;1-\gamma}s\sqrt{1+\frac{1}{n}} < X_{n+1} < \overline{X} + \overline{t}_{n-1;1-\gamma}s\sqrt{1+\frac{1}{n}}\right\} = \gamma. \tag{5.3.14}$$

Пример 5.3.13. Построим 95%-ные доверительные интервалы для прогноза месячных доходностей акций российских компаний из примеров 1.1.3, 1.2.10, 1.2.13, 1.4.10, 5.3.5 на следующий месяц.

Подставляя в формулу (5.3.14) значения

$$n = 56;\ \ \overline{t}_{56-1;1-0,95} = \overline{t}_{55;0,05} = \text{СТЬЮДЕНТ.ОБР.2Х(0,05; 55)} = 2{,}004044783;$$

$$\overline{x}_{\text{ВТБ}} = 0{,}006526318;\ \ s_{\text{ВТБ}} = 0{,}094495296;$$

$$\overline{x}_{\text{Газпром}} = -0{,}001153938;\ \ s_{\text{Газпром}} = 0{,}067803139;$$

$$\overline{x}_{\text{Сбербанк}} = 0{,}015489170;\ \ s_{\text{Сбербанк}} = 0{,}093977345,$$

получим для прогноза доходности акций ВТБ на следующий месяц 95%-ный доверительный интервал

$$\left(\overline{x}_{\text{ВТБ}} - \overline{t}_{55;0,05}s_{\text{ВТБ}}\sqrt{1+\frac{1}{n}};\ \overline{x}_{\text{ВТБ}} + \overline{t}_{55;0,05}s_{\text{ВТБ}}\sqrt{1+\frac{1}{n}}\right)$$

или

$$\left(0{,}00652632 - 2{,}00404478 \cdot 0{,}09449530\sqrt{1+\frac{1}{56}}; 0{,}00652632 + 2{,}00404478 \cdot 0{,}09449530\sqrt{1+\frac{1}{56}}\right)$$

или, окончательно,

$$\left(-0{,}184529834;\ 0{,}197582470\right).$$

Аналогично вычислим границы доверительных интервалов для прогнозов доходностей акций Газпрома:

$$(-0{,}138242316;\ 0{,}135934440)$$

и Сбербанка:

$$(-0{,}174519759;\ 0{,}205498098).$$

Замечаем, что доверительные интервалы для прогнозов гораздо шире, чем доверительные интервалы для соответствующих математических ожиданий!

Интервальные оценки математических ожиданий, полученные в примере 5.3.5, можно использовать при принятии решений о выборе активов для долгосрочных инвестиций (как было показано в примере 5.3.5, с 95%-ной надежностью на инвестициях в акции Сбербанка в лучшем случае можно за достаточно длительный период в среднем больше заработать, чем на инвестициях в акции ВТБ или Газпрома, а в худшем случае — меньше потерять).

Но если говорить об инвестициях на один месяц, то выбор актива гораздо сложнее: интервальные оценки прогнозов доходностей позволяют с надежностью 95% гарантировать, что вложения в акции Газпрома по сравнению с инвестициями в акции Сбербанка приведут к меньшим потерям в неблагоприятном случае (хотя и к меньшим доходам в благоприятном случае). Окончательный выбор зависит от склонности инвестора к риску. ❑

Задачи

368. Постройте 90%-ный и 95%-ный доверительные интервалы для срока службы купленной новой осветительной лампы из той же партии, из которой извлекалась выборка в условиях задачи 350.

369. По данным задачи 352 постройте и сравните 95%-ные доверительные интервалы для математического ожидания цены закрытия данной акции и для прогноза завтрашней цены закрытия акции.

370. В условиях примера 5.1.15 постройте 90%-ный, 95%-ный и 99%-ный доверительные интервалы для завтрашнего объема продаж и сравните их с полученными в задаче 354 доверительными интервалами для математического ожидания объема продаж.

371. Для акций из задачи 5 постройте 95%-ные доверительные интервалы для прогнозов на следующую неделю значений цен, объемов торгов, логарифмов объемов торгов и доходностей.

5.3.6. Односторонние доверительные интервалы

Во многих ситуациях бывает полезно строить *односторонние доверительные интервалы*, т. е. такие доверительные интервалы, одна из границ которых равна $-\infty$ или $+\infty$.

Учитывая, что функция Microsoft Excel $t_{k;\alpha}$ = СТЬЮДЕНТ.ОБР(<α>; <k>) вычисляет квантиль уровня α случайной величины T_k, распределенной по закону Стьюдента с k степенями свободы, т. е. такое число $t_{k;\alpha}$, что $\mathbf{P}\{T_k \leqslant t_{k;\alpha}\} = \alpha$, а статистика $\left(\overline{X} - \mathbf{E}(X)\right)\sqrt{n}/s$ распределена по закону Стьюдента с $n-1$ степенью свободы, можно записать:

$$\mathbf{P}\left\{T_{n-1} \leqslant t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{T_{n-1} < t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} < t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{-\frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} > -t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\mathbf{E}(X) - \bar{X} > -t_{n-1;\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{\mathbf{E}(X) > \bar{X} - t_{n-1;\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma.$$

Таким образом, односторонний доверительный интервал, определяющий с надежностью γ **нижнюю границу математического ожидания**, задается формулой

$$\mathbf{P}\left\{\bar{X} - t_{n-1;\gamma}\frac{s}{\sqrt{n}} < \mathbf{E}(X) < +\infty\right\} = \gamma. \tag{5.3.15}$$

Аналогично можно получить односторонний доверительный интервал, определяющий **верхнюю границу математического ожидания** с надежностью γ:

$$\mathbf{P}\left\{T_{n-1} \leqslant t_{n-1;1-\gamma}\right\} = 1-\gamma \Leftrightarrow \mathbf{P}\left\{T_{n-1} > t_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} > t_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{-\frac{(\bar{X} - \mathbf{E}(X))\sqrt{n}}{s} < -t_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\mathbf{E}(X) - \bar{X} < -t_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{\mathbf{E}(X) < \bar{X} - t_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma.$$

Учитывая четность плотности распределения случайной величины T_k, заметим, что $t_{n-1;1-\gamma} = -t_{n-1;\gamma}$, поэтому формулу одностороннего доверительного интервала с надежностью γ для верхней границы математического ожидания окончательно запишем в следующем виде:

$$\mathbf{P}\left\{-\infty < \mathbf{E}(X) < \bar{X} + t_{n-1;\gamma}\frac{s}{\sqrt{n}}\right\} = \gamma. \tag{5.3.16}$$

Поскольку статистика $\chi^2_{n-1} = (n-1)s^2/\sigma^2$ распределена по закону «Хи-квадрат» с $n-1$ степенью свободы, а функция Microsoft Excel $\chi^2_{k;\alpha} =$ $=$ ХИ2.ОБР(<α>; <k>) вычисляет квантиль уровня α случайной величины χ^2_k, т. е. такое число, что $\mathbf{P}\left\{\chi^2_k \leqslant \chi^2_{k;\alpha}\right\} = \alpha$, можно записать:

$$\mathbf{P}\left\{\chi^2_{n-1} \leqslant \chi^2_{n-1;1-\gamma}\right\} = 1-\gamma \Leftrightarrow \mathbf{P}\left\{\chi^2_{n-1} > \chi^2_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\frac{(n-1)s^2}{\sigma^2} > \chi^2_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{\sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;1-\gamma}}\right\} = \gamma.$$

Таким образом, получаем односторонний доверительный интервал, определяющий с надежностью γ **верхнюю границу генеральной дисперсии** σ^2 (рис. 5.3.4):

$$\mathbf{P}\left\{\sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;1-\gamma}}\right\} = \gamma. \tag{5.3.17}$$

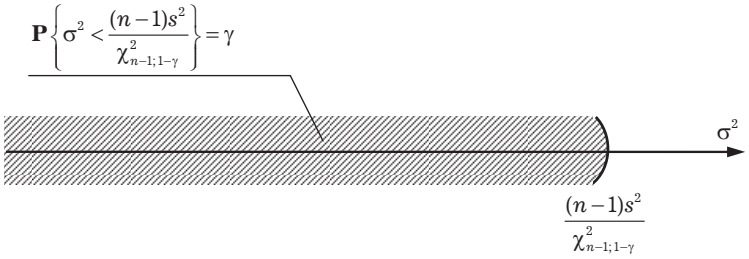


Рис. 5.3.4. *Односторонний доверительный интервал для дисперсии*

Пример 5.3.14. В условиях примера 5.3.9 определим с надежностью 0,9 верхние границы для генеральной дисперсии и для математического ожидания суммы ежемесячного пополнения случайно выбранного вклада.

Подставив в формулу (5.3.17) одностороннего доверительного интервала для генеральной дисперсии σ^2 значения $\gamma = 0{,}9$, $n = 100$, $s = 8$, $\chi^2_{n-1;\gamma} = \chi^2_{99;0,9} =$ $=$ ХИ2.ОБР(0,9; 99) $= 117{,}4068832$, получим, что с вероятностью 0,9 генеральная дисперсия ежемесячного пополнения окажется меньше чем

$$\frac{99 \cdot 8^2}{117{,}4068832} = 53{,}96617156 \text{ (тыс. руб.)}^2.$$

Перейдем к определению 90%-ной верхней границы математического ожидания ежемесячного пополнения вклада.

Подставляя в формулу (5.3.16) значения $\gamma = 0{,}9$, $n = 100$, $\bar{x} = 90$, $s = 8$, $t_{n-1;\gamma} = t_{99;0,9} =$ СТЬЮДЕНТ.ОБР(0,1; 99) $= 1{,}290161442$, найдем искомую 90%-ную верхнюю границу для математического ожидания ежемесячного пополнения:

$$\bar{X} - t_{n-1;1-\gamma}\frac{s}{\sqrt{n}} = 90 + 1{,}290161442 \cdot \frac{8}{\sqrt{100}} = 91{,}03212915 \text{ тыс. руб.} \ \square$$

Построим теперь односторонний доверительный интервал, определяющий с надежностью γ **верхнюю границу прогноза будущего значения** X_{n+1}:

$$\mathbf{P}\left\{T_{n-1} \leqslant t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{T_{n-1} < t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\frac{(X_{n+1} - \bar{X})\sqrt{n}}{s\sqrt{n+1}} < t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{X_{n+1} - \bar{X} < t_{n-1;\gamma}s\sqrt{\frac{n+1}{n}}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{X_{n+1} < \bar{X} + t_{n-1;\gamma}s\sqrt{1+\frac{1}{n}}\right\} = \gamma. \qquad (5.3.18)$$

Аналогично строится односторонний доверительный интервал, определяющий с надежностью γ **нижнюю границу прогноза будущего значения** X_{n+1}:

$$\mathbf{P}\left\{T_{n-1} \leqslant t_{n-1;1-\gamma}\right\} = 1-\gamma \Leftrightarrow \mathbf{P}\left\{T_{n-1} > t_{n-1;1-\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{T_{n-1} > -t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P}\left\{\frac{(X_{n+1} - \bar{X})\sqrt{n}}{s\sqrt{n+1}} > -t_{n-1;\gamma}\right\} = \gamma \Leftrightarrow \mathbf{P}\left\{X_{n+1} - \bar{X} > -t_{n-1;\gamma}s\sqrt{\frac{n+1}{n}}\right\} = \gamma \Leftrightarrow \quad (5.3.19)$$

$$\Leftrightarrow \mathbf{P}\left\{X_{n+1} > \bar{X} - t_{n-1;\gamma}s\sqrt{1+\frac{1}{n}}\right\} = \gamma.$$

Вспомним определение суммы под риском (**VaR**) из **п. 3.6.4:** 100γ%-ной суммой под риском для инвестиционной операции за период T называется такое число $\mathbf{VaR}_\gamma$, что с вероятностью γ убыток $(-X)$ от проведения операции в течение данного периода будет не больше, чем $\mathbf{VaR}_\gamma$: $\mathbf{P}\{-X \leqslant \mathbf{VaR}_\gamma\} = \gamma$.

Проиллюстрируем следующим примером, как можно получить **оценку показателя VaR** с помощью одностороннего доверительного интервала для прогноза доходности.

Пример 5.3.15. Найдем оценку 95%-ной суммы под риском на ближайший месяц для пакета акций ВТБ из примеров 1.1.3, 1.2.10, 1.2.13, 1.4.10, 5.3.5, 5.3.13, приобретаемого в данный момент за 1 млн руб.

Подставив в формулу (5.3.19) значения

$$n = 56;\ \ t_{n-1;\gamma} = t_{55;0,95} = \text{СТЬЮДЕНТ.ОБР(0,95; 55)} = 1{,}673033965;$$
$$\overline{x}_{\text{ВТБ}} = 0{,}006526318;\ \ s_{\text{ВТБ}} = 0{,}094495296,$$

построим односторонний доверительный интервал, определяющий с надежностью $\gamma = 0{,}95$ нижнюю границу прогноза будущего значения доходности акций ВТБ через месяц:

$$\mathbf{P}\left\{X_{n+1}^{(\text{ВТБ})} > 0{,}006526318 - 1{,}673033965 \cdot 0{,}094495296 \cdot \sqrt{1+\frac{1}{56}}\right\} = 0{,}95$$

или

$$\mathbf{P}\left\{X_{n+1} > -0{,}152972828\right\} = 0{,}95\,.$$

Итак, с надежностью 0,95 доходность от приобретения пакета акций ВТБ окажется больше, чем −0,152972828. Это означает, что если сегодня приобрести пакет акций ВТБ за 1 млн руб., то с 95%-ной надежностью можно гарантировать, что убыток от владения таким пакетом в течение недели не превысит $0{,}152972828 \cdot 1\,000\,000 = 152\,972{,}83$ руб., т. е. оценка 95%-ной суммы под риском на ближайший месяц для пакета акций ВТБ, приобретаемого в данный момент за 1 млн руб., составляет $\widehat{\mathbf{Var}}_{0,95}^{(\text{ВТБ})} = 152\,972$ руб. 83 коп.

Аналогичным образом рассчитаем оценки $\mathbf{VaR}_{0,95}$ на ближайший месяц для приобретаемого сегодня за 1 млн руб. пакета акций Газпрома и для приобретаемого в настоящий момент за 1 млн руб. пакета акций Сбербанка:

$$\widehat{\mathbf{Var}}_{0,95}^{(\text{Газпром})} = 115\,599 \text{ руб. } 24 \text{ коп.}; \qquad \widehat{\mathbf{Var}}_{0,95}^{(\text{Сбербанк})} = 143\,135 \text{ руб. } 73 \text{ коп.}$$

Видим, что наименьшему риску подвержены инвестиции в акции Газпрома, а наибольшему риску — инвестиции в акции ВТБ. ❑

Задачи

372. В условиях задачи 351 определите верхние границы для генеральной дисперсии и для математического ожидания суммарного пополнения вклада (с надежностью 0,9).

373. По каждой акции из задачи 5 рассчитайте оценку 95%-ной суммы под риском на ближайшую неделю для пакета акций, приобретаемого сегодня за 100 000 руб.

Глава 6. Проверка статистических гипотез

§ 6.1. Статистические гипотезы

6.1.1. Проверка гипотез и сопутствующие риски

Теория проверки гипотез создавалась и развивалась в работах К. Пирсона, Р. Фишера, Е. Неймана, А.Н. Колмогорова, Н.В. Смирнова, А. Вальда и других исследователей.

Статистической гипотезой называется любое высказывание относительно генеральной совокупности, проверяемое по выборочным данным. Гипотеза называется *параметрической*, если в ней содержится утверждение о значении какого-либо параметра генеральной совокупности, и *непараметрической* в противном случае.

Примером параметрической гипотезы является гипотеза о равенстве математического ожидания доходности акции конкретному значению 0,14 в предположении, что доходность имеет нормальный закон распределения. Еще один пример параметрической гипотезы — предположение о том, что вероятность наступления риска угона для определенной категории страхователей КАСКО равна 0,03.

В качестве примера непараметрической гипотезы можно привести предположение о том, что логарифм объема торгов акциями имеет нормальный закон распределения.

Параметрическую гипотезу называют *простой*, если в ней значение параметра приравнивается конкретному числу, или *сложной*, если значение параметра выбирается из какого-либо интервала.

Проверяемую гипотезу называют *нулевой* и обозначают H_0. Гипотезу H_1, логически противоречащую гипотезе H_0, называют *альтернативной*. Правило, по которому принимают решение «принять или отклонить гипотезу H_0», называют *критерием*.

Статистику

$$\psi = \psi(X_1, X_2, \ldots, X_n),$$

измеряющую согласие выборки $(X_1, X_2, \ldots, X_n)$ с гипотезой H_0, называют *статистикой критерия*.

Из множества S значений статистики ψ выделяется *критическая область* — такое подмножество R, что при условии истинности гипотезы H_0 вероятность $\mathbf{P}\{\psi \in R \,|\, H_0\} = \alpha$, где *уровень значимости* α — достаточно малое число (обычно α берут равным 0,05; 0,01; 0,005 или 0,001). Это подмножество R называется *критической областью*.

Общая схема проверки гипотез выглядит следующим образом.

1. Формулируется в виде статистической гипотезы задача исследования (т. е. определяется нулевая гипотеза H_0 и альтернативная гипотеза H_1).

2. Выбирается статистическая характеристика гипотезы — статистика критерия ψ.

3. Анализируются возможные ошибочные решения и их последствия.

4. Задается уровень значимости α и рассчитываются границы *критической области* R, в которую статистика ψ может попасть только с малой вероятностью α.

5. Если фактическое значение статистики ψ (вычисленное по конкретной выборке $x_1, x_2, \ldots, x_n$) попадает в критическую область, то нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1; в противном случае нулевая гипотеза H_0 не отвергается. Множество $S \setminus R$ называется *областью принятия гипотезы* H_0.

Если гипотеза H_0 отвергается, это значит, что выборочные данные противоречат ей, т. е. она неверна, если же гипотеза H_0 принимается, то это вовсе не означает, что H_0 является единственно верной гипотезой: просто гипотеза H_0 не противоречит результатам испытаний; однако таким же свойством, наряду с H_0, могут обладать и другие гипотезы. Таким образом, приняв или отклонив гипотезу H_0, можно оказаться правым, а можно совершить ошибку первого или второго рода (табл. 6.1.1).

Таблица 6.1.1

Риски при проверке гипотез

Принятая гипотеза / Истинная гипотеза	H_0	H_1
H_0	правильное решение, $\mathbf{P}\{H_0 \mid H_0\} = \mathbf{P}\{\psi \in (S \setminus R) \mid H_0\} = 1 - \alpha$ — *надежность*	ошибка первого рода, $\mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{\psi \in R \mid H_0\} = \alpha$ — *уровень значимости*
H_1	ошибка второго рода, $\mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{\psi \in (S \setminus R) \mid H_1\} = \beta$	правильное решение, $\mathbf{P}\{H_1 \mid H_1\} = \mathbf{P}\{\psi \in R \mid H_1\} = 1 - \beta$ — *мощность*

К одной и той же нулевой гипотезе H_0 можно предложить несколько различных альтернативных гипотез H_1, например, нулевой гипотезе H_0: $\mathbf{E}(X) = a_0$ о равенстве математического ожидания случайной величины конкретному значению a_0 могут соответствовать такие альтернативные гипотезы:

$$H_1 : \mathbf{E}(X) > a_0; \qquad H_1' : \mathbf{E}(X) < a_0; \qquad H_1'' : \mathbf{E}(X) \neq a_0.$$

Поэтому в каждом конкретном случае важно правильно выбирать альтернативную гипотезу, удобную для принятия содержательного решения.

На практике, как правило, в качестве нулевой гипотезы выбирают такую, которая соответствует «нормальному положению дел» (например, человек здоров; обвиняемый невиновен; деталь изготовлена с соблюдением стандарта; потенциальный заемщик способен погасить кредит; транзакция не является мошеннической и т. д). Соответственно, альтернативная гипотеза соответствует противоположной ситуации, которая обычно рассматриваетсяч как менее вероятная, но при этом требующая обязательной реакции (человек болен; обвиняемый виновен; деталь изготовлена с нарушением стандарта; потенциальный заемщик не способен погасить кредит; транзакция является мошеннической и т. д). Вследствие этого ошибку первого рода часто называют *ложной тревогой*.

Пример 6.1.1. В задачах кредитного скоринга в качестве нулевой гипотезы H_0 выбирается предположение о том, что заемщик выполнит условия кредитного договора (и ему можно выдать кредит), а в качестве альтернативной гипотезы H_1 — предположение, что заемщик не погасит кредит (поэтому выдавать ему кредит не следует).

Предоставление «хорошему» заемщику кредита в размере B_0 по ставке i приносит банку процентный доход $B_0 i$. Предоставление «плохому» заемщику кредита такого же размера приносит убыток в размере, близком к B_0 (что при нормальных процентных ставках, меньших 100%, намного больше $B_0 i$). Отказ в кредите приносит банку нулевой доход независимо от типа заемщика.

Задачи кредитного скоринга имеют две особенности. Во-первых, в выборках, которыми располагают банки, «хороших» клиентов всегда больше чем «плохих», а во-вторых, ошибка второго рода — классификация «плохого» заемщика как «хорошего» (т. е. предоставление кредита клиенту, который его не погасит) обходится дороже, чем ошибка первого рода — классификация «хорошего» заемщика как «плохого» (т. е. отказ «хорошему» заемщику в выдаче кредита). ❑

При проверке гипотез значение уровня значимости α задается лицом, принимающим решение, а вероятность ошибки второго рода β определяется уровнем значимости: меньшие значения уровня значимости α дают бóльшую надежность $1-\alpha$ — бóльшую уверенность в отсутствии ошибки в тех случаях, когда будет принята альтернативная гипотеза. Но это связано с увеличением риска не отвергнуть ложную нулевую гипотезу — риска совершения ошибки второго рода, что приводит к снижению мощности критерия $1-\beta$ (т. е. к увеличению вероятности ошибки второго рода β.

Можно, впрочем, уменьшить вероятность ошибки второго рода без увеличения вероятности ошибки первого рода: за счет увеличения объема выборки. Но, как правило, увеличивать объем выборки возможно лишь до определенного предела. Таким образом, выбор уровня значимости требует компромисса между надежностью, мощностью и объемом выборки.

Важно также отметить, что о мощности критерия можно говорить только по отношению к конкретной простой альтернативной гипотезе.

Пример 6.1.2. Товарищество собственников жилья, объединяющее собственников более 10 тыс. квартир в нескольких соседних домах некоторого микрорайона, рассматривает возможность заключения договора с новым охранным предприятием, предоставляющим услуги по более высоким ценам, чем охранное предприятие, с которым договор действует в настоящее время.

Были случайным образом выбраны $n = 500$ членов товарищества, которым был задан вопрос, согласны ли они на небольшое увеличение стоимости услуг по охране. Если более чем 275 из них согласятся на увеличение стоимости, то правление товарищества будет считать, что по крайней мере 60% всех членов товарищества тоже будут за увеличение.

Определим вероятность ошибки первого рода, если в действительности 60% членов товарищества будут за увеличение взноса;

Пусть p — вероятность того, что случайно выбранный член товарищества согласится на увеличение взноса.

Гипотезу $H_0 : p = 0{,}6$, соответствующую тому, что не менее 60% членов товарищества согласны с увеличением стоимости охраны, принимают, если $M > 275$, где M — число членов товарищества, согласных на увеличение взноса. Альтернативную гипотезу $H_1 : p < 0{,}6$, означающую, что менее 60% членов товарищества согласны с увеличением стоимости охраны, принимают, если $M \leqslant 275$. Вероятность ошибки первого рода

$$\alpha = \mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{M \leqslant 275 \mid p = 0{,}6\} = \mathbf{P}\{0 \leqslant M \leqslant 275 \mid p = 0{,}6\} =$$
$$= \text{БИНОМ.РАСП.ДИАП(500; 0,6; 0; 275)} = 0{,}013006432.$$

(Функция БИНОМ.РАСП.ДИАП(<n>; <p>; <k_1>; <k_2>) вычисляет вероятность попадания случайной величины $\mathrm{Bin}(n; p)$ в диапазон $[k_1; k_2]$.)

Итак, вероятность ошибки первого рода [т. е. вероятность принятия решения о том, что с увеличением стоимости охранных услуг согласны менее 60% всех членов товарищества, (вероятность принять гипотезу H_1) при условии, что не более 275 из 500 опрошенных членов товарищества высказались за увеличение стоимости, (при условии, что на самом деле верна гипотеза H_0)] достаточно мала — она составляет $\alpha = 1{,}3\%$.

Поэтому если по результатам опроса 500 членов товарищества окажется, что с повышением стоимости охранных услуг согласны не более 275 из 500 опрошенных, и гипотеза H_0 будет отвергнута, то очень маловероятно, что такое решение будет неверным.

Если правление товарищества совершит ошибку первого рода, то вопрос о смене охранного предприятия будет снят, но на самом деле вполне возможно, что с увеличением стоимости охраны 60% согласны не менее 60% членов товарищества. Если же правление совершит ошибку второго рода, то правление будет продолжать проработку вопроса о смене охранного предприятия, хотя на самом деле с повышением стоимости охранных услуг согласны менее 60% членов товарищества.

Теперь найдем вероятность ошибки второго рода при альтернативной гипотезе H_1: $p = 0{,}54$ о том, что в действительности лишь 54% всех членов товарищества будут согласны с запланированным увеличением взноса. Если альтернативное значение вероятности равно $p = 0{,}54$, то вероятность ошибки второго рода

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M > 275 \mid p = 0{,}54\} = \mathbf{P}\{276 \leqslant M \leqslant 500 \mid p = 0{,}54\} =$$
$$= \text{БИНОМ.РАСП.ДИАП(500; 0,54; 276; 500)} = 0{,}311166061.$$

Итак, вероятность ошибки второго рода, т. е. вероятность того, что более 275 из 500 опрошенных членов товарищества выскажутся за увеличение стоимости, (вероятность H_0) при условии, что на самом деле только 54% всех членов товарищества согласны с увеличением стоимости охранных услуг, (при условии H_1) относительно велика: она составляет $\beta = 31{,}1\%$.

Таким образом, вероятность ошибиться в случае принятия гипотезы H_0 достаточно высока!

Если же в качестве альтернативной рассматривается гипотеза H_1: $p = 0{,}5$ о том, что в действительность только 50% членов товарищества согласны с увеличением взноса, то вероятность ошибки второго рода будет гораздо меньше:

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M > 275 \mid p = 0{,}5\} = \mathbf{P}\{276 \leqslant M \leqslant 500 \mid p = 0{,}5\} =$$
$$= \text{БИНОМ.РАСП.ДИАП(500; 0,5; 276; 500)} = 0{,}011233102. \ \square$$

Во многих случаях масштаб негативных последствия от ошибок первого и второго рода различается:

- положить в больницу здорового человека менее опасно, чем не лечить больного;
- для общества значительно опаснее ошибка, связанная с осуждением невиновного, чем с оправданием виновного;
- отбраковка качественной детали ведет к финансовым потерям, тогда как установка некачественной детали (например, в автомобиль или самолет) может привести к катастрофе с человеческими жертвами;
- отказ в выдаче кредита надежному заемщику приводит к недополучению процентного дохода в размере некоторой части суммы кредита, а выдача кредита ненадежному заемщику — к убыткам в размере, сравнимом с полной суммой кредита;
- блокировка транзакции, которая на самом деле не является моженнической, приводит к определенным неудобствам клиента, тогда как подтверждение мошеннической транзакции — к финансовым убыткам.

Поэтому при проверке гипотез бывает полезно оценивать ожидаемые доходы и убытки.

Пример 6.1.3. Если крупная партия произведенных за день деталей содержит 3% брака, то ее относят к высшему сорту и продают за 200 000 руб. Если же в партии 10% брака, ее относят к первому сорту и продают за 150 000 руб.

Если партия была отнесена к высшему сорту ошибочно и продана за 200 000 руб., это влечет за собой возврат партии, ее переоценку и выплату неустойки, что приводит к суммарным потерям в размере 100 000 руб.

Для контроля качества отбирается 20 случайно выбранных деталей, и если среди них окажется не более одной дефектной, то партию относят к высшему сорту, в противном случае — к первому сорту.

Если очередная партия была отнесена к первому сорту, то производство останавливается на день и производится переналадка оборудования. Стоимость переналадки составляет 20 000 руб. (с учетом того недополученных доходов от продажи деталей, которые могли быть выпущены, если бы производство не останавливалось).

В данном случае нулевой гипотезой является гипотеза H_0: $p = 0{,}03$, а альтернативной гипотезой — гипотеза H_1: $p = 0{,}10$.

В качестве статистики критерия выбрано число M бракованных деталей среди 20 отобранных для проверки. Эта статистика может принимать значения из множества $S = \{0; 1; 2; ...; 20\}$. Критической областью в данном примере является множество $R = \{2; 3; 4; ...; 20\}$.

Ошибка первого рода происходит в том случае, когда ошибочно отвергается нулевая гипотеза, т. е. когда партия изделий, которая на самом деле содержит 3% брака, признается не соответствующей высшему сорту.

Вероятность ошибки первого рода

$$\alpha = \mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{M \in \{2; 3; 4; \ldots; 20\} \mid p = 0{,}03\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(20;\ 0{,}03;\ 2;\ 20) = 0{,}119838022.$$

Ошибка второго рода происходит, если нулевая гипотеза ошибочно принимается, т. е. если партия изделий, которая на самом деле содержит 10% брака, признается соответствующей высшему сорту.

Вероятность ошибки второго рода

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M \in \{0;1\} \mid p = 0{,}1\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(20;\ 0{,}1;\ 0;\ 1) = 0{,}391746998.$$

Если известно, что за прошлый год 90% партий было отнесено к высшему сорту, и 10% — к первому, то можно оценить вероятность выпуска партии высшего сорта как 0,9, вероятность выпуска партии первого сорта и переналадки станка как 0,1 и вычислить ожидаемый доход за день

$$\hat{\mathbf{E}}(\Pi) = \big((1-\alpha)\cdot 200\,000 + \alpha\cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(\beta\cdot(200\,000-100\,000)+(1-\beta)\cdot(150\,000-20\,000)\big)\cdot 0{,}1 =$$
$$= \big((1-0{,}119838022)\cdot 200\,000 + 0{,}119838022\cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(0{,}391746998\cdot 100\,000 + (1-0{,}391746998)\cdot 130\,000\big)\cdot 0{,}1 = 186\,432{,}05 \text{ руб.}$$

Предположим, что на завод пришел новый менеджер, который рассматривает варианты изменения правила принятия решения для того, чтобы увеличить доход.

Один из вариантов, предлагаемых новым менеджером, состоит в том, чтобы относить партию к высшему сорту, если из 20 случайно отобранных для контроля качества деталей окажется не более двух дефектных, и к первому сорту в противном случае.

При таком правиле принятия решения вероятность ошибки первого рода

$$\alpha = \mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{M \in \{3;4;5;\ldots;20\} \mid p = 0{,}03\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(20;\ 0{,}03;\ 3;\ 20) = 0{,}021008356,$$

становится меньше, но вероятность ошибки второго рода

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M \in \{0;1;2\} \mid p = 0{,}1\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(20;\ 0{,}1;\ 0;\ 2) = 0{,}676926805.$$

увеличивается.

При этом ожидаемый доход за день становится больше:

$$\hat{\mathbf{E}}(\Pi) = \big((1-\alpha)\cdot 200\,000 + \alpha\cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(\beta\cdot(200\,000-100\,000)+(1-\beta)\cdot(150\,000-20\,000)\big)\cdot 0{,}1 =$$
$$= \big((1-0{,}021008356)\cdot 200\,000 + 0{,}021008356\cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(0{,}676926805\cdot 100\,000 + (1-0{,}676926805)\cdot 130\,000\big)\cdot 0{,}1 = 190\,023{,}84 \text{ руб.}$$

Еще один вариант, который предлагает менеджер, заключается в увеличении размера контрольной партии до 100 деталей и отненении партии к высшему сорту в том и только в том случае, если из 100 случайно отобранных для проверки деталей окажется не более шести дефектных.

В этом случае вероятность ошибки первого рода

$$\alpha = \mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{M \in \{7;8;9;\ldots;100\} \mid p = 0{,}03\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(100;\ 0{,}03;\ 7;\ 100) = 0{,}031227509$$

оказывается чуть большей, чем при втором правиле принятия решения, но существенно меньшей, чем при первом правиле; вероятность ошибки второго рода

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M \in \{0;1;2;3;4;5;6\} \mid p = 0{,}1\} =$$
$$= \text{БИНОМ.РАСП.ДИАП}(100;\ 0{,}1;\ 0;\ 6) = 0{,}117155615$$

становится существенно меньшей, чем при первом и втором правилах.

При этом ожидаемый доход за день оказывается наибольшим из трех вариантов:

$$\hat{\mathbf{E}}(\Pi) = \big((1-\alpha)\cdot 200\,000 + \alpha \cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(\beta\cdot(200\,000 - 100\,000) + (1-\beta)\cdot(150\,000 - 20\,000)\big)\cdot 0{,}1 =$$
$$= \big((1-0{,}021008356)\cdot 200\,000 + 0{,}021008356\cdot 150\,000\big)\cdot 0{,}9 +$$
$$+\big(0{,}676926805\cdot 100\,000 + (1-0{,}676926805)\cdot 130\,000\big)\cdot 0{,}1 = 191\,243{,}30 \text{ руб.}$$

Таким образом, последний из вариантов, предлагаемых менеджером, представляется наилучшим: он позволяет увеличить средний доход при сокращении и ошибки первого рода, и ошибки второго рода. ❑

Следующий пример иллюстрирует, как меняется вероятность ошибки второго рода (и, соответственно, мощность критерия) при изменении гипотезы H_1.

Пример 6.1.4. Предположим, что у нас имеется монета с вероятностью выпадения герба, равной p, и мы хотим проверить гипотезу H_0: $p = 0{,}5$ при альтернативной гипотезе H_1: $p \neq 0{,}5$.

Подбросим нашу монету 10 раз, и если герб выпадет 4, 5 или 6 раз из 10, то примем гипотезу H_0, в противном случае примем гипотезу H_1.

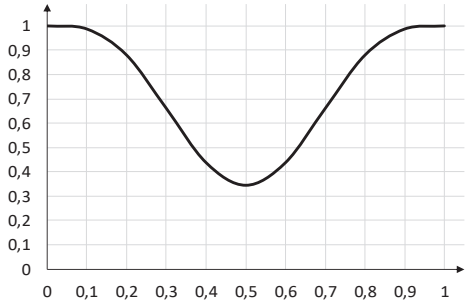
Найдем значение вероятности ошибки первого рода

$$\alpha = \mathbf{P}\{H_1 \mid H_0\} = \mathbf{P}\{M \notin \{4;5;6\} \mid p = 0{,}5\} = \text{ 1 – БИНОМ.РАСП.ДИАП}(10;\ 0{,}5;\ 4;\ 6).$$

Введем в ячейку B1 листа Microsoft Excel (рис. 6.1.1, *а*) формулу для вероятности ошибки первого рода («= 1 – БИНОМ.РАСП.ДИАП(10; 0,5; 4; 6)»), получим значение $\alpha = 0{,}34375$.

	A	B	C	D	E	F	G	H	I	J	K	L
1	alpha	0,34375										
2		= 1 – БИНОМ.РАСП.ДИАП(10; 0,5; 4; 6)										
3	p'	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
4	beta	0	0,0127861	0,1200095	0,3397972	0,5629575	0,65625	0,5629575	0,3397972	0,1200095	0,0127861	3,191E-62
5		= БИНОМ.РАСП.ДИАП(10; B3; 4; 6)										
6	w = 1 – beta	1	0,9872139	0,8799905	0,6602028	0,4370425	0,34375	0,4370424	0,6602028	0,8799905	0,9872139	1
7		= 1 – B4										

а)



б)

Рис. 6.1.1. *Функция мощности критерия в примере 6.1.4: вычисления* (*а*); *график* (*б*)

Значения вероятности ошибки второго рода при альтернативной вероятности выпадения герба, равной p', вычисляются по формуле

$$\beta = \mathbf{P}\{H_0 \mid H_1\} = \mathbf{P}\{M \in \{4;5;6\} \mid p = p'\} = \text{БИНОМ.РАСП.ДИАП(10; <}p'\text{>; 4; 6)}.$$

Заполним диапазон B3:L3 значениями $p' = 0;0{,}1;0{,}2;\ldots;1$, в ячейку B4 введем формулу «= БИНОМ.РАСП.ДИАП(10; B3; 4; 6)» для вычисления вероятности ошибки второго рода при альтернативной вероятности выпадения герба $p' = 0$ (содержащейся в ячейке B3), а в ячейку B6 введем формулу «= 1 – B4» для вычисления мощности $w = 1 - \beta$.

Чтобы получить значения вероятности ошибки первого рода β и мощности $w = 1 - \beta$ для всех $p' = 0;0{,}1;0{,}2;\ldots;1$, осталось протянуть формулу из ячейки B4 на диапазон B4:L4, а формулу из ячейки B6 на диапазон B6:L6 (рис. 6.1.1, *а*).

График функции мощности, т. е. зависимости мощности критерия $1-\beta$ от альтернативной вероятности выпадения герба p', построен на рис. 6.1.1, *б*. ❑

Рассмотрим еще один

Пример 6.1.5. Пусть у нас есть выборка объемом $n = 10$ элементов из генеральной совокупности, имеющей распределение Пуассона с параметром λ. Будем проверять гипотезу $H_0 : \lambda = 1$ при альтернативной гипотезе $H_1 : \lambda = 2$ с помощью критерия, который предписывает принимать H_0 в случае, если максимальный элемент выборки $x_{(\max)} \leqslant 1$, и принимать H_1 в противном случае.

Найдем мощность этого критерия:

$$w = 1-\beta = 1-\mathbf{P}\{H_0 \mid H_1\} = 1-\mathbf{P}\{X_{(\max)} \leqslant 1 \mid \lambda = 2\} = 1-\mathbf{P}\{(X_1 \leqslant 1)\cap(X_2 \leqslant 1)\cap\cdots\cap(X_n \leqslant 1) \mid \lambda = 2\} =$$

$$= 1 - \text{ПУАССОН.РАСП(1; 2; ИСТИНА)}\hat{}10 = 0{,}999878291. \text{ ❑}$$

Обсудим теперь **задачу построения наиболее мощного критерия**, которая решается в случае простых гипотез с помощью леммы Неймана — Пирсона.

Критерий называется *наиболее мощным*, если из всех возможных критериев с заданным уровнем значимости α он обладает наибольшей мощностью, т. е. если его критическая область R^* является такой, что

$$1-\beta = \mathbf{P}\{\psi \in R^* \mid H_1\} = \max_{R:\, \mathbf{P}\{\psi \in R \mid H_0\} = \alpha} \mathbf{P}\{\psi \in R \mid H_1\}.$$

Использование наиболее мощного критерия гарантирует при заданной вероятности α ошибки первого рода наименьшую, по сравнению с другими критериями, вероятность ошибки β второго рода, поскольку мощность критерия равна $(1-\beta)$.

Лемма Неймана — Пирсона. Существует такая константа C, зависящая только от α, что критическая область R^* наиболее мощного критерия

$$R^* = \tag{6.1.1}$$

$$= \left\{(x_1, x_2, \ldots, x_n) : L_0(x_1, x_2, \ldots, x_n) = 0 \cup \left(\frac{L_1(x_1, x_2, \ldots, x_n)}{L_0(x_1, x_2, \ldots, x_n)} > C \quad \text{при} \quad L_0(x_1, x_2, \ldots, x_n) \neq 0\right)\right\},$$

при этом константа C является решением уравнения

$$\mathbf{P}\left\{\frac{L_1(x_1, x_2, \ldots, x_n)}{L_0(x_1, x_2, \ldots, x_n)} > C \,\middle|\, H_0\right\} = \alpha. \tag{6.1.2}$$

Здесь $L_i(x_1, x_2, \ldots, x_n)$ — функция правдоподобия (см. **п. 5.2.2**) в предположении справедливости гипотезы H_i $(i = 0,1)$.

Пример 6.1.6. Рассмотрим случайную величину $X = \mathcal{N}(a;\sigma)$, причем будем считать, что числовое значение параметра σ известно, а числовое значение параметра a неизвестно. Построим наиболее мощный критерий для проверки гипотезы $H_0 : a = a_0$ при альтернативной гипотезе $H_1 : a = a_1 > a_0$.

Если верна гипотеза $H_0 : a = a_0$, т. е. $X = \mathcal{N}(a_0;\sigma)$, то функция правдоподобия в точке $(x_1, x_2, \ldots, x_n)$ равна

$$L_0(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - a_0)^2},$$

Если же верна гипотеза $H_1 : a = a_1 > a_0$, то функция правдоподобия

$$L_1(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - a_1)^2}.$$

При этом отношение правдоподобия

$$\frac{L_1(x_1, x_2, \ldots, x_n)}{L_0(x_1, x_2, \ldots, x_n)} = \mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}[(x_i - a_1)^2 - (x_i - a_0)^2]} = \mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}[x_i^2 - 2x_i a_1 + a_1^2 - x_i^2 + 2x_i a_0 - a_0^2]} = \mathrm{e}^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}[-2x_i(a_1 - a_0) + a_1^2 - a_0^2]} =$$

$$= \mathrm{e}^{\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}[2x_i(a_1 - a_0) - (a_1 + a_0)(a_1 - a_0)]} = \mathrm{e}^{\frac{2(a_1 - a_0)}{2\sigma^2}\sum\limits_{i=1}^{n}\left(x_i - \frac{a_0 + a_1}{2}\right)} = \mathrm{e}^{\frac{a_1 - a_0}{\sigma^2}\left(\sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n}\left(\frac{a_0 + a_1}{2}\right)\right)} = \mathrm{e}^{\frac{a_1 - a_0}{\sigma^2}\left(n\overline{x} - \frac{n(a_0 + a_1)}{2}\right)} = \mathrm{e}^{\frac{n(a_1 - a_0)}{\sigma^2}\left(\overline{x} - \frac{a_0 + a_1}{2}\right)}.$$

Поскольку при $a_1 > a_0$ это отношение является монотонно возрастающей функцией от $\overline{x}$, и так как $L_0(x_1, x_2, \ldots, x_n) \neq 0$, то неравенство

$$\frac{L_1(x_1, x_2, \ldots, x_n)}{L_0(x_1, x_2, \ldots, x_n)} > C$$

равносильно неравенству $\overline{x} > \widetilde{C}$, где C и $\widetilde{C}$ — некоторые константы. Поэтому соотношения (6.1.1) и (6.1.2) примут вид $R^* = \left\{(x_1, x_2, \ldots, x_n) : \overline{x} > \widetilde{C}\right\}$ и $\mathbf{P}\left\{\overline{X} > \widetilde{C} \,|\, H_0\right\} = \alpha$.

Известно, что если $X = N(a; \sigma)$, то

$$\frac{(\overline{X} - a)\sqrt{n}}{\sigma} = N(0; 1),$$

поэтому при выполнении гипотезы $H_0 : a = a_0$

$$\frac{(\overline{X} - a_0)\sqrt{n}}{\sigma} = N(0; 1),$$

и тогда

$$\alpha = \mathbf{P}\left\{\overline{X} > \widetilde{C} \,|\, H_0\right\} = 1 - \mathbf{P}\left\{\overline{X} \leqslant \widetilde{C} \,|\, X = N(a_0; \sigma)\right\} =$$

$$1 - \mathbf{P}\left\{\frac{(\overline{X} - a_0)\sqrt{n}}{\sigma} \leqslant \frac{(\widetilde{C} - a_0)\sqrt{n}}{\sigma} \,\middle|\, X = N(a_0; \sigma)\right\} = 1 - \mathbf{P}\left\{Z \leqslant \frac{(\widetilde{C} - a_0)\sqrt{n}}{\sigma} \,\middle|\, Z = N(0; 1)\right\},$$

откуда

$$\frac{(\widetilde{C} - a_0)\sqrt{n}}{\sigma} = z_{1-\alpha} \quad \text{или} \quad \widetilde{C} = a_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}},$$

где $z_{1-\alpha}$ = НОРМ.СТ.ОБР(<1 − α>) — квантиль уровня 1 − α случайной величины Z, распределенной по стандартному нормальному закону.

Таким образом, можно сформулировать наиболее мощный критерий проверки гипотезы $H_0 : a = a_0$ при альтернативе $H_1 : a = a_1 > a_0$:

$$\begin{cases} \text{если } \overline{x} \leqslant a_0 + z_{1-\alpha}\sigma/\sqrt{n}, \text{ то } H_0 \text{ принимают } (H_1 \text{ отклоняют}); \\ \text{если } \overline{x} > a_0 + z_{1-\alpha}\sigma/\sqrt{n}, \text{ то } H_0 \text{ отклоняют } (H_1 \text{ принимают}). \end{cases}$$

При использовании этого критерия вероятность ошибки первого рода равна α, а вероятность ошибки второго рода равна

$$\beta = \mathbf{P}\left\{H_0 | H_1\right\} = \mathbf{P}\left\{Z \in \left(S \setminus R^*\right) \,\middle|\, H_1\right\} = \mathbf{P}\left\{\overline{X} < \widetilde{C} \,|\, X = N(a_1; \sigma)\right\} = \mathbf{P}\left\{\overline{X} \leqslant \widetilde{C} \,|\, X = N(a_1; \sigma)\right\} =$$

$$= \mathbf{P}\left\{\frac{(\overline{X} - a_1)\sqrt{n}}{\sigma} \leqslant \frac{(\widetilde{C} - a_1)\sqrt{n}}{\sigma} \,\middle|\, X = N(a_1; \sigma)\right\} = \mathbf{P}\left\{Z \leqslant \frac{(\widetilde{C} - a_1)\sqrt{n}}{\sigma} \,\middle|\, Z = N(0; 1)\right\} =$$

$$= \mathbf{P}\left\{Z \leqslant \frac{(a_0 + z_{1-\alpha}\sigma/\sqrt{n} - a_1)\sqrt{n}}{\sigma} \,\middle|\, Z = N(0; 1)\right\} = \mathbf{P}\left\{Z \leqslant z_{1-\alpha} + \frac{(a_0 - a_1)\sqrt{n}}{\sigma} \,\middle|\, Z = N(0; 1)\right\}.$$

Из последнего выражения видно, что с ростом уровня значимости α (т. е. вероятности ошибки первого рода) вероятность β ошибки второго рода уменьшается (при неизменном объеме выборки n); с ростом объема n выборки вероятность β ошибки второго рода уменьшается (при неизменном уровне значимости α).

Нетрудно также получить формулу для объема n выборки, обеспечивающего при заданном уровне значимости α вероятность ошибки второго рода, меньшую β:

$$n \geqslant \frac{\sigma^2\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{(a_1-a_0)^2}. \quad \square \tag{6.1.3}$$

Задачи

374. Инвестиционная компания следит за тем, чтобы не инвестировать в слишком рискованные активы. Поэтому перед принятием решений об инвестициях обязательно проверяется гипотеза о том, что риск по активу является приемлемым, при альтернативной гипотезе, что риск по активу превышает допустимый предел. Сравните последствия от ошибок первого и второго рода.

375. Решение о досмотре багажа пассажира принимается на основании сигнала металодетектора. Сформулируйте гипотезы, которые целесообразно выбрать в качестве нулевой и альтернативной, сравните между собой последствия от ошибок первого и второго рода.

376. Станок производит детали, при этом 3%-ная доля брака среди всех произведенных деталей считается допустимой, 5%-ная доля брака — недопустимой. Для контроля качества каждый час отбирается 30 деталей, и если хотя бы две из них не удовлетворяют техническим условиям, станок останавливают и производят его наладку. Из опыта известно, что станок требует переналадки с вероятностью 0,05. При этом проверка каждой детали стоит 10 руб., пропущенная бракованная партия приводит к убыткам в размере 10 000 руб., а «ложная тревога» обходится в 2000 руб. Сформулируйте гипотезы, которые целесообразно выбрать в качестве нулевой и альтернативной, найдите вероятности ошибок первого и второго рода, а также размер ожидаемых потерь, связанных с ошибками первого и второго рода. Выясните, не лучше ли изменить правило принятия решения и останавливать станок для переналадки в случае, когда среди 30 отобранных деталей есть хотя бы одна бракованная.

377. Имеется игральный кубик с вероятностью выпадения единицы, равной p, и мы хотим проверить гипотезу H_0: $p = 1/6$ при альтернативной гипотезе H_1: $p \neq 1/6$. Монета подбрасывается 10 раз, и если единица выпадет 1 или 2 раза из 10, то принимается гипотеза H_0, в противном случае принимается гипотеза H_1. Найдите значение вероятности ошибки первого рода; значения вероятности ошибки второго рода при альтернативной вероятности выпадения единицы, равной $p' = 0; 0{,}1; 0{,}2; \ldots; 1$; значения мощности критерия при $p' = 0; 0{,}1; 0{,}2; \ldots; 1$. Постройте график функции мощности.

378. Имеется выборка объемом $n = 12$ элементов из генеральной совокупности, имеющей нормальный закон распределения с математическим ожиданием, равным a, и средним квадратическим отклонением, равным 1. Проверяется гипотеза H_0: $a = 0$ при альтернативной гипотезе H_1: $a = 1$. Критерий предписывает принимать гипотезу H_0 в случае, если максимальный элемент выборки $x_{(\max)} < 3{,}5$, и гипотезу H_1 в противном случае. Найдите вероятности ошибок первого и второго рода, надежность и мощность этого критерия.

379. Докажите формулу (6.1.3).

6.1.2. Проверка гипотез с помощью доверительных интервалов

В **§ 5.3** мы строили доверительные интервалы для того, чтобы с заданной надежностью оценивать неизвестные значения параметров.

С помощью доверительного интервала, накрывающего с надежностью γ значение некоторого параметра θ, можно на уровне значимости $\alpha = 1-\gamma$ проверить гипотезу о равенстве данного параметра конкретному значению θ_0.

Гипотеза $H_0 : \theta = \theta_0$ при альтернативной гипотезе $H_1 : \theta \neq \theta_0$ проверяется с помощью двустороннего доверительного интервала, определяемого формулой

$$\mathbf{P}\left\{\hat{\theta}_1 < \theta < \hat{\theta}_2\right\} = \gamma .$$

Вероятность же того, что θ_0 не накрывается доверительным интервалом $\left(\hat{\theta}_1; \hat{\theta}_2\right)$, равна $1-\gamma$, и поскольку при построении доверительных интервалов надежность γ выбирается близкой к единице, величина $\alpha = 1-\gamma$ близка к нулю и может быть использована в качестве уровня значимости.

Если гипотетическое значение параметра θ_0 накрывается доверительным интервалом $\left(\hat{\theta}_1; \hat{\theta}_2\right)$, то это не является чем-то необычным, поэтому не дает оснований отвергнуть гипотезу $H_0 : \theta = \theta_0$.

Если же гипотетическое значение параметра θ_0 не накрывается доверительным интервалом $\left(\hat{\theta}_1; \hat{\theta}_2\right)$, то это дает основания отвергнуть гипотезу $H_0 : \theta = \theta_0$ и принять гипотезу $H_1 : \theta \neq \theta_0$ в силу близкой к нулю вероятности события, заключающегося в том, что θ_0 не накрывается доверительным интервалом $\left(\hat{\theta}_1; \hat{\theta}_2\right)$. Это проиллюстрировано на рис. 6.1.2.

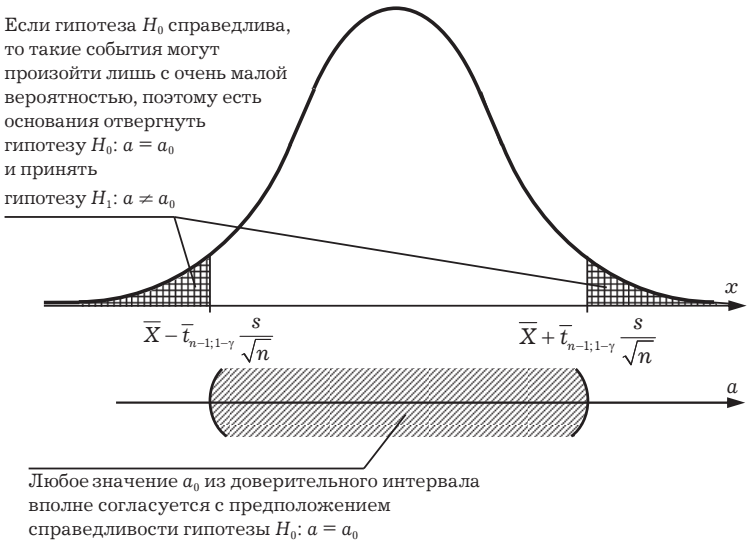


Рис. 6.1.2. *Использование доверительного интервала для проверки гипотезы о значении математического ожидания*

Гипотеза $H_0:\theta=\theta_0$ при альтернативных гипотезах $H_1':\theta<\theta_0$ и $H_1'':\theta>\theta_0$ проверяется аналогичным образом помощью односторонних доверительных интервалов, определяемых соответственно формулами $\mathbf{P}\{\hat{\theta}_1<\theta<+\infty\}=\gamma$ и $\mathbf{P}\{-\infty<\theta<\hat{\theta}_2\}=\gamma$: если значение θ_0 не накрывается интервалом $(\hat{\theta}_1;+\infty)$, то это дает основание отвергнуть гипотезу $H_0:\theta=\theta_0$ в пользу альтернативы $H_1':\theta<\theta_0$; если же значение θ_0 не накрывается интервалом $(-\infty;\hat{\theta}_2)$, то это дает основание отвергнуть гипотезу $H_0:\theta=\theta_0$ в пользу альтернативы $H_1'':\theta>\theta_0$.

Рассмотрим конкретный

Пример 6.1.7. Завод по производству мороженого фасует пломбир в брикеты по 250 г. Была отобрана контрольная партия из 100 брикетов, средняя масса которых оказалась равной $\bar{x}=246$ г, при этом стандартное отклонение оказалось равным $s=12$ г.

Будем проверять на уровне значимости $\alpha=0{,}05$ гипотезу $H_0:\mathbf{E}(X)=250$ г при альтернативной гипотезе $H_1:\mathbf{E}(X)\neq 250$ г.

Воспользуемся формулой (5.3.4) и построим соответствующий надежности $\gamma=1-\alpha=1-0{,}05=0{,}95$ двусторонний доверительный интервал

$$\left(\overline{X}-\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}};\overline{X}+\overline{t}_{n-1;1-\gamma}\frac{s}{\sqrt{n}}\right)$$

для генеральной средней массы брикета $\mathbf{E}(X)$:

$$\left(246-1{,}984217\cdot\frac{12}{\sqrt{100}};246+1{,}984217\cdot\frac{12}{\sqrt{100}}\right)$$

или

$$(243{,}6189397;\ 248{,}3810603).$$

(Здесь $\overline{t}_{100-1;1-0{,}95}=\overline{t}_{99;0{,}05}$ = СТЬЮДЕНТ.ОБР.2Х(0,05; 99) = 1,984217.)

Вероятность того, что построенный доверительный интервал накроет истинное значение математического ожидания a_0, равна $\gamma=0{,}95$, а вероятность того, что данный интервал не накроет a_0, равна $\alpha=0{,}05$ (рис. 6.1.2).

Таким образом, если предположить справедливость гипотезы $H_0:\mathbf{E}(X)=250$ г, то наблюдаемое нами событие, заключающееся в том, что построенный доверительный интервал не накрывает число $\mathbf{E}(X)$, имеет близкую к нулю вероятность $\alpha=0{,}05$.

Это дает основание отвергнуть гипотезу $H_0:\mathbf{E}(X)=250$ г в пользу гипотезы $H_1:\mathbf{E}(X)\neq 250$ г. ❑

Задачи

380. С помощью 95%-ного доверительного интервала для математического ожидания суммарного пополнения вклада за месяц, построенного в задаче **351**, проверьте на уровне значимости $\alpha=0{,}05$ гипотезу о том, что математическое ожидание суммарного пополнения вклада $\mathbf{E}(X)=50$ тыс. руб. при альтернативной гипотезе $\mathbf{E}(X)\neq 50$ тыс. руб.

381. С помощью 95%-ного и 99%-ного доверительных интервалов для вероятности невозврата кредита, построенных в задаче **357**, проверьте на 5%-ном и 1%-ном уровне значимости гипотезу о том, что вероятность невозврата кредита равна 0,170 (при альтернативе, что она не равна 0,170).

382. С помощью 90%-ного одностороннего доверительного интервала для математического ожидания суммарного пополнения вклада за месяц, построенного в задаче **372**, проверьте на уровне значимости $\alpha = 0{,}1$ гипотезу о том, что математическое ожидание суммарного пополнения вклада $\mathbf{E}(X) = 50$ тыс. руб. при альтернативной гипотезе $\mathbf{E}(X) > 50$ тыс. руб.

6.1.3. Наблюдаемый уровень значимости (*p*-значение)

Описанная в **п. 6.1.1** общая схема проверки гипотез предполагает, что уровень значимости α задается перед проверкой гипотезы, а затем проверяемая гипотеза отвергается в том и только в том случае, если фактическое значение статистики критерия попадает в критическую область, которая зависит от выбранного уровня значимости.

Если задать другое значение уровня значимости, то это приведет к изменению критической области, поэтому и результат проверки гипотезы может измениться.

Выбор уровня значимости довольно произволен, и для одной и той же выборке может оказаться так, что при одном значении уровня значимости проверяемая гипотеза отвергается, а при другом значении уровня значимости — не отвергается.

При проверке гипотез оказывается полезен подход, основанный на вычислении наблюдаемого уровня значимости.

Наблюдаемым уровнем значимости (или *p-значением*) нулевой гипотезы H_0, проверяемой по выборке $x_1, x_2, \ldots, x_n$ с помощью критерия $\psi(x_1, x_2, \ldots, x_n)$, имеющего критическую область R, называется наименьший уровень значимости, при котором нулевая гипотеза отклоняется:

$$p(x_1, x_2, \ldots, x_n) = \min\{\alpha : \psi(x_1, x_2, \ldots, x_n) \in R\}.$$

Еще одна интерпретация p-значения — это вероятность, с которой (при условии истинности нулевой гипотезы) могла бы реализоваться наблюдаемая выборка, или любая другая выборка с еще менее вероятным значением статистики ψ.

Случайная величина $p(x_1, x_2, \ldots, x_n)$ имеет равномерное распределение. Фактически, функция $p(x_1, x_2, \ldots, x_n)$ приводит значение статистики критерия ψ к шкале вероятности. Маловероятным значениям (хвостам распределения) статистики ψ соотвествуют p-значения, близкие к нулю или к единице.

Если известно p-значение $p(x_1, x_2, \ldots, x_n)$, то при всех значениях уровня значимости $\alpha > p(x_1, x_2, \ldots, x_n)$ основания отвергнуть гипотезу H_0 имеются, а при всех $\alpha \leqslant p(x_1, x_2, \ldots, x_n)$ основания отвергнуть гипотезу H_0 отсутствуют.

Поэтому на практике чем меньше p-значение, тем меньше вероятность ошибиться, отвергнув гипотезу H_0, и, соответственно, тем больше уверенность в том, что необходимо отвергнуть нулевую гипотезу.

Пример 6.1.8. В условиях примера 6.1.7 при уровне значимости $\alpha = 0{,}05$ гипотеза $H_0 : \mathbf{E}(X) = 250$ была отвергнута в пользу гипотезы $H_1 : \mathbf{E}(X) \neq 250$ г.

Будем задавать различные значения уровня значимости: $\alpha = 0{,}05$; 0,01; 0,002; 0,00125; 0,00120800418; 0,00120800417; 0,001. Вычисления в Microsoft Excel проиллюстрированы рис. 6.1.3.

	A	B	C	D	E	F	G	H
1	n	100						
2	x_cp	246						
3	s	12						
4	alpha	0,05	0,01	0,002	0,00125	0,00120800418	0,00120800417	0,001
5	t_n – 1; alpha	1,984217	2,626405	3,17460385	3,322736367	3,33333333231	3,33333333488	3,391528833
6		= СТЬЮДЕНТ.ОБР.2Х(B4; $B$1 – 1)						
7	Нижняя граница дов. интервала для E(X)	243,6189	242,8483	242,1904754	242,0127164	242,00000000123	241,99999999815	241,9301654
8		= $B$2 – B5*$B$3/КОРЕНЬ($B$1)						
9	Верхняя граница дов. интервала для E(X)	248,3811	249,1517	249,8095246	249,9872836	249,99999999877	250,00000000185	250,0698346
10		= $B$2 + B5*$B$3/КОРЕНЬ($B$1)						

Рис. 6.1.3. *Доверительные интервалы при различных значениях уровня значимости в примере 6.1.7*

Видим, что доверительные интервалы при $\alpha \geqslant 0{,}00120800418$ не накрывают число 250, поэтому при этих значениях уровня значимости есть основания отвергнуть гипотезу $H_0 : \mathbf{E}(X) = 250$ г и принять гипотезу $H_1 : \mathbf{E}(X) \neq 250$ г.

При уровне значимости, равном 0,00120800417 и 0,001, доверительные интервалы уже накрывают число 250, поэтому основания отвергнуть гипотезу $H_0 : \mathbf{E}(X) = 250$ отсутствуют.

Таким образом, при $\alpha < 0{,}00120800418$ нет оснований отвергнуть гипотезу $H_0 : \mathbf{E}(X) = 250$ г и принять гипотезу $H_1 : \mathbf{E}(X) \neq 250$ г.

Резюмируя, заключаем, что наблюдаемый уровень значимости в данном примере равен 0,00120800418. ❑

Замечание. Конечно, нельзя говорить, что p-значение равно вероятности того, что гипотеза H_0 верна: p-значение представляет собой вероятность того, что при истинности нулевой гипотезы могла бы реализоваться наблюдаемая выборка или любая другая выборка с еще менее вероятным значением статистики.

Задачи

383. Найдите фактический уровень значимости в условиях задачи **380**.
384. Найдите p-значение в условиях задачи **381**.
385. Найдите фактический уровень значимости в условиях задачи **382**.

§ 6.2. Проверка параметрических гипотез

6.2.1. Проверка гипотезы о числовом значении математического ожидания нормального закона распределения

В примере 6.1.6 мы построили наиболее мощный критерий для проверки гипотезы $H_0 : a = a_0$ о числовом значении математического ожидания при альтернативной гипотезе $H_1 : a = a_1 > a_0$ (для случая, когда рассматриваемая генеральная совокупность распределена по номальному закону с известным значением параметра σ и неизвестным значением параметра a).

Критическая область этого критерия задается соотношением

$$\overline{X} > a_0 + z_{1-\alpha}\sigma / \sqrt{n} \quad \text{или} \quad Z = \frac{(\overline{X} - a_0)\sqrt{n}}{\sigma} > z_{1-\alpha},$$

где $z_{1-\alpha}$ = НОРМ.СТ.ОБР(<1 – α>) — квантиль уровня $1-\alpha$ случайной величины Z, распределенной по стандартному нормальному закону.

При использовании этого критерия вероятность ошибки первого рода равна α, а вероятность ошибки второго рода равна

$$\beta = \mathbf{P}\left\{ Z \leqslant z_{1-\alpha} + \frac{(a_0 - a_1)\sqrt{n}}{\sigma} \,\middle|\, Z = \mathcal{N}(0;1) \right\}.$$

При этом объем выборки

$$n \geqslant \frac{\sigma^2 \left(z_{1-\alpha} + z_{1-\beta} \right)^2}{\left(a_1 - a_0 \right)^2}$$

гарантирует, что при заданном уровне значимости α вероятность ошибки второго рода будет меньше β.

В той же ситуации нормального закона распределения генеральной совокупности с известным σ и неизвестным a для проверки гипотезы $H_0 : a = a_0$ при альтернативе $H_1 : a = a_1 < a_0$ аналогичным образом можно построить критическую область наиболее мощного критерия:

$$Z = \frac{\left(\bar{X} - a_0 \right)\sqrt{n}}{\sigma} < -z_{1-\alpha}.$$

При проверке гипотезы $H_0 : a = a_0$ при альтернативе $H_1 : a = a_1 \neq a_0$ в качестве статистики критерия также рассматривается статистика

$$Z = \frac{\left(\bar{X} - a_0 \right)\sqrt{n}}{\sigma},$$

и критическая область определяется неравенством

$$|Z| = \left| \frac{\left(\bar{X} - a_0 \right)\sqrt{n}}{\sigma} \right| > z_{1-\alpha/2}.$$

Замечание. В случае альтернативы $H_1 : a = a_1 \neq a_0$ критерий уже не является наиболее мощным!

Пример 6.2.1. Руководство сети магазинов модной одежды должно принять решение, имеет ли смысл продлевать договор аренда помещения для магазина в одном из торговых центров.

Из опыта сети магазинов известно, что магазин прибылен, если среднее число посетителей в день превышает $a_0 = 400$ чел. Также известна дисперсия числа посетителей в день $\sigma^2 = 400$. При этом за последние $n = 365$ дней в данном магазине среднее число посетителей в день оказалось равно $\bar{X} = 404$ чел.

Будем проверять на 1%-ном уровне значимости гипотезу $H_0 : \mathbf{E}(X) = a_0 = 400$ при альтернативной гипотезе $H_0 : \mathbf{E}(X) > a_0$. В этом случае если у нас будут основания отвергнуть гипотезу H_0, то это позволит обоснованно продлить договор аренды. Если же не будет оснований отвергнуть гипотезу H_0, то это будет означать, что нет оснований считать магазин прибыльным.

Так как значение генеральной дисперсии $\sigma^2 = 400$ известно, то гипотезу H_0 отвергают и принимают гипотезу H_1, если наблюдаемое значение статистики

$$Z = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}$$

больше критической точки $z_{1-\alpha}$.

В нашем примере наблюдаемое значение статистики Z равно

$$\frac{(404 - 400)\sqrt{365}}{\sqrt{400}} = 3{,}820994635,$$

$$z_{1-\alpha} = \text{НОРМ.СТ.ОБР(1 – 0,01)} = 2{,}326347874,$$

т. е. наблюдаемое значение статистики Z попадает в критическую область ($Z > z_{1-\alpha}$), и это дает основания отвергнуть гипотезу H_0 на 1%-ном уровне значимости, считать магазин прибыльным и принять решение о продлении договора аренды.

При этом наблюдаемый уровень значимости

$$\mathbf{P}\{Z > 3{,}820994635\} = \text{1 – НОРМ.СТ.РАСП(3,820994635; ИСТИНА)} = 0{,}00006646.$$

Это означает, что при любом уровне значимости, большем 0,00006646, есть основания отвергнуть гипотезу H_0.

Таким образом, вероятность ошибиться, отвергнув гипотезу H_0 и продлив договор аренды, не превосходит очень маленького значения 0,006646 — скорее всего, если продлить договор аренды, магазин будет прибыльным! ❑

На практике, в отличие от разобранного примера 6.2.1, значение σ не бывает известно, поэтому рассмотренными критериями проверки гипотез можно пользоваться только при достаточно больших выборках. В общем же случае при неизвестном значении σ можно использовать тот факт, что статистика

$$T_{n-1} = \frac{(\bar{X} - a)\sqrt{n}}{s}$$

распределена по закону Стьюдента с $n-1$ степенью свободы.

Критическая область наиболее мощного критерия проверки гипотезы $H_0 : a = a_0$ при альтернативной гипотезе $H_1 : a = a_1 > a_0$ задается соотношением

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} > t_{n-1;1-\alpha},$$

где $t_{n-1;1-\alpha}$ = СТЬЮДЕНТ.ОБР(<1 – α>; <n – 1>) — квантиль уровня $1 - \alpha$ случайной величины, распределенной по Стьюдента с $n-1$ степенью свободы.

При проверке гипотезы $H_0 : a = a_0$ при альтернативной гипотезе $H_1 : a = a_1 < a_0$ критическая область наиболее мощного критерия определяется неравенством

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} < -t_{n-1;1-\alpha},$$

а в случае альтернативной гипотезы $H_1 : a = a_1 \neq a_0$ критическая область соответствует неравенству

$$|T_{n-1}| = \left|\frac{(\bar{X} - a_0)\sqrt{n}}{s}\right| > t_{n-1;1-\alpha/2}.$$

Замечание. В случае альтернативы $H_1 : a = a_1 \neq a_0$ критерий не является наиболее мощным!

Пример 6.2.2. Проверим с использованием статистики T_{n-1} гипотезу $H_0 : \mathbf{E}(X) = 250$ г при альтернативной гипотезе $H_1 : \mathbf{E}(X) \neq 250$ г. по данным примера 6.1.7: в контрольной партии из $n = 100$ брикетов средняя масса брикета оказалась равной $\bar{x} = 246$ г, при этом стандартное отклонение оказалось равным $s = 12$ г.

Уровень значимости, как и в примере 6.1.7, выберем равным $\alpha = 0{,}05$.

Модуль наблюдаемого значения статистики T_{n-1} равен

$$\left|\frac{(246-250)\sqrt{100}}{12}\right| = |-3{,}333333333| = 3{,}333333333,$$

критическая точка

$$t_{n-1;1-\alpha/2} = \text{СТЬЮДЕНТ.ОБР(1 – 0,05/2; 99)} = 1{,}984216952,$$

т. е. наблюдаемое значение статистики критерия T_{n-1} попадает в критическую область ($T_{n-1} > t_{n-1;1-\alpha/2}$), и это дает основания отвергнуть гипотезу H_0 на 5%-ном уровне значимости и считать, что фасовочный станок разладился.

При этом наблюдаемый уровень значимости

$$\mathbf{P}\left\{|T_{n-1}| > 3{,}333333333\right\} = \text{СТЬЮДЕНТ.РАСП.2Х(3,333333333; 99)} = 0{,}001208004.$$

(Функция Microsoft Excel СТЬЮДЕНТ.РАСП.2Х(<*t*>; <*k*>) возвращает «двустороннюю» вероятность $\mathbf{P}\left\{|T_k| > t\right\}$).

Это означает, что при любом уровне значимости, большем 0,001208004, есть основания отвергнуть гипотезу H_0.

Таким образом, вероятность ошибиться, отвергнув гипотезу H_0, не превосходит достаточно маленького значения 0,001208004 — скорее всего, мы не ошибемся, если остановим станок для переналадки! ❑

Пример 6.2.3. Производитель электроламп утверждает, что средний срок их службы составляет $a_0 = 1000$ ч. Проверим эту гипотезу на 5%-ном уровне значимости по выборке из $n = 25$ ламп, для которой $\bar{x} = 875$ ч, $s = 50$ ч.

Будем проверять на уровне значимости $\alpha = 5\% = 0{,}05$ гипотезу $H_0 : \mathbf{E}(X) = a_0$ о том, что математическое ожидание срока службы лампы равно $a_0 = 1000$ ч при альтернативной гипотезе $H_1 : \mathbf{E}(X) < a_0$ (состоящей в том, что на самом деле математическое ожидание срока службы лампы меньше, чем заявляет их производитель). Критическая область критерия определяется неравенством

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} < -t_{n-1;1-\alpha},$$

наблюдаемое значение статистики

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s}$$

оказалось равным

$$\frac{(875-1000)\sqrt{25}}{50} = -12{,}50,$$

а критическая точка

$$-t_{n-1;1-\alpha} = -\text{СТЬЮДЕНТ.ОБР}(1-0{,}05;\ 99) = -1{,}984216952.$$

Поскольку наблюдаемое значение статистики T_{n-1} оказалось меньше критической точки $-t_{n-1;1-\alpha}$, есть основания отвергнуть нулевую гипотезу: принимаем решение, что производитель завышает истинный срок службы ламп.

При этом наблюдаемый уровень значимости

$$\mathbf{P}\{T_{n-1} < -12{,}5\} = \text{СТЬЮДЕНТ.РАСП}(-12{,}5;\ 24;\ \text{ИСТИНА}) = 0{,}00000000000267. \ \square$$

Заметим, что когда мы отвергаем гипотезу $H_0 : \mathbf{E}(X) = a_0$ на каком-либо, сколь угодно маленьком, уровне з н а ч и м о с т и, это не обязательно означает, что значение математическогого ожидания генеральной совокупности с у щ е с т в е н н о отличается от a_0.

Рассмотрим соответствующий

Пример 6.2.4. Производитель планшетов провел исследование мнения потребителей и выяснил, что большинство из них недовольно слишком малым временем работы батареи при просмотре видео. Известно, что математическое ожидание времени работы батареи при просмотре видео на продаваемых сейчас планшетах равно $a_0 = 714$ мин.

Технические подразделения предложили новое решение, и провели тестирование n батарей нового типа. Оказалось, что времени работы этих n батарей при просмотре видео составило $\overline{x} = 715$ мин. при стандартном отклонени $s = 86$ мин.

Будем проверять гипотезу $H_0 : \mathbf{E}(X) = a_0$ о том, что математическое ожидание времени работы новых батарей при просмотре видео равно $a_0 = 714$ мин. (как раньше) при альтернативной гипотезе $H_1 : \mathbf{E}(X) > a_0$ (о том, что математическое ожидание времени работы при просмотре видео у новых батарей больше, чем у старых).

Если протестировано всего $n = 25$ планшетов, то наблюдаемое значение статистики

$$T_{n-1} = \frac{(\overline{X} - a_0)\sqrt{n}}{s}$$

равно

$$\frac{(715-714)\sqrt{25}}{86} = 0{,}058139535,$$

что соответствует наблюдаемому уровню значимости

$$\mathbf{P}\{T_{n-1} < 0{,}058139535\} = \text{СТЬЮДЕНТ.РАСП}(0{,}058139535;\ 24;\ \text{ИСТИНА}) = 0{,}477059428.$$

Естественно, при этом нет оснований отвергнуть гипотезу $H_0 : \mathbf{E}(X) = a_0$.

Но при $n = 25\,000$ наблюдаемое значение T_{n-1} увеличится до

$$\frac{(715-714)\sqrt{25\,000}}{86} = 1{,}838533523,$$

и соответствующий наблюдаемый уровень значимости будет равен

$$\mathbf{P}\{T_{n-1} < 1{,}838533523\} = \text{СТЬЮДЕНТ.РАСП}(1{,}838533523;\ 24\,999;\ \text{ИСТИНА}) = \\ = 0{,}032997841.$$

Видим, что при $n = 25\,000$ уже имеются основания отвергнуть гипотезу $H_0 : \mathbf{E}(X) = a_0$ на 3,3%-процентном уровне значимости и принять гипотезу $H_1 : \mathbf{E}(X) > a_0$ (о том, что математическое ожидание времени работы при просмотре видео у новых батарей больше, чем у старых), понимая, что вероятность ошибиться при этом будет меньше 3,3%.

Если же протестировано $n = 250\,000$ планшетов, то наблюдаемое значение T_{n-1} равно

$$\frac{(715-714)\sqrt{250\,000}}{86} = 5{,}813953488,$$

откуда p-значение

$$\mathbf{P}\{T_{n-1} < 5{,}813953488\} = \text{СТЬЮДЕНТ.РАСП}(5{,}813953488;\ 249\,999;\ \text{ИСТИНА}) = \\ = 0{,}00000000305441006.$$

При $n = 250\,000$ уже на любом разумном уровне значимости есть основания отвергнуть гипотезу $H_0 : \mathbf{E}(X) = a_0$ и считать, что математическое ожидание времени работы при просмотре видео у новых батарей больше, чем у старых.

Однако довольны ли пользователи таким увеличением срока работы бараreи? Скорее всего, нет. ❑

Поскольку статистики многих критериев проверки гипотез содержат в числителе квадратный корень из объема выборки (или другую возрастающую функцию n), то может показаться, что можно набрать побольше наблюдений, чтобы получить основания отвергнуть какую угодно гипотезу («подогнав ответ» под нужное «решение»).

Однако это не так: одной значимости различий для принятия решения мало, нужна еще и существенность различий.

В рассмотренном примере различия значимы (вероятность ошибиться, сказав, что различий нет, близка к нулю), но совсем не существенны (одну дополнительную минуту работы планшета пользователи, скорее всего, просто не заметят). Поэтому, скорее всего, требуется дальнейшая доработка планшета, чтобы пользователи почувствовали, что в новой версии планшета видео можно смотреть дольше, чем в старой.

Пример 6.2.5. В условиях примера 5.1.15, предположив нормальный закон распределения объема продаж, проверим на 10%-ном, 5%-ном и 1%-ном уровне значимости гипотезу о том, что математическое ожидание объема ежедневных продаж равно $a_0 = 51$ тыс. ден. ед. при альтернативе, что оно меньше a_0.

Проверяемая гипотеза имеет вид $H_0 : \mathbf{E}(X) = a_0$, в качестве альтернативной выбрана гипотеза $H_1 : \mathbf{E}(X) < a_0$, критическая область задается неравенством

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} < -t_{n-1;1-\alpha}.$$

Объем выборки $n = 100$, выборочное среднее $\bar{x} = \text{СРЗНАЧ(A2:A101)} = 49{,}596$, стандартное отклонение $s = \text{СТАНДОТКЛОН.В(A2:A101)} = 10{,}8541657$, поэтому наблюдаемое значение статистики T_{n-1} равно

$$\frac{(49{,}596-51)\sqrt{100}}{10{,}8541657}=-1{,}523924865.$$

Наблюдаемый уровень значимости

$$\mathbf{P}\{T_{n-1}<-1{,}523924865\}=$$

$$=\text{СТЬЮДЕНТ.РАСП}(-1{,}523924865;\ 100-1;\ \text{ИСТИНА})=0{,}065357815.$$

При $\alpha=0{,}1$ критическая точка

$$-t_{n-1;1-\alpha}=-\text{СТЬЮДЕНТ.ОБР}(1-0{,}1;\ 99)=-1{,}290161442,$$

поэтому на 10%-ном уровне значимости есть основания отвергнуть гипотезу H_0 о том, что математическое ожидание объема ежедневных продаж равно $a_0=51$ тыс. ден. ед. при альтернативе H_1, что оно меньше a_0.

При $\alpha=0{,}05$ критическая точка

$$-t_{n-1;1-\alpha}=-\text{СТЬЮДЕНТ.ОБР}(1-0{,}05;\ 99)=-1{,}660391156,$$

поэтому на 5%-ном уровне значимости нет оснований отвергнуть гипотезу H_0.

При $\alpha=0{,}01$ критическая точка

$$-t_{n-1;1-\alpha}=-\text{СТЬЮДЕНТ.ОБР}(1-0{,}01;\ 99)=-2{,}364605862,$$

поэтому и на 1%-ном уровне значимости нет оснований отвергнуть гипотезу H_0.

Если мы на 1%-ном уровне значимости не отвергли гипотезу H_0, то могли совершить ошибку второго рода. Пусть альтернативное значение математического ожидания равно $a_1=50$, тогда вероятность ошибки второго рода равна

$$\beta=\mathbf{P}\{H_0|H_1\}=\mathbf{P}\left\{Z\in(S\setminus R^*)\,\middle|\,H_1\right\}=\mathbf{P}\left\{\frac{(\overline{X}-a_0)\sqrt{n}}{s}\geqslant-t_{n-1;1-\alpha}\,\middle|\,T_{n-1}=\frac{(\overline{X}-a_1)\sqrt{n}}{s}\right\}=$$

$$=\mathbf{P}\left\{\frac{(\overline{X}-a_0)\sqrt{n}}{s}+\frac{(a_0-a_1)\sqrt{n}}{s}\geqslant-t_{n-1;1-\alpha}+\frac{(a_0-a_1)\sqrt{n}}{s}\,\middle|\,T_{n-1}=\frac{(\overline{X}-a_1)\sqrt{n}}{s}\right\}=$$

$$=\mathbf{P}\left\{\frac{(\overline{X}-a_1)\sqrt{n}}{s}\geqslant-t_{n-1;1-\alpha}+\frac{(a_0-a_1)\sqrt{n}}{s}\,\middle|\,T_{n-1}=\frac{(\overline{X}-a_1)\sqrt{n}}{s}\right\}=$$

$$=\mathbf{P}\left\{T_{n-1}\geqslant-t_{n-1;1-\alpha}+\frac{(a_0-a_1)\sqrt{n}}{s}\right\}=1-\mathbf{P}\left\{T_{n-1}\leqslant-t_{n-1;1-\alpha}+\frac{(a_0-a_1)\sqrt{n}}{s}\right\}.$$

Подставляя значения $\alpha=0{,}01$, $-t_{n-1;1-\alpha}=-2{,}364605862$, $a_0=51$, $a_1=50$, $n=100$, $s=10{,}85417$, получаем:

$$-t_{n-1;1-\alpha}+\frac{(a_0-a_1)\sqrt{n}}{s}=$$

$$=-2{,}364605862+(51-50)*\text{КОРЕНЬ}(100)/10{,}8541657=-1{,}443300963,$$

$$\beta=\mathbf{P}\{H_0|H_1\}=$$

$$=1-\text{СТЬЮДЕНТ.РАСП}(-1{,}443300963;\ 100-1;\ \text{ИСТИНА})=0{,}923953722$$ —

вероятность ошибиться весьма высока! ❑

Пример 6.2.6. Для акций из примеров 1.1.3, 1.2.10, 1.2.13, 1.4.10, 5.3.5, 5.3.13, 5.3.15, предположив нормальный закон распределения доходности, проверим

гипотезы о том, что математические ожидания месячной доходности каждой из акций равны 0,02 при альтернативных гипотезах, что они меньше, чем 0,02.

Проверяемая для каждой акции гипотеза имеет вид $H_0: \mathbf{E}(X) = a_0 = 0{,}02$, в качестве альтернативной выбрана гипотеза $H_1: \mathbf{E}(X) < a_0$, критическая область задается неравенством

$$T_{n-1} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} < -t_{n-1;1-\alpha},$$

Подставляя значения

$$n = 56;\ \ t_{56-1;1-0,05} = t_{55;0,95} = \text{СТЬЮДЕНТ.ОБР(0,95; 55)} = 1{,}673033965;$$

$$\bar{x}_{\text{ВТБ}} = 0{,}006526318;\ s_{\text{ВТБ}} = 0{,}094495296;$$

$$\bar{x}_{\text{Газпром}} = -0{,}001153938;\ s_{\text{Газпром}} = 0{,}067803139;$$

$$\bar{x}_{\text{Сбербанк}} = 0{,}015489170;\ \ s_{\text{Сбербанк}} = 0{,}093977345,$$

получаем:

$$T_{n-1}^{(\text{ВТБ})} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} = \frac{(0{,}006526318 - 0{,}02)\sqrt{56}}{0{,}094495296} = -1{,}067013982 > -t_{55;0,95};$$

$$T_{n-1}^{(\text{Газпром})} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} = \frac{(-0{,}001153938 - 0{,}02)\sqrt{56}}{0{,}067803139} = -2{,}334723432 < -t_{55;0,95};$$

$$T_{n-1}^{(\text{Сбербанк})} = \frac{(\bar{X} - a_0)\sqrt{n}}{s} = \frac{(0{,}015489170 - 0{,}02)\sqrt{56}}{0{,}093977345} = -0{,}359192573 > -t_{55;0,95}.$$

Поэтому для акций Газпрома на 5%-ном уровне значимости есть основания отвергнуть гипотезу о том, что математическое ожидание месячной доходности равно 0,02 при альтернативной гипотезе, что оно меньше, чем 0,02, а для акций ВТБ и Сбербанка нет оснований отвергнуть соответствующие гипотезы.

Наблюдаемые уровни значимости

$$\mathbf{P}\left\{T_{n-1}^{(\text{ВТБ})} < -1{,}067013982\right\} = 0{,}145313441;$$

$$\mathbf{P}\left\{T_{n-1}^{(\text{Газпром})} < -2{,}334723432\right\} = 0{,}011617999;$$

$$\mathbf{P}\left\{T_{n-1}^{(\text{Сбербанк})} < -0{,}359192573\right\} > 0{,}360413193. \square$$

Для удобства рассмотренные в этом и следующих пунктах критерии проверки гипотез о значениях параметров случайных величин собраны в таблицы, приведенные в **прил. 3**.

Задачи

386. Инвестиционный фонд объявил, что доходность по месячным вложениям в него превышает среднерыночную на 0,003. В течение последнего года средняя доходность по рынку составила 0,005, а средняя доходность по фонду составила $\bar{x} = 0{,}0065$ при стандартном отклонении $s = 0{,}019$. Проверьте на 5%-ном уровне значимости справедливость заявления фонда.

387. В условиях задачи **351** проверьте с помощью статистики T_{n-1} на уровне значимости $\alpha = 0{,}05$ гипотезу о том, что математическое ожидание суммарного пополнения вклада $\mathbf{E}(X) = 50$ тыс. руб. при альтернативных гипотезах:

а) $\mathbf{E}(X) \neq 50$ тыс. руб.; *б*) $\mathbf{E}(X) < 50$ тыс. руб.; *в*) $\mathbf{E}(X) > 50$ тыс. руб. В каждом случае вычислите также наблюдаемый уровень значимости.

388. Для каждой акции из задачи 5 проверьте на 5%-ном и 1%-ном уровне значимости гипотезу о том, что математическое ожидание недельной доходности равно нулю при альтернативной гипотезе о том, что оно больше нуля. В каждом случае вычислите также наблюдаемый уровень значимости.

6.2.2. Проверка гипотезы о числовом значении вероятности

Проверка гипотезы $H_0 : p = p_0$ о числовом значении вероятности некоторого события при большом объеме выборки основана на близости закона распределения статистики

$$Z = \frac{(\hat{p} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

к стандартному нормальному закону в предположении справедливости H_0.

Критическая область определяется альтернативной гипотезой:

$$\begin{cases} Z > z_{1-\alpha} & \text{при альтернативной гипотезе } H_1 : p > p_0; \\ Z < -z_{1-\alpha} & \text{при альтернативной гипотезе } H_1 : p < p_0; \\ |Z| > z_{1-\alpha/2} & \text{при альтернативной гипотезе } H_1 : p \neq p_0. \end{cases}$$

Пример 6.2.7. Банк рассылает своим клиентам смс с предложением обратиться за кредитом. Вероятность того, что клиент, получивший смс, обратится за кредитом, равна 0,08. После разработки нового рекламного предложения было разослано 1000 смс с новым рекламным текстом и получено 100 обращений.

Проверим на 10%-ном уровне значимости гипотезу H_0: $p = p_0 = 0{,}08$ при альтернативной гипотезе H_1: $p > p_0$.

Нулевая гипотеза отвергается, если наблюдаемое числовое значение статистики

$$Z = \frac{(\hat{p} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

окажется больше критической границы $z_{1-\alpha}$.

Наблюдаемое значение статистики Z равно

$$\frac{(0{,}1 - 0{,}08)\sqrt{1000}}{\sqrt{0{,}08 \cdot 0{,}92}} = 2{,}331262021,$$

что больше правой критической границы

$$z_{1-\alpha} = \text{НОРМ.СТ.ОБР}(1 - 0{,}1) = 1{,}281551566.$$

Значит, на 10%-ном уровне значимости есть основания отвергнуть нулевую гипотезу и принять альтернативную гипотезу (о большей эффективности нового рекламного предложения). При этом вероятность вынесения неправильного решения об эффективности новой формы — вероятность ошибки первого рода — равна 0,1. ❑

Пример 6.2.8. Менеджер по продажам утверждает, что он получает заказы в среднем не менее чем от 30% предполагаемых клиентов. За последний месяц у

менеджера было $n = 100$ потенциальных клиентов, из которых $m = 20$ сделали заказы. Проверим на 5%-ном уровне значимости гипотезу H_0: $p = p_0 = 0{,}3$ при альтернативной гипотезе H_1: $p < p_0$.

Нулевая гипотеза отвергается, если наблюдаемое числовое значение статистики

$$Z = \frac{(\hat{p} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

окажется меньше левой критической границы $-u_{(1-2\alpha)/2}$.

В данной задаче наблюдаемое значение статистики Z равно

$$\frac{(0{,}2-0{,}3)\sqrt{100}}{\sqrt{0{,}3\cdot 0{,}7}} = -2{,}182178902,$$

что меньше левой критической границы

$$-z_{1-\alpha} = -\text{НОРМ.СТ.ОБР}(1-0{,}05) = -1{,}644853627,$$

поэтому есть основания отвергнуть нулевую гипотезу на 5%-ном уровне значимости — менеджер преувеличивает. ❑

В примере 6.2.4 мы заметили, что неверно считать, что чем меньше p-значение, тем больше отличия реальной ситуации от гипотезы H_0. Приведем еще один

Пример 6.2.9. Пусть до проведения А-банком маркетинговой кампании среди своих существующих клиентов каждый месяц 3% вкладчиков уходили в другие банки, а после проведения маркетинговой кампании в течение первого месяца из 10 000 вкладчиков А-банка только 294 забрали свои сбережения (294/10 000 =0,0294). Тогда

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0{,}0294 - 0{,}03}{\sqrt{0{,}03(1-0{,}03)/10\,000}} = -0{,}35, \quad \mathbf{P}\{\mathcal{N}(0;1) < -0{,}35\} = 0{,}363,$$

т. е. вероятность ошибиться, отвергнув гипотезу H_0: $p = p_0$ (при альтернативной гипотезе H_1: $p < p_0$), составляет более 36% — нет никаких оснований отвергнуть гипотезу H_0 и считать, что маркетинговая кампания привела к уменьшению оттока клиентов.

Пусть в Б-банке не 10 000, а 1 000 000 вкладчиков, и при этом существует та же проблема оттока клиентов: среди существующих клиентов 3% вкладчиков уходят в другие банки в течение месяца. Б-банк провел аналогичную маркетинговую кампанию, и в первый месяц после нее 29 400 вкладчиков из 1 000 000 забрали свои сбережения (29 400/1 000 000 =0,0294). В этом случае

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0{,}0294 - 0{,}03}{\sqrt{0{,}03(1-0{,}03)/1\,000\,000}} = -3{,}52, \quad \mathbf{P}\{\mathcal{N}(0;1) < -3{,}52\} = 0{,}000218.$$

Вероятность ошибиться, отвергнув гипотезу H_0, уже пренебрежимо мала, и мы получаем основания отвергнуть гипотезу H_0 и считать, что маркетинговая кампания привела к уменьшению оттока клиентов. Но на самом деле отток уменьшился значимо, но несущественно: с 3% до 2,94%, всего на 0,06%. ❑

Задачи

389. Продюсер некоторой телепередачи утверждает, что она должна привлечь внимание не менее, чем трети телезрителей. При этом из 64 опрошен-

ных только 16 заявили о своем намерении посмотреть эту передачу. Оцените утверждение продюсера на 5%-ном уровне значимости.

390. Статистика страхового брокера утверждает, что только 3 из 10 визитов страхового агента к потенциальным страхователям заканчиваются заключением договора о страховании. Агент Иванов заключил $m = 40$ договоров за $n = 100$ встреч с потенциальными клиентами. Определите, случайны ли результаты Иванова, или они свидетельствуют о его высокой квалификации.

391. Количество бракованных деталей в партии не должно превышать 5%. В результате контроля 100 деталей из этой партии обнаружено 6 бракованных. Определите, можно ли на уровне значимости $\alpha = 0{,}05$ считать, что процент брака превосходит допустимый.

392. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,02. Среди случайно отобранных 480 изделий оказалось 10 дефектных. Определите, можно ли на 1%-ном уровне значимости принять партию. Вычислите вероятность вынесения ошибочного заключения, если альтернативное значение вероятности брака равно 0,022.

6.2.3. Проверка гипотезы о числовом значении дисперсии нормального закона распределения

Проверка гипотезы $H_0 : \sigma^2 = b_0$ о числовом значении дисперсии основана на том, что в предположении справедливости H_0 статистика

$$\chi^2_{n-1} = \frac{(n-1)s^2}{b_0}$$

имеет закон распределения «Хи-квадрат» с $n - 1$ степенью свободы.

Критическая область определяется альтернативной гипотезой:

$$\begin{cases} \chi^2_{n-1} > \chi^2_{n-1;1-\alpha} & \text{при альтернативе } H_1 : \sigma^2 > b_0; \\ \chi^2_{n-1} < \chi^2_{n-1;\alpha} & \text{при альтернативе } H_1 : \sigma^2 < b_0; \\ \chi^2_{n-1} \in \left(-\infty; \chi^2_{n-1;\alpha/2}\right) \cup \left(\chi^2_{n-1;1-\alpha/2}; +\infty\right) & \text{при альтернативе } H_1 : \sigma^2 \neq b_0. \end{cases}$$

Пример 6.2.10. Управляющий портфелем заботится о том, чтобы не осуществлять вложения в ценные бумаги с дисперсией годовой доходности, превышающей $b_0 = 0{,}04$. Выборка из $n = 51$ наблюдения за доходностью некоторого актива показала, что исправленная выборочная дисперсия $s^2 = 0{,}045$.

Проверим на 5%-ном уровне значимости гипотезу H_0: $\sigma^2 = b_0 = 0{,}04$ при альтернативной гипотезе H_1: $\sigma^2 > b_0$.

Нулевую гипотезу можно отвергнуть на уровне значимости α, если наблюдаемое числовое значение статистики

$$\chi^2_{n-1} = \frac{(n-1)s^2}{b_0}$$

окажется больше критической границы $\chi^2_{n-1;1-\alpha}$.

В данном случае наблюдаемое значение статистики χ^2_{n-1} равно

$$\frac{50\cdot 0{,}045}{0{,}04}=56{,}25,$$

а критическая точка $\chi^2_{51-1;1-0{,}05}$ = ХИ2.ОБР(1 – 0,05; 51 – 1) = 67,50480655, поэтому на 5%-ном уровне значимости нет оснований отвергнуть гипотезу о том, что доходность данного актива характеризуются дисперсией, не большей 0,04.

Если альтернативное значение генеральной дисперсии $b_1 = 0{,}045$, то вероятность ошибки второго рода (при альтернативной гипотезе H_1: $\sigma^2 = b_1 = 0{,}045$) будет равна

$$\beta = \mathbf{P}\{H_0 \mid H_1\} =$$

$$= \mathbf{P}\left\{\chi^2_{n-1} \leqslant \chi^2_{n-1;1-\alpha}\frac{b_0}{b_1}\right\} = \mathbf{P}\left\{\chi^2_{50} \leqslant \chi^2_{50;0{,}95}\frac{0{,}04}{0{,}045}\right\} = \mathbf{P}\left\{\chi^2_{50} \leqslant 67{,}50480655\cdot\frac{0{,}04}{0{,}045}\right\} =$$

= ХИ2.РАСП(67,50480655*0,04/0,045; 50; ИСТИНА) = 0,842848949. ❑

Задачи

393. По данным за последние $n = 12$ месяцев найдена средняя доходность по некоторой акции $\bar{x} = 65\%$ и соответствующее стандартное отклонение $\hat{\sigma} = 7\%$. Считая, что доходность акции распределена по нормальному закону, проверьте на уровне значимости $\alpha = 0{,}05$ гипотезы о том, что в генеральной совокупности средняя доходность равна 60%, а дисперсия равна $50(\%)^2$. Рассчитайте также вероятности неправильных выводов.

394. Типичное среднее квадратическое отклонение экзаменационных оценок по математике у студентов первого курса в прошлые годы было равным 0,8. Предполагается, что новый компьютерно-ориентированный курс математики сглаживает различия в оценках. Проверьте это предположение на 5%-ном уровне значимости, если исправленная выборочная дисперсия оценок 25 студентов, изучавших новый компьютерно-ориентированный курс, равна 0,49. Определите также вероятность ошибочного заключения.

395. Партия изделий принимается, если среднее квадратическое отклонение контролируемого размера не превышает 0,2 мкм. Стандартное отклонение, вычисленное по выборке из n изделий, оказалось равным 0,25 мкм. Определите, можно ли принять партию на 5%-ном уровне значимости, если: *а*) $n = 40$; *б*) $n = 4000$.

396. По данным за последние 12 мес. найдена средняя месячная доходность по некоторой акции А $\bar{x} = 0{,}0065$ и соответствующее стандартное отклонение $s = 0{,}0007$. Допустив, что доходность акции — случайная величина, распределенная по нормальному закону, проверьте на уровне значимости $\alpha = 0{,}05$ гипотезы о том, что в генеральной совокупности средняя доходность равна 0,0060, а дисперсия равна $0{,}0005^2$. Определите также вероятности неправильных выводов.

6.2.4. Проверка гипотезы о равенстве дисперсий двух генеральных совокупностей

Критерий проверки гипотезы о равенстве дисперсий $\sigma_1^2 = \mathbf{Var}(X^{(1)})$ и $\sigma_2^2 = \mathbf{Var}(X^{(2)})$ двух распределенных по нормальному закону случайных величин $X^{(1)}$ и $X^{(2)}$ основан на том, что в предположении справедливости гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ статистика

$$F_{n_1-1;\,n_2-1}=\frac{s_1^2}{s_2^2}$$

распределена по закону Фишера — Снедекора с n_1-1 и n_2-1 степенями свободы (здесь n_1, n_2 — объемы выборок, s_1^2, s_2^2 — исправленные выборочные дисперсии).

Критическая область определяется альтернативной гипотезой:

$$\begin{cases} F_{n_1-1;\,n_2-1}=\dfrac{s_1^2}{s_2^2}>f_{n_1-1;\,n_2-1;\,1-\alpha} & \text{при альтернативе } H_1:\sigma_1^2>\sigma_2^2; \\ F_{n_2-1;\,n_1-1}=\dfrac{s_2^2}{s_1^2}>f_{n_2-1;\,n_1-1;\,1-\alpha} & \text{при альтернативе } H_1:\sigma_1^2<\sigma_2^2; \\ F_{m_1-1;\,m_2-1}=\dfrac{\max\left\{s_1^2;\,s_2^2\right\}}{\min\left\{s_1^2;\,s_2^2\right\}}>f^2_{m_1-1;\,m_2-1;\,1-\alpha/2} & \text{при альтернативе } H_1:\sigma_1^2\neq\sigma_2^2. \end{cases}$$

В определении критической области для альтернативы $H_1:\sigma_1^2\neq\sigma_2^2$

$$m_1=\begin{cases} n_1 & \text{при } s_1^2\geqslant s_2^2; \\ n_2 & \text{при } s_1^2<s_2^2; \end{cases}\quad m_2=\begin{cases} n_2 & \text{при } s_1^2\geqslant s_2^2; \\ n_1 & \text{при } s_1^2<s_2^2. \end{cases}$$

Пример 6.2.11. По данным о доходностях двух акций за десять кварталов:

$q_i^{(1)}$	0,7	−1,6	−0,2	−1,2	−0,1	3,4	3,7	0,8	0,0	2,0
$q_i^{(2)}$	1,9	0,8	1,1	0,1	−0,1	4,4	5,5	1,6	4,6	3,4

проверим на 5%-ном уровне значимости, имеется ли различие между дисперсиями доходностей.

Рассчитаем исправленные выборочные дисперсии доходностей двух акций:

$$s_1^2=3{,}20,\quad s_2^2=4{,}009.$$

Наблюдаемое значение статистики

$$F_{n_2-1;n_1-1}=\frac{s_2^2}{s_1^2}$$

равно

$$\frac{4{,}009}{3{,}201}=1{,}252595036,$$

что меньше критической точки

$$f_{n_2-1;\,n_1-1;\,1-\alpha/2}=f_{9;\,9;\,1-0{,}05/2}=\text{F.ОБР(1 – 0,05/2; 10 – 1; 10 – 1)}=4{,}025994158,$$

поэтому на 5%-ном уровне значимости нет оснований отвергнуть гипотезу о равенстве генеральных дисперсий. ❑

Задачи

397. Для сравнения точности двух станков-автоматов взяты две выборки из 13 и 9 деталей соответственно. В каждой из этих выборок проведены измерения контролируемого размера. Исправленные выборочные дисперсии ока-

зались равными соответственно 130,6 и 124,8. Определите, можно ли на 5%-ном уровне значимости считать точность станков одинаковой.

398. Для каждой пары акций из задачи 5 проверьте на 5%-ном и 1%-ном уровне значимости гипотезу о том, что дисперсии доходностей акций равны между собой (при альтернативе, что не равны). В каждом случае вычислите также наблюдаемый уровень значимости.

6.2.5. Проверка гипотезы о равенстве математических ожиданий двух генеральных совокупностей

Критерий проверки гипотезы о равенстве математических ожижаний $\mathbf{E}(X^{(1)})$ и $\mathbf{E}(X^{(2)})$ двух распределенных по нормальному закону случайных величин $X^{(1)}$ и $X^{(2)}$ в случае, когда известно, что дисперсии случайных величин совпадают, основан на том факте, что в предположении справедливости гипотезы $H_0: \mathbf{E}(X^{(1)}) = \mathbf{E}(X^{(2)})$ статистика

$$T_{n_1+n_2-2} = \frac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

распределена по закону Стьюдента с $n_1 + n_2 - 2$ степенями свободы (здесь n_1, n_2 — объемы выборок, $\overline{X^{(1)}}, \overline{X^{(2)}}$ — выборочные средние, s_1^2, s_2^2 — исправленные выборочные дисперсии).

Критическая область имеет вид

$$\begin{cases} T_{n_1+n_2-2} > t_{1-\alpha;\, n_1+n_2-2} & \text{при альтернативной гипотезе } H_1: \mathbf{E}(X^{(1)}) > \mathbf{E}(X^{(2)}); \\ T_{n_1+n_2-2} < -t_{1-\alpha;\, n_1+n_2-2} & \text{при альтернативной гипотезе } H_1: \mathbf{E}(X^{(1)}) < \mathbf{E}(X^{(2)}); \\ \left|T_{n_1+n_2-2}\right| > t_{1-\alpha/2;\, n_1+n_2-2} & \text{при альтернативной гипотезе } H_1: \mathbf{E}(X^{(1)}) \neq \mathbf{E}(X^{(2)}). \end{cases}$$

Пример 6.2.12. Автомобильный журнал тестировал два вида зимних покрышек: А и В. Для 10 покрышек вида А средний пробег оказался равен $\overline{x^{(1)}} = 40\,000$ км при стандартном отклонении $s_1 = 5\,950$ км, для 12 покрышек вида В $\overline{x^{(2)}} = 38\,000$ км, $s_2 = 5\,150$ км. Известно, что генеральные совокупности имеют нормальный закон распределения, причем $\sigma_1 = \sigma_2$. Чтобы определить, существенно ли различие между стойкостью двух видов шин, проверим на 5%-ном уровне значимости гипотезу H_0: $a_1 = a_2$ при альтернативной гипотезе H_1: $a_1 \neq a_2$.

Поскольку известно, что средние квадратические отклонения стойкости шин равны, то нулевую гипотезу следует отвергнуть, если модуль наблюдаемого числового значения статистики

$$T_{n_1+n_2-2} = \frac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

окажется больше критической точки $t_{n_1+n_2-2;\,1-\alpha/2}$.

В данном случае $\left|T_{n_1+n_2-2}\right|$ оказался равен

$$\left|\frac{40000-38000}{\sqrt{\frac{9\cdot 5950^2+11\cdot 5150^2}{10+12-2}\left(\frac{1}{10}+\frac{1}{12}\right)}}\right|=0,805646149,$$

что меньше критической точки $t_{10+12-2;1-0,05/2}=2,085963447$, поэтому на 5%-ном уровне значимости нет оснований отвергнуть гипотезу H_0 (т. е. можно утверждать, что различия между стойкостью двух видов шин не являются существенными). ❑

Пример 6.2.13. По данным о доходностях двух акций из примера 6.2.11 проверим, имеется ли существенное различие между доходностями.

Рассчитаем выборочные средние доходности: $\overline{q^{(1)}}=0,75$, $\overline{q^{(2)}}=2,33$.

В примере 6.2.11 были найдены исправленные выборочные дисперсии доходностей $s_1^2=3,20$, $s_2^2=4,01$, а также была произведена проверка гипотезы о равенстве генеральных дисперсий при альтернативной гипотезе о том, что дисперсии не равны: оказалось, что на 5%-ном уровне значимости нет оснований отвергнуть гипотезу о равенстве генеральных дисперсий.

Это позволяет применить известную статистику для проверки гипотезы о равенстве генеральных средних (при альтернативной гипотезе о том, что генеральные средние не равны).

Модуль наблюдаемого значения статистики

$$T_{n_1+n_2-2}=\frac{\overline{Q^{(1)}}-\overline{Q^{(2)}}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

равный

$$\left|\frac{0,75-2,33}{\sqrt{\frac{9\cdot 3,20+9\cdot 4,01}{9+9-2}\left(\frac{1}{9}+\frac{1}{9}\right)}}\right|=\left|-1,664362163\right|=1,664362163,$$

оказался меньше критической точки

$$t_{n_1+n_2-2;1-\alpha/2}=t_{9+9-2;\ 1-0,05/2}=t_{16;\ 0,975}=2,119905299,$$

поэтому на 5%-ном уровне значимости нет оснований отвергнуть гипотезу о равенстве генеральных средних доходностей двух акций; делаем вывод о несущественном различии доходностей акций.

Заметим, что на 10%-ном уровне значимости то гипотеза о равенстве генеральных средних должна быть отвергнута, поскольку $\left|T_{n_1+n_2-2}\right|=1,86$ больше критической точки

$$t_{9+9-2;\ 1-0,1/2}=t_{16;\ 0,95}=1,745883676\,.$$

Иными словами, если мы готовы довольствоваться 90%-ной надежностью выводов, то можно признать, что доходности акций различаются, если же нам нужна надежность не менее 95%, то оснований утверждать, что между доходностями двух акций есть различия, у нас нет.

Вычислим наблюдаемый уровень значимости

$$p = \mathbf{P}\left\{\left|T_{n_1+n_2-2}\right| > 1{,}664362163\right\} =$$

$$= \text{СТЬЮДЕНТ.РАСП.2Х}(1{,}664362163;\ 16) = 0{,}115498132,$$

такова вероятность ошибиться, отвергнув гипотезу H_0.

Теперь проверим гипотезу из примера 6.2.11 и гипотезу из данного примера с использованием программы «Двухвыборочный F-тест для дисперсии» надстройки «Анализ данных» пакета Microsoft Excel. Введем данные в рабочий лист, как показано на рис. 6.2.4.

	A	B
1	Q(1)	Q(2)
2	0,7	1,9
3	–1,6	0,8
4	–0,2	1,1
5	–1,2	0,1
6	–0,1	–0,1
7	3,4	4,4
8	3,7	5,5
9	0,8	1,6
10	0,0	4,6

Рис. 6.2.4. *Числовые данные для программ «Двухвыборочный F-тест для дисперсии» и «Двухвыборочный t-тест с одинаковыми дисперсиями» в примере 6.2.13*

Воспользуемся программой «Двухвыборочный F-тест для дисперсии». Для этого выберем соответствующий пункт меню надстройки «Анализ данных». В появившемся окне ввода данных (рис. 6.2.5) укажем интервал переменной 1: $B$1:$B$11 и интервал переменной 2: $A$1:$A$11 (в которых находятся исходные данные — обратим внимание, что первой переменной должна быть та, которая обладает большей выборочной дисперсией, в данном случае $Q^{(2)}$). Отметим флажок «Метки», поскольку первые строки интервалов переменных содержат названия этих переменных.

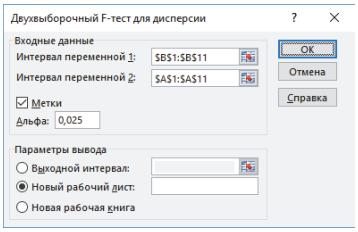


Рис. 6.2.5. *Окно ввода данных программы «Двухвыборочный F-тест для дисперсии»*

Особенно отметим, что в случае проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ о равенстве генеральных дисперсий при альтернативной гипотезе $H_1 : \sigma_1^2 > \sigma_2^2$ в поле «Альфа» нужно вводить уровень значимости гипотезы, а в случае, когда альтернативная гипотеза имеет вид $H_1 : \sigma_1^2 \neq \sigma_2^2$, в поле «Альфа» следует вводить половину уровня значимости. В данном случае в поле «Альфа» введем $\alpha/2 = 0{,}025$.

Укажем, что результаты работы программы необходимо вывести на новый рабочий лист. Результаты работы программы представлены на рис. 6.2.6.

Теперь воспользуемся программой «Двухвыборочный t-тест с одинаковыми дисперсиями». Для этого выберем соответствующий пункт меню надстройки «Анализ

данных». В появившемся окне ввода данных (рис. 6.2.7) укажем **интервал переменной 1: A1:A12** и **интервал переменной 2: B1:B12** (в которых находятся исходные данные). Укажем гипотетическую среднюю разность, равную нулю, отметим флажок «**Метки**», поскольку первые строки интервалов переменных содержат названия этих переменных. Зададим уровень значимости «**Альфа**» (по условию $\alpha = 0{,}05$). Укажем, что результаты работы программы необходимо вывести на **новый рабочий лист**. Результаты работы программы представлены на рис. 6.2.8.

Естественно, результаты работы программ «**Двухвыборочный F-тест для дисперсии**» и «**Двухвыборочный t-тест с одинаковыми дисперсиями**» совпали с полученными без использования этих программ. ❑

Двухвыборочный F-тест для дисперсии

	Q(2)	*Q(1)*
Среднее	2,33	0,75
Дисперсия	4,01	3,20
Наблюдения	10	10
df	9	9
F	1,252595036	
P(F<=f) одностороннее	0,371359966	
F критическое одностороннее	4,025994158	

Рис. 6.2.6. *Результаты работы программы «Двухвыборочный F-тест для дисперсии»*

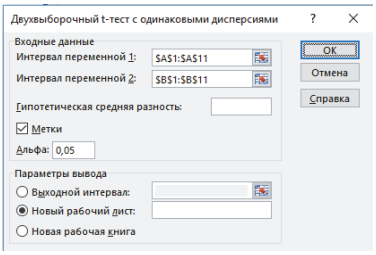


Рис. 6.2.7. *Окно ввода данных программы «Двухвыборочный t-тест с одинаковыми дисперсиями»*

Двухвыборочный t-тест с одинаковыми дисперсиями

	Q(1)	*Q(2)*
Среднее	0,75	2,33
Дисперсия	3,200555556	4,009
Наблюдения	10	10
Объединенная дисперсия	3,604777778	
Гипотетическая разность средних	0	
df	18	
t-статистика	–1,860813467	
P(T<=t) одностороннее	0,039593357	
t критическое одностороннее	1,734063607	
P(T<=t) двухстороннее	0,079186714	
t критическое двухстороннее	2,10092204	

Рис. 6.2.8. *Результаты работы программы «Двухвыборочный t-тест с одинаковыми дисперсиями» в примере 6.2.13*

В общем случае гипотезу $H_0 : \mathbf{E}(X^{(1)}) = \mathbf{E}(X^{(2)})$ (при неизвестных, но одинаковых дисперсиях) можно проверять, основываясь на наблюдаемом уровне значимости (p-значении), т. е. на вероятности того, что статистика $T_{n_1+n_2-2}$ окажется больше модуля наблюдаемого числового значения этой статистики («**P(T<=t) одностороннее**»), которая используется при проверке гипотезы H_0 при односторонних альтернативных гипотезах) или вероятности того, что модуль статистики $T_{n_1+n_2-2}$ окажет-

ся больше наблюдаемого числового значения этой статистики («P(T<= t) двусторонне»), которая используется при проверке гипотезы H_0 при двусторонней альтернативе). Если соответствующее типу альтернативной гипотезы p-значение оказывается не меньше принятого уровня значимости α, то гипотезу H_0 следует отвергнуть.

Замечание. В случае, когда оказывается, что дисперсии случайных величин $X^{(1)}$ и $X^{(2)}$ не совпадают (т. е. когда отвергается гипотеза о равенстве дисперсий), можно использовать **критерий Кохрана — Кокса:** если

$$\left|\frac{\overline{X^{(1)}}-\overline{X^{(2)}}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right| > \frac{\frac{s_1^2}{n_1}t_{n_1-1;1-\alpha}+\frac{s_2^2}{n_2}t_{n_2-1;1-\alpha}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

то гипотеза $H_0: \mathbf{E}(X^{(1)}) = \mathbf{E}(X^{(2)})$ отвергается и принимается гипотеза $H_1: \mathbf{E}(X^{(1)}) \neq \mathbf{E}(X^{(2)})$.

В Microsoft Excel этот критерий реализован в программе «Двухвыборочный t-тест с различными дисперсиями».

Задачи

399. Банк предлагает своим корпоративным клиентам два типа карточных счетов для начисления зарплаты сотрудникам: «Авиакарту», держатели которой при покупках копят мили, и «Сверхкарту», по которой до 5% ежемесячных трат возвращается держателям обратно на счет. За последний год средние годовые расходы по счетам 10 200 обладателей «Авиакарты» оказались равны 560 300 руб. при стандартном отклонении 172 545 руб., а средние годовые расходы по счетам 9 800 обладателей «Сверхкарты» составили 485 722 руб. при стандартном отклонении 192 545 руб. Можно ли утверждать, что математическое ожидание расходов по счетам у держателей «Авиакарты» больше, чем у держателей «Сверхкарты»?

400. У группы из $n = 8$ студентов-второкурсников проверялись знания английского языка по 25-балльной системе.

№ студента	1	2	3	4	5	6	7	8
Оценка до консультации	20	18	17	16	14	14	12	9
Оценка после консультации	18	22	15	17	8	20	9	7

Преподаватель считает, что консультация не помогла. Проверьте эту гипотезу на 1%-ном уровне значимости.

401. Проверьте на 1%-ном уровне значимости, имеется ли существенное различие между двумя видами снотворного по данным о 10 пациентах:

Пациент	Дополнительные часы сна	
	Препарат А	Препарат В
1	0,7	1,9
2	-1,6	0,8
3	-0,2	1,1
4	-1,2	0,1
5	-0,1	-0,1
6	3,4	4,4
7	3,7	5,5
8	0,8	1,6
9	0,0	4,6
10	2,0	3,4

402. Для каждой пары акций из задачи 5 проверьте на 5%-ном и 1%-ном уровне значимости гипотезу о том, что математические ожидания недельной доходности акций равны между собой (при альтернативе, что не равны). В каждом случае вычислите также наблюдаемый уровень значимости.

6.2.6. Проверка гипотезы о равенстве генеральных долей

Проверка гипотезы $H_0: p = p_0$ о числовом значении вероятности некоторого события при большом объеме выборки основана на близости закона распределения статистики

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

к стандартному нормальному закону в предположении справедливости H_0.

Критическая область определяется альтернативной гипотезой:

$$\begin{cases} Z > z_{1-\alpha} & \text{при альтернативной гипотезе } H_1: p_1 > p_2; \\ Z < -z_{1-\alpha} & \text{при альтернативной гипотезе } H_1: p_1 < p_2; \\ |Z| > z_{1-\alpha/2} & \text{при альтернативной гипотезе } H_1: p_1 \neq p_2. \end{cases}$$

Пример 6.2.14. Существует мнение, что женщины лучше возвращают кредиты, чем мужчины. Среди клиентов банка, у которых закончились сроки кредитных договоров, было 4000 женщин, из которых 3555 не допускали нарушений условий договоров, и 10000 мужчин, из которых договора не нарушали 7896 человек.

Таким образом, относительная частота отстутвий нарушений кредитных договоров составляет для женщий $\hat{p}_1 = 3555/4000 = 0{,}88875$, а для мужчин $\hat{p}_2 = 7896/10\,000 = 0{,}78960$. Определим, не является ли 10%-ная разница ошибкой выборки. Для этого проверим на 5%-ном уровне значимости гипотезу H_0: $p_1 = p_2$ при альтернативной гипотезе $H_1: p_1 \neq p_2$.

Модуль наблюдаемого значения статистики

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

оказался равен

$$\left| \frac{0{,}88875 - 0{,}78960}{\sqrt{\frac{3555 + 7896}{4000 + 10\,000}\left(1 - \frac{3555 + 7896}{4000 + 10\,000}\right)\left(\frac{1}{4000} + \frac{1}{10\,000}\right)}} \right| = 13{,}73346443,$$

что больше критической точки $z_{1-\alpha/2} = z_{0{,}975} = 1{,}959963985$. Поэтому на 5%-ном уровне значимости есть основания отвергнуть нулевую гипотезу: следует сделать вывод о том, что разница значима, она не является ошибкой выборки. ❑

Задачи

403. В прошлом году в университет поступило 3000 из 8000 юношей, подававших документы, и 1700 из 5000 девушек. Проверьте на 5%-ном уровне значимости, различаются ли у юношей и девушек вероятности поступления.

404. Из 200 кредитов, выданных в первом отделении банка, не были возвращены 40, а из 300 кредитов, выданных во втором отделении, не были возвращены 80. Можно ли на этом основании заключить, что вероятность выдачи кредита ненадежному клиенту во втором отделении выше, чем в первом? Проверьте соответствующую гипотезу на 1%-ном уровне значимости.

§ 6.3. Проверка непараметрических гипотез

6.3.1. Критерии согласия

Разобьем множество возможных значений случайной величины X на k разрядов (для непрерывной случайной величины роль разрядов играют интервалы значений, а для дискретной — отдельные возможные значения или их группы).

Выдвинем нулевую гипотезу $H_0: F_X(x) = F^{\text{теор.}}(x)$ (состоящую в том, что генеральная совокупность имеет функцию распределения $F^{\text{теор.}}(x)$) при альтернативной гипотезе $H_1: F_X(x) \neq F^{\text{теор.}}(x)$.

Одним из критериев согласия выборочного и теоретического распределений является *критерий* χ^2 (*критерий Пирсона*), который основывается на том, что статистика

$$\chi^2_{k-l-1} = \sum_{i=1}^{k} \frac{(n_i - np_i^{\text{теор.}})^2}{np_i^{\text{теор.}}} \tag{6.3.1}$$

(где k — число разрядов, l — число оцениваемых параметров, n_i — число попаданий элементов выборки в i-й разряд, n — общее число элементов выборки, $p_i^{\text{теор.}}$ — теоретическая вероятность попадания случайной величины X в i-й разряд при условии истинности нулевой гипотезы) не зависит от выдвинутой гипотезы и распределена по закону «Хи-квадрат» с $k-l-1$ степенью свободы.

Если выбрать уровень значимости α, то

$$\alpha = \mathbf{P}\left\{\chi^2_{k-l-1} > \chi^2_{k-l-1;1-\alpha}\right\}, \tag{6.3.2}$$

и критическая область определяется неравенством

$$\chi^2_{k-l-1} > \chi^2_{k-l-1;1-\alpha}. \tag{6.3.3}$$

Замечание. Критерий Пирсона можно использовать только в том случае, когда во всех разрядах $np_i^{\text{теор.}} \geqslant 5$, поэтому разряды, для которых это условие не выполняется, необходимо объединять (вообще говоря, не обязательно с соседними).

Пример 6.3.1. В примерах 5.1.15, 5.2.2 мы исследовали данные об объеме ежедневных продаж товара (в тыс. ден. ед.) за последние 100 дней (табл. 5.1.4).

Были вычислены выборочное среднее СРЗНАЧ(A2:A101) = 49,596 — несмещенная, состоятельная и эффективная оценка математического ожидания объе-

ма продаж, а также стандартное отклонение $s_X = $ СТАНДОТКЛОН(A2:A101) = = 10,8541657 — оценка генерального среднего квадратического отклонения.

На рис. 6.3.2, *а* был построен полигон частот, а на рис. 6.3.2, *б* — кумулята. Их внешний вид напоминает соответственно плотность и функцию нормального закона распределения.

Значения теоретической функции нормального закона распределения, в которой в качестве параметров a и σ были подставлены оценки $a = \bar{x} = 49{,}596$ и $\sigma = s_X = 10{,}85417$, представлены в табл. 6.3.1. Ее график построен на том же рис. 6.3.2, *б*, что и кумулята. Однако внешнее сходство не дает оснований для того, чтобы сделать вывод о том, насколько выборочные данные согласуются с предположением о нормальном законе распределения объема продаж.

Для того, чтобы сделать такой вывод, проверим на 5%-ном уровне значимости гипотезу $H_0: F_X(x) = F^{\text{теор.}}(x)$ при альтернативной гипотезе $H_1: F_X(x) \neq F^{\text{теор.}}(x)$. Воспользуемся для этого критерием «Хи-квадрат».

Перенесем из табл. 5.2.1 в табл. 6.3.1 интервалы $[a_i; a_{i+1})$, число n_i элементов выборки, попавших в соответствующий интервал, значения функции распределения в правых границах интервалов $F_{\mathcal{N}}(a_{i+1})$.

Первый интервал будем считать начинающимся в точке $-\infty$, а последний интервал — заканчивающимся в точке $+\infty$, для каждого из интервалов, начиная со второго, значение функции распределения в левой границе, очевидно, совпадает со значением функции распределения в правой границе предыдущего интервала.

Поскольку критерием «Хи-квадрат» можно пользоваться только в том случае, когда все $np_i^{\text{теор.}} \geqslant 5$, объединим первый и два последних разряда, для которых это условие не выполняется[1] (в итоговом разряде $np_i^{\text{теор.}}$ будет равно 10,8095). Новое число интервалов обозначим $k' = 6$. Дальнейшие вычисления сведены в табл. 6.3.1.

Окончательное значение $\chi^2_{k'-l-1}$ равно сумме величин в последнем столбце, т. е. 1,3495980.

Таблица 6.3.1

Расчет $\chi^2_{k'-l-1}$ в примере 6.3.1

$[a_i; a_{i+1})$	n_i	$F_{\mathcal{N}}(a_i)$	$F_{\mathcal{N}}(a_{i+1})$	$p_i^{\text{теор}} = F_{\mathcal{N}}(a_{i+1}) - F_{\mathcal{N}}(a_i)$	$np_i^{\text{теор}}$	n_i после объединения	$np_i^{\text{теор}}$ после объединения	$\frac{(n_i - np_i^{\text{теор}})^2}{np_i^{\text{теор}}}$
[24,2; 31,2)	3	0,000000	0,045054	0,045054	4,5054	3 + 7 + + 1 = 11	4,5054 + 4,8213 + + 1,4828 = 10,8095	0,0033573
[31,2; 38,2)	12	0,045054	0,146878	0,101824	10,1824	12	10,1824	0,3244490
[38,2; 45,2)	20	0,146878	0,342737	0,195859	19,5859	20	19,5859	0,0087552
[45,2; 52,2)	26	0,342737	0,594799	0,252062	25,2062	26	25,2062	0,0249986
[52,2; 59,2)	22	0,594799	0,811874	0,217075	21,7075	22	21,7075	0,0039413
[59,2; 66,2)	9	0,811874	0,936959	0,125085	12,5085	9	12,5085	0,9840966
[66,2; 73,2)	7	0,936959	0,985172	0,048213	4,8213			
[73,2; 80,2)	1	0,985172	1,000000	0,014828	1,4828			
$\chi^2_{k'-l-1}$								1,3495980

[1] В табл. 6.3.1 эти разряды выделены желтым.

Теперь найдем критическую точку $\chi^2_{k'-l-1;1-\alpha}$. Для этого, прежде всего, рассчитаем число степеней свободы: при расчете $\chi^2_{k'-l-1}$ использовалось $k'=6$ интервалов, у предполагаемого нормального распределения были неизвестны $l=2$ параметра, поэтому $k'-l-1=6-2-1=3$. С помощью Microsoft Excel находим

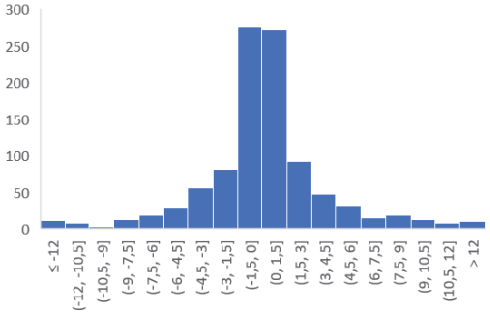
$$\chi^2_{k'-l-1;1-\alpha} = \chi^2_{6-2-1;1-0,05} = \chi^2_{3;0,95} = \text{ХИ2.ОБР(0,95; 3)} = 7{,}814727903.$$

Поскольку наблюдаемое значение $\chi^2_{k'-l-1}$ меньше критической точки $\chi^2_{k'-l-1;1-\alpha}$, нет оснований отвергнуть гипотезу о нормальном законе распределения объема продаж. ❑

Пример 6.3.2. Сгенерируем две независимые выборки по 1000 элементов: $Y = \text{Unif}(-1; 1)$, $Z = \text{Exp}(0{,}2)$. На их основе построим выборку $X = YZ$. Результаты вычислений в Microsoft Excel представлены на рис. 6.3.1, *а*.

	A	B	C	D	E	F	G	H	I	J
1	Unif(0;1)-№1		Y=Unif(-1;1)		Unif(0;1)-№2		Z = Exp(0,2)		X = YZ	
2	0,3132895	=СЛЧИС()	-0,37342	=2*A2-1	0,574533	=СЛЧИС()	2,77099	=-LN(E2)/0,2	-1,03475	=C2*G2
3	0,3349871	=СЛЧИС()	-0,33003	=2*A3-1	0,558411	=СЛЧИС()	2,913302	=-LN(E3)/0,2	-0,96146	=C3*G3
4	0,127391	=СЛЧИС()	-0,74522	=2*A4-1	0,843288	=СЛЧИС()	0,852232	=-LN(E4)/0,2	-0,63510	=C4*G4
5	...	...	...	...	...	...	...	...	...	...

а)



б)

	...	K	L	M	N	O	P	Q	R	S	T	U
1	...	X_среднее	0,115925		Лев. гран.	Прав. гран.	*Частота*	*Интегральный %*	F(a_j+1)	*p_j*	n*p_j	(n_j-np_j)^2/np_j
2	...	s_X	3,956181		-∞	-10,5	13	1,30%	0,010694	0,010694	10,69366	0,497418
3	...	n	1000		-10,5	-9	9	2,20%	0,024529	0,013836	13,83552	1,690017
4	...	x(max)	23,17509		-9	-7,5	10	3,20%	0,051015	0,026486	26,48571	10,26133
5	...	x(min)	-20,9916		-7,5	-6	15	4,70%	0,096441	0,045426	45,42627	20,37935
6	...	Delta	1,5		-6	-4,5	22	6,90%	0,166246	0,069805	69,80464	32,73828
7	...				-4,5	-3	62	13,10%	0,262351	0,096105	96,10488	12,10285
8	...	Хи-квадрат набл.	475,5003		-3	-1,5	103	23,40%	0,380898	0,118547	118,5474	2,039032
9	...				-1,5	0	297	53,10%	0,511914	0,131016	131,0163	210,2838
10	...	Число ст. свободы	13		0	1,5	248	77,90%	0,641646	0,129731	129,7315	107,8184
11	...	Уровень знач.	0,05		1,5	3	81	86,00%	0,75674	0,115094	115,094	10,09956
12	...	Хи-квадрат критич.	22,36203		3	4,5	58	91,80%	0,848224	0,091484	91,48418	12,25557
13	...				4,5	6	31	94,90%	0,913376	0,065152	65,15155	17,90177
14	...				6	7,5	17	96,60%	0,954946	0,041571	41,57068	14,52269
15	...				7,5	9	10	97,60%	0,978711	0,023765	23,76462	7,972554
16	...				9	10,5	5	98,10%	0,990883	0,012172	12,17176	4,225693
17	...				10,5	+∞	19	100,00%	1	0,009117	9,117412	10,71198

в)

Рис. 6.3.1. *Проверка гипотезы о согласии сгенерированной выборки с нормальным законом распределения в примере 6.3.2*

Построим гистограмму распределения признака X (рис. 6.3.1, *б*). В целом она напоминает гистограмму нормального закона распределения. Проверим на 5%-ном уровне значимости гипотезу о том, что случайная величина X распределена по нормальному закону. Для этого вначале построим интервальный вариационный ряд, затем с помощью программы **«Гистограмма»** надстройки **«Анализ данных»** Microsoft Excel найдем выборочные интервальные частоты, далее с помощью функции НОРМ.РАСП вычислим значения функции нормального закона распределениями (с параметрами, в качестве которых подставим оценки $\bar{x}$ и s_X) в правых концах интервалов. Теперь можно вычислить теоретические частоты попадания в интервалы как $np_j = n[F(a_{j+1}) - F(a_j)]$, наблюдаемое значение статистики $\chi^2_{k'-l-1}$, равное 475,5003, и критическую точку $\chi^2_{k'-l-1;1-\alpha} = 22{,}36203249$.

Наблюдаемое значение статистики «Хи-квадрат» значительно больше критической точки, что дает веские основания отвергнуть гипотезу о нормальном распределении признака X.

Подробные вычисления представлены в файле **«Гипотезы по сгенерированным выборкам.xlsx»**. ❑

Задачи

405. Проверьте на 5%-ном уровне значимости гипотезу о нормальном законе распределения роста студентов по данным примеров 1.1.2, 1.2.9, 5.1.14.

406. Для каждого из 100 сотрудников компании регистрировалось число обращений в службу ИТ-поддержки в течение года:

Число обращений в службу ИТ-поддержки	0	1	2	3	4 и более
Число компьютеров	50	25	18	7	0

Проверьте на 5%-ном уровне значимости гипотезу о том, что число обращений в службу ИТ-поддержки одним пользователем за год распределено по закону Пуассона.

407. Дано распределение числа экзаменов, успешно сданных 50 студентами в сессию (всего было 4 экзамена):

Число успешно сданных экзаменов	0	1	2	3	4
Число студентов	1	1	3	15	30

Проверьте на уровне значимости $\alpha = 0{,}1$ гипотезу о том, что число сданных экзаменов (из четырех) имеет биномиальный закон распределения.

408. По данным об акциях из задачи 5 проверьте на 5%-ном уровне значимости гипотезы о нормальном законе распределения доходностей.

409. Сгенерируйте выборку $X = \text{Unif}(-2; 2)$ объемом $n = 1000$ элементов, и на ее основе создайте выборку $Y = X^2$. Проверьте на 5%-ном уровне занчимости гипотезу о согласии распределения генеральной совокупности Y с законом распределения Unif(0; 4).

410. Сгенерируйте две независимые выборки $X^{(1)} = \text{Exp}(1)$, $X^{(2)} = \text{Exp}(1)$ одинакового объема $n = 1000$ элементов, а затем сформируйте выборку $X = X^{(1)} + X^{(2)}$. Проверьте на 5%-ном уровне значимости гипотезу о согласии распределения генеральной совокупности X с законом распредедения Exp(0,5).

6.3.2. Критерии независимости качественных признаков

Распределение «Хи-квадрат» проверяется для проверки самых разных типов гипотез.

Мы уже использовали критерии, основанные на этом распределении, для проверки гипотез о числовом значении дисперсии случайной величины, распределенной по нормальному закону, а также для проверки гипотез о согласии выборочного и теоретического законов распределений.

Теперь рассмотрим критерий «Хи-квадрат» для проверки гипотез о независимости двух качественных признаков.

Для проверки по таблице сопряженности (см. **п. 1.3.2**)

Признак 1 \ Признак 2	B_1	B_2	$\cdots$	B_l	Итого
A_1	m_{11}	m_{12}	$\cdots$	m_{1l}	$n_{1*}=\sum\limits_{j=1}^{l} n_{1j}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2l}	$n_{2*}=\sum\limits_{j=1}^{l} n_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k*}=\sum\limits_{j=1}^{l} n_{kj}$
Всего	$n_{*1}=\sum\limits_{i=1}^{k} n_{i1}$	$n_{*2}=\sum\limits_{i=1}^{k} n_{i2}$	$\cdots$	$n_{*l}=\sum\limits_{i=1}^{k} n_{il}$	$n_{**}=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} n_{ij}=\sum\limits_{i=1}^{k} n_{i*}=\sum\limits_{j=1}^{l} n_{*j}$

гипотезы H_0 о независимости двух качественных признаков рассчитывается статистика

$$\chi^2_{(k-1)(l-1)}=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(n_{ij}-n^*_{ij}\right)^2}{n^*_{ij}}=n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}}{n_{i*}n_{*j}}-1\right),$$

где n_{ij} — выборочная частота комбинации значений A_i и B_j признаков A и B, $n^*_{ij}=\frac{n_{i*}n_{*j}}{n_{**}}$ — теоретическая частота комбинации значений A_i и B_j.

В предположении справедливости гипотезы H_0 данная статистика распределена по закону «Хи-квадрат» с $(k-1)(l-1)$ степенями свободы. Поэтому при проверке гипотезы H_0 об отсутствии связи между признаками наблюдаемое числовое значение $\chi^2_{(k-1)(l-1)}$ этой статистики сравнивается с критической точкой $\chi^2_{(k-1)(l-1);1-\alpha}$, при $\chi^2_{(k-1)(l-1)}>\chi^2_{(k-1)(l-1);1-\alpha}$ гипотеза H_0 отвергается.

Можно использовать для проверки той же гипотезы и *критерий отношения правдоподобия*, статистика которого, определяемая как

$$^{\text{мп}}\chi^2_{(k-1)(l-1)}=2\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}\ln\frac{n_{ij}}{n^*_{ij}}=$$

$$=2\left(\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}\ln n_{ij}-\sum_{i=1}^{r}n_{i*}\ln n_{i*}-\sum_{j=1}^{s}n_{*j}\ln n_{*j}+n_{**}\ln n_{**}\right),$$

в предположении справедливости гипотезы H_0 также распределена по закону «Хи-квадрат» с $(k-1)(l-1)$ степенями свободы.

Замечание 1. Критерии χ^2 и ${}^{\text{мп}}\chi^2$ можно использовать только в том случае, когда во всех клетках таблицы сопряженности $np_i^{\text{теор.}} \geqslant 5$, поэтому клетки, для которых это условие не выполняется, необходимо объединять (вообще говоря, не обязательно с соседними).

Замечание 1. Значения статистик χ^2 и ${}^{\text{мп}}\chi^2$, как правило, близки друг к другу. Тем не менее, для большей уверенности целесообразно пользоваться обоими критериями.

Пример 6.3.3. В результате проведенного опроса выпускников факультета финансов о их планах относительно дальнейшей работы и фактической реализации этих планов получена таблица сопряженности (табл. 6.3.2).

Таблица 6.3.2

Планы выпускников относительно дальнейшей работы и их фактическая реализация

Планы опрошенных (X)	Фактическое распределение (Y)			
	1. Работают в банке	2. Работают в экономическом подразделении нефинансовой организации	3. Работают в финтех-стартапе	Итого
1. Собирались работать в банке	120	40	5	165
2. Собирались работать в экономическом подразделении нефинансовой организации	10	35	10	55
3. Собирались работать в финтех-стартапе	15	25	40	80
Итого	145	100	55	300

Проверим на уровне значимости $\alpha = 0{,}05$ гипотезу о независимости признаков X и Y.

Рассчитаем теоретические частоты n_{ij}^* двумерного распределения в предположении справедливости гипотезы H_0 о независимости рассматриваемых случайных величин: $n_{ij}^* = \frac{n_{i*}n_{*j}}{n}$.

Вначале вычислим маргинальные частоты

$$n_{1*} = 165,\ n_{2*} = 55,\ n_{3*} = 80,\ n_{*1} = 145,\ n_{*2} = 100,\ n_{*3} = 55,\ n = n_{**} = 300.$$

Так как число степеней свободы в нашем примере равно $(3-1)(3-1) = 4$, сделаем расчет лишь четырех независимых частот:

$$n_{11}^* = \frac{165 \cdot 145}{300} = 79{,}75,\quad n_{12}^* = \frac{165 \cdot 100}{300} = 55{,}00,$$

$$n_{21}^* = \frac{55 \cdot 145}{300} = 26{,}58,\quad n_{22}^* = \frac{55 \cdot 100}{300} = 18{,}3.$$

Оставшиеся частоты рассчитаем, используя равенства

$$\sum_{i=1}^{r} n_{ij}^* = \sum_{i=1}^{r} n_{ij} = n_{*j},\qquad \sum_{j=1}^{s} n_{ij}^* = \sum_{j=1}^{s} n_{ij} = n_{i*}:$$

$$n_{13}^* = 165 - (79{,}75 + 55{,}00) = 30{,}25,\ n_{23}^* = 55 - (28{,}58 + 18{,}30) = 10{,}10,$$

$$n_{31}^* = 145 - (79{,}75 + 26{,}58) = 38{,}70,\ n_{32}^* = 100 - (18{,}30 + 55{,}00) = 26{,}7,$$

$$n_{33}^* = 80 - (38{,}67 + 26{,}67) = 55 - (30{,}25 + 10{,}08) = 14{,}67.$$

Расчеты удобно оформить в виде табл. 6.3.3, в которой теоретические частоты n_{ij}^{*} размещены в правых верхних углах соответствующих клеток.

Таблица 6.3.3

Вычисление наблюдаемого значения статистики $\chi^2_{(k-1)(l-1)}$ в примере 6.3.3

X \ Y	1	2	3	n_{i*}
1	$\boxed{79,8}$ $n_{11}=120$	$\boxed{55,0}$ $n_{12}=40$	$\boxed{30,3}$ $n_{13}=5$	$n_{1*}=165$
2	$\boxed{26,6}$ $n_{21}=10$	$\boxed{18,3}$ $n_{22}=35$	$\boxed{10,1}$ $n_{23}=10$	$n_{2*}=55$
3	$\boxed{38,7}$ $n_{31}=15$	$\boxed{26,7}$ $n_{32}=25$	$\boxed{14,7}$ $n_{33}=40$	$n_{3*}=80$
n_{*j}	$n_{*1}=145$	$n_{*2}=100$	$n_{*3}=55$	$n=n_{**}=$ $=300$

Вначале воспользуется критерием «Хи- квадрат». Наблюдаемое числовое значение статистики $\chi^2_{(k-1)(l-1)}$ равно

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_{ij}^{*})^2}{n_{ij}^{*}}=\frac{(120-79,8)^2}{79,8}+\frac{(40-55)^2}{55}+\frac{(5-30,3)^2}{30,3}+\frac{(10-26,6)^2}{26,6}+$$

$$+\frac{(35-18,3)^2}{18,3}+\frac{(10-10,1)^2}{10,1}+\frac{(15-38,7)^2}{38,7}+\frac{(25-26,7)^2}{26,7}+\frac{(40-14,7)^2}{14,7}=129,326,$$

критическая точка $\chi^2_{4;\,0,05}=9,488$, поэтому гипотезу H_0 следует отвергнуть, т. е. на 5%-ном уровне значимости можно считать, что связь между планами относительно работы и их фактической реализацией существует.

Теперь воспользуемся критерием отношения правдоподобия. Наблюдаемое значение статистики ${}^{\text{мп}}\chi^2_4$ равно

$$2(120\cdot\ln 120+40\cdot\ln 40+5\cdot\ln 5+10\cdot\ln 10+35\cdot\ln 35+10\cdot\ln 10+$$
$$+15\cdot\ln 15+25\cdot\ln 25+40\cdot\ln 40-165\cdot\ln 165-55\cdot\ln 55-80\cdot\ln 80-$$
$$-145\cdot\ln 145-100\cdot\ln 100-55\cdot\ln 55+300\cdot\ln 300)=128,759,$$

и поскольку оно больше критической точки $\chi^2_{4;\,0,05}=9,488$, гипотезу H_0 следует отвергнуть на 5%-ном уровне значимости.

Значения статистик двух критериев оказались близки между собой, оба критерия рекомендовали отвергнуть гипотезу H_0 ❑

Пример 6.3.4. В 1955—1959 годах была осуществлена Русская антропологическая экспедиция под руководством В. В. Бунака, в ходе которой было исследовано 17 тысяч представителей русского населения РСФСР.

При этом по специальной шкале определялся и фиксировался цвет глаз людей. Были получены результаты, представленные в табл.6.3.4.

Таблица 6.3.4

Цвет глаз и пол по результатам Русской антропологической экспедиции[2]

Категория	Светлый тип	Переходный тип	Темный тип	Выборка
Мужчины	3917	4347	490	8754
Женщины	3397	4095	582	8074
В сумме	7314	8442	1072	16 828

Проверим гипотезу H_0 об отсутствии связи между цветом глаз и полом. Результаты вычислений в Microsoft Excel представлены на рис. 6.3.2.

	A	B	C	D	E
1	**Исходные данные**				
2	**Категория**	**Светлый тип**	**Переходный тип**	**Темный тип**	**Выборка**
3	**Мужчины**	3917	4347	490	**8754**
4	**Женщины**	3397	4095	582	**8074**
5	**В сумме**	**7314**	**8442**	**1072**	**16828**
6					
7	**Ожидаемые частоты**				
8	**Категория**	**Светлый тип**	**Переходный тип**	**Темный тип**	
9	**Мужчины**	3804,775137	4391,565724	557,6591395	
10	**Женщины**	3509,224863	4050,434276	514,3408605	
11					
12	**Ожидаемые частоты (формулы)**				
13	**Категория**	**Светлый тип**	**Переходный тип**	**Темный тип**	
14	**Мужчины**	=$E3*B$5/$E$5	=$E3*C$5/$E$5	=$E3*D$5/$E$5	
15	**Женщины**	=$E4*B$5/$E$5	=$E4*C$5/$E$5	=$E4*D$5/$E$5	
16					
17	**Расчет наблюдаемого значения Хи-квадрат**				
18	**Категория**	**Светлый тип**	**Переходный тип**	**Темный тип**	
19	**Мужчины**	3,310161441	0,45225413	8,208883953	
20	**Женщины**	3,588946402	0,490343406	8,900244009	
21					
22	**Расчет наблюдаемого значения Хи-квадрат (формулы)**				
23	**Категория**	**Светлый тип**	**Переходный тип**	**Темный тип**	
24	**Мужчины**	=(B3-B9)^2/B9	=(C3-C9)^2/C9	=(D3-D9)^2/D9	
25	**Женщины**	=(B4-B10)^2/B10	=(C4-C10)^2/C10	=(D4-D10)^2/D10	
26					
27	**Результаты**		**Формулы**		
28	**Хи-квадрат_наблюдаемое**	24,950833340848	=СУММ(B19:D20)		
29	**Число степеней свободы**	2	=(3-1)*(2-1)		
30	**Уровень значимости**	0,05			
31	**Хи-квадрат_критическое**	5,991464547108	=ХИ2.ОБР(1-B30;B29)		
32	**p-значение (первый способ)**	0,000003819402	=1-ХИ2.РАСП(B28;B29;ИСТИНА)		
33	**p-значение (второй способ)**	0,000003819402	=ХИ2.РАСП.ПХ(B28;B29)		
34					
35	**Уровень значимости (альфа)**	0,000003819402			
36	**Хи-квадрат_критическое**	24,950833340834	=ХИ2.ОБР(1-B35;B29)		

Рис. 6.3.2. *Проверка гипотезы об отсутствии связи между цветом глаз и полом в примере 6.3.3*

Наблюдаемое значение статистики $\chi^2_{(k-1)(l-1)} = \chi^2_{(3-1)(2-1)} = 24{,}950833340848$, соответствующая 5%-ному уровню значимости критическая точка $\chi^2_{(3-1)(2-1);1-0,05} = 5{,}991464547108$, откуда следует, что гипотеза H_0 об отсутствии связи между цветом глаз и полом отвергается на 5%-ном уровне значимости.

Наблюдаемый уровень значимости (p-значение) можно рассчитать с помощью формулы «=1-ХИ2.РАСП(B28;B29;ИСТИНА)» или с помощью эквивалент-

[2] Источник: *Бунак В. В.* Происхождение и этническая история русского народа по антропологическим данным / Труды института этнографии имени Н. Н. Миклухо-Маклая Академии наук СССР. — М.: Наука, 1965. — 416 с.

ной формулы «=ХИ2.РАСП.ПХ(B28;B29)» — обе они дают результат 0,000003819402. Для проверки найдем критическую точку $\chi^2_{(3-1)(2-1);1-0,000003819402} = 24,950833340834$ — она, действительно, совпадает с наблюдаемым значением статистики «Хи-квадрат».

Результаты сохранены в файле «Цвет глаз и пол.xlsx» ❑

Пример 6.3.5. В файле «Опрос о кредитных картах.xlsx» представлены результаты опроса 100 клиентов банка о том, какой фактор является для них наиболее важным при выборе кредитной карты (рис. 6.3.3, *а*).

Чтобы по этим данным построить таблицу сопряженности, воспользуемся инструментарием сводных таблиц Microsoft Excel (см. **п. 1.3.2**). Построим сводную таблицу, нажав кнопку «Сводная таблица» на вкладке «Вставка» и выбрав далее признак «Пол» в качестве строк и признак «Предпочитаемый фактор» в качестве столбцов. В качестве значений выберем «Количество по полю Предпочитаемый фактор».

Дальнейшие вычисления наблюдаемого значения статистики «Хи-квадрат» и соответствующей критической точки проиллюстрированы рис. 6.3.3, *б*.

	A	B
1	**Пол**	**Предпочитаемый фактор**
2	женщина	низкая стоимость годового обслуживания
3	мужчина	сниженная процентная ставка
4	мужчина	сниженная процентная ставка
5	мужчина	начисление процентов при положительном остатке
6	женщина	120-дневная отсрочка взимания процентов

а)

	...	D	E	F	G	H	I
1	...	Количество по полю Предпочитаемый фактор	Названия столбцов				
2	...	**Названия строк**	**120-дневная отсрочка взимания процентов**	**начисление процентов при положительном остатке**	**низкая стоимость годового обслуживания**	**сниженная процентная ставка**	**Общий итог**
3	...	женщина	15	10	10	5	40
4	...	мужчина	10	25	5	20	60
5	...	**Общий итог**	**25**	**35**	**15**	**25**	**100**
6	...						
7	...	Теоретические частоты					
8	...	**Названия строк**	**120-дневная отсрочка взимания процентов**	**начисление процентов при положительном остатке**	**низкая стоимость годового обслуживания**	**сниженная процентная ставка**	**Общий итог**
9	...	женщина	10	14	6	10	40
10	...	мужчина	15	21	9	15	60
11	...	**Общий итог**	**25**	**35**	**15**	**25**	**100**
12	...						
13	...	(n_ij - n*_ij)^2/n*_ij					
14	...	**Названия строк**	**120-дневная отсрочка взимания процентов**	**начисление процентов при положительном остатке**	**низкая стоимость годового обслуживания**	**сниженная процентная ставка**	**Общий итог**
15	...	женщина	2,5	1,142857143	2,666666667	2,5	40
16	...	мужчина	1,666666667	0,761904762	1,777777778	1,666666667	60
17	...	**Общий итог**	**25**	**35**	**15**	**25**	**100**
18	...						
19	...	Хи-квадрат наблюдаемое	14,68253968				
20	...	Число степеней свободы	3				
21		Уровень значимости	0,05				
22		Хи-квадрат критическое	7,814727903				

б)

Рис. 6.3.3. *Проверка гипотезы об отсутствии связи между полом и наиболее важным фактором при выборе кредитной карты в примере 6.3.3*

Наблюдаемое значение статистики $\chi^2_{(k-1)(l-1)} = \chi^2_{(4-1)(2-1)} = 14{,}68253968$, соответствующая 5%-ному уровню значимости критическая точка $\chi^2_{(4-1)(2-1);1-0{,}05} = 7{,}814727903$, откуда следует, что гипотеза H_0 об отсутствии связи между полом и предпочитаемым фактором при выборе кредитной карты отвергается на 5%-ном уровне значимости.

Все вычисления сохранены в файле «Опрос о кредитных картах.xlsx». ❑

Задачи

411. Найдите наблюдаемый уровень значимости (p-значение) обоих критериев в примере 6.3.2.

412. В файле «Отношение к трейлеру.xlsx» представлены результаты опроса 200 кинозрителей об их отношении к рекламному трейлеру нового фильма. Проверьте на 5%-ном уровне значимости гипотезу об отсутствии связи между полом и отношениек к трейлеру. Рассчитайте наблюдаемый уровень значимости.

413. В файле «Приобретение автомобилей.xlsx» содержатся данные о 344 семьях, собирающихся приобрести автомобили, и для каждой семьи указано, предпочитает ли она внедорожник, сколько в ней детей (больше двух или не больше двух), какие у семьи доходы (высокие или низкие). Проверьте на 5%-ном уровне значимости гипотезы: *а*) об отсутствии связи между количеством детей и типом предпочитаемого автомобиля; *б*) об отсутствии связи между уровнем доходов и типом предпочитаемого автомобиля.

414. Известно распределение 1 725 школьников по качеству одежды и умственным способностям:

Качество одежды \ Умственные способности	плохие	ниже средних	средние	выше средних	высокие	превосходные	И т о г о
очень хорошее	33	48	113	209	194	39	636
хорошее	41	100	202	255	138	15	751
удовлетворительное	39	58	70	61	33	4	265
плохое	17	13	22	10	10	1	73
В с е г о	130	219	407	535	375	59	1725

Выясните, существует ли связь между качеством одежды и умственными способностями.

6.3.3. Критерии сдвига

Для проверки гипотезы $H_0: F_{X^{(1)}}(x) = F_{X^{(2)}}(x)$, утверждающей, что обе выборки взяты из генеральных совокупностей, имеющих одинаковые распределения (при альтернативе $H_1: F_{X^{(1)}}(x) \neq F_{X^{(2)}}(x)$ или $H_1': F_{X^{(1)}}(x) > F_{X^{(2)}}(x)$ или $H_1'': F_{X^{(1)}}(x) < F_{X^{(2)}}(x)\big)$, можно использовать *критерий Манна — Уитни*, который заключается в следующем.

Элементы обеих выборок объединяют и располагают в порядке неубывания (т. е. в вариационный ряд). Каждому элементу объединенной выборки ставится в соответствие его номер в получившемся ряду — *ранг* (в случае *связанных рангов*, когда несколько элементов объединенной выборки совпадают, им присваивается одинаковый ранг, равный среднему арифметическому их порядковых номеров в общем ряду). Статистика

$$U = R_{\min} - \frac{n_{\min}(n_{\min}+1)}{2},$$

где $R_{\min}$ — меньшая из сумм рангов элементов первой выборки и рангов элементов второй выборки в объединенном вариационном ряду, $n_{\min}$ — объем соответствующей выборки, имеет довольно сложный закон распределения, и критические точки этого закона распределения, как правило, существуют в виде таблиц.

Но можно воспользоваться нормальным приближением, поскольку при достаточно больших n_1 и n_2 статистика

$$Z=\frac{U-\frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{12}\left(n_1+n_2+1-\frac{\sum_{k=1}^{m} t_k(t_k^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right)}}$$

в случае справедливости гипотезы H_0 распределена приблизительно по стандартному нормальному закону (здесь m — число групп связанных рангов, принадлежащих разным выборкам, t_k — объем k-й группы).

Гипотеза $H_0: F_{X^{(1)}}(x)=F_{X^{(2)}}(x)$ при альтернативе $H_1: F_{X^{(1)}}(x)\neq F_{X^{(2)}}(x)$ отвергается на уровне значимости α, если наблюдаемое значение

$$Z\notin\left(z_{\alpha/2};z_{1-\alpha/2}\right). \tag{6.3.4}$$

Гипотеза $H_0: F_{X^{(1)}}(x)=F_{X^{(2)}}(x)$ при альтернативе $H_1': F_{X^{(1)}}(x)>F_{X^{(2)}}(x)$ отвергается на уровне значимости α, если

$$Z<z_{\alpha}. \tag{6.3.5}$$

Гипотеза $H_0: F_{X^{(1)}}(x)=F_{X^{(2)}}(x)$ при альтернативе $H_1'': F_{X^{(1)}}(x)<F_{X^{(2)}}(x)$ отвергается на уровне значимости α, если

$$Z>z_{1-\alpha}. \tag{6.3.6}$$

Критерий Манна — Уитни является примером к р и т е р и я с д в и г а. В этом непараметрическом критерии, в отличие от критерия Стьюдента, никакие предположения относительно закона распределения не делаются.

Пример 6.3.6. Имеются данные о количестве страховых полисов КАСКО, проданных за последний месяц сотрудниками страховой компании — восемью женщинами и девятью мужчинами

Женщины	11	12	16	13	18	15	13	14	
Мужчины	7	10	14	15	12	16	14	12	12

На основании визуального анализа этих данных создается впечатление, что женщинам лучше удается продавать страховки, чем мужчинам. Проверим это предположение на 5%-ном уровне значимости.

Будем проверять гипотезу при альтернативе $H_1: F_{X^{(1)}}(x)<F_{X^{(2)}}(x)$ $\left(\text{т. е. } X^{(1)}>X^{(2)}\right)$.

Составим объединенный вариационный ряд, в котором элементы первой выборки выделим жирным:

Элементы	7	10	**11**	**12**	12	12	12	**13**	**13**	**14**	14	14	15	**15**	**16**	16	**18**
Ранги	1	2	3	5,5	5,5	5,5	5,5	8,5	8,5	11	11	11	13,5	13,5	15,5	15,5	17

Считаем последовательно. Сумма рангов элементов первой выборки в объединенном вариационном ряду равна 3 + 5,5 + 8,5 + + 8,5 + 11 + 13,5 + + 15,5 + 17 = 82,5, для второй выборки эта сумма составляет 1 + 2 + 5,5 + 5,5 + 5,5 + 11 + 11 + 13,5 + 15,5 = 70,5; наименьшее из этих чисел $R_{\min}=\min\{82{,}5;\,70{,}5\}=70{,}5$, при этом $n_{\min}=9$ (это число элементов второй выборки, соответствующей наименьшей из ранговых сумм).

Статистика

$$U=R_{\min}-\frac{n_{\min}(n_{\min}+1)}{2}=70{,}5-\frac{9\cdot 10}{2}=25{,}5,$$

Значение статистики Z оказалось равным

$$\frac{25{,}5-36}{\sqrt{6\left(18-\dfrac{4(4^2-1)+2(2^2-1)+3(3^2-1)+2(2^2-1)+2(2^2-1)}{17\cdot16}\right)}}=$$

$$=\frac{-10{,}5}{\sqrt{6\left(18-\dfrac{4\cdot15+2\cdot3+3\cdot8+2\cdot3+2\cdot3}{17\cdot16}\right)}}=\frac{-10{,}5}{\sqrt{6\left(18-\dfrac{102}{17\cdot16}\right)}}=\frac{-10{,}5}{\sqrt{105{,}75}}=-1{,}02105494,$$

и оно меньше односторонней критической точки $z_{1-\alpha}=z_{0{,}95}=1{,}644854$, поэтому нет оснований отвергнуть гипотезу H_0 — различие в объемах продаж, совершаемых женщинами и мужчинами, не существенно. ❑

Замечание. Строго говоря, объемы выборок в примере 6.3.6 нельзя считать большими. Однако в данном примере учебного характера прямое исследование положения наблюдаемого значения статистики U относительно критической области критерия также не дает оснований отвергнуть H_0, показывая p-значение, равное 0,84705883. В общем же случае нормальное приближение работает достаточно хорошщо, начиная с $n_1 \geqslant 20, n_2 \geqslant 20$.

Задачи

415. Решите задачи 400—401 с помощью критерия Манна — Уитни.

416. С помощью критерия Манна — Уитни проверьте на 5%-ном уровне значимости гипотезу об одинаковой с вероятностной точки зрения эффективности двух методов (A и B) повышения выпуска продукции по следующим данным (в первой строке приведены проценты прироста выпуска продукции при использовании метода A, во второй — метода B):

A	0,1	0,1	0,5	0,8	0,9	0,9	1,0	1,1
B	0,4	0,4	0,8	1,8	0,8	1,3		

В качестве альтернативной выберите гипотезу о том, что метод B эффективнее метода A.

417. Сгенерируйте четыре независимые выборки одинакового объема — по 1000 элементов: $Y_1 = \mathrm{Unif}(0;1)$, $Y_2 = \mathrm{Unif}(0;1)$, $Z_1 = \mathcal{N}(0;1), Z_2 = \mathcal{N}(0;1)$. Сформируйте выборки $X^{(1)}$, $X^{(2)}$ по правилу: $X^{(1)} = Y_1^{Z_1}, X^{(2)} = Y_2^{Z_2}$. Проверьте на 5%-ном уровне занчимости гипотезу $H_0: F_{X^{(1)}}(x) = F_{X^{(2)}}(x)$ при альтернативе $H_1: F_{X^{(1)}}(x) \neq F_{X^{(2)}}(x)$.

418. Сгенерируйте выборку $X = \mathrm{Unif}(0;4)$ объемом $n = 1000$ элементов, и на ее основе создайте две выборки по следующему правилу: если $X < 1$, то $X^{(1)} = \mathrm{Exp}(1)$, $X^{(2)} = \mathcal{N}(0;1)$, иначе $X^{(1)} = \mathcal{N}(0;1)$, $X^{(2)} = \mathrm{Exp}(1)$. Проверьте на 5%-ном уровне значимости гипотезу $H_0: F_{X^{(1)}}(x) = F_{X^{(2)}}(x)$ при альтернативе $H_1: F_{X^{(1)}}(x) \neq F_{X^{(2)}}(x)$.

§ 6.4. Дисперсионный анализ

6.4.1. Однофакторный дисперсионный анализ

Дисперсионный анализ[3] предназначен для поиска зависимостей в экспериментальных данных путем исследования значимости различий в средних значениях. Суть метода заключается в выделении из общей вариации (общего разброса) признака Y частей, соответствующих влиянию факторов, поддающихся измерению.

В отличие от t-критерия, рассмотренного в **п. 6.2.5**, дисперсионный анализ позволяет сравнивать средние значения в трех и более группах.

Суть дисперсионного анализа сводится к изучению влияния одного или нескольких независимых признаков, обычно именуемых *факторами*, на результативный признак Y путем разложения общей вариации (дисперсии, разброса значений) изучаемого признака на части, соответствующие раз-

[3] В литературе также встречается обозначение *ANOVA* (от англ. *ANalysis Of VAriance*).

дельному и совместному влиянию факторов, и статистическом изучении этих частей с целью выяснения приемлемости гипотез о существовании этих влияний.

При этом в качестве результативного рассматривается некоторый непрерывный количественный признак, а в качестве факторов выступают номинальные качественные признаки, которые могут иметь две или более градации — два или более уровня.

В зависимости от типа и количества признаков различают:

- однофакторный и многофакторный дисперсионный анализ (один или несколько независимых признаков);
- одномерный и многомерный дисперсионный анализ (один или несколько результативных признаков);
- дисперсионный анализ с повторными измерениями (для зависимых выборок);
- дисперсионный анализ с постоянными факторами, случайными факторами, с факторами обоих типов.

По цели исследования выделяют следующие типы моделей дисперсионного анализа:

- *детерминированный тип* (М1) — здесь уровни всех контролируемых факторов заранее фиксированы и проверяют именно их влияние на случайную величину Y;
- *случайный тип* (М2) — здесь уровни каждого фактора получены как случайная выборка из генеральной совокупности уровней фактора;
- *смешанный тип* (М3) — здесь уровни одних факторов заранее фиксированы, а уровни других представляют собой случайную выборку.

Модель однофакторного дисперсионного анализа имеет следующий вид:

$$Y_k^{(i)} = \theta^{(0)} + \theta^{(i)} + \varepsilon_k^{(i)}; \quad i = 1,2,\ldots,\nu; \quad k = 1,2,\ldots,n_i, \tag{6.4.1}$$

где $Y_k^{(i)}$ — значение случайной величины Y, принимаемое при уровне $A^{(i)}$ фактора A в k-м наблюдении, $\theta^{(0)} = \mathbf{E}(Y)$, $\theta^{(i)}$ — эффект влияния уровня $A^{(i)}$ фактора A на наблюдение $Y_k^{(i)}$, $\varepsilon_k^{(i)}$ — случайный эффект влияния прочих неконтролируемых факторов.

К этой модели предъявляются следующие *требования:*

- все $n_1 + n_2 + \cdots + n_\nu$ случайных величин $\varepsilon_k^{(i)}$ должны быть независимыми;
- все $n_1 + n_2 + \cdots + n_\nu$ случайных величин $\varepsilon_k^{(i)}$ должны иметь одинаковое нормальное распределение $\varepsilon_k^{(i)} = \mathcal{N}\left(0; \sigma_{\text{ост.}}\right)$.

При этом в модели М1 все $\theta^{(i)}$ — детерминированные величины и $\sum_{i=1}^{\nu} n_i \theta^{(i)} = 0$, а в модели М2 $\theta^{(i)}$ — случайные величины $\theta^{(i)} = \mathcal{N}\left(0; \sigma_\theta\right)$, причем все $\theta^{(i)}$ независимы между собой и со всеми $\varepsilon_k^{(i)}$.

По наблюдениям вычисляются *общая вариация*

$$SS_{\text{общ.}} = \sum_{i=1}^{\nu} \sum_{k=1}^{n_i} (Y_k^{(i)} - \overline{Y})^2 \tag{6.4.2}$$

результативного признака Y и две ее составляющие — *межгрупповая вариация*

$$SS_A = \sum_{i=1}^{\nu} n_i (\overline{Y}^{(i)} - \overline{Y})^2 \tag{6.4.3}$$

и *внутригрупповая вариация*

$$SS_{\text{ост.}} = \sum_{i=1}^{\nu} \sum_{k=1}^{n_i} (Y_k^{(i)} - \overline{Y}^{(i)})^2. \tag{6.4.4}$$

Здесь $\overline{Y}=\dfrac{\sum\limits_{i=1}^{\nu} n_i \overline{Y}^{(i)}}{\sum\limits_{i=1}^{\nu} n_i}$ — общее среднее, $\overline{Y}^{(i)}=\dfrac{\sum\limits_{k=1}^{n_i} Y_k^{(i)}}{n_i}$ — групповое среднее.

Справедливо соотношение

$$SS_{\text{общ.}} = SS_A + SS_{\text{ост.}}. \tag{6.4.5}$$

Для выяснения того, влияет ли фактор A на результативный признак:

- в модели М1 проверяют гипотезу $H_0: \theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(\nu)} = 0$, состоящую в том, что эффекты влияния зафиксированных уровней фактора A равны нулю, иначе — фактор A не влияет на изучаемую случайную величину;
- в модели М2 проверяют гипотезу $H_0: \sigma_\theta^2 = 0$, состоящую в том, что (с учетом нулевых математических ожиданий) все эффекты влияния зафиксированных уровней фактора A $\theta^{(i)} = 0$, иначе — фактор A не влияет на изучаемую случайную величину.

Критерии проверки этих и других гипотез, а также оценки параметров модели 6.4.1 приведены в табл. 6.4.1.

Таблица 6.4.1

Модели однофакторного дисперсионного анализа

М1	**М2**
$\theta^{(i)}$ — детерминированные величины $(i=1,2,\ldots,\nu)$, при этом $\sum\limits_{i=1}^{\nu}\theta^{(i)} n_i = 0$	$\theta^{(i)}$ — случайные величины, $\theta^{(i)} = \mathcal{N}(0;\sigma_\theta)$
$\widehat{\mathbf{E}(Y)} = \overline{Y}$	
$s^2_{\text{ост.}} = \dfrac{SS^2_{\text{ост.}}}{n-\nu}$ — несмещенная оценка дисперсии $\sigma^2_{\text{ост.}}$ (6.4.6)	
$\hat{\theta}^{(i)} = \overline{Y^{(i)}} - \overline{Y}$ $(i=1,2,\ldots,\nu)$	$s^2_\theta = \left(\dfrac{SS^2_A}{\nu-1} - SS^2_{\text{ост.}}\right)\dfrac{n(\nu-1)}{n^2 - \sum\limits_{i=1}^{\nu} n_i^2}$ — несмещенная оценка дисперсии σ^2_θ
$H_0: \theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(\nu)} = 0$ принимают, если	$H_0: \sigma^2_\theta = 0$ принимают, если
$\dfrac{SS^2_A/(\nu-1)}{SS^2_{\text{ост}}/(n-\nu)} < f_{\nu-1;n-\nu;\alpha}$ (6.4.7)	
Если условие (6.4.7) не выполняется, то гипотезу H_0 отвергают и вычисляют *коэффициент детерминации*	
$\hat{\eta}^2_{Y\|A} = \dfrac{SS_A}{SS_{\text{общ.}}}$ — (6.4.8)	
долю вариации наблюдаемого признака, объясняемую влиянием фактора A	
$H_0^{(ij)}: \theta^{(i)} = \theta^{(j)}$ при $H_1^{(ij)}: \theta^{(i)} \neq \theta^{(j)}$ принимают, если	
$T^{(ij)}_{n-\nu} = \dfrac{\overline{Y}^{(i)} - \overline{Y}^{(j)}}{s_{\text{ост.}}}\sqrt{\dfrac{n_i n_j}{n_i + n_j}} < t_{n-\nu;\alpha}$ (6.4.9)	
$H_0^{(0)}: \theta^{(0)} = a_0$ при $H_1^{(0)}: \theta^{(0)} \neq a_0$ принимают, если	
$\dfrac{\lvert\overline{Y} - a_0\rvert}{s_{\text{ост.}}}\sqrt{n} < t_{n-\nu;\alpha}$	$\dfrac{\overline{Y} - a_0}{\sqrt{SS^2_A/(\nu-1)}}\sqrt{n} < t_{\nu-1;\alpha}$

Пример 6.4.1. Известны данные о количестве голосов Y, отданных за кандидата в трех однотипных случайно выбранных избирательных округах, в которых использовались разные виды агитации ($A^{(1)}$ — личные встречи кандидата с избирателями, $A^{(2)}$ — раздача листовок с программой кандидата, $A^{(3)}$ — выступление кандидата по телевидению):

Номер избирательного участка	$A^{(1)}$	$A^{(2)}$	$A^{(3)}$
1	280	300	150
2	250	250	240
3	200	210	170
4	290	310	200
5		270	150
6		300	

Детерминированная модель однофакторного дисперсионного анализа количества голосов, поданных за кандидата в избирательном округе, имеет следующий вид:

$$Y_k^{(i)} = \theta^{(0)} + \theta^{(i)} + \varepsilon_k^{(i)}; \quad i = 1, 2, 3; \quad k = 1, 2, \ldots, n_i; \quad n_1 = 4, \quad n_2 = 6, \quad n_3 = 5,$$

где $\theta^{(0)} = \mathbf{E}(Y)$, $\theta^{(i)}$ — неслучайный эффект влияния уровня $A^{(i)}$ фактора A на наблюдение $Y_k^{(i)}$, $\varepsilon_k^{(i)}$ — случайный эффект влияния прочих неконтролируемых факторов.

К этой модели предъявляются следующие требования:

- все случайные величины $\varepsilon_k^{(i)}$ или, иначе, все случайные величины $Y_k^{(i)}$ должны быть независимыми;
- $\varepsilon_k^{(i)} = \mathcal{N}\left(0; \sigma_{\text{ост}}\right)$ или, иначе, $Y_k^{(i)} = \mathcal{N}\left(\theta^{(0)} + \theta^{(i)}; \sigma_{\text{ост}}\right)$, т. е. при каждом уровне фактора наблюдения должны проводиться в одинаковых («нормальных») вероятностных условиях с дисперсией, не изменяющейся при переходе от одного уровня фактора к другому;
- $\sum_{i=1}^{3} n_i \theta^{(i)} = 0$.

Проверим гипотезу $H_0: \theta^{(1)} = \theta^{(2)} = \theta^{(3)} = 0$, состоящую в том, что эффекты влияния зафиксированных уровней фактора A равны нулю — количество голосов, отданных за кандидата при каждом виде агитации, не зависит от вида агитации.

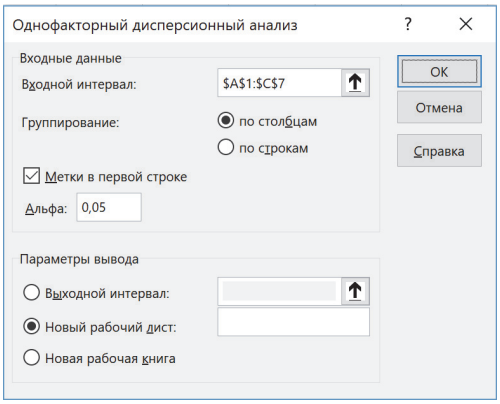
Введем исходные данные в рабочий лист Microsoft Excel (рис. 6.4.1, а).

Для исследования модели воспользуемся программой «Однофакторный дисперсионный анализ». Для этого выберем соответствующий пункт меню надстройки «Анализ данных». В появившемся окне ввода данных (рис. 6.4.1, *б*) укажем входной интервал A1:C7, в который мы ввели исходные данные (с заголовками столбцов — обозначениями уровней фактора, поэтому отметим флажок «Метки в первой строке»). Зададим уровень значимости «Альфа» (по условию $\alpha = 0{,}05$). Укажем, что данные сгруппированы по столбцам, а результаты работы программы необходимо вывести на новый рабочий лист. Результаты работы программы представлены на рис. 6.4.1, *в*.

Выполнение первого требования, предъявляемого к результатам наблюдений, определяется организацией выборов (число голосов, отданных за кандидата на любом избирательном участке при правильной организации выборов, очевидно, не зависит от числа голосов, отданных за того же кандидата на двух оставшихся участках).

	A	B	C
1	A(1)	A(2)	A(3)
2	280	300	150
3	250	250	240
4	200	210	170
5	290	310	200
6		270	150
7		300	

а)



б)

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
A(1)	4	1020	255	1633,333		
A(2)	6	1640	273,3333	1466,667		
A(3)	5	910	182	1470		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-значение*	*F критическое*
Между группами	24326,67	2	12163,33	8,06	0,006	3,89
Внутри групп	18113,33	12	1509,44			
Итого	42440	14				

в)

Рис. 6.4.1. *Исходные данные* (*а*), *окно ввода данных* (*б*) *и результаты работы* (*в*) *программы «Однофакторный дисперсионный анализ»*

Проверим на 5%-ном уровне значимости гипотезу о равенстве групповых генеральных дисперсий количества голосов, предположив нормальность распределения случайных величин $Y_k^{(i)}$ и их независимость. Для этого применим критерий Бартлетта.

Число наблюдений на каждом уровне n_i («**Счет**»), несмещенные оценки групповых средних $\overline{y}^{(i)} = \sum_{k=1}^{n_i} y_k^{(i)} \Big/ n_i$ («**Средние**») и групповых дисперсий $s_i^2 = \sum_{k=1}^{n_i} \left(y_k^{(i)} - \overline{y}_k^{(i)}\right)^2 \Big/ (n_i - 1)$ («**Дисперсии**») рассчитаны программой ($i = 1, 2, 3$).

Для расчета числового значения статистики Бартлетта

$$\Phi_{\nu-1} = q\left((n_1-1)\ln\frac{s_{\text{ост}}^2}{s_1^2} + (n_2-1)\ln\frac{s_{\text{ост}}^2}{s_2^2} + (n_3-1)\ln\frac{s_{\text{ост}}^2}{s_3^2}\right),$$

где $\nu = 3$ — число уровней фактора,

$$q = \frac{1}{1 + \frac{1}{3(\nu-1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \frac{1}{n_3 - 1} - \frac{1}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}\right)},$$

$$s^2_{\text{ост}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)},$$

вычислим последовательно:

$$q = \frac{1}{1 + \frac{1}{3(3-1)}\left(\frac{1}{4-1} + \frac{1}{6-1} + \frac{1}{5-1} - \frac{1}{(4-1) + (6-1) + (5-1)}\right)} = 0{,}896;$$

$$s^2_{\text{ост}} = \frac{(4-1)1633{,}33 + (6-1)1466{,}67 + (5-1)1470}{(4-1) + (6-1) + (5-1)} = 1509{,}44,$$

тогда числовое значение статистики Бартлетта $\Phi_{\nu-1}$ будет равно

$$0{,}896\left((4-1)\ln\frac{1509{,}44}{16{,}33{,}33} + (6-1)\ln\frac{1509{,}44}{1466{,}67} + (5-1)\ln\frac{1509{,}44}{1470}\right) = 0{,}012.$$

Статистика Бартлетта $\Phi_{\nu-1}$ в предположении справедливости гипотезы о равенстве групповых генеральных дисперсий имеет распределение, близкое к $\chi^2_{\nu-1}$. При $\alpha = 0{,}05$ критическая точка $\chi^2_{\nu-1;\alpha} = \chi^2_{2;0{,}05} = 5{,}991$. Поскольку наблюдаемое числовое значение статистики Бартлетта оказалось меньше критической точки, нет оснований отвергнуть проверяемую гипотезу. Таким образом, второе требование, предъявляемое к результатам наблюдений, выполнено.

Таблица **«Дисперсионный анализ»**, полученная в результате работы программы (рис. 6.4.1, *в*), представляет собой дисперсионную таблицу. Расшифруем ее отдельные ячейки в табл. 6.4.2.

Таблица 6.4.2

Дисперсионная таблица в примере 6.4.1

Источник вариации результативного признака Y	Показатель вариации (SS)	Число степеней свободы (df)	Оценка дисперсии $\sigma^2_{\text{ост.}}$ (MS)	$F_{\nu-1;\,n-\nu} = \frac{s_A^2}{s^2_{\text{ост}}}$	p-значение	$f_{\nu-1;\,n-\nu;\,\alpha}$
Фактор A	$SS_A = \sum_{i=1}^{\nu} n_i\left(\overline{y}^{(i)} - \overline{y}\right)^2 = 24\,326{,}67$	$\nu - 1 = 2$	$s_A^2 = \frac{SS_A}{\nu-1} = 12\,163{,}33$	8,06	0,006	3,89
Остаточные факторы	$SS_{\text{ост.}} = \sum_{i=1}^{\nu}\sum_{k=1}^{n_i}\left(y_k^{(i)} - \overline{y}^{(i)}\right)^2 = 18\,113{,}33$	$n - \nu = 12$	$s^2_{\text{ост}} = \frac{SS_{\text{ост}}}{n-\nu} = 1509{,}44$			
Общая вариация	$SS_{\text{итог.}} = \sum_{i=1}^{\nu}\sum_{k=1}^{n_i}\left(y_k^{(i)} - \overline{y}\right)^2 = 42\,440{,}00$	$n - 1 = 14$				

В этой таблице

$$\overline{y}^{(i)} = \frac{\sum_{k=1}^{n_i} y_k^{(i)}}{n_i}, \quad \overline{y} = \frac{\sum_{i=1}^{\nu} n_i \overline{y}^{(i)}}{\sum_{i=1}^{\nu} n_i}.$$

Проверка гипотезы H_0: $\theta^{(1)} = \theta^{(2)} = \theta^{(3)} = 0$ производится на основе анализа статистики $F_{\nu-1;\, n-\nu} = s_A^2 / s_{\text{ост}}^2$; , имеющей (в предположении справедливости H_0) распределение Фишера — Снедекора с $\nu - 1 = 2$ и $n - \nu = 12$ степенями свободы. В данном случае наблюдаемое значение этой статистики оказалось равным 8,06, и оно больше критической точки $f_{2;\,12;\,0,05} = 3{,}89$, откуда следует, что гипотеза H_0 отвергается на 5%-ном уровне значимости [значение F статистики $s_A^2 / s_{\text{ост}}^2$ и критическая точка $f_{2;\,12;\,0,05}$ (F критическое) приводятся в результатах работы программы (рис. 6.4.1, *в*, табл. 6.4.2)].

Гипотезу H_0 можно проверить и так: если p-значение (равное рассчитанному уровню значимости гипотезы H_0) оказывается не меньше принятого уровня значимости α (в данном случае $\alpha = 0{,}05$), гипотезу H_0 принимают, а если p-значение оказывается меньше α, гипотезу H_0 отвергают. P-значение приводятся в результатах работы программы (рис. 6.4.1, *в*, табл. 6.4.2). В данном случае p-значение равно $p = 0{,}006 < \alpha = 0{,}05$, значит, гипотезу H_0 следует отвергнуть на 5%-ном уровне значимости. Заметим, что если p-значение оказывается достаточно малым, то в Microsoft Excel оно может быть представлено в экспоненциальном формате, например, **6,04E–03** = $6{,}04 \cdot 10^{-3} = 0{,}00604$.

Таким образом, вид агитации существенно влияет на количество голосов, отданных за кандидата. Оценим силу этого влияния с помощью коэффициента детерминации

$$\hat{\eta}^2(Y \mid A) = \frac{SS_A}{SS_{\text{итог}}} = \frac{24\,326{,}67}{42\,440{,}00} = 0{,}57 \ —$$

такова (57%) доля общей вариации наблюдаемого числа голосов, отданных за кандидата, обусловленная изменчивостью фактора A — вида агитации. Влиянием неконтролируемых факторов обусловлено 43% вариации числа голосов.

Нетрудно убедиться в том, что

$$SS_A = n\hat{\sigma}_Y^2 \hat{\eta}^2(Y \mid A), \ SS_{\text{ост}} = n\hat{\sigma}_Y^2 \left(1 - \hat{\eta}^2(Y \mid A)\right),$$

а статистика

$$\frac{s_A^2}{s_{\text{ост}}^2} = \frac{\hat{\eta}^2(Y \mid A) / (\nu - 1)}{\left(1 - \hat{\eta}^2(Y \mid A)\right) / (n - \nu)}.$$

Попарное сравнение видов агитации по влиянию на количество поданных голосов проведем при помощи проверки гипотез $H_0^{(i;j)}$: $\theta^{(i)} = \theta^{(j)}$ при альтернативах $H_1^{(i;j)}$: $\theta^{(i)} \neq \theta^{(j)}$ [$(i;\, j) = (1;2), (1;3), (2;3)$]. Вычислим значения статистики

$$T_{n-\nu}^{(i;\,j)} = \frac{\overline{Y}^{(i)} - \overline{Y}^{(j)}}{s_{\text{ост}}} \sqrt{\frac{n_i n_j}{n_i + n_j}},$$

имеющей (в предположении справедливости $H_0^{(i;j)}$) распределение Стьюдента с $n - \nu = 12$ степенями свободы (здесь $s_{\text{ост.}} = \sqrt{s_{\text{ост.}}^2} = \sqrt{1509{,}44} = 38{,}85$).

В условиях данного примера оказалось, что наблюдаемое значение статистики $T_{12}^{(1;2)}$ равно

$$\frac{255{,}00-273{,}33}{38{,}85}\sqrt{\frac{4\cdot 6}{4+6}}=-0{,}73,$$

значение статистики $T_{12}^{(1;3)}$ равно

$$\frac{255{,}00-182{,}00}{38{,}85}\sqrt{\frac{4\cdot 5}{4+5}}=2{,}80,$$

а значение статистики $T_{12}^{(2;3)}$ равно

$$\frac{273{,}33-182{,}00}{38{,}85}\sqrt{\frac{6\cdot 5}{6+5}}=3{,}88;$$

критическая точка распределения Стьюдента на 5%-ном уровне значимости $t_{n-\nu;\alpha}=t_{12;0{,}05}=2{,}18$.

Гипотеза $H_0^{(1;2)}\colon \theta^{(1)}=\theta^{(2)}$ не отвергается (поскольку модуль значения статистики $T_{12}^{(1;2)}$, равный $|0{,}73|<t_{12;\,0{,}05}$), т. е. эффекты влияния первого и второго уровней фактора (личные встречи кандидата с избирателями, раздача листовок с программой кандидата) на результативный признак Y (число голосов, поданных за кандидата) одинаковы, а гипотезы $H_0^{(1;3)}\colon \theta^{(1)}=\theta^{(3)}$ и $H_0^{(2;3)}\colon \theta^{(2)}=\theta^{(3)}$ отвергаются (так как $|2{,}80|>t_{12;\,0{,}05}$, $|3{,}88|>t_{12;\,0{,}05}$), т. е. при переходе с первого (личные встречи кандидата с избирателями) и второго уровня (раздача листовок с программой кандидата) на третий (выступление кандидата по телевидению) эффект влияния изменяется.

Оценим параметры детерминированной модели однофакторного дисперсионного анализа. Оценка $\hat{\theta}^{(0)}$ параметра $\theta^{(0)}$ равна общему среднему $\hat{\theta}^{(0)}=\overline{y}=\sum_{i=1}^{3}n_i\overline{y}^{(i)}\Big/(n_1+n_2+n_3)=238{,}00$. Оценка $\hat{\theta}^{(i)}$ эффекта $\theta^{(i)}$ влияния i-го уровня фактора ($i=1, 2, 3$) равна $\left(\overline{y}^{(i)}-\overline{y}\right)$, т. е. $\hat{\theta}^{(1)}=255{,}00-238{,}00=17{,}00$; $\hat{\theta}^{(2)}=273{,}33-238{,}00=35{,}33$; $\hat{\theta}^{(3)}=182{,}00-238{,}00=-56{,}00$. Оценка параметра $\sigma_{\text{ост}}$ равна $s_{\text{ост.}}=\sqrt{s^2_{\text{ост.}}}=\sqrt{1509{,}44}=38{,}85$. ❑

Задачи

419. Исследователь хочет выяснить, отличаются ли по эффективности четыре метода изучения некоторой темы ($A^{(1)}$ — тема изучается по краткому конспекту, $A^{(2)}$ — при ее изучении учащиеся конспектируют учебник, $A^{(3)}$ — работают с программированным учебником, $A^{(4)}$ — изучают материал на обучающей машине). Для этого случайным образом были образованы четыре равные группы испытуемых и после обучения их соответствующими методами им было предложено по 100 вопросов по этой теме. Количества правильных ответов, данные испытуемыми, таковы:

$A^{(1)}$	26	34	46	48	42	49	74	61	51	53
$A^{(2)}$	51	50	33	28	47	50	48	60	71	42
$A^{(3)}$	52	64	39	54	58	53	77	56	63	59
$A^{(4)}$	41	49	56	64	72	65	63	87	77	62

Выясните, отличаются ли методы по эффективности (примите уровень значимости равным 5%).

420. Дисперсия каждой из следующих групп:

Первая группа	13	11	10	9	7
Вторая группа	28	26	25	24	22

равна 4. Определите, будет ли дисперсия 10 значений, полученных путем объединения групп, меньше, больше или равна 4, не вычисляя эту дисперсию.

421. Первой группе испытуемых, случайно отобранных из студентов университета, сообщалось, что их мнения обычно расходятся с мнением студентов университета в целом, второй группе сообщалось, что их мнение в среднем соответствует мнению всех, третьей группе сообщалось о сильном соответствии их мнений и мнения университета в целом. Затем испытуемых просили высказать суждения по 18 актуальным вопросам (смертная казнь, контроль рождаемости и т. д.), и за оценку конформизма испытуемого принималось число ответов из 18, которое совпадало с мнением учащихся в целом. По следующим данным:

Первая группа	10	12	11	8	9	8
Вторая группа	15	12	15	11	14	
Третья группа	15	13	18	16	14	15

требуется: *а*) выяснить на 5%-ном уровне значимости, влияет ли предварительная информация о степени соответствия мнений на оценку конформизма; *б*) при подтверждении влияния оценить степень влияния; *в*) проверить на 5%-ном уровне значимости существенность различия оценок конформизма каждых двух групп; *г*) проверить на 5%-ном уровне значимости гипотезу $H_0 : \mathbf{E}(Y) = 1{,}4$; *д*) с надежностью 0,95 найти интервальную оценку дисперсии $\sigma^2_{\text{ост.}}$.

422. Школьная администрация хочет установить, влияет ли время дня на результат 50-минутной контрольной работы. Работа может начаться в любую минуту с 9 часов и закончиться не позже 15 часов. Так как моментов начала работы довольно много — 310, то было решено случайным образом из 310 пятидесятиминутных периодов отобрать четыре и в эти периоды провести работу по проверке чтения, включающую 44 вопроса. Количество правильных ответов у 18 случайно выбранных учеников таковы:

11:09 — 11:59	10:14 — 11:04	13:31 — 14:21	9:50 — 10:40
35	36	28	27
32	32	27	16
41	44	23	40
36	42	15	32
35			36

Выясните на 5%-ном уровне значимости, влияет ли время дня на результат работы. Найдите оценки параметров модели дисперсионного анализа, соответствующей условиям задачи.

423. Из совокупности экспертов извлечена случайная выборка в десять чел. Каждый эксперт оценивал независимую случайную выборку, состоящую из 20 детей, по семибалльной шкале «*эмоциональной устойчивости*». Получены следующие суммы квадратов: $SS^2_A = 94{,}32$, $SS^2_{\text{ост}} = 1831{,}6$. Сравните внутригрупповую и межгрупповую дисперсии, и сделайте вывод, кто в среднем отличается больше: дети, которых оценивает один и тот же эксперт, или эксперты.

6.4.2. Двухфакторный дисперсионный анализ

Детерминированная модель двухфакторного дисперсионного анализа с повторениями (модель М1) имеет следующий вид:

$$Y_k^{(i,j)} = \theta^{(0)} + \theta_A^{(i)} + \theta_B^{(j)} + \theta_{AB}^{(i,j)} + \varepsilon_k^{(i,j)}; i = 1,2,\ldots,\nu_A; j = 1,2,\ldots,\nu_B; k = 1,2,\ldots,m,$$

где $\theta^{(0)} = \mathbf{E}(Y)$, $\theta_A^{(i)}$, $\theta_B^{(j)}$ и $\theta_{AB}^{(i,j)}$ — неслучайные эффекты влияния на наблюдение $Y_k^{(i,j)}$ уровней $A^{(i)}$, $B^{(j)}$ факторов A и B и взаимодействия этих уровней, $\varepsilon_k^{(i,j)}$ — случайный эффект влияния прочих неконтролируемых факторов.

К этой модели предъявляются следующие требования:

- все $n = m\nu_A\nu_B$ случайных величин $\varepsilon_k^{(i,j)}$ или, иначе, все $n = m\nu_A\nu_B$ наблюдений $Y_k^{(i,j)}$ должны быть независимыми;
- $\varepsilon_k^{(i,j)} \sim \mathcal{N}(0;\sigma_{\text{ост}})$ или, иначе, $Y_k^{(i,j)} \sim \mathcal{N}\left(\theta^{(0)} + \theta_A^{(i)} + \theta_B^{(j)} + \theta_{AB}^{(i,j)};\sigma_{\text{ост}}\right)$, т. е. при каждой комбинации уровней факторов наблюдения должны проводиться в одинаковых («нормальных») вероятностных условиях с дисперсией, не изменяющейся при переходе от одной комбинации уровней факторов к другой.

По наблюдениям $Y_k^{(i,j)}$ вычисляют общую вариацию результативного признака Y и четыре ее составляющие, отражающие влияние соответственно фактора A, фактора B, их взаимодействия и остаточных факторов:

$$SS_{\text{общ}} = \sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\sum_{k=1}^{m}\left(\overline{Y}_k^{(i,j)} - \overline{Y}\right)^2, \tag{6.4.10}$$

$$SS_A = m\nu_B\sum_{i=1}^{\nu_A}\left(\overline{Y}_A^{(i)} - \overline{Y}\right)^2, \tag{6.4.11}$$

$$SS_B = m\nu_A\sum_{j=1}^{\nu_B}\left(\overline{Y}_B^{(j)} - \overline{Y}\right)^2, \tag{6.4.12}$$

$$SS_{AB} = m\sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\left(\overline{Y}^{(i,j)} - \overline{Y}_A^{(i)} - \overline{Y}_B^{(j)} + \overline{Y}\right)^2, \tag{6.4.13}$$

$$SS_{\text{ост}} = \sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\sum_{k=1}^{m}\left(Y_k^{(i,j)} - \overline{Y}^{(i,j)}\right)^2. \tag{6.4.14}$$

Здесь

$$\overline{Y}^{(i,j)} = \frac{\sum_{k=1}^{m} Y_k^{(i,j)}}{m}, \quad i = 1,2,\ldots,\nu_A, \quad j = 1,2,\ldots,\nu_B, \tag{6.4.15}$$

$$\overline{Y}_A^{(i)} = \frac{\sum_{j=1}^{\nu_B}\sum_{k=1}^{m} Y_k^{(i,j)}}{m\nu_B}, \quad i = 1,2,\ldots,\nu_A, \tag{6.4.16}$$

$$\overline{Y}_B^{(j)} = \frac{\sum_{i=1}^{\nu_A}\sum_{k=1}^{m} Y_k^{(i,j)}}{m\nu_A}, \quad j = 1,2,\ldots,\nu_B, \tag{6.4.17}$$

$$\overline{Y} = \frac{\sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\sum_{k=1}^{m} Y_k^{(i,j)}}{m\nu_A\nu_B}. \tag{6.4.18}$$

Справедливо тождество

$$SS_{\text{общ}} = SS_A + SS_B + SS_{AB} + SS_{\text{ост}}. \tag{6.4.19}$$

Вводятся следующие обозначения:

$$s_A^2 = \frac{SS_A}{\nu_A - 1}, \tag{6.4.20}$$

$$s_B^2 = \frac{SS_B}{\nu_B - 1}, \tag{6.4.21}$$

$$s_{AB}^2 = \frac{SS_{AB}}{(\nu_A - 1)(\nu_B - 1)}, \tag{6.4.22}$$

$$s_{\text{ост}}^2 = \frac{SS_{\text{ост}}}{n - \nu_A\nu_B}. \tag{6.4.23}$$

Гипотеза $H_A: \theta_{A^{(1)}} = \theta_{A^{(2)}} = \cdots = \theta_{A^{(\nu_A)}} = 0$ об отсутствии влияния на результативный признак Y фактора A проверяется на основе анализа статистики

$$F_{\nu_A - 1;\, n - \nu_A\nu_B} = \frac{s_A^2}{s_{\text{ост}}^2}. \tag{6.4.24}$$

Если гипотеза H_A верна, то статистика $F_{\nu_A - 1;\, n - \nu_A\nu_B}$ имеет распределение Фишера — Снедекора с $(\nu_A - 1)$ и $(n - \nu_A\nu_B)$ степенями свободы; поэтому гипотеза H_A принимается, если

$$F_{\nu_A - 1;\, n - \mu_A\nu_B} < f_{\nu_A - 1;\, n - \mu_A\nu_B;\, \alpha}\ . \tag{6.4.25}$$

Аналогичным образом проверяются гипотезы $H_B: \theta_{B^{(1)}} = \theta_{B^{(2)}} = \cdots = \theta_{B^{(\nu_B)}} = 0$ (об отсутствии влияния на результативный признак Y фактора B) и H_{AB} (об отсутствии влияния на результативный признак Y взаимодействия факторов A и B).

Пример 6.4.2. При исследовании зависимости средней оценки Y по математической статистике в группе от метода обучения ($A^{(1)}$ — традиционный классический, $A^{(2)}$ — компьютерный, $A^{(3)}$ — комбинированный), будущего направления подготовки ($B^{(1)}$ — «Менеджмент», $B^{(2)}$ — «Социология») и их взаимодействия было выделено случайным образом 18 групп, которые приписывались в равных количествах шести комбинациям методов и специальностей. Знания оценивались тестом, состоящим из 120 вопросов. Известны сведения о среднем числе правильных ответов в группах:

A \ B	$B^{(1)}$			$B^{(2)}$		
$A^{(1)}$	63	63	64	72	73	75
$A^{(2)}$	65	68	69	79	79	80
$A^{(3)}$	79	79	80	105	104	104

Детерминированная модель двухфакторного дисперсионного анализа (с повторениями) средней оценки по математической статистике в группе имеет следующий вид:

$$Y_k^{(i;j)} = \theta^{(0)} + \theta_A^{(i)} + \theta_B^{(j)} + \theta_{AB}^{(i;j)} + \varepsilon_k^{(i;j)};\quad i = 1, 2, 3;\quad j = 1, 2;\quad k = 1, 2, 3,$$

где $\theta^{(0)} = \mathbf{E}(Y)$, $\theta_A^{(i)}$, $\theta_B^{(j)}$ и $\theta_{AB}^{(i;j)}$ — неслучайные эффекты влияния на наблюдение $Y_k^{(i;j)}$ уровней $A^{(i)}$, $B^{(j)}$ факторов A и B и взаимодействия этих уровней, $\varepsilon_k^{(i;j)}$ — случайный эффект влияния прочих неконтролируемых факторов.

К этой модели предъявляются следующие требования:

- все $n = 3 \cdot 3 \cdot 2 = 18$ случайных величин $\varepsilon_k^{(i;j)}$ или, иначе, все 18 наблюдений $Y_k^{(i;j)}$ должны быть независимыми;
- $\varepsilon_k^{(i;j)} = \mathcal{N}(0; \sigma_{\text{ост}})$ или, иначе, $Y_k^{(i;j)} = \mathcal{N}\left(\theta^{(0)} + \theta_A^{(i)} + \theta_B^{(j)} + \theta_{AB}^{(i;j)}; \sigma_{\text{ост}}\right)$, т. е. при каждой комбинации уровней факторов наблюдения должны проводиться в одинаковых («нормальных») вероятностных условиях с дисперсией, не изменяющейся при переходе от одной комбинации уровней факторов к другой;
- $\sum_{i=1}^{3} \theta_A^{(i)} = 0$, $\sum_{j=1}^{2} \theta_B^{(j)} = 0$, $\sum_{i=1}^{3} \theta_{AB}^{(i;j)} = 0$ $(j = 1, 2, \ldots, \nu_B)$, $\sum_{j=1}^{2} \theta_{AB}^{(i;j)} = 0$ $(i = 1, 2, \ldots, \nu_A)$.

Введем исходные данные в рабочий лист Microsoft Excel (рис. 6.4.2, *а*).

Для исследования модели воспользуемся программой «Двухфакторный дисперсионный анализ с повторениями», выбрав соответствующий пункт меню надстройки «Анализ данных». В появившемся окне ввода данных (рис. 6.4.2, *б*) укажем входной интервал A1:C10, в который мы ввели исходные данные (с заголовками групп строк и столбцов — обозначениями уровней факторов), число строк для выборки (число наблюдений при каждой комбинации уровней факторов — в данном случае 3), уровень значимости «Альфа» (по условию $\alpha = 0{,}05$). Укажем, что результаты работы программы необходимо вывести на новый рабочий лист.

Результаты работы программы представлены на рис. 6.4.2, *в*.

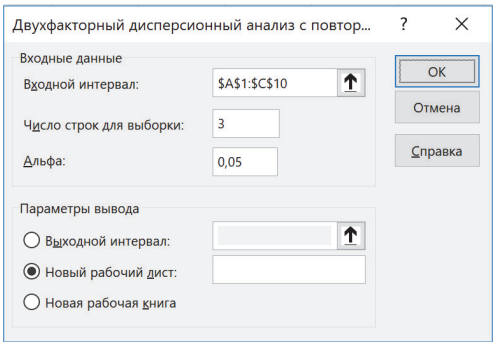
Таблица «Дисперсионный анализ», полученная в результате работы программы (рис. 6.4.2, *в*), представляет собой дисперсионную таблицу. В этой таблице «Выборка» — это фактор A, «Столбцы» — это фактор B, «Взаимодействие» — это взаимодействие факторов A и B, «Внутри» — это неконтролируемые факторы, «SS» — сумма квадратов, «df» — число степеней свободы, «MS» — средняя сумма квадратов, равная отношению SS к df, «F» — числовое значение статистики F, соответствующей проверяемой гипотезе, «P-значение» — это наблюдаемый уровень значимости, «*F* критическое» — $100\alpha\%$-ная критическая точка распределения Фишера — Снедекора с соответствующими числами степеней свободы. Расшифруем отдельные ячейки дисперсионной таблицы в табл. 6.4.3. В этой таблице

$$\overline{y}^{(i;j)} = \frac{\sum_{k=1}^{m} y_k^{(i;j)}}{m},\quad i = 1, 2, \ldots, \nu_A,\quad j = 1, 2, \ldots, \nu_B;$$

$$\overline{y}_A^{(i)} = \frac{\sum_{j=1}^{\nu_B} \sum_{k=1}^{m} y_k^{(i;j)}}{m\nu_B},\quad i = 1, 2, \ldots, \nu_A;\quad \overline{y}_B^{(j)} = \frac{\sum_{j=1}^{\nu_A} \sum_{k=1}^{m} y_k^{(i;j)}}{m\nu_A},\quad j = 1, 2, \ldots, \nu_B;\quad \overline{y} = \frac{\sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} \sum_{k=1}^{m} y_k^{(i;j)}}{m\nu_A\nu_B}.$$

	A	B	C
1		*B*(1)	*B*(2)
2	*A*(1)	63	72
3		63	73
4		64	75
5	*A*(2)	65	79
6		68	79
7		69	80
8	*A*(3)	79	105
9		79	104
10		80	104

а)



б)

Двухфакторный дисперсионный анализ с повторениями						
ИТОГИ	*B*(1)	*B*(2)	Итого			
A(1)						
Счет	3	3	6			
Сумма	190	220	410			
Среднее	63,33	73,33	68,33			
Дисперсия	0,33	2,33	31,07			
A(2)						
Счет	3	3	6			
Сумма	202	238	440			
Среднее	67,33	79,33	73,33			
Дисперсия	4,33	0,33	45,07			
A(3)						
Счет	3	3	6			
Сумма	238	313	551			
Среднее	79,33	104,33	91,83			
Дисперсия	0,33	0,33	187,77			
Итого						
Счет	9	9				
Сумма	630	771				
Среднее	70,00	85,67				
Дисперсия	53,25	203,50				
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-значение*	*F критическое*
Выборка	1839,00	2	919,50	689,625	4,12E–13	3,885
Столбцы	1104,50	1	1104,50	828,375	1,92E–12	4,747
Взаимодействие	199,00	2	99,50	74,625	1,7E–07	3,885
Внутри	16,00	12	1,33			
Итого	3158,5	17				

в)

Рис. 6.4.2. *Исходные данные* (*а*), *окно ввода данных* (*б*) *и результаты работы* (*в*) *программы «Двухфакторный дисперсионный анализ с повторениями»*

Таблица 6.4.3

Дисперсионная таблица в примере 6.4.2

Источник вариации признака Y	Показатель вариации (SS)	Число степеней свободы (df)	Оценка дисперсии $\sigma^2_{\text{ост}}$ (MS)
Фактор A (выборка)	$SS_A = m\nu_B \sum_{i=1}^{\nu_A} \left(\overline{y}_A^{(i)} - \overline{y}\right)^2 = 1839$	$\nu_A - 1 = 2$	$s_A^2 = \frac{SS_A}{\nu_A - 1} = 919{,}50$
Фактор B (столбцы)	$SS_B = m\nu_A \sum_{j=1}^{\nu_B} \left(\overline{y}_B^{(j)} - \overline{y}\right)^2 = 1104{,}5$	$\nu_B - 1 = 1$	$s_B^2 = \frac{SS_B}{\nu_B - 1} = 1104{,}50$
Взаимодействие факторов A и B	$SS_{AB} = m\sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B} \left(\overline{y}^{(i;j)} - \overline{y}_A^{(i)} - \overline{y}_B^{(j)} + \overline{y}\right)^2 = 199$	$(\nu_A - 1)(\nu_B - 1) = 2$	$s_{AB}^2 = \frac{SS_{AB}}{(\nu_A - 1)(\nu_B - 1)} = 9$
Остаточные факторы (внутри)	$SS_{\text{ост}} = \sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\sum_{k=1}^{m} \left(y_k^{(i;j)} - \overline{y}^{(i;j)}\right)^2 = 16$	$n - \nu_A\nu_B = 12$	$s_{\text{ост}}^2 = \frac{SS_{\text{ост}}}{n - \nu_A\nu_B} = 1{,}33$
Общая вариация	$SS_{\text{итог}} = \sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}\sum_{k=1}^{m} \left(\overline{y}_k^{(i;j)} - \overline{y}\right)^2 = 3158{,}5$	$n - 1 = 17$	

В условиях задачи $\nu_A = 3$, $\nu_B = 2$, $m = 3$.

Проверим на 5%-ном уровне значимости гипотезу $H_A: \theta_{A^{(1)}} = \theta_{A^{(2)}} = \theta_{A^{(3)}} = 0$ об отсутствии влияния на среднюю оценку Y фактора A — метода обучения. Наблюдаемое значение статистики

$$F_{\nu_A - 1;\, n - \nu_A\nu_B} = \frac{s_A^2}{s_{\text{ост}}^2} = \frac{MS_{\text{выборка}}}{MS_{\text{внутри}}}$$

равно $919{,}50/1{,}33 = 689{,}625$.

Если гипотеза H_A верна, то статистика $F_{\nu_A - 1;\, n - \nu_A\nu_B}$ имеет распределение Фишера — Снедекора с $\nu_A - 1 = 2$ и $n - \nu_A\nu_B = 12$ степенями свободы. При проверке гипотезы H_A **P-значение** (которое приводится в результатах работы программы в таблице «**Дисперсионный анализ**» на рис. 3.2.6) равно вероятности $p = \mathbf{P}\{F_{2;\,12} > 689{,}25\} = 4{,}12 \cdot 10^{-13}$ (т. е. наблюдаемому уровню значимости гипотезы H_A), и гипотеза H_A отвергается, поскольку $p < \alpha$.

Другой способ получить тот же вывод — сравнить наблюдаемое значение статистики $F_{\nu_A - 1;\, n - \nu_A\nu_B}$ («**F**» из таблицы «**Дисперсионный анализ**» на рис. 3.2.6) с соответствующей критической точкой $f_{\nu_A - 1;\, n - \nu_A\nu_B;\, \alpha}$ («**F критическое**» из таблицы «**Дисперсионный анализ**»): гипотеза H_A отвергается на 5%-ном уровне значимости, так как наблюдаемое значение статистики $F_{2;\,12}$ (в данном случае 689,625) больше критической точки $f_{0{,}05;\,2;\,12}$ (в данном случае 3,885).

Аналогичным образом отвергаются гипотезы $H_B: \theta_{B^{(1)}} = \theta_{B^{(2)}} = 0$ (об отсутствии влияния на среднюю оценку Y по математической статистике фактора B — будущей специальности) и H_{AB} (об отсутствии влияния на среднюю оценку Y взаимодействия метода обучения и будущей специальности). Таким образом, метод обучения, будущая специальность и их взаимодействие влияют на среднюю оценку по математической статистике в группе. Оценим силу этого влияния, вычислив соответствующие коэффициенты детерминации.

Поскольку коэффициент детерминации

$$\hat{\eta}^2(Y \mid A) = \frac{SS_A}{SS_{\text{итог}}} = \frac{1839}{3158{,}5} = 0{,}58,$$

то 58% общей вариации средней оценки Y обусловлено изменчивостью фактора A — метода обучения.

Так как

$$\hat{\eta}^2(Y \mid B) = \frac{SS_B}{SS_{\text{итог}}} = \frac{1104{,}5}{3158{,}5} = 0{,}35,$$

то 35% общей вариации средней оценки Y обусловлено изменчивостью фактора B — будущей специальности.

Ввиду того, что

$$\hat{\eta}^2(Y \mid AB) = \frac{SS_{AB}}{SS_{\text{итог}}} = \frac{199}{3158{,}5} = 0{,}06,$$

6% общей вариации средней оценки Y обусловлено взаимодействием факторов A и B.

Влиянием неконтролируемых факторов обусловлен $100 - 58 -- 35 - 6 = 1\%$ вариации средней оценки по математической статистике.

Оценки параметров детерминированной модели двухфакторного дисперсионного анализа с повторениями приведены в табл. 6.4.4. ❑

Таблица 6.4.4

Оценки параметров детерминированной модели двухфакторного дисперсионного анализа с повторениями в примере 6.4.2

Оценки	Формулы и числовые значения
$\hat{\theta}^{(0)}$	$\overline{y} = 77{,}83$
$\hat{\theta}_A^{(1)}$	$\overline{y}_A^{(1)} - \overline{y} = 68{,}33 - 77{,}83 = -9{,}50$
$\hat{\theta}_A^{(2)}$	$\overline{y}_A^{(2)} - \overline{y} = 73{,}33 - 77{,}83 = -4{,}50$
$\hat{\theta}_A^{(3)}$	$\overline{y}_A^{(3)} - \overline{y} = 91{,}83 - 77{,}83 = 14{,}00$
$\hat{\theta}_B^{(1)}$	$\overline{y}_B^{(1)} - \overline{y} = 70{,}00 - 77{,}83 = -7{,}83$
$\hat{\theta}_B^{(2)}$	$\overline{y}_B^{(2)} - \overline{y} = 85{,}67 - 77{,}83 = 7{,}84$
$\hat{\theta}_{AB}^{(1,1)}$	$\overline{y}_{AB}^{(1;1)} - \overline{y}_A^{(1)} - \overline{y}_B^{(1)} + \overline{y} = 63{,}33 - 68{,}33 - 70{,}00 + 77{,}83 = 2{,}83$
$\hat{\theta}_{AB}^{(1,2)}$	$\overline{y}_{AB}^{(1;2)} - \overline{y}_A^{(1)} - \overline{y}_B^{(2)} + \overline{y} = 73{,}33 - 68{,}33 - 85{,}67 + 77{,}83 = -2{,}84$
$\hat{\theta}_{AB}^{(2,1)}$	$\overline{y}_{AB}^{(2;1)} - \overline{y}_A^{(2)} - \overline{y}_B^{(1)} + \overline{y} = 67{,}33 - 73{,}33 - 70{,}00 + 77{,}83 = 1{,}83$
$\hat{\theta}_{AB}^{(2,2)}$	$\overline{y}_{AB}^{(2;2)} - \overline{y}_A^{(2)} - \overline{y}_B^{(2)} + \overline{y} = 79{,}33 - 73{,}33 - 85{,}67 + 77{,}83 = -1{,}84$
$\hat{\theta}_{AB}^{(3,1)}$	$\overline{y}_{AB}^{(3;1)} - \overline{y}_A^{(3)} - \overline{y}_B^{(1)} + \overline{y} = 79{,}33 - 91{,}83 - 70{,}00 + 77{,}83 = -4{,}67$
$\hat{\theta}_{AB}^{(3,2)}$	$\overline{y}_{AB}^{(3;2)} - \overline{y}_A^{(3)} - \overline{y}_B^{(2)} + \overline{y} = 104{,}33 - 91{,}83 - 85{,}67 + 77{,}83 = 4{,}66$
$s_{\text{ост.}}$	$\sqrt{s_{\text{ост}}^2} = \sqrt{1{,}33} = 1{,}15$

Задачи

424. При исследовании зависимости размера запрашиваемого кредита Y от профессии ($A^{(1)}$ — менеджеры, $A^{(2)}$ — продавцы, $A^{(3)}$ — инженеры), пола ($B^{(1)}$ — женщины, $B^{(2)}$ — мужчины) и их взаимодействия было выделено случайным образом 18 групп, которые приписывались в равных количествах шести комбинациям профессий и пола. Известны данные о средних размерах запрашиваемых кредитов в группах:

A \ B	$B^{(1)}$			$B^{(2)}$		
$A^{(1)}$	650 000	600 000	650 000	620 000	600 000	590 000
$A^{(2)}$	620 000	630 000	600 000	580 000	580 000	580 000
$A^{(3)}$	521 000	520 000	$(50+i)\cdot 10^4$	510 000	520 000	500 000

Проверьте гипотезу о равенстве групповых генеральных дисперсий.

425. В условиях предыдущей задачи запишите детерминированную модель двухфакторного дисперсионного анализа (с повторениями) размера запрашиваемого кредита Y и предъявляемые к модели требования. Затем постройте дисперсионную таблицу; на 5%-ном уровне значимости проверьте гипотезы об отсутствии влияния на среднюю оценку: профессии; пола; комбинации профессии и пола. При отклонении каких-либо из перечисленных гипотез рассчитайте соответствующий коэффициент детерминации. Оцените параметры модели двухфакторного дисперсионного анализа (с повторениями).

ГЛАВА 7. ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

§ 7.1. Модели линейной регрессии

7.1.1. Парная линейная регрессия

Некоторые величины в экономике можно контролировать непосредственно, их в любой момент можно не только измерить, но и изменить. К таким показателям относятся, например, численность сотрудников предприятия, стоимость основных производственных фондов, объем расходов на рекламу и т. п. Чтобы уменьшить численность сотрудников на 20 чел., директору предприятия достаточно издать соответствующий приказ о сокращении штатов; чтобы увеличить в следующем месяце расходы на рекламу, нужно предусмотреть такое увеличение в месячном бюджете и т. д.

Другие величины непосредственно изменить не получается: если руководитель предприятия издаст приказ об увеличении месячной прибыли (или объема продаж) на 50%, то даже если сотрудники будут стараться, вовсе не обязательно, что эта самая прибыль (объем продаж) увеличится.

Такими показателями, тем не менее, можно и нужно управлять, но опосредованно, через какие-то другие величины. Например, чтобы увеличить объем выпуска, скорее всего, потребуется увеличить численность сотрудников, приобрести дополнительные производственные мощности, возможно, каким-то образом усовершенствовать технологии. На увеличение объема продаж, скорее всего, может повлиять снижение цены, оптимизация рекламы и маркетинга.

В предыдущем абзаце рассуждения носили качественный характер, мы не называли никаких чисел, а все утверждения сопровождали словами «скорее всего, возможно».

Но в нашей профессиональной деятельности мы не можем удовлетвориться лишь качественными и приблизительными рассуждениями, нас интересуют и вопросы точного, количественного характера, например:

- «Верно ли, что объем продаж зависит от величины расходов на рекламу?»;
- «На сколько млн. руб. увеличится объем продаж, если затраты на рекламу возрастут на 1 млн. руб.?»;
- «На сколько процентов увеличится объем продаж, если затраты на рекламу возрастут на один процент?» и т. п.

Для ответов на подобные вопросы на основании числовой информации о поведении анализируемых признаков применяются модели регрессии.

Начнем с примера.

Пример 7.1.1. Известны данные за пять месяцев о месячных расходах на рекламу товара x (млн. руб.) и месячных объемах продаж этого товара Y (млн. руб.); эти данные представлены в табл. 7.1.1. Требуется отметить точки (x_i, Y_i) на графике и на основании визуального анализа этого графика выдвинуть предположения о характере зависимости объема продаж Y от расходов на рекламу x.

Введем обозначения: x_i — расходы на рекламу в i-м месяце, а Y_i — объем продаж в i-м месяце ($i = 1, 2, 3, 4, 5$). Отметим точки (x_i, Y_i) на графике (рис. 7.1.1).

Таблица 7.1.1

Цена товара и его объем продаж за пять месяцев

Месяц	Расходы на рекламу, x (млн. руб.)	Объем продаж, Y (млн. руб.)
1	3	3
2	5	7
3	7	11
4	6	14
5	9	15

Признак x можно контролировать непосредственно: фирма может установить любой желаемый размер расходов на рекламу (естественно, в пределах бюджета). Признак же Y непосредственному контролю не поддается, но на основании визуального анализа рис. 7.1.1 можно предположить, что, возможно, признак Y зависит от признака x: чем больше расходы на рекламу x, тем (обычно) больше и объем продаж Y. ❑

Как спрогнозировать объем продаж в следующем месяце?

Самый простой способ — считать, что в следующем месяце объем продаж будет таким же, как в нынешнем месяце: если обозначить прогноз на $(n+1)$-й месяц через $\hat{Y}_{n+1}$, то получим:

$$\hat{Y}_{n+1} = Y_n .$$

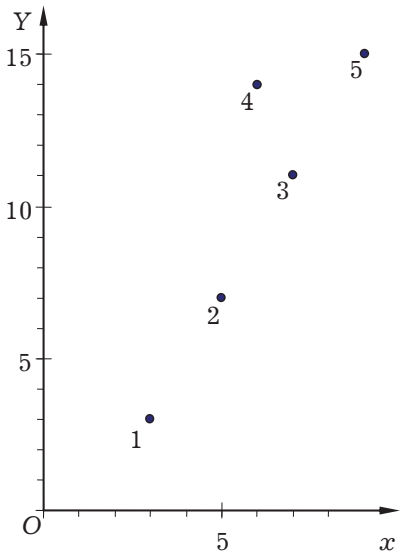


Рис. 7.1.1. *Исходные данные в примере 7.1.1*

Очевидно, в данном случае такой прогноз будет очень грубым, неточным.

Можно в качестве прогноза на $(n+1)$-й месяц взять среднее значение $\bar{Y}$ показателя Y за n месяцев:

$$\hat{Y}_{n+1} = \bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n},$$

но и такой прогноз представляется не очень хорошим.

Предлагается следующий подход к анализу ситуации. Предположим, что есть факторный признак x (контролируемый), и есть результа-

тивный признак Y, зависящий от факторного признака x, причем эта зависимость является не детерминированной, а вероятностной: на Y, кроме x, влияет множество других факторов, которые нами не рассматриваются и механизм действия которых нам неизвестен или сложен для исследования. Если бы этих случайных факторов не было, то зависимость Y от x была бы функциональной: каждому значению x соответствовало бы единственное значение $\hat{Y} = f(x)$.

Рассмотрим простейший случай линейной функции:

$$\hat{Y} = a_0 + a_1 x\,,$$

где a_0 и a_1 — некоторые числа. График этой функции для ситуации, рассмотренной в примере 7.1.1, построен жирной линией на рис. 7.1.2.

Всюду далее значения непосредственно контролируемых признаков будем обозначать строчными латинскими буквами (например, x), значения признаков, которыми нельзя управлять непосредственно — заглавными латинскими буквами (например, Y), а значения прогнозов таких признаков — «крышкой» (например, $\hat{Y}$). Нижний индекс у таких переменных означает номер наблюдения (например, в условиях примера 7.1.1 x_i — значение факторного признака x в i-м месяце, Y_i — значение результативного признака Y в i-м месяце, $\hat{Y}_i$ — прогноз значения признака Y в i-м месяце).

Если бы случайные факторы на признак Y не действовали, то все точки располагались бы строго на этой жирной прямой. Влияние неучтенных случайных факторов привело к тому, что объем продаж в первом, втором, третьем и пятом месяцах оказался меньше, а в четвертом месяце — больше, чем при отсутствии влияния случайных факторов на величину Y. Например, если бы случайные факторы на Y не действовали, то объем продаж в четвертом квартале при расходах на рекламу, равных $x_4 = 6$ млн. руб., был бы равен $\hat{Y}_4 = a_0 + a_1 x_4 \approx 10$ млн. руб., тогда как из-за действия случайных факторов в реальности объем продаж в четвертом квартале оказался выше: $Y_4 = 14$ млн. руб. Такие отклонения, вызванные влиянием случайных факторов, на рис. 7.1.2 отмечены пунктирными стрелками.

Как же построить такую прямую, иными словами, как же подобрать значения числовых коэффициентов a_0 и a_1? На рис. 7.1.3 построено три таких прямых, из которых жирная прямая нам, очевидно, «нравится больше других». Анализируя этот факт, понимаем, что нам хотелось бы построить прямую $\hat{Y}_x = a_0 + a_1 x$ так, чтобы она «проходила как можно ближе» к точкам (x_i, Y_i). Формализуем это выражение: суммарное расстояние от реально наблюдавшихся точек (x_i, Y_i) до прогнозов при отсутствии влияния случайных факторов (т. е. до точек на прямой $(x_i, \hat{y}_i)$, где $\hat{Y}_i = \hat{Y}_{x=x_i} = a_0 + a_1 x_i$) должно быть как можно меньше. Вспомним, что расстояния между точками (x_i, Y_i) и $(x_i, \hat{Y}_i)$, соответствующие вертикальным отклонениям, отмеченным на рис. 7.1.4 (как и на рис. 7.1.2) пунктирными стрелками, равны соответственно $|Y_i - \hat{Y}_i|$, и получим условие для определения коэффициентов a_0 и a_1:

$$\sum_{i=1}^{n} |Y_i - \hat{Y}_i| \to \min\,. \qquad (7.1.1)$$

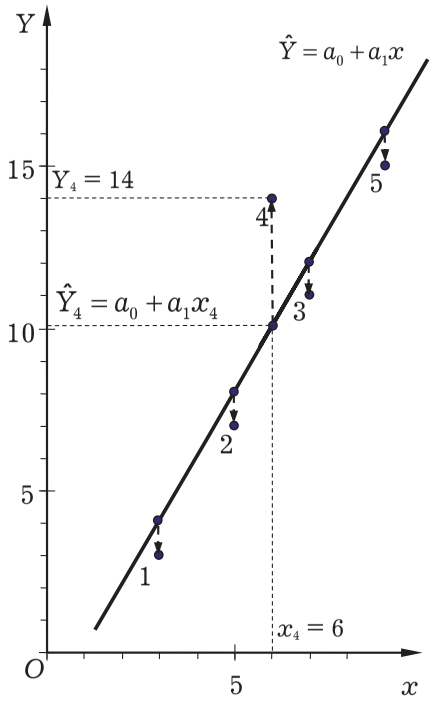


Рис. 7.1.2. *Влияние на Y фактора x (жирная линия) и случайных факторов (пунктирные стрелки)*

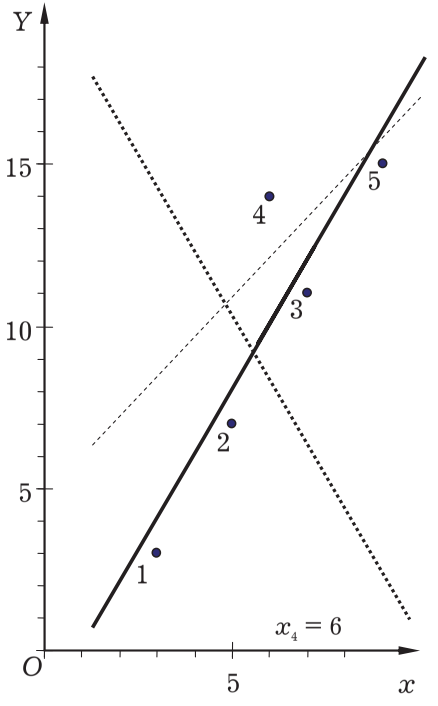


Рис. 7.1.3. *Три варианта прямых для объяснения влияния x на Y*

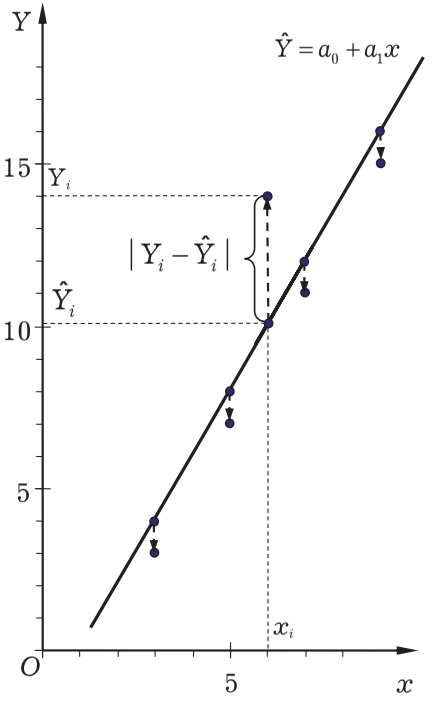


Рис. 7.1.4. *Расчет отклонений точек (x_i, Y_i) от прямой $\hat{Y} = a_0 + a_1 x$*

Определение коэффициентов a_0 и a_1 таким способом (7.1.1), который носит название **метода наименьших модулей,** возможно, но гораздо больше практически полезных результатов мы получим, если будем минимизировать не сумму расстояний, а сумму квадратов расстояний между точками (x_i, Y_i) и $(x_i, \hat{Y}_i)$:

$$T(a_0, a_1) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \to \min. \tag{7.1.2}$$

Такой метод подбора коэффициентов a_0 и a_1 (7.1.2) носит название **метода наименьших квадратов.** Перепишем формулу (7.1.2), подставив значения $\hat{Y}_i = a_0 + a_1 x_i$:

$$T(a_0, a_1) = \sum_{i=1}^{n} \left(Y_i - a_0 - a_1 x_i\right)^2 \to \min. \tag{7.1.3}$$

Итак, требуется подобрать значения a_0 и a_1 так, чтобы выражение $T(a_0, a_1)$ принимало наименьшее значение. Вспомним, что в точке максимума функции ее частные производные обязаны обращаться в нуль.

Замечание. Здесь a_0 и a_1 мы рассматриваем как переменные, тогда как все x_i и Y_i — это известные числа.

Вычислим частные производные функции $T(a_0, a_1)$ по переменным a_0 и a_1:

$$\frac{\partial T}{\partial a_0} = \sum_{i=1}^{n} \left[2\left(Y_i - a_0 - a_1 x_i\right)(-1)\right] = -2\sum_{i=1}^{n} \left(Y_i - a_0 - a_1 x_i\right) =$$

$$= -2\sum_{i=1}^{n} Y_i + 2\sum_{i=1}^{n} a_0 + 2\sum_{i=1}^{n} a_1 x_i = -2\sum_{i=1}^{n} Y_i + 2na_0 + 2a_1 \sum_{i=1}^{n} x_i,$$

$$\frac{\partial T}{\partial a_1} = \sum_{i=1}^{n} \left[2\left(Y_i - a_0 - a_1 x_i\right)(-x_i)\right] = -2\sum_{i=1}^{n} \left[\left(Y_i - a_0 - a_1 x_i\right) x_i\right] =$$

$$= -2\sum_{i=1}^{n} x_i Y_i + 2\sum_{i=1}^{n} a_0 x_i + 2\sum_{i=1}^{n} a_1 x_i^2 = -2\sum_{i=1}^{n} x_i Y_i + 2a_0 \sum_{i=1}^{n} x_i + 2a_1 \sum_{i=1}^{n} x_i^2.$$

Приравняв эти частные производные нулю, получим:

$$\begin{cases} -2\sum\limits_{i=1}^{n} Y_i + 2na_0 + 2a_1 \sum\limits_{i=1}^{n} x_i = 0, \\ -2\sum\limits_{i=1}^{n} x_i Y_i + 2a_0 \sum\limits_{i=1}^{n} x_i + 2a_1 \sum\limits_{i=1}^{n} x_i^2 = 0, \end{cases}$$

откуда [после деления обоих уравнений на $(-2n)$]

$$\begin{cases} \dfrac{\sum\limits_{i=1}^{n} Y_i}{n} = a_0 + a_1 \dfrac{\sum\limits_{i=1}^{n} x_i}{n}, \\ \dfrac{\sum\limits_{i=1}^{n} x_i Y_i}{n} = a_0 \dfrac{\sum\limits_{i=1}^{n} x_i}{n} + a_1 \dfrac{\sum\limits_{i=1}^{n} x_i^2}{n}. \end{cases}$$

Вспоминая понятие выборочного среднего, последнюю систему перепишем в виде

$$\begin{cases} \overline{Y} = a_0 + a_1 \overline{x}, \\ \overline{xY} = a_0 \overline{x} + a_1 \overline{x^2}. \end{cases}$$

Выразим из первого уравнения

$$a_0 = \overline{Y} - a_1 \overline{x}$$

и подставим во второе уравнение:

$$\overline{xY} = (\overline{Y} - a_1 \overline{x})\overline{x} + a_1 \overline{x^2}.$$

Наша система уравнений для определения значений a_0 и a_1 приняла такой вид (мы поменяли уравнения местами):

$$\begin{cases} \overline{xY} - \overline{x}\,\overline{Y} = a_1 \left[\overline{x^2} - (\overline{x})^2 \right], \\ a_0 = \overline{Y} - a_1 \overline{x}. \end{cases}$$

Теперь мы можем выразить

$$\hat{a}_1 = \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - (\overline{x})^2}, \tag{7.1.4}$$

при этом

$$\hat{a}_0 = \overline{Y} - \hat{a}_1 \overline{x}. \tag{7.1.5}$$

Замечание 1. Среднее значение произведения $\overline{xY}$ и произведение средних значений $\overline{x}\,\overline{Y}$ — это разные числа, точно также разными числами являются $\overline{x^2}$ и $(\overline{x})^2$.

Замечание 2. В формулах (7.1.4)—(7.1.5) мы неслучайно отметили значения коэффициентов a_1 и a_0 крышкой: чему эти коэффициенты равны на самом деле, мы не знаем, при этом $\hat{a}_1$ и $\hat{a}_0$ — это наши о ц е н к и (или прогнозы) истинных значений коэффициентов a_1 и a_0. Функция $\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x$, где $\hat{a}_1$ и $\hat{a}_0$ вычислены по формулам (7.1.4)—(7.1.5), называется *оценкой линейной функции регрессии* результативного признака Y на факторный признак x.

С помощью алгебраических преобразований формулу (7.1.4) можно записать в следующем виде:

$$\hat{a}_1 = \frac{\overline{(x - \overline{x})(Y - \overline{Y})}}{\overline{(x - \overline{x})^2}}. \tag{7.1.6}$$

Доказательство. Преобразуем числитель формулы (7.1.4):

$$\overline{xY} - \overline{x}\,\overline{Y} = \overline{xY} - \overline{x}\,\overline{Y} - \overline{x}\,\overline{Y} + \overline{x}\,\overline{Y} = \overline{xY} - \overline{\overline{x}\,Y} - \overline{x\,\overline{Y}} + \overline{\overline{x}\,\overline{Y}} = \overline{xY - \overline{x}Y - x\overline{Y} + \overline{x}\,\overline{Y}} = \overline{(x - \overline{x})(Y - \overline{Y})}.$$

Здесь мы учли, что среднее значение — это число, а любое число можно вносить под знак среднего значения. Полностью аналогично можно получить, что $\overline{(x^2)} - (\overline{x})^2 = \overline{(x - \overline{x})^2}$, что и доказывает формулу (7.1.6). ❑

Замечание. При практических расчетах рекомендуется пользоваться формулой (7.1.6), а не (7.1.4), поскольку формула (7.1.6) обладает существенно меньшей вычислительной погрешностью.

Пример 7.1.2. Построим оценку линейной функции регрессии объема продаж Y на расходы на рекламу x в условиях примера 7.1.1.

Вычисление оценки коэффициента a_1 удобно свести в табл. 7.1.2. Первые три столбца этой таблицы повторяют табл. 7.1.1.

Таблица 7.1.2

Расчет оценки коэффициента a_1

i	x_i	Y_i	$x_i-\overline{x}$	$Y_i-\overline{Y}$	$(x_i-\overline{x})(Y_i-\overline{Y})$	$(x_i-\overline{x})^2$
1	3	3	−3	−7	21	9
2	5	7	−1	−3	3	1
3	7	11	1	1	1	1
4	6	14	0	4	0	0
5	9	15	3	5	15	9
	$\overline{x}=6$	$\overline{Y}=10$			$\overline{(x-\overline{x})(Y-\overline{Y})}=8$	$\overline{(x-\overline{x})^2}=4$

Под двойной чертой вычисляем

$$\overline{x}=\frac{\sum_{i=1}^{5}x_i}{5}=\frac{3+5+7+6+9}{5}=\frac{30}{5}=6,\quad \overline{Y}=\frac{\sum_{i=1}^{5}Y_i}{5}=\frac{3+7+11+14+15}{5}=\frac{50}{5}=10.$$

В следующих столбцах вычисляем отклонения значений x_i и Y_i от своих средних, произведение этих отклонений и квадраты значений $(x_i-\overline{x})$.

Далее находим

$$\overline{(x-\overline{x})(Y-\overline{Y})}=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(Y_i-\overline{Y})}{5}=\frac{21+3+1+0+15}{5}=\frac{40}{5}=8,$$

$$\overline{(x-\overline{x})^2}=\frac{\sum_{i=1}^{5}(x_i-\overline{x})^2}{5}=\frac{9+1+1+0+9}{5}=\frac{20}{5}=4.$$

Теперь можно по формуле (7.1.6) найти оценку коэффициента a_1:

$$\hat{a}_1=\frac{\overline{(x-\overline{x})(Y-\overline{Y})}}{\overline{(x-\overline{x})^2}}=\frac{8}{4}=2,$$

а по формуле (7.1.5) — оценку коэффициента a_0:

$$\hat{a}_0=\overline{Y}-\hat{a}_1\overline{x}=10-2\cdot 6=-2.$$

Таким образом, оценка линейной функции регрессии объема продаж Y на расходы на рекламу x имеет следующий вид:

$$\hat{Y}_x=\hat{a}_0+\hat{a}_1x=-2+2x.$$

График этой функции — жирная линия на рис. 7.1.2—7.1.4.

Теперь можно спрогнозировать, каким окажется объем продаж, если расходы на рекламу будут равны $x = 3{,}5$ или $x = 8$ при усредненном влиянии случайных факторов на объем продаж:

$$\hat{Y}_{x=3,5}=-2+2\cdot 3{,}5=5,\quad \hat{Y}_{x=3,5}=-2+2\cdot 8=14.\ \square$$

Конечно, вручную специалисты — практики не считают, для этого есть компьютерные программы, например, в пакете Microsoft Excel есть надстройка «Анализ данных».

Пример 7.1.3. Построим оценку линейной функции регрессии объема продаж Y на расходы на рекламу x в условиях примера 7.1.1 с помощью программы «Регрессия» надстройки «Анализ данных» пакета Microsoft Excel.

Введем исходные данные о расходах на рекламу и объемах продаж, представленные в табл. 7.1.1, на рабочий лист Microsoft Excel (рис. 7.1.5).

Для вычисления оценок коэффициентов $\hat{a}_0$ и $\hat{a}_1$ воспользуемся программой **«Регрессия»** надстройки **«Анализ данных»** Microsoft Excel.

В окне ввода исходных данных программы **«Регрессия»** (рис. 7.1.6) укажем **входные интервалы** результативного признака Y (**B1:B6**) и факторного признака x (**A1:A6**). Установим флажок **«Метки»** (указав, что в первой строке находятся названия переменных) и очистим флажок **«Константа — ноль»** (чтобы в уравнении присутствовал свободный член a_0). Укажем, что результаты работы программы необходимо вывести на **новый рабочий лист**. Отметим также флажок **«График подбора»**, чтобы в результатах работы программы **«Регрессия»** получить изображение исходных данных и график оценки функции регрессии.

Результаты работы программы **«Регрессия»** представлены на рис. 7.1.7. Из полученных результатов, кроме графика, который, конечно, похож на графики, построенные на рис. 7.1.2—7.1.4, мы можем интерпретировать лишь те результаты, которые выделены зеленым пунктиром на рис. 7.1.7: в строке **«Наблюдения»** таблицы **«Регрессионная статистика»** стоит число наблюдений: в данном случае имеем 5 наблюдений (5 месяцев), а в столбце **«Коэффициенты»** — оценки коэффициентов a_0 и a_1: $\hat{a}_0 = -2$ (в строке **«Y-пересечение»**) и $\hat{a}_1 = 2$ (в строке **«x»**).

Таким образом, получена точно такая же оценка линейной функции регрессии объема продаж Y на расходы на рекламу x, что и вручную (см. пример 7.1.2): $\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x = -2 + 2x$. ❑

	A	B
1	*x*	*Y*
2	3	3
3	5	7
4	7	11
5	6	14
6	9	15

Рис. 7.1.5. *Числовые данные для программы «Регрессия»*

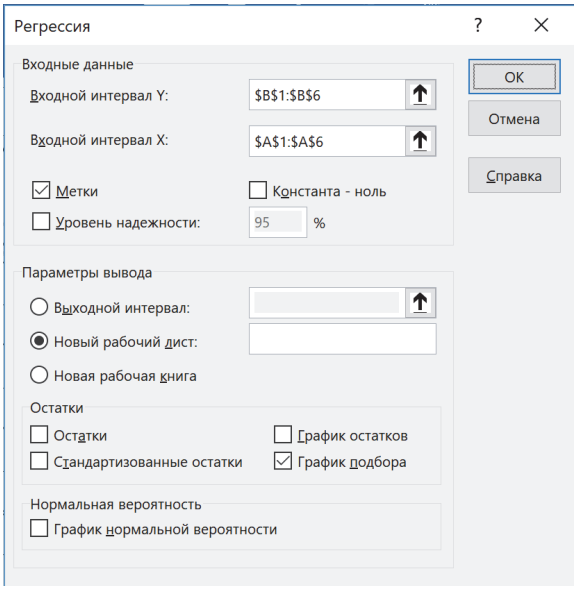


Рис. 7.1.6. *Окно ввода данных программы «Регрессия»*

ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,894					
R-квадрат	0,8					
Нормированный R-квадрат	0,733					
Стандартная ошибка	2,582					
Наблюдения	5					
Дисперсионный анализ						
	df	SS	MS	F	Значимость F	
Регрессия	1	80	80	12	0,041	
Остаток	3	20	6,667			
Итого	4	100				
	Коэффициенты	Стандартная ошибка	t-статистика	P-значение	Нижние 95%	Верхние 95%
Y-пересечение	–2	3,651	–0,548	0,622	–13,621	9,621
x	2	0,577	3,464	0,041	0,163	3,837
ВЫВОД ОСТАТКА						
Наблюдение	Предсказанное Y	Остатки				
1	4	–1				
2	8	–1				
3	12	–1				
4	10	4				
5	16	–1				

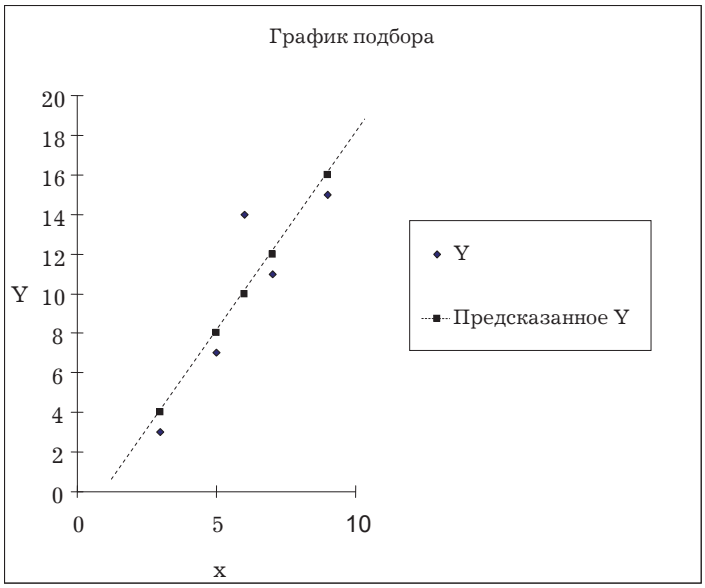


Рис. 7.1.7. *Результаты работы программы «Регрессия»*

Модель парной линейной регрессии признака Y записывается следующим образом:

$$Y = a_0 + a_1 x + \varepsilon.$$

В этой модели x — это факторный признак (независимая переменная), соответствующая контролируемому фактору, Y — результативный признак (зависимая переменная), который складывается из детерминированной составляющей $\hat{Y}_x = a_0 + a_1 x$, соответствующей линейному влиянию на Y фактора x, и случайной составляющей ε, отражающей влияние на Y всех прочих факторов (помимо фактора x), считающихся случайными.

Оценка параметров a_0 и a_1 производится с помощью метода наименьших квадратов [по формулам (7.1.5)—(7.1.6)] на основании выборочных данных, т. е. конкретных значений признаков x и Y либо на n объектах, либо в n различных моментов времени. Выборочные данные обычно представляют в табличном виде (см. табл. 7.1.3).

Таблица 7.1.3

Выборка для оценки параметров модели парной линейной регрессии

x	Y
x_1	Y_1
x_2	Y_2
⋮	⋮
x_n	Y_n

Число пар (x_i, Y_i) называется при этом *объемом выборки*. Если задана выборка объема n, то при справедливости модели имеют место n соотношений:

$$Y_i = a_0 + a_1 x_i + \varepsilon_i; \quad i = 1, 2, \ldots, n.$$

При этом все ε_i (случайные эффекты влияния на результативный признак неконтролируемых факторов) рассматриваются как случайные величины, причем предполагается выполнение следующих условий:

- все случайные величины ε_i имеют одинаковое нормальное распределение $\varepsilon_i = \mathcal{N}(0, \sigma_{ELR})$ или, иначе, все наблюдения Y_i имеют нормальное распределение $Y_i = \mathcal{N}(a_0 + a_1 x_i, \sigma_{ELR})$, $i = 1, 2, \ldots, n$.
- все случайные величины ε_i независимы или, иначе, все наблюдения Y_i независимы, $i = 1, 2, \ldots, n$.

Параметр a_1 показывает, на сколько единиц измерения изменится в среднем зависимая переменная Y, если независимая переменная увеличится на единицу.

Действительно, значению x независимой переменной соответствует в среднем значение $\hat{Y}_x = a_0 + a_1 x$ зависимой переменной (так как $\mathbf{E}(Y \mid x = x_i) = a_0 + a_1 x_i$), а значению $(x + 1)$ независимой переменной соответствует в среднем значение

$$\hat{Y}_{x+1} = a_0 + a_1(x+1) = a_0 + a_1 x + a_1 = \hat{Y}_x + a_1$$

зависимой переменной, т. е. на a_1 единиц большее, чем $\hat{Y}_x$.

Часто удобнее иметь дело не с вопросом «На сколько единиц измерения изменится Y, если x увеличится на одну единицу измерения?», а с вопросом «На сколько процентов изменится Y, если x увеличится на один процент?». Ответим на этот вопрос в условиях модели парной линейной регрессии.

Пусть значению x независимой переменной соответствует в среднем значение $\hat{Y}_x = a_0 + a_1 x$ зависимой переменной. Если x увеличить на 1% от среднего значения x (т. е. до $x' = x + 0{,}01\bar{x}$) то новому значению независимой переменной ($x' = x + 0{,}01\bar{x}$) будет соответствовать в среднем значение

$$\hat{Y}_{x'} = a_0 + a_1 x' = a_0 + a_1(x + 0{,}01\bar{x}) = a_0 + a_1 x + 0{,}01 a_1 \bar{x} =$$
$$= \hat{Y}_x + 0{,}01 \frac{a_1 \bar{x}}{\bar{Y}} \bar{Y} = \hat{Y}_x + e_{Y|x} \cdot 0{,}01 \bar{Y}$$

зависимой переменной, т. е. на

$$e_{Y|x} = \frac{a_1 \bar{x}}{\bar{Y}} \tag{7.1.7}$$

процентов от среднего значения $\bar{Y}$ большее, чем $\hat{Y}_x$.

Коэффициент $e_{Y|x}$, вычисляемый по формуле (7.1.1), называется *коэффициентом средней эластичности* признака Y по признаку x, а оценка коэффициента средней эластичности Y по x получается из формулы (7.1.1) заменой a_1 на $\hat{a}_1$:

$$\hat{e}_{Y|x} = \frac{\hat{a}_1 \bar{x}}{\bar{Y}}.$$

Так, в условиях примера 7.1.1 увеличение расходов на рекламу x на 1 млн. руб. сопровождается увеличением объема продаж в среднем примерно на $\hat{a}_1 = 2$ млн. руб. Увеличение расходов на рекламу на 1% от средней величины таких расходов ($\bar{Y}$=10 млн. руб.) сопровождается увеличением объема продаж в среднем примерно на

$$\hat{e}_{Y|x} = \frac{\hat{a}_1 \bar{x}}{\bar{Y}} = \frac{2 \cdot 6}{10} = 1{,}2$$

процента от $\bar{Y}$.

Замечание. Коэффициенты эластичности (и их оценки) уже измерены в процентах и, конечно, будет грубой ошибкой сказать, что увеличение расходов на рекламу на 1% от средней величины таких расходов сопровождается увеличением объема продаж в среднем на 120% от $\bar{Y}$!

Обратим внимание, что в модели парной линейной регрессии три неизвестных параметра: a_0, a_1 и σ_{ELR}. Оценки параметров a_0 и a_1 вычисляются, как уже говорилось, по формулам (7.1.5)—(7.1.6), а оценка параметра σ_{ELR} обозначается s_{ELR}, вычисляется по формуле

$$s_{ELR} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}} \tag{7.1.8}$$

и называется *стандартной ошибкой регрессии*. Значение s_{ELR} приводится в результатах работы программы **«Регрессия»** в таблице **«Регрессионная статистика»** под заголовком **«Стандартная ошибка»**.

Как следует из рис. 7.1.7, в условиях примера 7.1.1 стандартная ошибка регрессии приблизительно равна 2,58.

Замечание 1. Индекс «*ELR*» в обозначении стандартной ошибки регрессии означает «*Equation of Linear Regression*» — уравнение линейной регрессии.

Замечание 2. Доказательство формулы (7.1.2) приводится в большинстве стандартных курсов эконометрики.

Стандартная ошибка используется для построения различных доверительных интервалов и проверки статистических гипотез о модели регрессии.

Величины

$$SS_{\text{итог.}}=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2,\quad SS_{\text{регр.}}=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2 \text{ и } SS_{\text{ост.}}=\sum_{i=1}^{n}\left(\hat{Y}_i-Y_i\right)^2,$$

называются соответственно *полной* ($SS_{\text{итог.}}$), *объясненной* ($SS_{\text{регр.}}$) и *остаточной* ($SS_{\text{ост.}}$) *суммами квадратов отклонений* значений признака Y от своего среднего значения $\bar{Y}$ («*SS*» означает «*Sum of Squares*» — сумма квадратов).

Можно показать, что

$$SS_{\text{итог.}}=SS_{\text{регр.}}+SS_{\text{ост.}}$$

(доказательство этого факта можно найти в стандартных курсах эконометрики), поэтому полная сумма квадратов отклонений характеризует степень вариации результативного признака Y (т. е. разброса значений показателя Y вокруг своего среднего значения $\bar{Y}$), объясненная сумма характеризует часть полной суммы квадратов отклонений, которая связана с линейным влиянием факторного признака x, а остаточная сумма — часть полной суммы квадратов отклонений, которая связана с влиянием случайных факторов.

При этом величина

$$\hat{R}^2=\frac{SS_{\text{регр.}}}{SS_{\text{итог.}}}$$

равная доле общей вариации результативного признака Y, обусловленной линейным влиянием факторного признака x, называется *коэффициентом линейной детерминации*.

Значения величин $SS_{\text{регр.}}, SS_{\text{ост.}}$ и $SS_{\text{итог.}}$ приводятся в результатах работы программы **«Регрессия»** в столбце **«SS»** в строках **«Регрессия»**, **«Остаток»** и **«Итого»** соответственно, а коэффициент линейной детерминации $\hat{R}^2$ выводится в результатах работы программы **«Регрессия»** в таблице **«Регрессионная статистика»** под заголовком **«R-квадрат»** (см. рис. 7.1.7, на котором приведены числовые результаты компьютерных расчетов по модели линейной регрессии из примеров 7.1.1—7.1.3).

Как следует из рис. 7.1.7, в условиях примера 7.1.1 $SS_{\text{регр.}}=80$, $SS_{\text{итог.}}=100$,

$$\hat{R}^2=\frac{SS_{\text{регр.}}}{SS_{\text{итог.}}}=0,8=80\%,$$

поэтому вариация объема продаж Y на 80% обусловлена линейным влиянием расходов на рекламу x.

Вычисляя по выборке оценки $\hat{a}_0$ и $\hat{a}_1$ параметров a_0 и a_1, мы отдаем себе отчет, что эти оценки в любом случае остаются всего лишь приближенными значениями данных параметров, а (неизвестные нам) истинные значения a_0 и a_1, вообще говоря, отличаются соответственно от $\hat{a}_0$ и $\hat{a}_1$.

Так, мы отмечали ранее, что в условиях примера 7.1.1 увеличение расходов на рекламу x на 1 млн. руб. сопровождается увеличением объема продаж в среднем примерно на $\hat{a}_1 = 2$ млн. руб. (примерно — потому что $\hat{a}_1$ — это приближенное значение параметра a_1, а истинное значение этого параметра нам неизвестно).

Если на самом деле истинное значение параметра a_1 окажется равным не 2 млн. руб., а нулю, то модель парной линейной регрессии принимает вид

$$Y = a_0 + a_1 x + \varepsilon = a_0 + 0x + \varepsilon = a_0 + \varepsilon,$$

что означает отсутствие линейной регрессионной зависимости Y от x — в этом случае все выводы, полученные с помощью модели регрессии, окажутся необоснованными!

Обсудим проверку гипотезы о незначимости модели линейной регрессии (т. е. об отсутствии линейной регрессионной зависимости признака Y от признака x).

Предположим, что на самом деле нет линейной регрессионной зависимости результативного признака Y от факторного признака x, т. е. в линейной функции регрессии $\hat{Y}_x = a_0 + a_1 x$ коэффициент a_1 равен нулю (несмотря на то, что его оценка $\hat{a}_1$ отлична от нуля).

Рассмотрим гипотезу $H_0 : a_1 = 0$ при альтернативной гипотезе $H_1 : a_1 \neq 0$.

Предположим, что гипотеза H_0 верна, тогда величина

$$T_{n-2} = \hat{a}_1 \frac{\sqrt{\sum_{j=1}^{n} (x_j - \overline{x})^2}}{s_{ELR}} \tag{7.1.9}$$

имеет распределение Стьюдента с $n - 2$ степенями свободы (это доказывается в большинстве курсов машинного обучения и эконометрики).

Пусть $t_{\alpha;\, n-2}$ — двусторонняя критическая точка распределения Стьюдента с $n - 2$ степенями свободы, соответствующая вероятности α, т. е. такое число, что

$$\mathbf{P}\left\{ |T_k| > t_{\alpha;k} \right\} = \alpha. \tag{7.1.10}$$

Замечание. Для вычисления $t_{\alpha;\, n-2}$ в Microsoft Excel можно воспользоваться функцией

$$t_{\alpha;k} = \text{СТЬЮДЕНТ.ОБР.2Х}(<\alpha>; <k>).$$

Если выбрать уровень значимости α, близкий к нулю, то величина T_{n-2} (7.1.1) в случае, когда справедлива гипотеза H_0, лишь с очень малой вероятностью α может выйти за границы отрезка

$$\left[-t_{\alpha;\, n-2}, t_{\alpha;\, n-2} \right]$$

[согласно формуле (7.1.2)].

Таким образом, **критерий** проверки гипотезы $H_0 : a_1 = 0$ (об отсутствии регрессионной зависимости Y от x) при альтернативной гипотезе $H_1 : a_1 \neq 0$ (о наличии такой зависимости) состоит в следующем:

- если окажется, что произойдет маловероятное событие, состоящее в том, что

$$|T_{n-2}| > t_{\alpha;n-2},$$

то это будет основанием отвергнуть гипотезу H_0 и принять гипотезу H_1;

- если же окажется, что

$$|T_{n-2}| \leqslant t_{\alpha;n-2},$$

то это будет означать, что нет оснований отвергнуть гипотезу H_0.

Значение величины T_{n-2} (7.1.9) приводится в результатах работы программы «Регрессия» в таблице, выделенной красной сплошной линией на рис. 7.1.7, на пересечении столбца «t-статистика» и строки «x».

В той же таблице на пересечении столбца «P-значение» и строки «x» приводится *рассчитанный уровень значимости* гипотезы H_0, т. е. вероятность ошибиться, отвергнув гипотезу H_0.

Замечание. Если P-значение оказывается достаточно малым, то в Microsoft Excel оно может быть представлено в экспоненциальном формате, например, 6,04E–03 = $6{,}04 \cdot 10^{-3} = 0{,}00604$.

Если P-значение окажется не больше принятого уровня значимости α, то это будет означать, что вероятность ошибиться, отвергнув гипотезу H_0, не превосходит α, т. е. пренебрежимо мала, значит, есть основания отвергнуть гипотезу H_0; если же P-значение окажется больше принятого уровня значимости α, то это будет означать, что мы имеем недостаточно оснований, чтобы отвергнуть гипотезу H_0.

Пример 7.1.4. Проверим гипотезу о незначимости модели парной линейной регрессии объема продаж Y на расходы на рекламу x в условиях примера 7.1.1.

Как следует из рис. 7.1.7, в условиях примера 7.1.1 значение величины T_{n-2} равно 3,464, а P-значение равно 0,041. Двусторонняя критическая точка распределения Стьюдента с $5 - 2 = 3$ степенями свободы, соответствующая уровню значимости $\alpha = 0{,}05$ ($t_{0{,}05;3}$), вычисленная в Microsoft Excel как

$$t_{0{,}05;3} = \text{СТЬЮДЕНТ.ОБР.2Х(0,05; 3)},$$

равна 3,182.

Поскольку значение t-статистики (3,464) оказалось больше критической точки (3,182), мы имеем основания отвергнуть на 5%-ном уровне значимости гипотезу $H_0 : a_1 = 0$ (об отсутствии регрессионной зависимости Y от x) и принять альтернативную гипотезу $H_1 : a_1 \neq 0$ (о наличии такой зависимости).

Тот же вывод проще сделать, если заметить, что P-значение, равное 0,041, оказалось меньше принятого уровня значимости $\alpha = 0{,}05$ — это означает, что маловероятно ошибиться, отвергнув гипотезу $H_0 : a_1 = 0$, значит, есть основания эту гипотезу отвергнуть. ❑

При отклонении гипотезы $H_0 : a_1 = 0$ (об отсутствии регрессионной зависимости Y от x) можно сделать лишь тот вывод, что истинное значение a_1 не равно нулю. Но практически наверняка истинное значение a_1 не совпадает в точности с оценкой $\hat{a}_1$. Важно знать, насколько далеко от $\hat{a}_1$ находится a_1.

Если в условиях примера 7.1.1 истинное значение параметра a_1 окажется равным не 2 млн. руб., а 1,9 млн. руб., то большой беды, скорее всего, не случится. А если значение a_1 равно на самом деле 0 млн. руб. (или, того хуже, −10 млн. руб.)?

Вспомним, что вместо того, чтобы в качестве оценки некоторого неизвестного параметра θ брать какое-то конкретное число $\hat{\theta}$ (практически гарантиро-

ванно не равное истинному значению этого параметра), можно построить интервальную оценку параметра θ, т. е. интервал $\left(\hat{\theta}_1, \hat{\theta}_2\right)$, который с заранее заданной надежностью γ (близкой к единице) накрывает истинное значение параметра θ:

$$\mathbf{P}\left\{\hat{\theta}_1 < \theta < \hat{\theta}_2\right\} = \gamma.$$

Интервальные оценки параметров a_1 и a_0, соответствующие надежности γ, таковы (приведем соответствующие формулы без доказательств, которые можно найти в стандартных курсах эконометрики):

$$\hat{a}_1 - t_{1-\gamma;n-2}\frac{s_{ELR}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}} < a_1 < \hat{a}_1 + t_{1-\gamma;n-2}\frac{s_{ELR}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}},$$

$$\hat{a}_0 - t_{1-\gamma;n-2}\frac{s_{ELR}\sqrt{\sum_{i=1}^{n}x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i-\overline{x})^2}} < a_0 < \hat{a}_0 + t_{1-\gamma;n-2}\frac{s_{ELR}\sqrt{\sum_{i=1}^{n}x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i-\overline{x})^2}}.$$

Здесь $t_{1-\gamma;n-2}$ — двусторонняя критическая точка распределения Стьюдента с $n-2$ степенями свободы, соответствующая вероятности $1-\gamma$.

Интервальные оценки параметров a_0 и a_1, соответствующие надежности $\gamma = 0{,}95$, приводятся в результатах работы программы **«Регрессия»** (см. рис. 7.1.7): нижние границы интервалов приводятся в столбце **«Нижние 95%»**, а верхние границы интервалов — в столбце **«Верхние 95%»**.

Как следует из рис. 7.1.7, в условиях примера 7.1.1 с вероятностью 0,95 можно ожидать, что каждый миллион рублей, дополнительно вложенный в рекламу, приведет к увеличению объема продаж в среднем от 0,163 млн. руб. до 3,837 млн. руб.

Если требуется получить интервальные оценки с большей (или меньшей), чем 0,95, надежностью, то нужно в окне ввода параметров программы «Регрессия» задать нужный **уровень надежности** в процентах (см. рис. 7.1.6).

Точечным прогнозом $\mathbf{E}(Y \mid x')$ генерального среднего значения результативного признака Y при значении факторного признака, равном $x = x'$, будет величина

$$\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x.$$

В условиях примера 7.1.1 точечные прогнозы генерального среднего объема продаж $\mathbf{M}(Y \mid x')$ при расходах на рекламу, равных $x' = 1;\ 4;\ 7;\ 10$, таковы:

$$\hat{Y}_{x=1} = -2 + 2 \cdot 1 = 0;$$

$$\hat{Y}_{x=4} = -2 + 2 \cdot 4 = 6;$$

$$\hat{Y}_{x=7} = -2 + 2 \cdot 7 = 12;$$

$$\hat{Y}_{x=10} = -2 + 2 \cdot 10 = 18$$

соединим эти точечные прогнозы прямой линией на рис. 7.1.8.

Интервальная оценка генерального среднего $\mathbf{E}(Y \mid x')$, соответствующая надежности γ, задается формулой

$$\hat{y}_{x'} - t_{1-\gamma;n-2} s_{ELR} \sqrt{\frac{1}{n} + \frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}} < \mathbf{E}(Y|x') < \hat{y}_{x'} + t_{1-\gamma;n-2} s_{ELR} \sqrt{\frac{1}{n} + \frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}.$$

В условиях примера 7.1.1 при $\gamma = 0,95$

$$-9,895 < \mathbf{E}(Y|x'=1) < 9,895,$$
$$0,803 < \mathbf{E}(Y|x'=4) < 11,197,$$
$$7,891 < \mathbf{E}(Y|x'=7) < 16,109,$$
$$9,783 < \mathbf{E}(Y|x'=10) < 26,217,$$

эти интервальные оценки изображены на рис. 7.1.8 тонкими кривыми.

Интервальный прогноз значения результативного признака Y при значении факторного признака, равном $x = x'$, соответствующий надежности γ, задается формулой

$$\hat{y}_{x'} - t_{1-\gamma;n-2} s_{ELR} \sqrt{1 + \frac{1}{n} + \frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}} < Y|x' < \hat{y}_{x'} + t_{1-\gamma;n-2} s_{ELR} \sqrt{1 + \frac{1}{n} + \frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}.$$

В условиях примера 7.1.1 при $\gamma = 0,95$

$$-12,862 < Y|x'=1 < 12,862,$$
$$-3,723 < Y|x'=4 < 15,723,$$
$$2,813 < Y|x'=7 < 21,187,$$
$$6,379 < Y|x'=10 < 29,621.$$

эти интервальные оценки изображены на рис. 7.1.8 жирными кривыми.

Расчет интервальных оценок для $\mathbf{E}(Y|x')$ и $Y|x'$ в Microsoft Excel проиллюстрирован рис. 7.1.9.

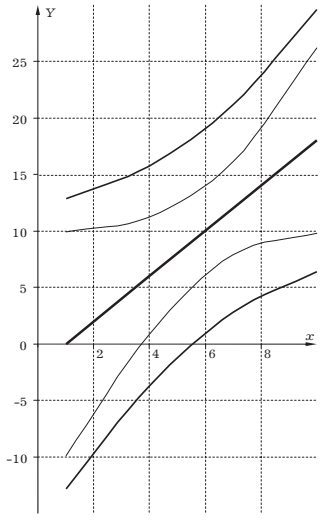


Рис. 7.1.8. *Точечный и интервальный прогнозы в модели парной линейной регрессии*

	A	B	C	D	E	F	G	H	I	J	K	L
1	x	Y		a_0=	–2	$\overline{x}$=	=СРЗНАЧ(A2:A6)	$\sum_{i=1}^{n}(x_i-\overline{x})^2$=	=(СЧЁТ(A2:A6)-1)*F2	s_{ELR}=	2,582	
2	3	3		a_1=	2	s_X^2 =	=ДИСП(A2:A6)	$t_{58;\,0{,}05}$=	=СТЬЮДЕНТ.ОБР.2Х(0,05;3)			
3	5	7										
4	7	11		x'	$\hat{Y}_{x'}$	$(x'-\overline{x})^2$	$t_{n-2;\,1-\gamma}s_{ELR}\sqrt{\frac{1}{n}+\frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}$	Нижняя граница $\mathbf{M}(Y\|x')$	Верхняя граница $\mathbf{M}(Y\|x')$	$t_{1-\gamma;\,n-2}s_{ELR}\sqrt{1+\frac{1}{n}+\frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}$	Нижняя граница $(Y\|x')$	Верхняя граница $(Y\|x')$
5	6	14		1	=$E$1+$E$2*D5	=(D5-$G$1)^2	=$I$2*$K$1*КОРЕНЬ(1/5+F5/$I$1)	=E5-G5	=E5+G5	=$I$2*$K$1*КОРЕНЬ(1+1/5+F5/$I$1)	=E5-J5	=E5+J5
6	9	15		4	=$E$1+$E$2*D6	=(D6-$G$1)^2	=$I$2*$K$1*КОРЕНЬ(1/5+F6/$I$1)	=E6-G6	=E6+G6	=$I$2*$K$1*КОРЕНЬ(1+1/5+F6/$I$1)	=E6-J6	=E6+J6
7	...	...		7	=$E$1+$E$2*D7	=(D7-$G$1)^2	=$I$2*$K$1*КОРЕНЬ(1/5+F7/$I$1)	=E7-G7	=E7+G7	=$I$2*$K$1*КОРЕНЬ(1+1/5+F7/$I$1)	=E7-J7	=E7+J7
8	...	...		10	=$E$1+$E$2*D8	=(D8-$G$1)^2	=$I$2*$K$1*КОРЕНЬ(1/5+F8/$I$1)	=E8-G8	=E8+G8	=$I$2*$K$1*КОРЕНЬ(1+1/5+F8/$I$1)	=E8-J8	=E8+J8

а)

	A	B	C	D	E	F	G	H	I	J	K	L
1	x	Y		a_0=	–2	$\overline{x}$=	6,00	$\sum_{i=1}^{n}(x_i-\overline{x})^2$=	20,00	s_{ELR}=	2,582	
2	3	3		a_1=	2	s_X^2 =	5,00	$t_{58;\,0{,}05}$ =	3,18			
3	5	7										
4	7	11		x'	$\hat{Y}_{x'}$	$(x'-\overline{x})^2$	$t_{n-2;\,1-\gamma}s_{ELR}\sqrt{\frac{1}{n}+\frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}$	Нижняя граница $\mathbf{M}(Y\|x')$	Верхняя граница $\mathbf{M}(Y\|x')$	$t_{1-\gamma;\,n-2}s_{ELR}\sqrt{1+\frac{1}{n}+\frac{(x'-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}}$	Нижняя граница $(Y\|x')$	Верхняя граница $(Y\|x')$
5	6	14		1	0,000	25	9,894670214	–9,895	9,895	12,86175892	–12,862	12,862
6	9	15		4	6,000	4	5,196935401	0,803	11,197	9,722575865	–3,723	15,723
7	...	...		7	12,000	1	4,108538180	7,891	16,109	9,186970658	2,813	21,187
8	...	...		10	18,000	16	8,217076359	9,783	26,217	11,62070083	6,379	29,621

б)

Рис. 7.1.9. *Расчет доверительных интервалов для* $\mathbf{E}(Y\,|\,x')$ *и* $Y\,|\,x'$*: формулы Microsoft Excel* (*а*)*; результаты расчетов* (*б*)

Замечаем, что при данных значениях расходов на рекламу ($x'=1;\ 4;\ 7;\ 10$) нет никакой гарантии даже в том, что объем продаж окажется больше этих расходов: например, даже при $x'=10$ объем продаж с вероятностью 0,95 окажется в границах (6,379, 29,621)!

Первое изложение метода наименьших квадратов было опубликовано в 1805 г. А.М. Лежандром. В 1809 г. К.Ф. Гаусс опубликовал теорию расчета орбит небесных тел. В этой работе он утверждал, что использовал метод наименьших квадратов с 1795 г. Это, естественно, привело к спорам с А.М. Лежандром о приоритете открытия. Однако, по мнению Гаусса, он пошел дальше Лежандра, поскольку рассматривал метод наименьших квадратов с позиций теории вероятностей и в связи с нормальным законом распределения. В 1821 г. К.Ф. Гаусс показал, что оценки метода наименьших квадратов оптимальны в классе линейных несмещенных оценок. В 1900 г. этот же результат был независимо получен А.А. Марковым. Работы А.А. Маркова, опубликованные в начале XX в., позволили встроить метод наименьших квадратов в теорию статистического оценивания параметров, в которой он является важной и естественной частью.

Термин «регрессия» был предложен в XIX в. Ф. Гальтоном, который в 1886—1889 гг. изучил 205 пар родителей и 930 их взрослых детей и опубликовал ряд статей, в которых сформулировал «закон регрессии к среднему»: «для многих непрерывных признаков, таких как рост и интеллект, взрослое потомство данного родителя отклоняется в меньшей степени от среднего значения для данной популяции, чем родитель, то есть, потомки «регрессируют» к среднему для популяции».

Для Ф. Гальтона термин «регрессия» имел чисто биологический смысл, а статистическая теория регрессионного анализа была разработана в развитие результатов К.Ф. Гаусса и А.А. Маркова Дж.У. Юлом, К. Пирсоном и Р. Фишером в конце XIX — начале XX в.

Задачи

426. Средние месячные значения фондового индекса x и котировок акции E за шесть месяцев таковы:

x	3	4	5	10	15	20
E	12	10	7	5	4	2

Оцените параметры модели линейной регрессии котировок акции на значения индекса. На одном рисунке постройте график оценки функции регрессии и отметьте точки наблюдения.

427. Известны данные об оценке социального статуса Y шести людей и их среднемесячного дохода x (тыс. ден. ед.):

x	30	70	110	140	150
Y	3	5	7	6	9

Проведите регрессионный анализ зависимости оценки социального статуса Y от среднемесячного дохода x, предположив линейность функции регрессии. Изобразите на графике наблюдения, прямую регрессии и интервальные прогнозы оценки социального статуса при различных размерах дохода. Убедитесь в том, что прямая регрессии проходит через точку $(\overline{x};\overline{Y})$.

428. Оцените коэффициенты линейной функции спроса (описывающей зависимость сбыта от цены товара) по следующим данным:

Цена товара, руб.	540	500	550	509	600
Сбыт, шт.	57	60	58	51	48

7.1.2. Множественная линейная регрессия

В предыдущем параграфе рассматривалось исследование линейной регрессионной зависимости результативного признака Y от единственного факторного признака x:

$$Y = a_0 + a_1 x + \varepsilon.$$

Однако гораздо чаще встречаются ситуации, в которых результативный признак зависит не от одного фактора, а от многих.

Так, например, объем выпуска фирмы, вообще говоря, зависит от количества используемых ресурсов: труда (численности занятых) $x^{(1)}$, капитала (стоимости фондов) $x^{(2)}$ и др. Объем продаж розничного магазина, зависит, вообще говоря, от уровня цен $x^{(1)}$, расходов на рекламу $x^{(2)}$, площади торговых помещений $x^{(3)}$, расстояния от центра города $x^{(4)}$, расстояния от метро $x^{(5)}$ и др. Объем спроса на товар зависит от его цены, цен товаров-заменителей, цен дополнительных товаров, сезона и др.

Модель множественной линейной регрессии признака Y записывается следующим образом:

$$Y = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x^{(m)} + \varepsilon.$$

В этой модели $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ — это факторные признаки (независимые переменные), Y — результативный признак (зависимая переменная), который складывается из детерминированной составляющей

$$\hat{Y}_x = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x^{(m)},$$

соответствующей линейному влиянию на Y факторов $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, и случайной составляющей $\varepsilon = \mathcal{N}(0; \sigma_{ELR})$, отражающей влияние на Y всех прочих (помимо $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$) факторов, считающихся случайными.

Выборка в случае m факторных признаков имеет вид табл. 7.1.4, а метод наименьших квадратов для оценки параметров a_0, a_1, a_2, ..., a_m состоит в минимизации функции нескольких переменных:

$$T(a_0, a_1, a_2, \ldots, a_m) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \tag{7.1.11}$$

$$= \sum_{i=1}^{n} \left(Y_i - a_0 - a_1 x_i^{(1)} - a_2 x_i^{(2)} - \cdots - a_m x_i^{(m)}\right)^2 \to \min.$$

Таблица 7.1.4

Выборка для оценки параметров модели множественной линейной регрессии

$x^{(1)}$	$x^{(2)}$	$\cdots$	$x^{(m)}$	Y
$x_1^{(1)}$	$x_1^{(2)}$	$\cdots$	$x_1^{(m)}$	Y_1
$x_2^{(1)}$	$x_2^{(2)}$	$\cdots$	$x_2^{(m)}$	Y_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$x_n^{(1)}$	$x_n^{(2)}$	$\cdots$	$x_n^{(m)}$	Y_n

Обозначим

$$\mathbf{X}=\begin{pmatrix}1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)}\\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)}\end{pmatrix}$$

матрицу зафиксированных значений факторных признаков (введение столбца из единиц обусловлено наличием в уравнении регрессии свободного члена a_0),

$$\mathbf{Y}=\begin{pmatrix}Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix} —$$

вектор-столбец наблюдений результативного признака Y,

$$\hat{\mathbf{A}}=\begin{pmatrix}\hat{a}_0\\ \hat{a}_1\\ \vdots\\ \hat{a}_m\end{pmatrix} —$$

оценку вектора параметров модели

$$\mathbf{A}=\begin{pmatrix}a_0\\ a_1\\ \vdots\\ a_m\end{pmatrix},$$

s_{ELR} — оценку параметра σ_{ELR} (стандартную ошибку).

Необходимое условие максимума функции (7.1.11) состоит в равенстве нулю частных производных:

$$\frac{\partial T}{\partial a_j}=-2\sum_{i=1}^{n}x_i^{(j)}\left(Y_i-a_0-a_1x_i^{(1)}-a_2x_i^{(2)}-\cdots-a_mx_i^{(m)}\right)^2=0,\quad j=1,2,\ldots,m$$

или

$$\sum_{i=1}^{n}x_i^{(j)}Y_i=\left(\sum_{i=1}^{n}x_i^{(j)}x_i^{(0)}\right)a_0+\left(\sum_{i=1}^{n}x_i^{(j)}x_i^{(1)}\right)a_1+\left(\sum_{i=1}^{n}x_i^{(j)}x_i^{(2)}\right)a_2+\cdots+\left(\sum_{i=1}^{n}x_i^{(j)}x_i^{(m)}\right)a_m,$$
$$j=1,2,\ldots,m.$$

Последнюю группу равенств можно переписать в виде

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})_0\,a_0+(\mathbf{X}^{\mathrm{T}}\mathbf{X})_1\,a_1+(\mathbf{X}^{\mathrm{T}}\mathbf{X})_2\,a_2+\cdots+(\mathbf{X}^{\mathrm{T}}\mathbf{X})_m\,a_m$$

или, окончательно, в виде

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{A}.$$

Отсюда можно выразить вектор оценок параметров:

$$\hat{\mathbf{A}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

Функция

$$\mathbf{E}\left(Y \middle| x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right) = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x_i^{(m)}$$

называется *линейной функцией множественной регрессии*, а функция

$$\hat{Y}_{\mathbf{x}} = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_2 x^{(2)} + \cdots + \hat{a}_m x^{(m)},$$

где $\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m$ — оценки коэффициентов a_0, a_1, a_2, ..., a_m (полученные с помощью метода наименьших квадратов), — *оценкой линейной функции множественной регрессии.*

Смысл коэффициентов a_j ($j = 1, 2, \ldots, m$) аналогичен смыслу коэффициента a_1 в модели парной линейной регрессии: коэффициент a_j показывает, на сколько единиц измерения изменится в среднем результативный признак Y, если факторный признак $x^{(j)}$ увеличить на единицу его измерения.

Замечание. Коэффициент a_0 формально имеет смысл значения результативного признака при нулевых значениях факторных признаков, однако экономической интерпретации коэффициент a_0 не имеет, так как при нулевых значениях $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ предположения, лежащие в основе модели, как правило, не удовлетворяются.

Точно так же, как и в случае парной линейной регрессии, можно ввести коэффициенты средней эластичности

$$e_{Y|x^{(i)}} = \frac{a_i \overline{x^{(i)}}}{\overline{Y}},\ i = 1, 2, \ldots, m$$

и их оценки

$$\hat{e}_{Y|x^{(i)}} = \frac{\hat{a}_i \overline{x^{(i)}}}{\overline{Y}},\ i = 1, 2, \ldots, m.$$

Решение задачи (7.1.11) реализовано в программе **«Регрессия»** надстройки **«Анализ данных»** пакета Microsoft Excel. Работа с этой надстройкой в случае нескольких факторных признаков производится точно так же, как и в случае одного факторного признака.

Пример 7.1.5. Изучается линейная (в среднем) зависимость результативного признака Y — ожидаемой продолжительности жизни мужчины (в годах) от пяти факторных признаков — регрессоров $x^{(1)}$ — среднего числа детей в семье, $x^{(2)}$ — ВВП на душу населения (в долл. США по покупательной способности валют), $x^{(3)}$ — плотности населения (в чел. на кв. км), $x^{(4)}$ — процента грамотных и $x^{(5)}$ — рождаемости на 1000 чел. (см. табл. 7.1.5 и файл «Демография.xlsx»).

Построим оценку линейной функции регрессии ожидаемой продолжительности жизни мужчины (в годах) на $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ и $x^{(5)}$.

Рассчитаем оценки $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5$ и s_{ELR} параметров модели линейной регрессии. Для этого воспользуемся программой **«Регрессия»**, выбрав соответствующий пункт меню надстройки **«Анализ данных»** Microsoft Excel.

Введем исходные данные в рабочий лист Microsoft Excel (рис. 7.1.10).

В окне ввода исходных данных программы **«Регрессия»** (рис. 7.1.11) укажем **входные интервалы** результативного признака Y (**B1:B53**) и факторных признаков $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$, $x^{(5)}$ (**C1:G53**). Установим флажок **«Метки»** (указав, что в первой строке находятся названия переменных), очистим флажок **«Константа — ноль»** (чтобы в уравнении присутствовал свободный член a_0), **уровень надежности** $(1-\alpha)$ указывать не будем (по умолчанию он равен 95%). Укажем, что результаты работы программы необходимо вывести на **новый рабочий лист**. Укажем также, что необходимо вывести **остатки**.

Результаты работы программы **«Регрессия»** представлены на рис. 7.1.12.

Таблица 7.1.5

Данные мировой демографической статистики

№ п/п	Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	№ п/п	Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
1	Австралия	74	1,90	16 848	2,3	100	15	27	Марокко	66	3,83	1062	63,0	50	29
2	Австрия	73	1,50	18 396	94,0	99	12	28	Мексика	69	3,20	3604	46,0	87	28
3	Аргентина	68	2,80	3408	12,0	95	20	29	Нидерданды	75	1,58	17 245	366,0	99	13
4	Бангладеш	53	4,70	202	800,0	35	35	30	Новая Зеландия	73	2,03	14 381	13,0	99	16
5	Беларусь	66	1,88	6500	50,0	99	13	31	Норвегия	74	2,00	17 755	11,0	99	13
6	Бельгия	73	1,70	17 912	329,0	99	12	32	ОАЭ	70	4,50	14 193	32,0	68	28
7	Бразилия	57	2,70	2354	18,0	81	21	33	Польша	69	1,94	4429	123,0	99	14
8	Буркина-Фасо	47	6,94	357	36,0	18	47	34	Португалия	71	1,50	9000	108,0	85	12
9	Великобритания	74	1,83	15 974	237,0	99	13	35	Россия	64	1,83	6680	8,8	99	13
10	Вьетнам	63	3,33	230	218,0	88	27	36	Саудовская Аравия	66	6,67	6651	7,7	62	38
11	Гаити	43	5,94	383	231,0	53	40	37	Северная Корея	67	2,40	1000	189,0	99	24
12	Германия	73	1,47	17 539	227,0	99	11	38	Сингапур	73	1,88	14 990	4456,0	88	16
13	Гондурас	65	4,90	1030	46,0	73	35	39	США	73	2,06	23 474	26,0	97	15
14	Гонконг	75	1,40	14 641	5494,0	77	13	40	Таиланд	65	2,10	1800	115,0	93	19
15	Египет	60	3,77	748	57,0	48	29	41	Турция	69	3,21	3721	79,0	81	26
16	Замбия	44	6,68	573	11,0	73	46	42	Украина	65	1,82	2340	87,0	97	12
17	Индия	58	4,48	275	283,0	52	29	43	Филиппины	63	3,35	867	221,0	90	27
18	Ирландия	73	1,99	12 170	51,0	98	14	44	Финляндия	72	1,80	15 877	39,0	100	13
19	Испания	74	1,40	13 047	77,0	95	11	45	Франция	74	1,80	18 944	105,0	99	13
20	Италия	74	1,30	17 500	188,0	97	11	46	Чили	71	2,50	2591	18,0	93	23
21	Канада	74	1,80	19 904	2,8	97	14	47	Швейцария	75	1,60	22 384	170,0	99	12
22	Китай	67	1,84	377	124,0	78	21	48	Швеция	75	2,10	16 900	19,0	99	14
23	Колумбия	69	2,47	1538	31,0	87	24	49	Эфиопия	51	6,81	122	47,0	24	45
24	Коста-Рика	76	3,10	2031	64,0	93	26	50	ЮАР	62	4,37	3128	35,0	76	34
25	Куба	74	1,90	1382	99,0	94	17	51	Южная Корея	68	1,65	6627	447,0	96	16
26	Малайзия	66	3,51	2995	58,0	78	29	52	Япония	76	1,55	19 860	330,0	99	11

	A	B	C	D	E	F	G
1	Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
2	Австралия	74	1,9	16 848	2,3	100	15
3	Австрия	73	1,5	18 396	94	99	12
4	Аргентина	68	2,8	3408	12	95	20
5	Бангладеш	53	4,7	202	800	35	35
6	Беларусь	66	1,9	6500	50	99	13
7	Бельгия	73	1,7	17 912	329	99	12
8	Бразилия	57	2,7	2354	18	81	21
9	Буркина-Фасо	47	6,9	357	36	18	47
10	Великобритания	74	1,8	15 974	237	99	13
11	М	М	М	М	М	М	М

Рис. 7.1.10. *Числовые данные для программы «Регрессия»*

Оценки $\hat{a}_0 = 62{,}29, \hat{a}_1 = -3{,}37, \hat{a}_2 = 0{,}000375, \hat{a}_3 = 0{,}000215, \hat{a}_4 = 0{,}088, \hat{a}_5 = 0{,}193$ параметров a_0, a_1, a_2, a_3, a_4, a_5 содержатся в результатах работы программы **«Регрессия»** (рис. 7.1.12) в выделенной таблице в столбце **«Коэффициенты»** под заголовками **«Y-пересечение»**, **«x(1)»**, **«x(2)»**, **«x(3)»**, **«x(4)»**, **«x(5)»** соответственно.

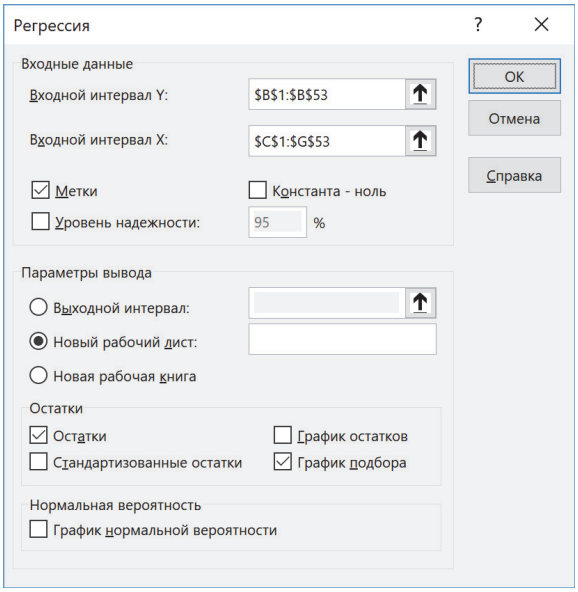


Рис. 7.1.11. *Окно ввода данных программы «Регрессия»*

ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,86043					
R-квадрат	0,74034					
Нормированный R-квадрат	0,71212					
Стандартная ошибка	4,38278					
Наблюдения	52					
Дисперсионный анализ						
	df	SS	MS	F	Значимость F	
Регрессия	5	2519,3760	503,87520	26,23148	1,99E-12	
Остаток	46	883,6045	19,20879			
Итого	51	3402,9810				
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	62,29155	7,31952	8,51034	5,29E-11	47,55813	77,02498
x(1)	-3,374930	1,86564	-1,80899	0,07699	-7,13027	0,38041
x(2)	0,000375	0,00013	2,87452	0,00611	0,00011	0,00064
x(3)	0,000215	0,00070	0,30707	0,76018	-0,00119	0,00162
x(4)	0,087983	0,05802	1,51644	0,13625	-0,02880	0,20477
x(5)	0,192832	0,31039	0,62125	0,53750	-0,43195	0,81762
ВЫВОД ОСТАТКА						
Наблюдение	Предсказанное Y	Остатки				
1	73,88634	0,11366				
2	75,16983	-2,16983				
3	66,33689	1,66311				
4	56,50543	-3,50543				
⋮	⋮	⋮				

Рис. 7.1.12. *Результаты работы программы «Регрессия»*

Таким образом, оценка линейной функции регрессии такова:

$$\hat{Y}_{\mathbf{x}} = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_1 x^{(2)} + \hat{a}_1 x^{(3)} + \hat{a}_1 x^{(4)} + \hat{a}_1 x^{(5)} =$$

$$= 62{,}29 - 3{,}37x^{(1)} + 0{,}000375x^{(2)} + 0{,}000215x^{(3)} + 0{,}088x^{(4)} + 0{,}193x^{(5)}. \; \square$$

В таблице **«Дисперсионный анализ»** (в результатах работы программы **«Регрессия»** — см. рис. 7.1.12) в столбце **«df»** приводятся количества степеней свободы m, $n-m$, $n-1$ соответственно случайных величин

$$SS_{\text{регр.}}=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2,\ \ SS_{\text{ост.}}=\sum_{i=1}^{n}\left(\hat{Y}_i-Y_i\right)^2,\ SS_{\text{итог.}}=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2,$$

значения которых приводятся в столбце **«SS»**; в столбце **«MS»** приводятся значения величин

$$MS_{\text{регр.}}=SS_{\text{регр.}}/m,\quad MS_{\text{ост.}}=SS_{\text{ост.}}/(n-m-1).$$

В условиях примера 7.1.5 $m=5$, $n-m=46$, $n-1=51$, значения величин $SS_{\text{регр.}}$, $SS_{\text{ост.}}$ и $SS_{\text{итог.}}$, как видно из рис. 7.1.12, равны соответственно 2519,38, 883,60 и 3402,98, а значения величин $MS_{\text{регр.}}$ и $MS_{\text{ост.}}$ равны соответственно 503,88 и 19,21.

Точно так же, как и в случае парной линейной регрессии, величина

$$\hat{R}^2=\frac{SS_{\text{регр.}}}{SS_{\text{итог.}}}$$

равна доле общей вариации результативного признака Y, обусловленной линейным влиянием факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, эта величина называется *коэффициентом множественной линейной детерминации*.

В таблице **«Регрессионная статистика»** (рис. 7.1.12) приводятся:

- оценка коэффициента множественной линейной детерминации

$$\hat{R}^2(Y\,|\,x^{(1)}, x^{(2)}, \ldots, x^{(m)})$$

(**R-квадрат**);

- оценка нормированного коэффициента линейной детерминации

$$\tilde{R}^2(Y\,|\,x^{(1)}, x^{(2)}, \ldots, x^{(m)})$$

(**нормированный R-квадрат**) — в отличие от коэффициента $\hat{R}^2$, который при включении в имеющуюся линейную регрессионную модель дополнительного регрессора всегда увеличивается, нормированный коэффициент детерминации $\tilde{R}^2$ может и увеличиваться, и уменьшаться; чем больше $\tilde{R}^2$, тем более адекватно уравнение регрессии,

- стандартная ошибка регрессии s_{ELR} (**стандартная ошибка**).

В условиях примера 7.1.5

$$\hat{R}^2(Y\,|\,x^{(1)}, x^{(2)}, \ldots, x^{(m)})=0{,}74\ —$$

судя по наблюдениям, 74% вариации продолжительности жизни мужчины обусловлено линейным влиянием на нее среднего числа детей в семье, величины ВВП на душу населения, плотности населения, процента грамотных и рождаемости;

$$\tilde{R}^2(Y\,|\,x^{(1)}, x^{(2)}, \ldots, x^{(m)})=0{,}71;\quad s_{ELR}=4{,}38.$$

В таблице **«Вывод остатка»**, фрагмент которой приведен на рис. 7.1.12, содержится **предсказанное Y** — это $\hat{Y}_i$, рассчитанные по построенному уравнению регрессии, и **остатки** — это разности $\left(Y_i-\hat{Y}_i\right)$. Зная эти остатки, можно рассчитать *среднюю абсолютную ошибку аппроксимации*

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|.$$

и *среднюю относительную ошибку аппроксимации*

$$\delta = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i}.$$

В условиях примера 7.1.5 средняя абсолютная и средняя относительная ошибки аппроксимации равны соответственно $\Delta \approx 2{,}801$ г. и $\delta \approx 0{,}045 = 4{,}5\%$.

Нижние и верхние границы интервальных оценок параметров a_1, a_2, a_3, a_4, a_5 приводятся в результатах работы программы **«Регрессия»** (рис. 7.1.12) в столбцах **«Нижние 95%»** и **«Верхние 95%»** и соответствующей строке выделенной таблицы.

В условиях примера 7.1.5 интервальные оценки параметров a_1, a_2, a_3, a_4, a_5 таковы:

$$(-7{,}1303;\ 0{,}3804),\quad (0{,}0001;\ 0{,}0006),\quad (-0{,}0011;\ 0{,}0016),$$
$$(-0{,}0288;\ 0{,}2048),\quad (-0{,}4320;\ 0{,}8176).$$

Модель множественной линейной регрессии называется *значимой*, если хотя бы один из коэффициентов при факторных признаках (a_1, a_2, ..., a_m) отличен от нуля и *незначимой* в противном случае.

Проверка гипотезы H_0: $a_1 = a_2 = \cdots = a_m = 0$ (о незначимости модели множественной линейной регрессии) производится на основе анализа статистики

$$F_{m;\,n-m-1} = \frac{MS_{\text{регр}}}{MS_{\text{ост}}} = \frac{SS_{\text{регр}}/m}{SS_{\text{ост}}/(n-m-1)} = \frac{\hat{R}^2/m}{(1-\hat{R}^2)/(n-m-1)},$$

имеющей (в предположении справедливости H_0) распределение Фишера — Снедекора с m и $(n-m-1)$ степенями свободы.

Значение статистики $F_{m;\,n-m-1}$ сравнивается с критической точкой $f_{\alpha;\,m;\,n-m-1}$, соответствующей выбранному уровню значимости α. Для расчета критической точки $f_{\alpha;\,m;\,n-m-1}$ в Microsoft Excel можно воспользоваться функцией

$$f_{\alpha;k_1;k_2} = \text{F.ОБР.ПХ}(<\alpha>;\ <k_1>;\ <k_2>).$$

В условиях примера 7.1.5 наблюдаемое значение статистики $F_{5;\,46}$ равно 26,23148, что больше критической точки $f_{0{,}05;\,5;\,46} = 2{,}41736$, поэтому гипотеза H_0 отвергается на 5%-ном уровне значимости.

Гипотезу H_0 можно проверить и так: если **значимость F** (рассчитанный уровень значимости гипотезы H_0) оказывается больше принятого уровня значимости, то гипотезу H_0 принимают (и говорят, что уравнение регрессии статистически незначимо, не адекватно), а если **значимость F** оказывается меньше α, гипотезу H_0 отвергают (уравнение значимо, адекватно).

Наблюдаемое значение статистики $F_{\alpha;\,m;\,n-m-1}$ и рассчитанный уровень значимости гипотезы H_0 приводятся в результатах работы программы **«Регрессия»** (рис. 7.1.12) в таблице **«Дисперсионный анализ»** (столбцы **«F»** и **«Значимость F»**).

В условиях примера 7.1.5 **значимость F** равна $2\cdot10^{-13}$ — на уровне значимости $\alpha = 0{,}05$ уравнение следует признать значимым.

Может оказаться так, что результативный признак Y зависит хотя бы от одного из факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, но не от всех

$x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, (в этом случае говорят о значимости модели регрессии в целом и незначимости отдельных факторных признаков).

Проверка гипотезы $H_0^{(j)}: a_j = 0$ (о незначимости j-го факторного признака) при альтернативной гипотезе $H_1^{(j)}: a_j \neq 0$ ($j = 1, 2, \ldots, m$) производится на основе анализа статистики $T_{n-m-1}^{(j)}$, значение которой приводится в результатах работы программы **«Регрессия»** (рис. 7.1.3) в столбце **«t-статистика»** и соответствующей строке выделенной таблицы.

При выполнении гипотезы $H_0^{(j)}$ статистика $T_{n-m-1}^{(j)}$ имеет распределение Стьюдента с $(n - m - 1)$ степенью свободы. Область отклонения гипотезы $H_0^{(j)}$ (на уровне значимости α) такова: $\left|T_{n-m-1}^{(j)}\right| > t_{\alpha; n-m-1}$.

В той же таблице в столбце **«P-значение»** приводятся рассчитанные уровни значимости гипотез $H_0^{(j)}$ — это вероятности

$$p_j = 2\mathbf{P}\{T_{n-m-1}^{(j)} > |t^{(j)}|\}$$

(гипотезу $H_0^{(j)}$ отвергают при альтернативе $H_1^{(j)}$, если $p_j < \alpha$).

В условиях примера 7.1.5 значение статистики $T_{46}^{(1)}$ равно −1,80899, статистики $T_{46}^{(2)}$ — 2,87452, статистики $T_{46}^{(3)}$ — 0,30707, статистики $T_{46}^{(4)}$ — 1,51644, статистики $T_{46}^{(5)}$ — 0,62125. Так как критическая точка $t_{0,05;\,46} = 2{,}01290$, то только гипотеза $H_0^{(2)}: a_2 = 0$ отвергается (оценка $\hat{a}_2$ параметра a_2 значима), а гипотезы $H_0^{(1)}: a_1 = 0, H_0^{(3)}: a_3 = 0, H_0^{(4)}: a_4 = 0, H_0^{(5)}: a_5 = 0$ не отвергаются (оценки $\hat{a}_1, \hat{a}_3, \hat{a}_4, \hat{a}_5$ параметров a_1, a_3, a_4, a_5 незначимы).

Так как $p_1 = 0{,}07699$, $p_2 = 0{,}00611$, $p_3 = 0{,}76018$, $p_4 = 0{,}13625$, $p_5 = 0{,}53750$, то только гипотеза $H_0^{(2)}: a_2 = 0$ отвергается, а гипотезы $H_0^{(1)}: a_1 = 0, H_0^{(3)}: a_3 = 0$, $H_0^{(4)}: a_4 = 0, H_0^{(5)}: a_5 = 0$ не отвергаются.

Эти же гипотезы можно проверить при помощи интервальных оценок параметров уравнения регрессии. Если доверительный интервал (с надежностью $1 - \alpha$) для коэффициента a_j не накрывает число нуль, то это дает основание на уровне значимости α отвергнуть гипотезу $H_0^{(j)}: a_j = 0$.

В условиях примера 7.1.5 только гипотеза $H_0^{(2)}: a_2 = 0$ отвергается, а гипотезы $H_0^{(1)}: a_1 = 0, H_0^{(3)}: a_3 = 0, H_0^{(4)}: a_4 = 0, H_0^{(5)}: a_5 = 0$ не отвергаются, поскольку только интервал (0,0001; 0,0006) не накрывает нуль.

При наличии в уравнении регрессии хотя бы одного незначимого коэффициента модель нельзя считать приемлемой для практического использования.

Чтобы получить «хорошую», приемлемую модель, можно воспользоваться процедурой пошагового исключения незначимых регрессоров.

На каждом шаге этой процедуры из модели исключается тот регрессор, при котором коэффициент незначим, а соответствующая этому коэффициенту величина P-значения является наибольшей (или, иначе, значение модуля соответствующей t-статистики является наименьшим). Процедуру пошагового исключения регрессоров нужно продолжать до тех пор, пока не будет получено значимое уравнение регрессии со значимыми коэффициентами.

Замечание. Если после исключения регрессора уравнение становится статистически незначимым или остается значимым, но его **нормированный *R*-квадрат** значительно уменьшается, то этот регрессор следует «возвратить» в уравнение и исключить очередной регрессор, коэффициент при котором незначим (конечно, при наличии такого регрессора).

Пример 7.1.6. В построенном в примере 7.1.5 уравнении регрессии большинство коэффициентов оказались незначимыми, и такое уравнение нельзя считать приемлемым. Построим на основании данных примера 7.1.5 значимую модель регрессии со всеми значимыми коэффициентами.

Вначале исключим из уравнения регрессор $x^{(3)}$, при котором коэффициент незначим, а соответствующая этому коэффициенту абсолютная величина значения статистики $T_{46}^{(3)}$, равная 0,31, является наименьшей (рассчитанный уровень значимости $p_3 = 0{,}760$ является наибольшим). Оценка линейной функции регрессии будет такой:

$$\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_1 x^{(2)} + \hat{a}_1 x^{(4)} + \hat{a}_1 x^{(5)} =$$
$$= 62{,}93 - 3{,}52x^{(1)} + 0{,}000383x^{(2)} + 0{,}082x^{(4)} + 0{,}206x^{(5)}.$$

стандартная ошибка $s_{ELR} = 4{,}34$, средняя относительная ошибка аппроксимации $\delta \approx 4{,}6\%$, оценка коэффициента множественной линейной корреляции равна 0,86, оценка коэффициента множественной линейной детерминации равна 0,74, оценка нормированного коэффициента множественной линейной детерминации равна 0,72.

Гипотеза H_0 о том, что все параметры при регрессорах одновременно равны нулю, отвергается на 5%-ном уровне значимости, поскольку значимость F (равная $3{,}3 \cdot 10^{-13}$) меньше принятого уровня значимости $\alpha = 0{,}05$.

Так как $p_1 = 0{,}055$, $p_2 = 0{,}004$, $p_4 = 0{,}136$, $p_5 = 0{,}502$, только гипотеза $H_0^{(2)} : a_2 = 0$ отвергается, а гипотезы $H_0^{(1)} : a_1 = 0$, $H_0^{(4)} : a_4 = 0$, $H_0^{(5)} : a_5 = 0$ не отвергаются.

Теперь исключим из уравнения регрессор $x^{(5)}$, при котором коэффициент незначим, а соответствующая этому коэффициенту абсолютная величина значения статистики $T_{46}^{(3)}$, равного 0,68, является наименьшей (рассчитанный уровень значимости $p_3 = 0{,}502$ является наибольшим). Оценка линейной функции регрессии будет такой:

$$\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_1 x^{(2)} + \hat{a}_1 x^{(4)} = 64{,}88 - 2{,}42x^{(1)} + 0{,}000328x^{(2)} + 0{,}079x^{(4)}.$$

стандартная ошибка $s_{ELR} = 4{,}32$, средняя относительная ошибка аппроксимации $\delta \approx 4{,}6\%$, оценка коэффициента множественной линейной корреляции равна 0,86, оценка коэффициента множественной линейной детерминации равна 0,74, оценка нормированного коэффициента множественной линейной детерминации равна 0,72.

Гипотеза H_0 о том, что все параметры при регрессорах одновременно равны нулю, отвергается на 5%-ном уровне значимости, поскольку значимость F (равная $5{,}7 \cdot 10^{-14}$) меньше принятого уровня значимости $\alpha = 0{,}05$.

Так как $p_1 = 0{,}002$, $p_2 = 0{,}001$, $p_4 = 0{,}146$, гипотезы $H_0^{(1)} : a_1 = 0$, $H_0^{(2)} : a_2 = 0$ отвергаются, а гипотеза $H_0^{(4)} : a_4 = 0$ не отвергается.

Исключим регрессор $x^{(4)}$, при котором коэффициент незначим, оценка линейной функции регрессии будет такой:

$$\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_1 x^{(2)} = 73{,}68 - 3{,}22x^{(1)} + 0{,}000349x^{(2)}.$$

стандартная ошибка $s_{ELR} = 4{,}37$, средняя относительная ошибка аппроксимации $\delta \approx 4{,}8\%$, оценка коэффициента множественной линейной корреляции равна 0,85, оценка коэффициента множественной линейной детерминации равна 0,73, оценка нормированного коэффициента множественной линейной детерминации равна 0,71.

Гипотеза H_0 о том, что все параметры при регрессорах одновременно равны нулю, отвергается на 5%-ном уровне значимости, поскольку значимость F (равная $1{,}8 \cdot 10^{-14}$) оказалась меньше принятого уровня значимости $\alpha = 0{,}05$. Так как $p_1 = 1{,}5 \cdot 10^{-8}$, $p_2 = 7{,}9 \cdot 10^{-4}$, гипотезы $H_0^{(1)} : a_1 = 0, H_0^{(2)} : a_2 = 0$ отвергаются.

Результаты пошагового исключения регрессоров сведены в табл. 7.1.6.

Наилучшим уравнением является полученное на четвертом шаге (см. табл. 7.1.6), поскольку и само уравнение, и все его коэффициенты значимы.

Судя по этому уравнению, более 70% дисперсии продолжительности жизни мужчины (Y) связано с линейным влиянием среднего числа детей в семье $x^{(1)}$ и ВВП на душу населения $x^{(2)}$ (так как $\hat{R}^2 = 0{,}73$).

Увеличение среднего числа детей в семье $x^{(1)}$ на единицу [при неизменном значении $x^{(2)}$] сопровождается наибольшим изменением средней продолжительности жизни мужчины (уменьшением на 3,22 года); увеличение $x^{(1)}$ на единицу сопровождается и наибольшим максимально возможным с 95%-ной вероятностью изменением результативного признака (уменьшением средней продолжительности жизни мужчины на 4,18 года), так как 95%-ные интервальные оценки параметров a_1 и a_2 таковы: (−4,18; −2,27) и (0,000153; 0,000544).

Анализ коэффициентов эластичности

$$\hat{e}_{Y|x^{(1)}} = \hat{a}_1 \frac{\overline{x^{(1)}}}{\overline{y}} = -3.22 \cdot \frac{2{,}83}{67{,}48} = -0{,}135, \quad \hat{e}_{Y|x^{(2)}} = \hat{a}_2 \frac{\overline{x^{(2)}}}{\overline{y}} = 0{,}000349 \cdot \frac{8421{,}9}{67{,}48} = 0{,}044$$

показывает, что увеличение среднего числа детей в семье $x^{(1)}$ на 1% (при неизменном значении $x^{(2)}$ сопровождается наибольшим процентным изменением средней продолжительности жизни мужчины — ее уменьшением на 0,135%; увеличение среднего числа детей $x^{(1)}$ на 1% сопровождается и наибольшим максимально возможным с 95%-ной вероятностью процентному изменением средней продолжительности жизни мужчины — ее уменьшением на

$$\left| -4{,}18 \cdot \frac{2{,}83}{67{,}48} \right| = 0{,}175\%. \ \square$$

Точечный прогноз результативного признака Y при значениях факторных признаков, равных $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, определяется как

$$\hat{Y}_x = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x^{(m)}.$$

В условиях примера 7.1.1 рассчитанное по уравнению число $\hat{Y}_i$ — это точечная оценка генерального среднего значения продолжительности жизни мужчины при условии, что значения факторных признаков (среднего числа детей в семье $x^{(1)}$ и ВВП на душу населения $x^{(2)}$) зафиксированы на каких-то уровнях, а именно $x^{(1)} = x_i^{(1)}, x^{(2)} = x_i^{(2)}$. Например, точечная оценка генерального среднего значения продолжительности жизни мужчины при значениях регрессоров на первом объекте (в Австралии) равна

Таблица 7.1.6

Пошаговое исключение незначимых регрессоров

Шаг	Уравнение, интервальные оценки коэффициентов, наблюдаемые значения t-статистики, P-значения	$\hat{R}^2$	$\tilde{R}^2$	s_{ELR}	δ	F	$f_{0,05;\,m;\,n-m-1}$
1	$\hat{Y}_{\mathbf{x}} = \underset{(47,56;\,77,02)}{62,29} - \underset{(-7,13;\,0,38)}{3,37}x^{(1)} + \underset{(0,000112;\,0,000637)}{0,000375}x^{(2)} + \underset{(-0,001193;\,0,001623)}{0,000215}x^{(3)} + \underset{(-0,029;\,0,205)}{0,088}x^{(4)} + \underset{(-0,432;\,0,818)}{0,193}x^{(5)}$ $(t_{0,05;\,46}=2,013)$ $-1,81$ $2,87$ $0,31$ $1,51$ $0,62$ $0,076$ $0,006$ $0,760$ $0,136$ $0,538$	0,74	0,71	4,38	4,5%	26,23	2,417
2	$\hat{Y}_{\mathbf{x}} = \underset{(48,95;\,76,91)}{62,92} - \underset{(-7,11;\,0,078)}{3,52}x^{(1)} + \underset{(0,000128;\,0,000638)}{0,000383}x^{(2)} + \underset{(-0,026;\,0,191)}{0,082}x^{(4)} + \underset{(-0,407;\,0,818)}{0,206}x^{(5)}$ $(t_{0,05;\,47}=2,011)$ $-1,97$ $3,02$ $1,51$ $0,676$ $0,055$ $0,004$ $0,136$ $0,502$	0,74	0,72	4,34	4,6%	33,41	2,570
3	$\hat{Y}_{\mathbf{x}} = \underset{(52,24;\,77,52)}{64,88} - \underset{(-3,86;\,-0,96)}{2,42}x^{(1)} + \underset{(0,000133;\,0,000524)}{0,000328}x^{(2)} + \underset{(-0,029;\,0,187)}{0,079}x^{(4)}$ $(t_{0,05;\,48}=2,011)$ $-3,35$ $3,38$ $1,47$ $0,002$ $0,001$ $0,146$	0,74	0,72	4,32	4,6%	44,90	2,798
4	$\hat{Y}_{\mathbf{x}} = \underset{(69,59;\,77,77)}{73,68} - \underset{(-4,18;\,-2,27)}{3,22}x^{(1)} + \underset{(0,000153;\,0,000544)}{0,000349}x^{(2)}$ $(t_{0,05;\,49}=2,009)$ $-6,77$ $3,58$ $0,000000015$ $0,00079$	0,73	0,71	4,37	4,8%	64,70	3,187

$$\hat{Y}_1 = 73{,}68 - 3{,}22 \cdot 1{,}9 + 0{,}000349 \cdot 16\,848 = 73{,}43,$$

а реальная продолжительность жизни мужчины в Австралии равна $Y_1 = 74$, остаток $Y_1 - \hat{Y}_1 = 0{,}57$. В тех странах, в которых остатки $Y_i - \hat{Y}_i$ положительны, продолжительность жизни мужчины выше среднего уровня, а в тех странах, где остатки отрицательны — ниже среднего уровня. Так, например, в Австралии $Y_1 - \hat{Y}_1 = 0{,}57$, а в Австрии $Y_2 - \hat{Y}_2 = -2{,}25$.

Для учета влияния на результативный признак качественных (а не количественных) факторов используются так называемые **фиктивные переменные** (или **индикаторы**). Рассмотрим конкретный пример.

Пример 7.1.7. Табл. 7.1.7 содержит данные об объемах продаж (в тыс. руб.), совершенных 12 менеджерами по продажам: 5 женщинами и 7 мужчинами, и средними баллами этих менеджеров, полученными за время обучения в вузе (по диплому).

Таблица 7.1.7

Данные об объемах продаж менеджеров и средних баллах их дипломов

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12
Объем продаж, Y	46	98	100	121	145	168	190	210	223	281	292	314
Средний балл диплома, $x^{(1)}$	3,47	3,79	3,79	3,62	3,83	3,93	3,67	3,78	3,91	3,99	4,12	4,82
Пол, $x^{(2)}$	0 (ж)	0 (ж)	0 (ж)	0 (ж)	1 (м)	0 (ж)	1 (м)	1 (м)	1 (м)	1 (м)	1 (м)	1 (м)

Выясним, как объем продаж зависит от пола и от среднего балла диплома о высшем образовании.

Введем фиктивную переменную $x^{(2)}$, равную единице для мужчин и нулю для женщин.

Результатом множественного линейного регрессионного анализа, проведенного с помощью Microsoft Excel, стало уравнение

$$\hat{Y}_x = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_1 x^{(2)} = -386{,}4 + 132{,}5x^{(1)} + 90{,}5x^{(2)}.$$

Это уравнение и оба коэффициента при факторных признаках значимы, нормированный коэффициент детерминации равен 0,81.

Анализ коэффициентов уравнения показывает, что объем продаж менеджера зависит от его оценок в вузе и пола. При этом если бы средний балл диплома менеджера был на единицу выше, то его объем продаж был бы в среднем выше на $\hat{a}_1 = 132{,}5$ тыс. руб.

Смысл же коэффициента при фиктивной переменной таков: объем продаж мужчин в среднем на 90,5 тыс. руб. выше, чем у женщин с таким же средним баллом в дипломе. ❑

Задачи

429. Известны данные о курсе доллара $x^{(1)}$, фондовом индексе $x^{(2)}$ и котировке акции Y за последние 10 дней:

$x^{(1)}$	68,75	68,70	68,54	68,90	68,88	68,35	67,98	68,10	68,05	67,90
$x^{(2)}$	4,0	4,2	4,7	5,1	4,9	4,6	4,8	4,3	4,4	4,5
Y	100	112	108	106	103	101	100	103	102	100

Постройте модель линейной регрессии, проведите исключение незначимых регрессоров, интерпретируйте коэффициенты линейной функции регрессии и коэффициенты эластичности, а затем спрогнозируйте котировку акции, если курс доллара составит 67,00 руб., а значение фондового индекса окажется при этом равным 5,0.

430. По данным примера 1.1.1 (файл «Кредиты.xlsx»), предварительно обработав выбросы (в задаче **32**) и пропущенные значения (в задачах **34—40**), удалив повторяющиеся строки (в задаче **42**) и преобразовав качественные признаки в индикаторы, оцените линейную функцию регрессии размера кредита на срок кредита (краткосрочный или долгосрочный), величину кредитного рейтинга, годовой доход, стаж на текущем месте работы, наличие недвижимости (в собственности, в ипотеке или в аренде), цель кредита и, срок кредитной истории. Проведите процедуру пошагового исключения незначимых факторных признаков. Проведите интерпретацию оценок коэффициентов регрессии и коэффициентов эластичности.

§ 7.2. Модели классификации с обучением

7.2.1. Логистическая регрессия

Модели регрессии используются для прогнозирования значений непрерывных признаков: зная набор значений факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, с помощью модели регрессии можно дать точечный прогноз значения результативного признака Y при данных значениях $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, интервальный прогноз генерального среднего значения результативного признака Y при данных значениях $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, интервальный прогноз значения результативного признака Y при известных значениях $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$.

При этом подразумевается, что признак Y является количественным (вообще говоря, непрерывным).

В задачах классификации с обучением требуется, зная набор значений факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, давать прогноз значению дискретного признака Y, который может принимать k значений, соответствующих некоторым заранее известным классам.

Например, когда пользователь смартфона или планшета использует систему распознавания рукописного ввода и пальцем рисует на панели некоторый символ, эта система решает задачу классификации: это буква «А» или это буква «Б» или это буква «В» и т. д.?

В задачах бинарной классификации классов ровно два.

Бинарная классификация используется очень широко. Например, принимая решение о выдаче кредита или об отказе в выдаче кредита, банк решает задачу кредитного скоринга: по известным из заявления о предоставлении кредита значениям признаков $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ прогнозирует значение *бинарного* (т. е. принимающего два значения) признака Y, равного нулю для «хороших» клиентов (которые вернут кредит) и единице для «плохих».

Страховые компании по каждому договору страхования транспортных средств на основании информации о страхователе и его автомобиле прогнозируют, произойдет ли по данному договору страховой случай ($Y = 1$) или не произойдет ($Y = 0$).

Телекоммуникационные компании постоянно решают задачу прогнозирования оттока клиентов: для каждого клиента на основа-

нии известных данных об использовании клиентом голосовых услуг и услуг передачи данных, а также о выставленных счетах и их оплате оператор связи прогнозирует значение признака Y, равного нулю для тех клиентов, кто не собирается менять сотового оператора, и единице для тех, кто планирует перейти к конкурентам.

Промышленные предприятия решают задачу мониторинга и превентивного ремонта оборудования: зная показания нескольких датчиков, подключенных к станку, за последнее время, требуется спрогнозировать, будет ли станок в ближайшую неделю (месяц, день) работать в штатном режиме ($Y = 0$) или выйдет из строя ($Y = 1$).

Торговые организации решают задачу прогнозирования отклика на рекламу: купит ли клиент (относительно которого известны прошлые покупки и ряд других признаков) рекламируемый товар, если ему сделать специальное предложение, ($Y = 1$) или не купит ($Y = 0$).

Замечание. Как правило, нулем кодируют обычное, более часто встречающееся и не требующее специальной реакции, событие (клиент вернет кредит; по договору страхования не наступит в страховой случай; клиент сотового оператора не уйдет к конкуренту; станок будет работать в штатном режиме; покупатель не отреагирует на специальное предложение), а единицей — событие, требующее особого внимания (клиент не вернет кредит; по договору страхования наступит в страховой случай; клиент сотового оператора уйдет к конкуренту; станок выйдет из строя; покупатель отреагирует на специальное предложение).

Пример 7.2.1. Для двадцати заемщиков банка, у которых закончился срок погашения кредита, известны значения признака x — срока, в течение которого клиент обслуживается банком (в мес.) и значения признака Y, равного нулю для тех клиентов, которые полностью погасили кредиты, и единице для тех, которые не выполнили условия кредитного договора (табл. 7.2.1).

Таблица 7.2.1

Данные о сроке обслуживания клиентов и результатах погашения кредитов

x	39	36	60	21	27	12	51	57	54	42	18	9	48	66	6	30	24	33	15	21
Y	1	0	1	1	1	0	1	1	1	0	0	0	1	1	0	0	0	1	0	0

Попробуем использовать для решения задачи классификации заемщиков на «хороших» и «плохих» модель парной линейной регрессии $Y = a_0 + a_1 x + \varepsilon$.

Введем исходные данные о сроке обслуживания клиентов и результатах погашения кредитов, представленные в табл. 7.2.1, на рабочий лист Microsoft Excel и запустим программу «Регрессия» надстройки «Анализ данных» Microsoft Excel, отметив флажок «График подбора».

В результате получим оценки параметров $\hat{a}_0 = -0{,}15393527$, $\hat{a}_1 = 0{,}01954963$. График подбора, построенный программой «Регрессия», представлен на рис. 7.2.1.

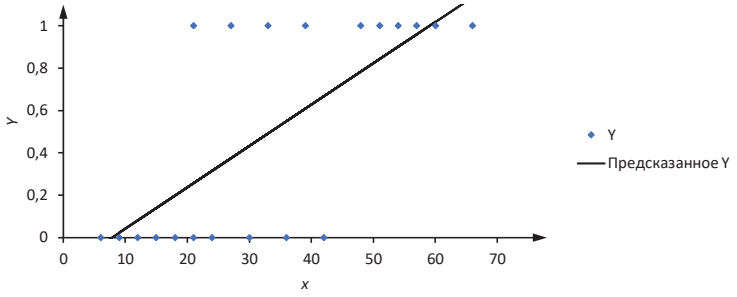


Рис. 7.2.1. *График подбора в примере 7.2.1*

На графике подбора видно, что большинство реальных точек находятся далеко от линии регрессии, поэтому такую модель нецелесообразно использовать для бинарной классификации. ❑

Основной недостаток модели линейной регрессии с точки зрения ее использования для бинарной классификации состоит в том, что в модели линейной регрессии результативный признак Y может принимать любые вещественные значения от $-\infty$ до $+\infty$, и это не поддается никакой разумной интерпретации, учитывая, что в задачах бинарной классификации на самом деле Y — это индикатор события, который может принимать всего два значения: нуль и единицу. Кроме того, очевидно, что так как $Y \in \{0;1\}$, этот признак нельзя считать распределенным по нормальному закону, как это предполагается в модели линейной регрессии.

Чтобы преодолеть этот недостаток, можно попытаться спрогнозировать не номер класса, а вероятность отнесения наблюдения к первому классу.

Пусть $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ — это факторные признаки (независимые переменные), Y — бинарный результативный признак, который имеет смысл индикатора события и может принимать одно из двух значений: 0 или 1.

Пусть $\mathbf{P}\{Y_i = 1\}$ — вероятность того, что i-е наблюдение относится к первому классу, $\mathbf{P}\{Y_i = 0\} = 1 - \mathbf{P}\{Y_i = 1\}$ — вероятность того, что i-е наблюдение относится к нулевому классу. Очевидно, математическое ожидание индикатора события по формуле (3.2.7) равно вероятности этого события, т. е. $\mathbf{E}(Y_i) = \mathbf{P}\{Y_i = 1\}$.

Предположим теперь, что условная вероятность $\mathbf{P}\left\{Y_i = 1 \middle| x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\}$ того, что i-е наблюдение относится к первому классу при данных значениях признаков $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}$ представляет собой некоторую функцию, зависящую от $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}$, а также от некоторых параметров $a_0, a_1, a_2, \ldots, a_m$:

$$\mathbf{P}\left\{Y_i = 1 \middle| x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\} = p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right). \qquad (7.2.1)$$

При этом, очевидно,

$$\mathbf{P}\left\{Y_i = 0 \middle| x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\} = 1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right).$$

При этом

$$\mathbf{P}\left\{Y_i = y_i \middle| x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\} = \begin{cases} \mathbf{P}\left\{Y_i = 1 \mid x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\}, & y_i = 1, \\ \mathbf{P}\left\{Y_i = 0 \middle| x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\}, & y_i = 0 \end{cases} =$$

$$= \begin{cases} p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right), & y_i = 1, \\ 1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right), & y_i = 0 \end{cases} =$$

$$= \begin{cases} \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^1, & y_i = 1, \\ \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^1, & y_i = 0 \end{cases} =$$

$$= \begin{cases} \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^1 \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^0, & y_i = 1, \\ \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^0 \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^1, & y_i = 0 \end{cases} =$$

$$= \begin{cases} \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{y_i} \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{1-y_i}, & y_i = 1, \\ \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{y_i} \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{1-y_i}, & y_i = 0 \end{cases} =$$

$$= \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{y_i} \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{1-y_i}.$$

Составим функцию правдоподобия [см. формулу (5.2.9)], равную вероятности получения и м е н н о набора $y_1, y_2, \ldots, y_n$ при извлечении выборки объемом n из генеральной совокупности, подчиняющейся закону распределения (7.2.1):

$$L\left(a_0, a_1, a_2, \ldots, a_m\right) = \prod_{i=1}^{n} \mathbf{P}\left\{Y_i = y_i \mid x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}\right\} = \tag{7.2.2}$$

$$= \prod_{i=1}^{n} \left[p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{y_i} \left[1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)\right]^{1-y_i}.$$

Если, например, $p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right) = p = \text{const}$, то

$$L(p) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i},$$

$$l(p) = \ln L(p) = \ln \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i} = \sum_{i=1}^{n} \ln\left(p^{y_i} (1-p)^{1-y_i}\right) = \sum_{i=1}^{n} \ln p^{y_i} + \sum_{i=1}^{n} \ln (1-p)^{1-y_i} =$$

$$= \sum_{i=1}^{n} y_i \ln p + \sum_{i=1}^{n} (1 - y_i) \ln (1-p), \tag{7.2.3}$$

$$l'(p) = \frac{d}{dp}\left\{\sum_{i=1}^{n} \left[y_i \ln p + (1 - y_i) \ln (1-p)\right]\right\} = \sum_{i=1}^{n} \left[\frac{y_i}{p} - \frac{1 - y_i}{1-p}\right] = \sum_{i=1}^{n} \frac{y_i (1-p) - (1-y_i) p}{p(1-p)} =$$

$$= \sum_{i=1}^{n} \frac{y_i - y_i p - p + y_i p}{p(1-p)} = \frac{1}{p(1-p)} \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p\right) = \frac{1}{p(1-p)} \left(\sum_{i=1}^{n} y_i - np\right),$$

$$l''(p) = \frac{d}{dp}\left\{\frac{1}{p(1-p)} \left(\sum_{i=1}^{n} y_i - np\right)\right\} = -\frac{n}{p(1-p)} < 0.$$

Поэтому функция правдоподобия в случае $p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right) =$ $= p = \text{const}$ достигает максимума при

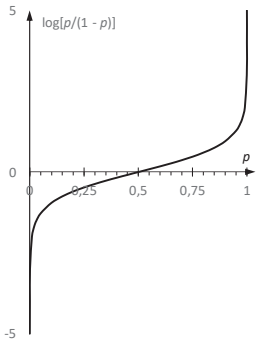
$$\hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}. \tag{7.2.4}$$

Рассмотрим *модель логистической регрессии*:

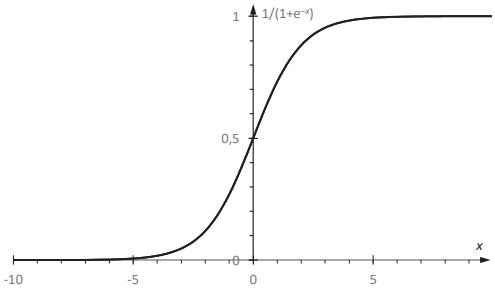
$$\log \frac{p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)}{1 - p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right)} = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_m x_i^{(m)}. \tag{7.2.5}$$

Функция $\text{logit}(p) = \ln \dfrac{p}{1-p}$ называется *логит-преобразованием*, ее график представлен на рис. 7.2.2, *а*. При изменении p от 0 до 1 функция $\text{logit}(p)$ возрастает от $-\infty$ до $+\infty$, поэтому в уравнении

$$U_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_m x_i^{(m)}$$



а)



б)

Рис. 7.2.2. *Графики логит-преобразования* $\operatorname{logit}(p)=\ln\dfrac{p}{1-p}$ (*а*)

и логистической функции $\operatorname{logit}^{-1}(x)=\dfrac{\mathrm{e}^x}{1+\mathrm{e}^x}=\dfrac{1}{1+\mathrm{e}^{-x}}$ (*б*)

левая часть

$$U_i=\log\frac{p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}$$

может быть любым числом от $-\infty$ до $+\infty$.

Хотя левая часть в модели (7.2.5) теперь может быть любым вещественным числом, ее закон распределения отличается от нормального, и метод наименьших квадратов для оценки параметров $a_0,a_1,a_2,\ldots,a_m$ не подойдет.

Чтобы оценить параметры модели (7.2.5), лучше воспользоваться методом максимального правдоподобия, т. е. максимизировать функцию правдоподобия (7.2.2), предварительно выразив p из уравнения (7.2.5):

$$\log\frac{p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}=a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}\Leftrightarrow$$

$$\Leftrightarrow\frac{p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}=\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}\Leftrightarrow$$

$$\Leftrightarrow-\frac{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)-1}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}=\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}\Leftrightarrow$$

$$\Leftrightarrow -1+\frac{1}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}=\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}\Leftrightarrow$$

$$\Leftrightarrow \frac{1}{1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)}=1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}\Leftrightarrow$$

$$\Leftrightarrow 1-p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)=\frac{1}{1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}}\Leftrightarrow$$

$$\Leftrightarrow p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)=1-\frac{1}{1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}}=$$

$$=\frac{1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}-1}{1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}}=\frac{\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}}{1+\mathrm{e}^{a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}}}=\frac{1}{1+\mathrm{e}^{-\left(a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}\right)}}.$$

Функция

$$\mathrm{logit}^{-1}(x)=\frac{\mathrm{e}^x}{1+\mathrm{e}^x}=\frac{1}{1+\mathrm{e}^{-x}},$$

обратная к логит-преобразованию, называется *логистической функцией* или *сигмоидом* (ее график представлен на рис. 7.2.2, *б*). При изменении x от $-\infty$ до $+\infty$ $\mathrm{logit}^{-1}(x)$ возрастает от 0 до 1.

Максимум функции правдоподобия (7.2.2) или ее логарифма аналитически найти не удастся, но можно использовать численные методы оптимизации. Вычислительно проще, конечно, максимизировать не функцию функции правдоподобия (7.2.2), а ее логарифм (7.2.3), в котором p представляет собой функцию

$$p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)=\frac{1}{1+\mathrm{e}^{-\left(a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}\right)}}.$$

В пакете Microsoft Excel есть надстройка «Поиск решения», которая позволяет решать задачи оптимизации. Если в правой части вкладки «Данные» есть кнопка «Поиск решения», то это означает, что данная надстройка в Microsoft Excel установлена. Если же кнопка «Поиск решения» отсутствует, то нужно выбрать в окне управления параметрами Excel (меню «Файл» — «Параметры») пункт «Надстройки» (мы обсуждали управление надстройками в **п. 1.2.3**, см. рис. 7.2.2, *а*), далее в нижней строке «Управление» выбрать «Надстройки Excel», нажать кнопку «Перейти...», в появившемся окне (рис. 7.2.2, *б*) отметить флажок «Поиск решения» и нажать «ОК» — после этого кнопка «Поиск решения» должна появиться на вкладке «Данные».

Обсудим содержательную интерпретацию коэффициентов модели логистической регрессии, которая следует из записи модели в форме

$$\ln\frac{p}{1-p}=a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_mx^{(m)}:$$

коэффициент a_j ($j=1, 2, \ldots, m$) показывает, на сколько единиц увеличится величина

$$\ln\frac{p}{1-p}$$

при увеличении $x^{(j)}$ на единицу.

Поскольку функция

$$\operatorname{logit}(p)=\ln\frac{p}{1-p}$$

является монотонно возрастающей, то положительность коэффициента a_j указывает на то, что при увеличении $x^{(j)}$ вероятность отнесения наблюдения к первому классу возрастает, а отрицательность a_j — на то, что при увеличении x вероятность отнесения наблюдения к первому классу возрастает.

Если же записать модель логистической регрессии в эквивалентной форме

$$\frac{p}{1-p}=\mathrm{e}^{a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_mx^{(m)}},$$

то можно заметить, что при увеличении $x^{(j)}$ на единицу (при неизменных остальных факторных признаках) величина $p/(1-p)$ увеличится в e^{a_j} раз:

$$\frac{\dfrac{p\left(x^{(j)}+1\right)}{1-p\left(x^{(j)}+1\right)}}{\dfrac{p\left(x^{(j)}\right)}{1-p\left(x^{(j)}\right)}}=\frac{\mathrm{e}^{a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_j\left(x^{(j)}+1\right)+\cdots+a_mx^{(m)}}}{\mathrm{e}^{a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_jx^{(j)}+\cdots+a_mx^{(m)}}}=\frac{\mathrm{e}^{a_j}\,\mathrm{e}^{a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_jx^{(j)}+\cdots+a_mx^{(m)}}}{\mathrm{e}^{a_0+a_1x^{(1)}+a_2x^{(2)}+\cdots+a_jx^{(j)}+\cdots+a_mx^{(m)}}}=\mathrm{e}^{a_j}.$$

Если p — вероятность некоторого события, то величина $p/(1-p)$ называется *шансом* данного события. Если вероятность представляет собой число из отрезка $[0;1]$, то шансы заполняют луч $[0;+\infty)$. При этом шансы, близкие к нулю (бесконечности), соответствуют вероятностям, близким к нулю (соответственно, единице). Например, вероятности, равной $0{,}1$, соответствует шанс $0{,}1/(1-0{,}1)=$ $=1/9$; вероятности $0{,}25$ — шанс $0{,}25/(1-0{,}25)=1/3$, вероятности $0{,}75$ — шанс $0{,}75/(1-0{,}75)=3$, вероятности $0{,}9$ — шанс $0{,}9/(1-0{,}9)=9$.

Замечание. Сегодня обычно степень возможности наступления событий анализируют в терминах вероятностей. Однако в некоторых сферах такой анализ принято проводить в терминах шансов — например, именно шансами оперируют букмекеры, принимающие ставки на спортивные соревнования.

Итак, теперь можно окончательно прояснить смысл коэффициентов логистической регрессии: при увеличении $x^{(j)}$ на единицу (при неизменных остальных факторных признаках) шанс отнесения наблюдения к первому классу увеличивается в e^{a_j} раз.

Чтобы спрогнозировать, к какому классу относится объект с известными значениями факторных признаков $x^{(1)},x^{(2)},\ldots,x^{(m)}$, естественно использовать **отсечение по порогу,** относя объект к нулевому классу при $p\left(x^{(1)},x^{(2)},\ldots,x^{(m)}\right)\leqslant 0{,}5$ и к первому классу при $p\left(x^{(1)},x^{(2)},\ldots,x^{(m)}\right)>0{,}5$.

Для проверки гипотезы $H_0\colon a_1=a_2=\cdots=a_m=0$ используется критерий отношения правдоподобия, основанный на том, что удвоенная разность максимального значения логарифма функции правдоподобия (7.2.3) при

$$p\left(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(m)};a_0,a_1,a_2,\ldots,a_m\right)=\frac{1}{1+\mathrm{e}^{-\left(a_0+a_1x_i^{(1)}+a_2x_i^{(2)}+\cdots+a_mx_i^{(m)}\right)}}$$

и значения логарифма функции правдоподобия при

$$p\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}; a_0, a_1, a_2, \ldots, a_m\right) = \hat{p} = \text{const}$$

в случае справедливости гипотезы H_0 распределена по закону χ^2_m.

Аналогичный критерий можно использовать для сравнения двух моделей — с меньшим количеством факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ (сокращенной модели) и с большим количеством факторных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}, x^{(p+1)}, x^{(p+2)}, \ldots, x^{(m)}$ (расширенной). Для проверки гипотезы $H_0: a_{p+1} = a_{p+2} = \cdots = a_m = 0$ вычисляется отношение правдоподобия

$$2\left(\ln L^*_{\text{расшир.}} - \ln L^*_{\text{сокр.}}\right) =$$

$$= 2\left(\ln L^*\left(a_0, a_1, a_2, \ldots, a_p, a_{p+1}, a_{p+2}, \ldots, a_m\right) - \ln L^*\left(a_0, a_1, a_2, \ldots, a_p\right)\right),$$

которая при справедливости гипотезы $H_0: a_{p+1} = a_{p+2} = \cdots = a_m = 0$ распределена по закону χ^2_{m-p}.

Пример 7.2.2. В условиях примера 7.2.1 оценим параметры a_0 и a_1 модели логистической регрессии

$$p(x_i; a_0, a_1) = \frac{1}{1 + e^{-(a_0 + a_1 x_i)}},$$

которую используем для прогнозирования вероятности того, что заемщик, который обслуживается в банке x мес., не вернет кредит.

Введем исходные данные о сроке обслуживания клиентов x и результатах погашения кредитов Y из табл. 7.2.1 в ячейки A1:B21 рабочего листа Microsoft Excel (рис. 7.2.3, *а*).

Ячейки E1 и E2 отведем соответственно под оценки параметров a_0 и a_1. В ячейке E5 рассчитаем оценку (7.2.4) вероятности невозврата кредита.

В ячейках G2:G21 рассчитаем для каждого наблюдения значения функции

$$u(x_i) = a_0 + a_1 x_i,$$

в ячейках H2:H21 — значения

$$p(x_i; a_0, a_1) = \frac{1}{1 + e^{-(a_0 + a_1 x_i)}} = \frac{1}{1 + e^{-u(x_i)}}.$$

Теперь можно вычислить в ячейках I2:I21 выражения

$$y_i \ln p(x_i; a_0, a_1)$$

из первой суммы в формуле (7.2.3), а в ячейках J2:J21 — произведения

$$(1 - y_i)\ln\left[1 - p(x_i; a_0, a_1)\right]$$

из второй суммы в (7.2.3), а затем вычислить значение логарифма функции правдоподобия в ячейке O1.

Если вероятность возврата кредита $p(x_i; a_0, a_1)$ не зависит от x_i, т. е. если $a_1 = 0$, то

$$p(x_i; a_0, a_1) = \frac{1}{1 + e^{-(a_0 + 0 x_i)}} = \frac{1}{1 + e^{-a_0}} = \text{const}.$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	x	Y		a0			u(x_i)=a0+a1*x_i	p(x_i)=1/(1+exp(u_i)	y_i*ln(p(x_i)	(1-y_i)*ln(1-p(x_i))	y_i*ln(p^)	(1-y_i)*ln(1-p^)		ln L	=СУММ(I2:J21)
2	39	1		a1			=$E$1+$E$2*A2	=1/(1+EXP(-G2))	=B2*LN(H2)	=(1-B2)*LN(1-H2)	=B2*LN($E$4)	=(1-B2)*LN(1-$E$4)		ln L(при p = const)	=СУММ(K2:L21)
3	36	0					=$E$1+$E$2*A3	=1/(1+EXP(-G3))	=B3*LN(H3)	=(1-B3)*LN(1-H3)	=B3*LN($E$4)	=(1-B3)*LN(1-$E$4)			
4	60	1		p^	=СРЗНАЧ(B2:B21)		=$E$1+$E$2*A4	=1/(1+EXP(-G4))	=B4*LN(H4)	=(1-B4)*LN(1-H4)	=B4*LN($E$4)	=(1-B4)*LN(1-$E$4)		Отношение правдоподобия	=2*(O1-O2)
5	21	1					=$E$1+$E$2*A5	=1/(1+EXP(-G5))	=B5*LN(H5)	=(1-B5)*LN(1-H5)	=B5*LN($E$4)	=(1-B5)*LN(1-$E$4)		Число степеней свободы	1
6	27	1					=$E$1+$E$2*A6	=1/(1+EXP(-G6))	=B6*LN(H6)	=(1-B6)*LN(1-H6)	=B6*LN($E$4)	=(1-B6)*LN(1-$E$4)		P-значение	=1-ХИ2.РАСП(O4;O5;ИСТИНА)
7	12	0					=$E$1+$E$2*A7	=1/(1+EXP(-G7))	=B7*LN(H7)	=(1-B7)*LN(1-H7)	=B7*LN($E$4)	=(1-B7)*LN(1-$E$4)			
8	51	1					=$E$1+$E$2*A8	=1/(1+EXP(-G8))	=B8*LN(H8)	=(1-B8)*LN(1-H8)	=B8*LN($E$4)	=(1-B8)*LN(1-$E$4)		x	
9	57	1					=$E$1+$E$2*A9	=1/(1+EXP(-G9))	=B9*LN(H9)	=(1-B9)*LN(1-H9)	=B9*LN($E$4)	=(1-B9)*LN(1-$E$4)		P^{Y = 1|x}	=1/(1+EXP(-E1-E2*O8))
⋮	⋮	⋮					⋮	⋮	⋮	⋮	⋮	⋮			
21	21	0					=$E$1+$E$2*A21	=1/(1+EXP(-G21))	=B9*LN(H21)	=(1-B21)*LN(1-H21)	=B21*LN($E$4)	=(1-B21)*LN(1-$E$4)			

а)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	x	Y		a0	-4,07771343		u(x_i)=a0+a1*x_i	p(x_i)=1/(1+exp(u_i)	y_i*ln(p(x_i)	(1-y_i)*ln(1-p(x_i))	y_i*ln(p^)	(1-y_i)*ln(1-p^)		ln L	-8,029878464
2	39	1		a1	0,125387119		0,812384211	0,6926173	-0,36727762	0	-0,6931472	0		ln L(при p = const)	-13,86294361
3	36	0					0,436222854	0,6073586	0	-0,934858667	0	-0, 6931472			
4	60	1		p^	0,5		3,445513712	0,9690971	-0,0313905	0	-0,6931472	0		Отношение правдоподобия	11,66613029
5	21	1					-1,444583932	0,1908365	-1,65633822	0	-0,6931472	0		Число степеней свободы	1
6	27	1					-0,692261217	0,3335302	-1,09802173	0	-0,6931472	0		P-значение	0,000636483
7	12	0					-2,573068003	0,070892	0	-0,07353025	0	-0, 6931472			
8	51	1					2,31702964	0,9102776	-0,09400563	0	-0,6931472	0		x	6
9	57	1					3,069352354	0,9556107	-0,04540466	0	-0,6931472	0		P^{Y = 1|x}	0,034710336
10	54	1					2,693190997	0,9366237	-0,06547372	0	-0,6931472	0			
11	42	0					1,188545569	0,7664808	0	-1,454491147	0	-0, 6931472			
12	18	0					-1,820745289	0,1393445	0	-0,150060934	0	-0, 6931472			
13	9	0					-2,949229361	0,0497729	0	-0,05105432	0	-0, 6931472			
14	48	1					1,940868283	0,8744475	-0,13416302	0	-0,6931472	0			
15	66	1					4,197836426	0,9851944	-0,01491625	0	-0,6931472	0			
16	6	0					-3,325390718	0,0347103	0	-0,035327053	0	-0, 6931472			
17	30	0					-0,31609986	0,4216265	0	-0,547535485	0	-0, 6931472			
18	24	0					-1,068422575	0,2557032	0	-0,295315376	0	-0, 6931472			
19	33	1					0,060061497	0,5150109	-0,66356729	0	-0,6931472	0			
20	15	0					-2,196906646	0,1000286	0	-0,105392313	0	-0, 6931472			
21	21	0					-1,444583932	0,1908365	0	-0,211754286	0	-0, 6931472			

б)

Рис. 7.2.3. *Оценка параметров логистической регрессии и проверка гипотезы о ее значимости в примере 7.2.2: формулы Microsoft Excel (а); результаты расчетов (б)*

Для случая $a_1 = 0$ выражения

$$y_i \ln \hat{p}$$

из первой суммы в формуле (7.2.3) и выражения

$$(1 - y_i)\ln[1 - \hat{p}]$$

из второй суммы в (7.2.3), в которые подставлена оценка (7.2.4) из E5, вычислим соответственно в ячейках K2:K21 и L2:L21, а в ячейке O2 рассчитаем по формуле (7.2.3) максимальное значение логарифма функции правдоподобия для случая $p = \text{const}$.

Далее нам нужно решить задачу

$$\ln L(a_0, a_1) \to \max,$$
$$a_0, a_1 \in \mathbb{R}.$$

Поскольку целевая функция $\ln L(a_0, a_1)$ является нелинейной, в данной задаче может быть несколько локальных максимумов, и поскольку нас интересует глобальный максимум функции $\ln L(a_0, a_1)$, который, очевидно, не меньше максимума $\ln L(p = \text{const})$, целесообразно добавить в данную оптимизационную задачу ограничение $\ln L(a_0, a_1) \geqslant \ln L(p = \text{const})$:

$$\ln L(a_0, a_1) \to \max,$$
$$\ln L(a_0, a_1) \geqslant \ln L(p = \text{const}), \qquad (7.2.6)$$
$$a_0, a_1 \in \mathbb{R}.$$

Воспользуемся для решения задачи (7.2.6) надстройкой «Поиск решения» пакета Microsoft Excel. Нажмем кнопку «Поиск решения», расположенную в правой части вкладки «Данные» и в появившемся окне ввода параметров (рис. 7.2.4) укажем, что требуется оптимизировать целевую функцию $\ln L(a_0, a_1)$, расположенную в ячейке $O$1, изменяя ячейки переменных $E$1:$E$2, которые соответствуют параметрам a_0 и a_1. Нажав кнопку «Добавить», добавим ограничение $O$1 >= $O$2 [т. е. $\ln L(a_0, a_1) \geqslant \ln L(p = \text{const})$].

Убедимся, что флажок «Сделать переменные без ограничений неотрицательными» очищен — ведь a_0 и a_1 могут быть произвольными вещественными числами, в том числе и отрицательными.

Выберем в качестве метода решения вариант «Поиск решения нелинейных задач методом ОПГ», чтобы воспользоваться методом обобщенного приведенного градиента (ОПГ), наиболее подходящим для данной задачи. Через короткое время после нажатия кнопки «Найти решение» появится окно, сообщающее, что решение найдено; в этом окне достаточно нажать кнопку «ОК».

Теперь в ячейках E1 и E2 появились оценки параметров $\hat{a}_0 = -4{,}07771343$ и $\hat{a}_1 = 0{,}125387119$. Таким образом, модель имеет вид

$$\ln \frac{p}{1-p} = -4{,}07771343 + 0{,}125387119\,x$$

или

$$p = \frac{1}{1 + \mathrm{e}^{4{,}07771343 - 0{,}125387119x}}.$$

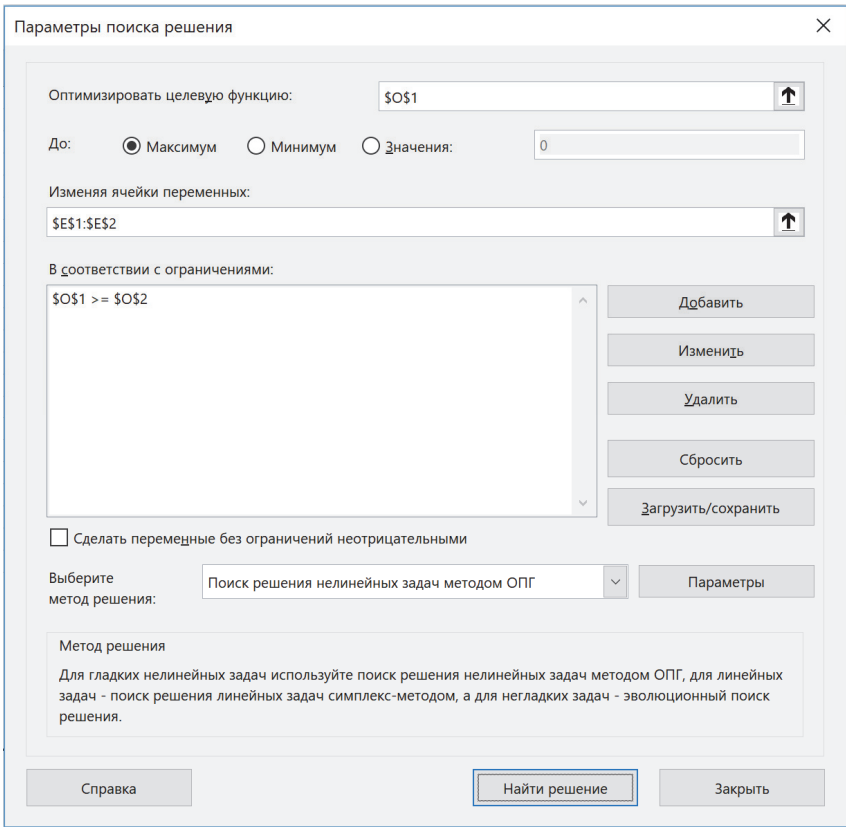


Рис. 7.2.4. *Окно ввода параметров надстройки «Поиск решения»*

Отношение правдоподобия вычислим в ячейке O4. Так как в модели всего один регрессор, то в ячейку O5 введем число степеней свободы, равное единице, и вычислим в ячейке O6 наблюдаемый уровень значимости гипотезы H_0: $a_1 = 0$. Поскольку он оказался равен 0,000636483, есть основания отвергнуть гипотезу H_0 и считать, что вероятность дефолта зависит от срока обслуживания клиента при любом уровне значимости, большем 0,000636483.

Положительность коэффициента a_1 указывает на то, что чем больше срок обслуживания клиента, тем выше вероятность того, что данный клиент не вернет кредит. Более точно, увеличение срока обслуживания на один месяц сопровождается в среднем увеличением шанса[1] невозврата кредита в $e^{0,125387119} = 1,1335872$ раза.

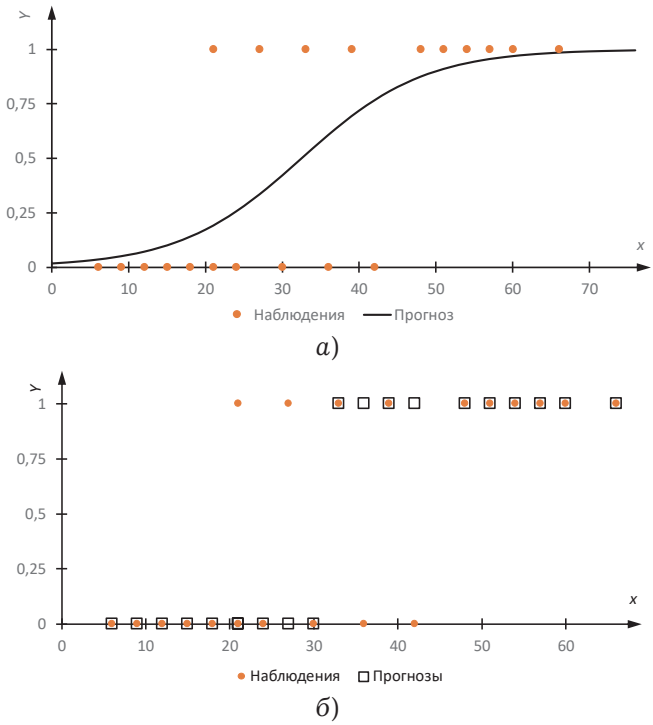
Значения

$$p(x_i) = \frac{1}{1 + e^{4,07771343 - 0,125387119 x_i}}$$

для каждого из наблюдений вычислены в ячейках H2:H21, график функции $p(x_i)$ построен на рис. 7.2.5, *а*, а оранжевыми точками отмечены наблюдения.

В ячейке O9 теперь можно вычислить прогноз $p(x)$ — вероятности того, что заемщик, который обслуживается в банке x мес. (где x берется из ячейки O8), вернет кредит.

[1] Напомним, что шансом события, которое может наступить с вероятностью p, называется число $p/(1 - p)$.



а)

б)

Рис. 7.2.5. *Исходные данные и прогнозы по модели логистической регрессии в примере 7.2.1: функция логистической регрессии (а) и прогнозы по порогу $p = 0,5$ (б)*

На рис. 7.2.5, *б* оранжевыми точками отмечены наблюдения, а квадратными рамками обозначены прогнозы, полученные с помощью отсечения (заемщик считается «хорошим» при прогнозе вероятности дефолта $p(x) \leqslant 0,5$ и «плохим» при $p(x) > 0,5$). Видно, что четыре оранжевые точки не обведены рамками: два «плохих» заемщика были признаны «хорошими», а два «хороших» заемщика — «плохими». В целом же 16 из 20 наблюдений были классифицированы правильно, а 4 из 20 — неправильно. Таким образом, доля правильных прогнозов составила $16/20 = 80\%$. ❑

В частном случае модель логистической регрессии была впервые использована в 1944 г. Дж. Берксоном. Общая теория логистической регрессии была разработана Д. Коксом в 1958 г.

Задачи

431. В некоторой группе заемщиков шанс невозврата кредита равен 0,37. Какова вероятность невозврата кредита случайно выбранным заемщиком из этой группы?

432. В некоторой группе абонентов вероятность перехода к другому сотовому оператору равна 0,16. Найдите шанс перехода к другому оператору для клиентов из этой группы.

433. Табл. 7.2.2 (файл «Подписка на онлайн-издание.xlsx») содержит данные об использовании и продлении подписки на онлайн-издание. Предварительно преобразовав качественные признаки в индикаторы, постройте модель логистической регрессии для классификации клиентов на тех, кто продлит под-

писку (класс 0) и тех, кто не продлит (класс 1). Проведите процедуру пошагового исключения незначимых факторных признаков. Интерпретируйте оценки коэффициентов логистической регрессии.

Таблица 7.2.2

Данные об использовании и продлении подписки на онлайн-издание

Номер договора	Возраст	Среднее число страниц, просмотренных в неделю	Число страниц, просмотренных в последнюю неделю	Продлил подписку
1	18-24	63,16	12	0
2	18-24	1,45	35	1
3	18-24	98,84	116	1
4	18-24	28,13	37	1
5	18-24	6,95	8	1
6	18-24	16,66	17	0
7	18-24	27,11	32	0
8	18-24	14,09	37	1
9	18-24	12,29	6	1
10	18-24	13,59	29	0
11	25-29	54,84	27	1
12	25-29	34,76	5	1
13	25-29	115,74	109	1
14	25-29	51,08	25	1
15	25-29	30,78	13	0
16	25-29	51,11	16	1
17	25-29	31,53	25	1
18	25-29	38,48	2	0
19	25-29	73,90	58	1
20	25-29	4,95	15	0
21	30-39	39,71	25	1
22	30-39	60,56	34	1
23	30-39	24,10	4	1
24	30-39	93,38	80	1
25	30-39	23,11	2	0
26	30-39	40,72	45	1
27	30-39	19,43	20	0
28	30-39	80,84	52	1
29	30-39	24,86	35	1
30	30-39	34,28	30	1
31	40-49	40,47	14	0
32	40-49	12,72	24	0
33	40-49	16,72	13	0
34	40-49	16,74	30	0
35	40-49	28,36	31	0
36	40-49	33,25	44	1
37	40-49	17,52	23	0
38	40-49	16,09	30	0
39	40-49	18,46	34	0
40	40-49	8,33	20	0
41	50-59	92,88	48	1
42	50-59	50,17	24	1
43	50-59	55,88	64	1
44	50-59	7,75	29	1
45	50-59	8,72	16	0
46	50-59	26,32	20	1
47	50-59	20,83	9	1
48	50-59	3,33	24	0
49	50-59	91,60	80	1
50	50-59	17,27	21	1
51	60+	76,01	117	1
52	60+	99,10	118	1
53	60+	4,53	6	1
54	60+	12,24	1	0
55	60+	26,61	11	0
56	60+	51,27	52	1
57	60+	9,32	29	1
58	60+	28,66	5	0
59	60+	114,47	123	1
60	60+	11,36	11	0

434. Предположим, что мы исследовали набор данных, содержащий для нескольких сотен студентов, изучающих учебную дисциплину «Анализ данных», значения признаков $x^{(1)}$ (количество часов, потраченных на выполнение домашних заданий), $x^{(2)}$ (средний балл школьного аттестата) и Y (единица, если студент получил оценку «отлично» по дисциплине «Анализ данных», и нуль в противном случае). Мы построили модель логистической регрессии, в которой $\hat{a}_0 = -6$, $\hat{a}_1 = 0,05$, $\hat{a}_2 = 1$ (коэффициенты при $x^{(1)}$ и $x^{(2)}$ оказались значимы). Оцените: *а*) вероятность того, что студент, потративший на выполнение домашних заданий 40 ч и имеющий средний балл школьного аттестата, равный 3,5, получит «отлично» по дисциплине «Анализ данных»; *б*) вероятность того, что студент, потративший на выполнение домашних заданий 60 ч и имеющий средний балл школьного аттестата, равный 3,5, получит «отлично» по дисциплине «Анализ данных»; *в*) количество

часов, которые должен потратить студент со средним баллом школьного аттестата, равным 3,5, на выполнение домашних заданий, чтобы с надежностью не менее 50% получить «отлично»; *г*) количество часов, которые должен потратить студент со средним баллом школьного аттестата, равным 3,5, на выполнение домашних заданий, чтобы с надежностью не менее 75% получить «отлично».

435. По данным примера 1.1.1 (файл **«Кредиты.xlsx»**), предварительно обработав выбросы (в задаче **32**) и пропущенные значения (в задачах **34—40**), удалив повторяющиеся строки (в задаче **42**) и преобразовав качественные признаки в индикаторы, постройте модель логистической регрессии для классификации заемщиков по статусу кредита. Проведите процедуру пошагового исключения незначимых факторных признаков. Проведите интерпретацию оценок коэффициентов логистической регрессии.

7.2.2. Деревья решений

Модели линейной и логистической регрессии очень популярны в практических задачах. Это связано с тем, что регрессионные модели хорошо изучены, их коэффициенты имеют четкую практическую интерпретацию, алгоритмы работают очень быстро, в том числе на больших и очень больших выборках.

Во многих случаях, например, в задачах кредитного скоринга, анализа оттока клиентов и т. п. крайне полезны на практике значения вероятностей отнесения объектов к классам, которые получаются в результате прогнозирования по модели логистической регрессии.

Однако предположения, которые лежат в основе регрессионных моделей (линейность зависимости результативного признака от факторов, независимость и одинаковое нормальное распределение остатков) как правило в практических задачах не выполняются, в связи с чем качество прогнозов регрессионных моделей оказывается хуже, чем у других методов, например, чем у деревьев и лесов решений.

Дерево решений — это иерархическая древовидная структура классифицирующих правил типа «Если ..., то ...». Дерево состоит из *узлов* и *ветвей*, соединяющих узлы.

Начальный узел, в который не входит ни одна ветвь, называется *корнем дерева*. Конечные узлы, из которых не выходят ветви, и которые соответствуют конкретным значениям прогнозов результативного признака, называются *листьями*.

Каждому узлу, не являющимся листом, соответствует вопрос относительно некоторого признака X, подразумевающий несколько вариантов ответа, и в зависимости от выбранного варианта ответа осуществляется переход к вершине следующего уровня.

Для принятия решения, к какому классу отнести некоторый объект, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня.

Дерево начинают строить по шагам сверху вниз. На каждом шаге строят ветвление по очередному признаку, рассматривая, насколько данный признак неоднороден относительно результативного признака. Если на какой-то ветви набирается явное большинство одного из значений результативного признака, то эту ветвь завершают соответствующим листом, если же явного большинства не набирается, то производят дальнейшее ветвление по следующему признаку.

Чтобы понять, как строятся деревья, рассмотрим

Пример 7.2.3. В табл. 7.2.3 для 12 клиентов банка приведены значения результативного признака «Кредит погашен» и пяти факторных признаков: «Наличие детей», «Образование», «Недвижимость», «Профессия», «Возраст». Построим дерево решений для задачи кредитного скоринга, в которой требуется определить правило принятия решения о выдаче кредита или отказе в выдаче кредита на основании значений названных факторных признаков.

Таблица 7.2.3

Данные о клиентах в примере 7.2.3

Номер договора	Наличие детей	Образование	Недвижимость	Профессия	Возраст	Кредит погашен
1	нет	высшее	в собственности	менеджер	40–60	погашен
2	нет	высшее	отсутствует	продавец	25–39	не погашен
3	нет	среднее	отсутствует	инженер	40–60	погашен
4	есть	высшее	отсутствует	продавец	25–39	погашен
5	есть	среднее	в собственности	менеджер	25–39	не погашен
6	нет	высшее	в ипотеке	предприниматель	40–60	погашен
7	нет	среднее	отсутствует	инженер	18–24	не погашен
8	нет	высшее	в ипотеке	продавец	40–60	погашен
9	есть	среднее	отсутствует	инженер	25–39	не погашен
10	есть	высшее	в собственности	предприниматель	25–39	не погашен
11	нет	среднее	отсутствует	продавец	18–24	не погашен
12	есть	высшее	отсутствует	инженер	25–39	погашен

С какого признака лучше всего начать построение дерева? Из 12 клиентов шестеро вернули кредит, а шестеро — нет. Рассмотрим ветвление по признаку «Возраст». Оба клиента, возраст которых находится в границах от 18 до 24 полных лет, кредит не вернули (это клиенты под номерами 7 и 11). Все четверо клиентов, возраст которых находится в границах от 40 до 60 лет, вернули кредит (это клиенты под номерами 1, 3, 6, 8). Наконец, среди клиентов, возраст которых находится в границах от 25 до 39 лет, двое вернули кредит (4 и 12), а четверо не вернули (2, 5, 9, 10). Таким образом, если производить ветвление по признаку «Возраст», то ветвь «18—24» целесообразно закончить листом «не погашен», ветвь «40—60» — листом «погашен», а по ветви «25—39» потребуется произвести дальнейшее ветвление. Таким образом, ветвление по признаку «Возраст» дает возможность для значительной части клиентов сразу сказать, вернут они кредит или нет. Это ветвление проиллюстрировано рис. 7.2.6, *а*.

Если же на первом шаге произвести ветвление по признаку «Профессия», то мы увидим, что половина из менеджеров вернула кредит, а половина не вернула, половина из предпринимателей вернула кредит, а половина не вернула, половина из продавцов вернула кредит, а половина не вернула, и точно так же половина из инженеров вернула кредит, а половина не вернула (рис. 7.2.6, *б*).

Поскольку ветвление по признаку «Профессия» не приводит к повышению нашей информированности о том, кто вернет кредит, а кто нет, проводить

ветвление по этому признаку на первом шаге нецелесообразно. Предоставляем читателю самостоятельно убедиться, что по остальным признакам также нет смысла производить ветвление на первом шаге.

Аналогичным образом на втором шаге целесообразно провести ветвление по признаку «Образование» (рис. 7.2.7), на третьем шаге — по признаку «Профессия» (рис. 7.2.8), на четвертом шаге — по признаку «Количество детей (рис. 7.2.9). Окончательно мы получим дерево, представленное на рис. 7.2.10 и пригодное для автоматического принятия решений. ❑

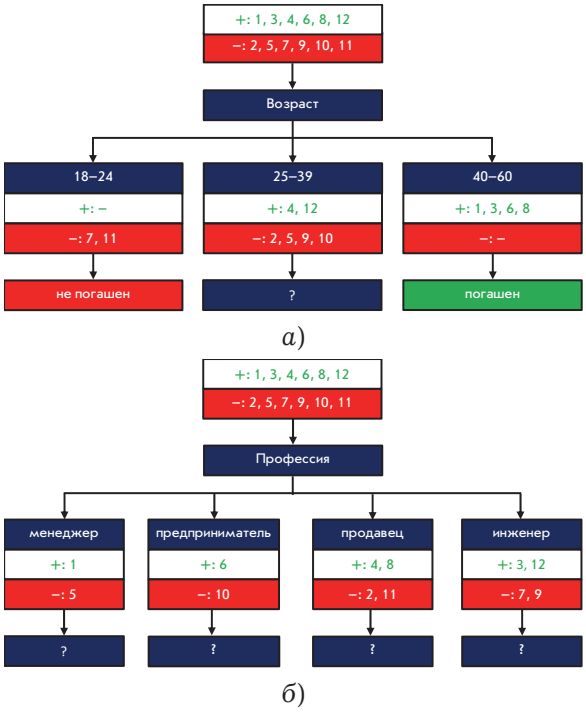


Рис. 7.2.6. *Первое ветвление в примере 7.2.3: а) по возрасту; б) по профессии*

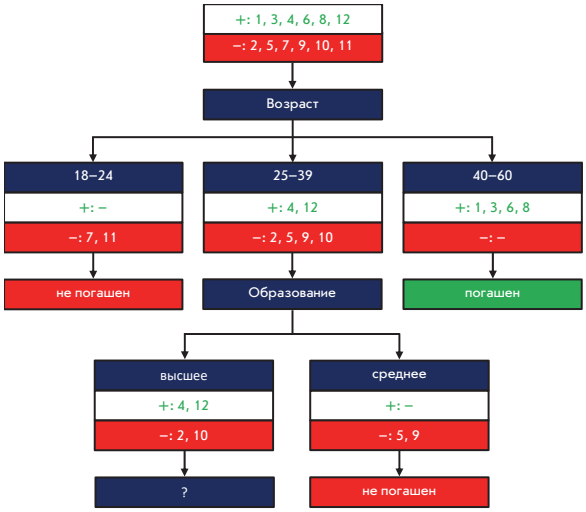


Рис. 7.2.7. *Второе ветвление в примере 7.2.3*

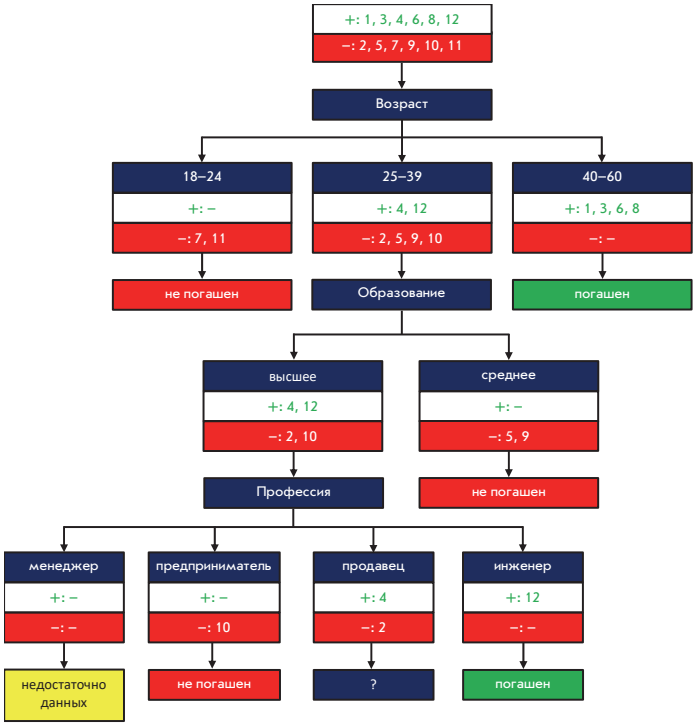


Рис. 7.2.8. *Третье ветвление в примере 7.2.3*

Общий алгоритм построения деревьев решений имеет следующий вид:

- на каждом шаге для текущего узла вычисляется критерий неоднородности текущего множества объектов в узле;
- для всех факторов и всех возможных разбиений вычисляются изменения критерия неоднородности и выбирается наилучшее разбиение (т. е. минимизирующее выбранный критерий);
- проверяется выполнение критерия остановки (если предусмотрен критерий остановки);
- производится «стрижка дерева» (англ. *pruning*), т. е. объединение некоторых листьев с целью улучшения обучающей способности (если «стрижка» предусмотрена).

В качестве **критериев неоднородности** деревьев в задачах классификации обычно выбирают *долю ошибок* (*misclassification error*)

$$1-\hat{p}_{k^*}^{(m)},$$

индекс загрязненности Джини (*Gini impurity*)

$$\sum_{k=0}^{1}\hat{p}_k^{(m)}\left(1-\hat{p}_k^{(m)}\right)$$

или *кросс-энтропию* (*cross-entropy*)

$$-\sum_{k=0}^{1}\hat{p}_k^{(m)}\log_2\hat{p}_k^{(m)}.$$

Здесь m — номер текущего узла, $k^* \in \{0;1\}$ — номер класса, к которому относится большинство объектов в текущем узле; $\hat{p}_k^{(m)}$ — доля объектов в данном листе, относящихся к классу k.

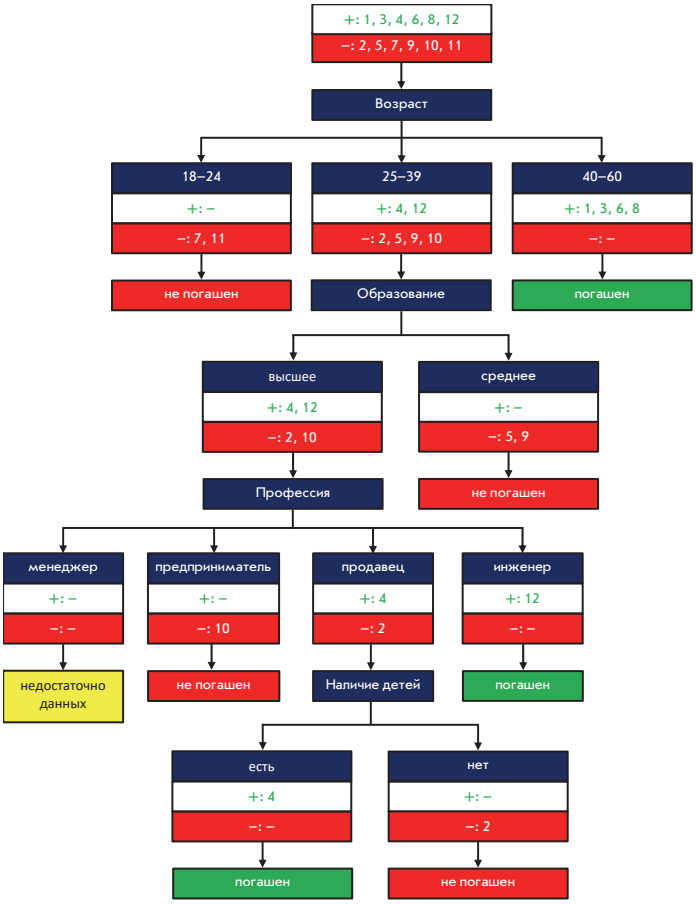


Рис. 7.2.9. *Четвертое ветвление в примере 7.2.3*

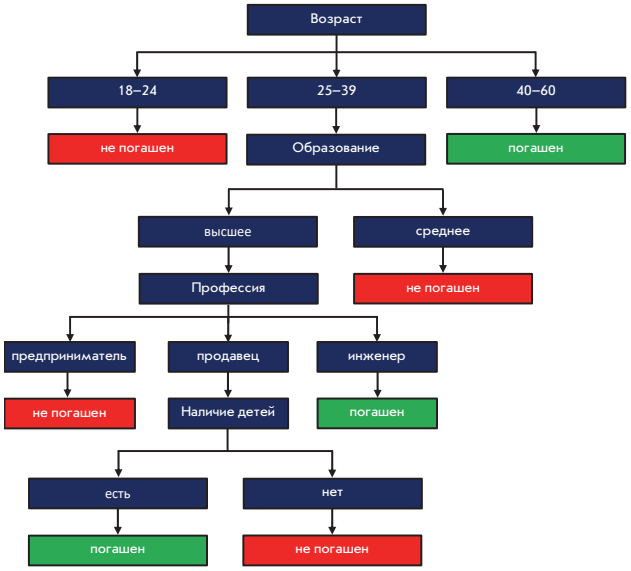


Рис. 7.2.10. *Итоговое дерево решений в примере 7.2.3*

Пример 7.2.4. В примере 7.2.3 в качестве критерия неоднородности использовалась доля ошибок. Так, на первом шаге перед ветвлением половина объектов представляла класс «погашен», а половина — класс «не погашен», и доля ошибок, таким образом, составляла $1/2$.

После ветвления по признаку «Возраст» на ветви «18—24» оба объекта «отдают голоса» за класс «не погашен» — таким образом, доля ошибок равна нулю, на ветви «40—60» четыре объекта из четырех «голосуют» за класс «погашен» — доля ошибок равна нулю, а на ветви «25—39»два объекта из шести голосуют за класс «погашен», а четыре — за класс «не погашен», так что на этой ветви (дадим ей номер $m = 3$) $k^* = 1$ («не погашен»), $1 - \hat{p}_k^{(m)} = 1 - 4/6 = 2/6 = 1/3$. Собирая все вместе с весами, соответствующими долям элементов на каждой ветви, получаем долю ошибок

$$\frac{2}{12} \cdot 0 + \frac{4}{12} \cdot 0 + \frac{6}{12} \cdot \frac{1}{3} = \frac{1}{6}.$$

Какова бы была доля ошибок, если бы мы провели ветвление по признаку «Профессия»? Один из двух менеджеров вернул кредит, а другой не вернул, один из двух предпринимателей вернул кредит, а другой не вернул, двое из четверых продавцов вернула кредит, а двое не вернули, и точно так же двое из четверых инженеров вернули кредит, а двое не вернули. Поэтому на каждой ветви доля ошибок составляет $1/2$, а доля ошибок на всем разбиении по ветвям равна

$$\frac{2}{12} \cdot \frac{1}{2} + \frac{2}{12} \cdot \frac{1}{2} + \frac{4}{12} \cdot \frac{1}{2} + \frac{4}{12} \cdot \frac{1}{2} = \frac{1}{2}.$$

Итак, если произвести ветвление на первом шаге по признаку «Возраст», то неоднородность (доля ошибок) уменьшится с 1/2 до 1/6, если же произвести ветвление по признаку «Профессия», то неоднородность так и останется равной 1/2. Поэтому, конечно же, для ветвления нужно выбрать признак «Возраст»! ❑

Замечание. Аналогично можно использовать деревья решений и для решения задач регрессии: в этом случае в качестве критерия неоднородности используют дисперсию или другую меру разброса Y.

В качестве **критерия остановки** можно выбрать, например, достижение максимальной глубины дерева, достижение максимального количества узлов, достижение минимального количества объектов в узле. Можно использовать и другие критерии, например, останавливаться в случае, когда улучшения по критерию неоднородности становятся слишком маленькими. Правильный выбор критерия остановки может существенно улучшить качество классификации, однако применение критериев остановки требует существенных затрат.

В табл. 7.2.4 проводится сравнение наиболее популярных алгоритмов построения деревьев решений.

Популярность деревьев решений связана с их наглядностью. Но, к сожалению, алгоритмы классификации, основанные на деревьях решений, являются слабыми, очень чувствительными к шумам: небольшое изменение обучающей выборки может привести к кардинальной перестройке дерева. Для усиления алгоритмов классификации используют различные техники, наиболее популярными из которых являются бэггинг (от англ. *Bootstrap AGGregatING* — агрегирование результатов по повторным выборкам) и бустинг (от англ. *boosting* — усиление).

Техника **бэггинга** была предложена в 1994 г. Л. Брейманом. Основная идея бэггинга состоит в том, что строится много деревьев, для каждого из которых выбирается некоторая повторная выборка (бутстреп) из генеральной совокупности объектов.

В 2000 г. Л. Брейман предложил особенную модификацию бэггинга, ставшую очень популярной — *метод случайного леса* (*random forest*), основанный

на построении большого числа деревьев решений, каждое из которых строится по повторной выборке (бутстрепу), получаемой из исходной обучающей выборки. При этом при построении каждого дерева на стадиях ветвления используется только фиксированное число случайно отбираемых признаков и строится полное дерево (т. е. каждый лист содержит наблюдения только одного класса). После этого результат определяется голосованием отдельных деревьев, входящих в лес, если решается задача классификации, или усреднением прогнозов отдельных деревьев, если решается задача регрессии.

Таблица 7.2.4

Популярные алгоритмы построения деревьев решений

Название	*Авторы*	*Назначение*	*Типы факторных признаков*	*Критерии неоднородности*	*Способ ветвления*	*Критерии остановки*
ID3	Р. Куинлан (1979)	только классификация	только категориальные	кросс-энтропия	разбиение признаков по всем значениям	все объекты лежат в одном классе или объекты невозможно разделить
C4.5	Р. Куинлан (1993)	только классификация	любые	кросс-энтропия	разбиение числовых признаков по границе, категориальных признаков по всем значениям	все объекты лежат в одном классе или объекты невозможно разделить или критерий неоднородности не изменяется
CART	Л. Брейман, Дж. Фридман, Ч.Дж. Стоун, Р.А. Олшен (1974—1984)	классификация и регрессия	любые	индекс загрязненности Джини для классификации; дисперсия для регрессии	разбиение числовых признаков по границе, категориальных признаков по всем значениям	критерий неоднородности не изменяется

Основная идея **бустинга** — формирование сильного (т. е. эффективного) алгоритма классификации на основе слабого алгоритма. Бустинг — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. В процессе формирования сильного алгоритма слабый алгоритм «доучивается» за счет того, что перераспределяются веса примеров из обучающей выборки: в случае верного распознавания вес снижается, в случае неверного — повышается. Первый вариант бустинга был предложен Р. Шапире в 1989 г.

В 1995 г. Й. Фройндом и Р. Шапире был разработан *AdaBoost* (адаптивный бустинг) — алгоритм, усиливающий классификаторы, объединяя их в «комитеты». Этот алгоритм является адаптивным в том смысле, что каждый следующий комитет классификаторов строится по объектам, неверно классифицированным предыдущими комитетами.

Прокомментируем рис. 7.2.11, иллюстрирующий бустинг алгоритмов классификации. Бустинг происходит по шагам, причем на каждом шаге t результаты предсказания взвешиваются с весами, сгенерированными по итогам предыдущего шага $t - 1$: Объекты, классифицированные верно, получают более низкие веса, а те, которые классифицированы неправильно — более высокие. В начале первого шага все объекты имеют одинаковые веса (это иллюстрируется одинаковым размером кружков на рис. 7.2.11, *а*).

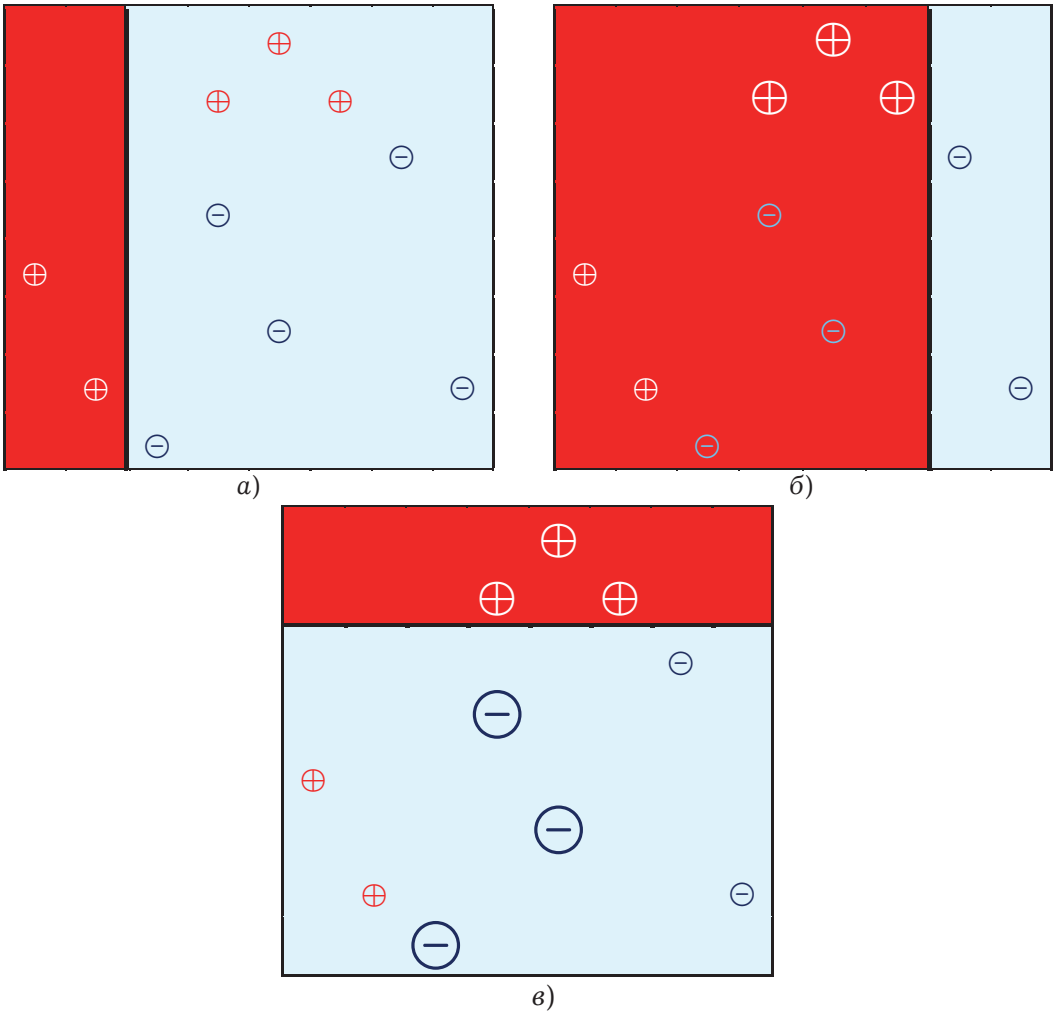


Рис. 7.2.11. *Бустинг*: *первый шаг* (*а*); *второй шаг* (*б*); *третий шаг* (*в*)

Алгоритм классификации, построенный на первом шаге, верно классифицировал все точки, за исключением трех «⊕» — веса этих неверно классифицированных объектов на втором шаге увеличиваются (рис. 7.2.11, *б*). Классификатор, построенный на втором шаге, неверно распознал три «⊖», значит, необходимо увеличить веса этих трех «⊖» (рис. 7.2.11, *в*). В результате большого числа итераций все модели получают веса в зависимости от точности, и создается сводный результат.

Бустинг алгоритмов регрессии строится аналогично.

В последнее время стали популярны алгоритмы градиентного бустинга *XGBoost* (предложен Т. Ченом в 2014 г.), *LightGBM* (выпущенный корпорацией *Microsoft* в январе 2017 г.) и *CatBoost* (реализованный компанией «Яндекс» в апреле 2017 г.). Эти алгоритмы используют идею численной оптимизации критерия неоднородности,[1] при этом деревья строится по набору данных, который на каждом шаге модифицируется: на первом шаге дерево строится по исходным при-

[1] Который растет в направлении, противоположном градиенту — отсюда название этой группы методов.

знакам, на следующих шагах деревья строятся по остаткам — разностям между исходными данными и предсказаниями модели с предыдущего шага. Далее собирается линейная комбинация прогнозов по построенным деревьям.

Опыт использования алгоритмов *XGBoost*, *LightGBM* и *CatBoost* в реальных задачах, которые мы решали в департаменте анализа данных, принятия решений и финансовых технологий Финуниверситета, свидетельствует о том, что результаты, получаемые с помощью этих трех алгоритмов, как правило, имеют одинаково хорошее качество.

Стрижка дерева подразумевает построение излишне обученного дерева (например, такого дерева, в каждом листе которого ровно по одному объекту) с последующей оптимизацией структуры дерева путем объединения листьев. На практике методы стрижки используются достаточно редко. Связано это с тем, что деревья сами по себе являются слабыми алгоритмами и представляют интерес только в бэггинге и бустинге. При этом при использовании бэггинга смысла стричь деревья нет, так как алгоритмы бэггинга (например, случайные леса) предполагают, что входящие в ансамбль деревья должны быть переобучены, а при бустинге деревья не стригут, так как деревья в бустинге имеют, как правило, очень небольшую глубину.

Более подробно способы построения эффективных алгоритмов классификации на основе деревьев решений изучаются в курсах машинного обучения.

Задачи

436. Продолжите пример 7.2.4 и вычислите, как изменялась доля ошибок на каждом шаге ветвления.

437. Проверьте, уменьшаются ли индекс загрязненности Джини и кросс-энтропия в процессе четырех шагов ветвления, описанные в примере 7.2.3.

438. Постройте дерево решений для классификации заемщиков по статусу погашения кредита на основании данных из табл. 7.2.5. В качестве критерия неоднородности используйте долю ошибок.

Таблица 7.2.5

*Данные о заемщиках в задаче **438***

Номер договора	Семейное положение	Наличие детей	Стаж работы на последнем месте	Образование	Недвижимость	Наличие подтверждения уровня доходов	Профессия	Средний месячный доход	Возраст	Кредит погашен
1	в браке	нет	<= 2	высшее	в собственности	есть	менеджер	41–70	40–60	погашен
2	в браке	нет	<= 2	высшее	отсутствует	нет	продавец	100–130	25–39	не погашен
3	не в браке	нет	> 2	среднее	отсутствует	нет	инженер	41–70	40–60	погашен
4	в браке	есть	<= 2	высшее	отсутствует	нет	продавец	71–100	25–39	погашен
5	в браке	есть	<= 2	среднее	в собственности	нет	менеджер	>130	25–39	не погашен
6	не в браке	нет	> 2	высшее	в ипотеке	есть	предприниматель	41–70	40–60	погашен
7	не в браке	нет	> 2	среднее	отсутствует	есть	инженер	41–70	18–24	не погашен
8	не в браке	нет	<= 2	высшее	в ипотеке	есть	продавец	41–70	40–60	погашен
9	не в браке	есть	> 2	среднее	отсутствует	есть	инженер	>130	25–39	не погашен
10	в браке	есть	> 2	высшее	в собственности	нет	предприниматель	71–100	25–39	не погашен
11	не в браке	нет	<=2	среднее	отсутствует	нет	продавец	41–70	18–24	не погашен
12	в браке	есть	> 2	высшее	отсутствует	нет	инженер	100–130	25–39	погашен

§ 7.3. Модели кластерного анализа

7.3.1. Иерархические агломеративные кластер-процедуры

Очень часто при исследовании больших совокупностей объектов необходимо выявить объекты, б л и з к и е между собой в определенном смысле.

Например, производители товаров и услуг массового потребления не ориентируются, как правило, на каждого конкретного потенциального покупателя, а сегментируют рынки, разделяя покупателей на группы (или классы), внутри которых покупатели похожи.

Сущность задач кластерного анализа заключается в разбиении большого числа объектов, характеризующихся многомерными признаками, на *кластеры* — однородные группы, внутри которых объекты между собой похожи.

Целью кластерного анализа является исследование внутренней структуры изучаемой системы n объектов, каждый из которых характеризуется m-мерным вектором признаков, «сжатие» этой системы без существенной потери содержащейся в ней информации путем выявления классов сходных между собой объектов и отождествления объектов внутри каждого класса.

Кластерный анализ решает задачу классификации объектов п р и о т с у т с т в у ю щ е й а п р и о р н о й и н ф о р м а ц и и о н а б л ю д е н и я х в н у т р и к л а с с о в, в отличие от рассмотренных в **пп. 7.3.1, 7.3.2** моделей классификации с обучением, в которых предполагается наличие такой информации. Впервые термин «кластерный анализ» (от англ. *cluster* — гроздь, сгусток, пучок) был предложен в 1939 г. Р. Трионом.

Среди **прикладных задач**, решаемых с помощью кластерного анализа, можно отметить следующие:

- проведение классификации объектов с учетом большого числа признаков, отражающих их природу;
- выявление структуры изучаемой совокупности объектов;
- снижение объема выборки путем отождествления каждого класса объектов с его типичным представителем;
- снижение размерности пространства путем отождествления близких признаков (в этом случае классифицируются не объекты, а признаки, этот прием перемены объектов и признаков местами часто приводит к интересным результатам);
- построение отдельных регрессионных моделей в каждом классе объектов; если дисперсия остатков регрессии сильно изменяется от класса к классу (такое явление называется г е т е р о с к е д а с т и ч н о с т ь ю), то построение регрессии для каждого класса может помочь избавиться от гетероскедастичности.

Наиболее трудным в задаче классификации является определение **меры однородности объектов.**

Пусть каждый из исходных n объектов с номерами 1, 2, ..., n задается как точка

$$\mathbf{x}_k = (x_{k1}, x_{k2}, ..., x_{km}) \qquad (k = 1, 2, ..., n)$$

в m-мерном пространстве признаков X_1, X_2, ..., X_m. Совокупность этих точек можно трактовать как выборку объема n из многомерной генеральной совокупности $\mathbf{X} = (X_1, X_2, ..., X_m)$.

Замечание. Результаты классификации зависят от выбора масштаба и единиц изменения признаков. Чтобы исправить такое положение, прибегают к стандартной нормировке признаков:

$$x_{ki} \to z_{ki} = \frac{(x_{ki} - \overline{x}_{\cdot i})}{\hat{\sigma}_{X_i}}.$$

Однако эта операция, уравнивая разделительные возможности всех признаков, может привести и к нежелательным последствиям.

Естественный способ определения однородности объектов заключается в вычислении **расстояний** между ними и отнесении объектов, расстояния между которыми в определенном смысле малы, к одному и тому же кластеру, а объектов, расстояния между которыми велики, к разным кластерам.

Наиболее распространенной мерой однородности объектов является *евклидово расстояние*

$$l_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \quad (i = 1,2,\ldots,n,\ j = 1,2,\ldots,n), \qquad (7.3.1)$$

которое обычно используется для количественных признаков, однородных по своему содержательному смыслу и одинаково «весомых» с точки зрения решения вопроса об отнесении объекта к тому или иному кластеру, и *расстояние Хэмминга*

$$l_{ij} = \sum_{k=1}^{m} | x_{ik} - x_{jk} | \quad (i = 1,2,\ldots,n,\ j = 1,2,\ldots,n), \qquad (7.3.2)$$

которое используется для дихотомических признаков, будучи равным числу несовпадений значений соответствующих признаков на данных объектах.

Замечание 1. Расстояние Хэмминга применимо к любым качественным признакам, так как любому нечисловому признаку можно поставить во взаимно однозначное соответствие набор индикаторов (фиктивных переменных).

Замечание 2. Расстояние Хэмминга можно использовать и для количественных признаков, при этом его называют *расстоянием городских кварталов* (или, в шутку, *дистанцией манхэттенского таксиста*), в этом случае l_{ij} — это не кратчайшее расстояние между двумя точками, а «путь, который должен преодолеть таксист, чтобы проехать от одной точки до другой по городским улицам, пересекающимся, как в Манхеттене, под прямым углом»).

Замечание 3. В кластерном анализе используются и другие формулы для вычисления расстояний между объектами.

Для оценки однородности объектов в пространстве $m = k + (m - k)$ разнотипных признаков (где k — число дихотомических признаков, к которым могут быть сведены и ранговые, и номинальные признаки, а $(m - k)$ — число количественных признаков), обычно используют не расстояние между объектами, а меру их **сходства.**

Сходство i-го и j-го объектов, например, можно рассчитать по формуле

$$r_{ij} = \frac{k}{m} r'_{ij} + \frac{(m-k)}{m} r''_{ij} \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n),$$

где

$$r'_{ij} = \frac{a_{ij}}{k}$$

(a_{ij} — число дихотомических признаков, значение которых совпадают у объектов с номерами i и j),

$$r''_{ij} = \frac{\overline{l}}{\overline{l} + l_{ij}}$$

(l_{ij} — евклидово расстояние в пространстве количественных признаков, $\overline{l}$ — среднее значение расстояния по всем парам объектов).

К мерам однородности (расстоянию l_{ij} и сходству r_{ij}) предъявляются следующие требования:

- *симметрии:*

$$l_{ij} = l_{ji}, \quad r_{ij} = r_{ji} \quad (i = 1, 2, ..., n,\ j = 1, 2, ..., n);$$

- *максимального сходства объекта с самим собой:*

$$r_{ii} = \max_{1 \leqslant j \leqslant n} r_{ij} \quad (i = 1, 2, ..., n);$$

- *монотонного убывания r_{ij} по l_{ij}:*

$$l_{kp} \geqslant l_{ij} \ \Leftrightarrow \ r_{kp} \leqslant r_{ij} \quad (i = 1, 2, ..., n,\ j = 1, 2, ..., n).$$

Замечание. Если известен способ вычисления расстояний l_{ij}, то очень легко ввести естественную меру сходства:

$$r_{ij} = \frac{\overline{l}}{\overline{l} + l_{ij}} \quad (i = 1, 2, ..., n,\ j = 1, 2, ..., n).$$

Точно так же можно определить расстояние, зная сходство.

Пусть известна матрица

$$\mathbf{L} = (l_{ij}) \in \mathbb{R}^{m \times m}$$

расстояний между n объектами и некоторое их разбиение

$$\mathfrak{R}(p) = \{\mathcal{K}_1, \mathcal{K}_2, ..., \mathcal{K}_p\}$$

на p кластеров.

Основным понятием *кластер-процедур* является расстояние

$$\rho_{st} = \rho(\mathcal{K}_s, \mathcal{K}_t)$$

между кластерами $\mathcal{K}_s$ и $\mathcal{K}_t$ ($s = 1, 2, ..., p,\ t = 1, 2, ..., p$).

Наиболее распространены следующие **способы измерения расстояний между кластерами:**

- *метод ближнего соседа*:

$$\rho_{st} = \min_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} \quad (s = 1, 2, ..., p,\ t = 1, 2, ..., p); \tag{7.3.3}$$

- *метод дальнего соседа*:

$$\rho_{st} = \max_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} \quad (s = 1, 2, ..., p,\ t = 1, 2, ..., p); \tag{7.3.4}$$

- *метод средней связи*:

$$\rho_{st} = \frac{\sum\limits_{i \in \mathcal{K}_s} \sum\limits_{j \in \mathcal{K}_t} l_{ij}}{n_s n_t} \quad (s = 1, 2, ..., p,\ t = 1, 2, ..., p), \tag{7.3.5}$$

где n_s и n_t — количество объектов соответственно в кластерах $\mathcal{K}_s$ и $\mathcal{K}_t$;

- *центроидный метод* (в котором измеряется расстояние между «центрами тяжести» кластеров):

$$\rho_{st} = l\left(\bar{x}_{\mathcal{K}_s}, \bar{x}_{\mathcal{K}_t}\right) \quad (s = 1, 2, ..., p,\ t = 1, 2, ..., p), \tag{7.3.6}$$

где

$$\bar{x}_{\mathcal{K}_s} = \frac{\sum\limits_{i \in \mathcal{K}_s} x_i}{n_s} —$$

среднее арифметическое векторных наблюдений x_i при $i \in \mathcal{K}_s$.

Рассмотренные методы вычисления расстояний между кластерами относятся к группе иерархических (т. е. деревообразующих) агломеративных (т. е. объединяющих) методов.

Иерархические агломеративные методы — это многошаговые методы классификации, работающие по следующему **алгоритму**:

- на нулевом шаге за разбиение принимается исходная совокупность n элементарных кластеров, матрица расстояний между которыми

$$(\rho_{ij}) = (l_{ij}) \in \mathbb{R}^{n \times n};$$

- на каждом следующем шаге происходит объединение (в соответствии с эвристическим или экстремальным подходом) двух кластеров $\mathcal{K}_s$ и $\mathcal{K}_t$, сформированных на предыдущем шаге, в один кластер

$$\mathcal{K}_{s \oplus t} = \mathcal{K}_s \cup \mathcal{K}_t,$$

при этом размерность матрицы расстояний уменьшается, по сравнению с размерностью этой матрицы на предыдущем шаге, на единицу.

При использовании агломеративных методов рассчитать расстояние $\rho_{i,s \oplus t}$ между кластерами $\mathcal{K}_i$ и $\mathcal{K}_{s \oplus t}$ $(i \neq s,\ i \neq t)$ можно, используя соответствующую методу формулу расстояния между кластерами, однако менее трудоемки расчеты по формуле, предложенной в 1967 г. Г.Н. Лансом и У.Т. Вильямсом:

$$\rho_{i,s \oplus t} = \alpha\rho_{is} + \beta\rho_{it} + \gamma\rho_{st} + \delta\left|\rho_{is} - \rho_{it}\right|, \tag{7.3.7}$$

в которой значения коэффициентов α, β, γ, δ зависят от используемого метода определения расстояний между кластерами (табл. 7.3.1).

Таблица 7.3.1

Применение формулы Ланса — Уильямса

Метод	α	β	γ	δ	$\rho_{i,s \oplus t}$
ближнего соседа	0,5	0,5	0	−0,5	$\min(\rho_{is}, \rho_{it}) = \min\limits_{p \in \mathcal{K}_i, j \in \mathcal{K}_{s \oplus t}} l_{pj}$
дальнего соседа	0,5	0,5	0	0,5	$\max(\rho_{is}, \rho_{it}) = \max\limits_{p \in \mathcal{K}_i, j \in \mathcal{K}_{s \oplus t}} l_{pj}$
средней связи	$\frac{n_s}{n_s + n_t}$	$\frac{n_s}{n_s + n_t}$	0	0	$\frac{n_s\rho_{is} + n_t\rho_{it}}{n_s + n_t} = \frac{1}{n_i n_{s \oplus t}} \sum\limits_{p \in \mathcal{K}_i} \sum\limits_{j \in \mathcal{K}_{s \oplus t}} l_{pj}$
центроидный	$\frac{n_s}{n_s + n_t}$	$\frac{n_s}{n_s + n_t}$	$-\alpha\beta$	0	$\frac{n_s\rho_{is} + n_t\rho_{it}}{n_s + n_t} - \frac{n_s n_t \rho_{st}}{(n_s + n_t)^2} = l\left(\bar{x}_{\mathcal{K}_i}, \bar{x}_{\mathcal{K}_{s \oplus t}}\right)$

В последнем столбце табл. 7.3.1 для каждого метода определения расстояний между кластерами слева приведена формула подсчета $\rho_{i,\ s \oplus t}$, следующая

из (7.3.11), а справа — формула, следующая из определения соответствующего метода, рассмотренного в предыдущем пункте.

Пример 7.3.1. Проиллюстрируем иерархические агломеративные кластер-процедуры на примере классификации пяти точек:

$$(4; 3), \quad (7; 4), \quad (2; 0), \quad (2; 1), \quad (0; 4),$$

изображенных на рис. 7.3.1.

Эти точки соответствуют результативности работы пяти менеджеров банка: первая координата соответствует числу выданных ипотечных кредитов за неделю, а вторая координата — числу выданных за тот же период автокредитов.

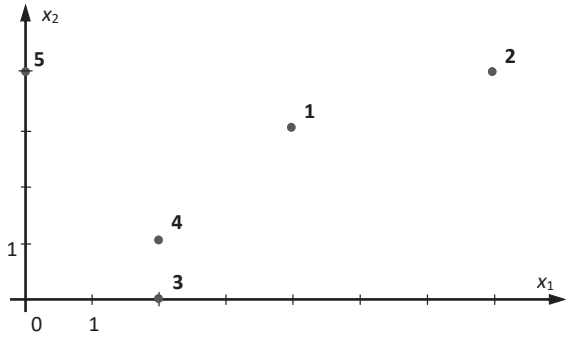


Рис. 7.3.1. *Число выданных ипотечных и автокредитов пятью менеджерами банка*

За метрику расстояний примем квадрат евклидова расстояния. Матрица квадратов евклидовых расстояний между точками приведена в табл. 7.3.2.

Таблица 7.3.2

Матрица квадратов евклидовых расстояний между объектами в примере 7.3.1

	1	2	3	4	5
1	0	10	13	8	17
2	10	0	41	34	49
3	13	41	0	**1**	20
4	8	34	**1**	0	13
5	17	49	20	13	0

Проведем вначале классификацию **по методу ближнего соседа.** Начальное разбиение таково: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому переходим к следующему разбиению: $\{1, 2, 3 \oplus 4, 5\}$.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3 \oplus 4, 5\} \to \{1 \oplus (3 \oplus 4), 2, 5\} \to$$
$$\to \{1 \oplus (3 \oplus 4) \oplus 2, 5\} \to \{1 \oplus 2 \oplus 3 \oplus 4 \oplus 5\}$$

представлена на рис. 7.3.2, *а*, а на рис. 7.3.2, *б* изображена соответствующая дендрограмма (слева проставлены расстояния, при которых происходит переход к новому разбиению).

Теперь проведем классификацию **по методу дальнего соседа.** Начальное разбиение состоит из пяти исходных объектов: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому, как и в методе ближнего соседа, переходим к следующему разбиению: $\{1, 2, 3 \oplus 4, 5\}$.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3 \oplus 4, 5\} \to \{1 \oplus 2, 3 \oplus 4, 5\} \to$$
$$\to \{1 \oplus 2, 3 \oplus 4 \oplus 5\} \to \{1 \oplus 2 \oplus 3 \oplus 4 \oplus 5\}$$

представлена на рис. 7.3.3.

Если число кластеров заранее известно, то классификацию заканчивают, как только будет сформировано разбиение с нужным числом кластеров. При неизвестном числе кластеров правило остановки связывают с понятием п о р о г а, т. е. некоторого расстояния l_0, определяемое условиями конкретной задачи.

	1	2	3⊕4	5
1	0	10	**8**	17
2	10	0	34	49
3⊕4	**8**	34	0	13
5	17	49	13	0

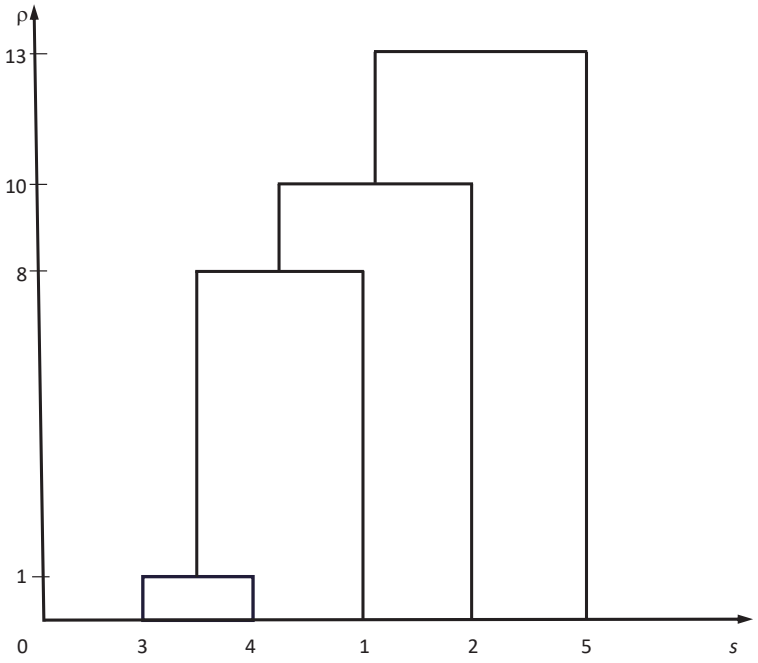
→

	1⊕(3⊕4)	2	5
1⊕(3⊕4)	0	**10**	13
2	**10**	0	49
5	13	49	0

→

	1⊕(3⊕4)⊕2	5
1⊕(3⊕4)⊕2	0	**13**
5	**13**	0

а)



б)

Рис. 7.3.2. *Классификация по методу ближнего соседа в примере 7.3.1: последовательность разбиений* (*а*) *и дендрограмма* (*б*)

	1	2	3⊕4	4
1	0	**10**	13	17
2	**10**	0	41	49
3⊕4	13	41	0	20
5	17	49	20	0

→

	1⊕2	3⊕4	5
1⊕2	0	41	49
3⊕4	41	0	**20**
5	49	**20**	0

→

	1⊕2	(3⊕4)⊕5
1⊕2	0	**49**
(3⊕4)⊕5	**49**	0

Рис. 7.3.3. *Классификация по методу дальнего соседа в примере 7.3.1*

Например, если в качестве порога выбрать

$$l_0 = \overline{l} = \sum_{j>i} \frac{l_{ij}}{n(n-1)}$$

(в условиях примера $\overline{l} = 20{,}6$), то метод ближнего соседа объединит все пять объектов в один кластер, а остальные три метода дадут два кластера: 3⊕4⊕5 и 1⊕2. ❑

Замечание. Поскольку метод ближнего соседа объединяет кластеры, в которых расстояние между ближайшими элементами минимально по сравнению с другими кластерами, то два объекта попадают в один кластер, если существует соединяющая их цепочка ближайших друг к другу объектов (так называемый ц е п о ч е ч н ы й э ф ф е к т). Поэтому метод ближнего соседа иногда называют методом одиночной связи. Для устранения цепочечного эффекта можно, например, ввести ограничение на максимальное расстояние между объектами кластера: в первый кластер включить два наиболее близких объекта, затем в этот кластер включить объект, который имеет минимальное расстояние с одним их объектов кластера, а его расстояние до другого объекта кластера не больше некоторого числа l_0 и т. д.; формирование первого кластера продолжают до тех пор, когда нельзя будет найти объект, расстояние от которого до любого объекта кластера не превзойдет l_0; формирование второго и последующих кластеров осуществляется из оставшихся объектов аналогичным образом.

В методе дальнего соседа объединяются кластеры, в которых минимально расстояние между самыми далекими друг от друга объектами. Это означает, что все остальные объекты в полученном после объединения кластере связаны друг с другом еще теснее, чем «соседи». Поэтому метод дальнего соседа называют иногда методом полной связи.

Задачи

439. Постройте матрицу сходства между объектами в примере 7.3.1.

440. Постройте дендрограмму для разбиения по методу дальнего соседа в примере 7.3.1.

441. Проведите классификацию объектов из примера 7.3.1 по методу средней связи.

442. Проведите классификацию объектов из примера 7.3.1 по центроидному методу.

7.3.2. Метод K-средних

Иерархические кластер-процедуры практически реализуемы лишь в задачах классификации не более нескольких десятков наблюдений. К решению задач с бóльшим числом наблюдений применяют **последовательные кластер-процедуры** — итерационные алгоритмы, на каждом шаге которых используется одно наблюдение (или небольшая часть исходных наблюдений) и результаты разбиения на предыдущем шаге.

Идею этих процедур поясним на примере *метода K-средних* с заранее заданным числом k классов:

- на нулевом шаге за центры искомых k кластеров принимают случайно выбранные k наблюдений — точки $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_k$; каждому кластеру присваивают единичный вес;
- на первом шаге находят расстояния от точки $\mathbf{x}_{k+1}$ до центров кластеров, построенных на предыдущем шаге, и точку $\mathbf{x}_{k+1}$ относят к кластеру, расстояние до которого минимально, после чего рассчитывают новый центр тяжести этого кластера (как взвешенное среднее по каждому показателю) и вес

кластера увеличивают на единицу; все остальные кластеры остаются неизменными (с прежними центрами и весами);

- на втором шаге аналогичную процедуру выполняют для точки $\mathbf{x}_{k+2}$ и т. д.

При достаточно большом числе n классифицируемых объектов или достаточно большом числе итераций пересчет центров тяжести практически не приводит к их изменению.

Если в какой-то точке не удается, рассмотрев все $n = k + (n - k)$ точек, достичь практически не изменяющихся центров тяжести, то либо, используя получившееся разбиение n точек на k кластеров в качестве начального, применяют изложенную процедуру к точкам $\mathbf{x}_1$, $\mathbf{x}_2$, ...; либо в качестве начального разбиения принимают различные комбинации k точек из исходных n, а в качестве окончательного берут наиболее часто встречающееся финальное разбиение.

Пример 7.3.2. Проиллюстрируем метод K-средних на примере классификации пяти точек:

$$(4; 3), \quad (7; 4), \quad (2; 0), \quad (2; 1), \quad (0; 4),$$

из примера 7.3.1 (см. рис. 7.3.1) по двум кластерам ($k = 2$).

Примем за центр первого кластера точку $\mathbf{x}_1(4; 3)$, а за центр второго кластера точку $\mathbf{x}_2(7; 4)$. Веса обоих кластеров на нулевом шаге равны единице, так как оба кластера содержат по одному объекту.

К какому из двух кластеров отнести точку $\mathbf{x}_3(2; 0)$? Расстояние от точки $\mathbf{x}_1$ до центра первого кластера равно $\rho(\mathbf{x}_1; \mathcal{K}_1) = \sqrt{(2-4)^2 + (0-3)^2} = \sqrt{13}$, а до второго кластера — $\rho(\mathbf{x}_1; \mathcal{K}_2) = \sqrt{(2-7)^2 + (0-4)^2} = \sqrt{41}$.

Так как точка $\mathbf{x}_1$ ближе к первому кластеру, чем ко второму, включаем ее в первый кластер, увеличивая на единицу его вес (теперь он равен 2) и пересчитывая центр кластера:

$$\left(\frac{2+4}{2}; \frac{0+3}{2}\right).$$

Дальнейшие вычисления сведены в табл. 7.3.3 и проиллюстрированы рис. 7.3.4.

Итак, в результате получилось два кластера: первый кластер имеет центр в точке (2; 2) и вес 4, второй кластер имеет центр в точке (7; 4) и вес 1.

Рис. 7.3.4 иллюстрирует изменения центров кластеров в ходе итераций метода K-средних. ❑

Метод K-средних был предложен в 1950-х гг. Г. Штейнгаузом и почти одновременно С. Ллойдом.

Таблица 7.3.3

Кластеризация по методу K-средних в примере 7.3.2

Шаги	Кластеры			
	$\mathcal{K}_1$		$\mathcal{K}_2$	
	центр	вес	центр	вес
0	(4; 3)	1	(7; 4)	1
1	(3; 1,5)	2	(7; 4)	1
2	(2,6666667; 1,3333333)	3	(7; 4)	1
3	(2; 2)	4	(7; 4)	1

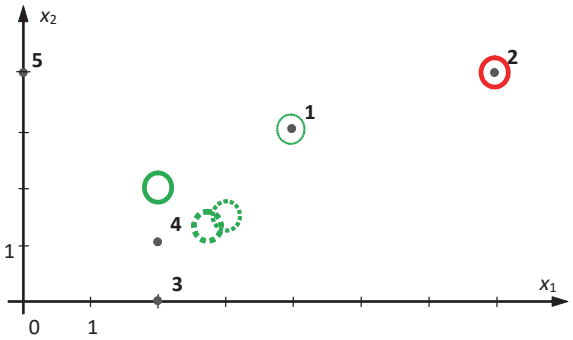


Рис. 7.3.4. *Классификация по методу K-средних в примере 7.3.2*

Задачи

443. Проведите классификацию объектов из примера 7.3.2 методом *K*-средних, выбрав в качестве начального разбиения точки $\mathbf{x}_2(7; 4)$ и $\mathbf{x}_5(0; 4)$.

444. Проведите классификацию пяти объектов:

$$(2; 4),\ (8; 6),\ (-2; -2),\ (-2; 0),\ (-6; 6)$$

иерархическими агломеративными методами и методом *K*-средних.

§ 7.4. Принципы машинного обучения

7.4.1. Парадигма машинного обучения

Под *машинным обучением* понимается разработка моделей (алгоритмов, роботов), предсказывающих результат на основании входных данных. Процесс машинного обучения начинается со сбора известных данных, затем на основании этих известных данных строится (или, как принято говорить, обучается) модель машинного обучения, которая затем используется для того, чтобы предсказывать неизвестное.

В зависимости от **практических задач, для которых предназначены алгоритмы машинного обучения,** они делятся на четыре класса (рис. 7.4.6).



Рис. 7.4.6. *Обучение с учителем и без учителя*

В задачах регрессии требуется, зная набор значений факторных признаков $x^{(1)}$, $x^{(1)}$, ..., $x^{(m)}$, предсказывать значение результативного признака Y. Например, по оси абсцисс на графике «Регрессия», представленном на рис. 7.4.6, может быть отложен возраст клиента, а по оси ординат — размер кредитного лимита. Точки при этом соответствуют известным данным, а красная пунктирная и зеленая сплошная линии — двум функциям, использующимся для прогноза. Замечаем, что более сложная функция (зеленая сплошная линия) дает лучшие прогнозы, но не идеальные.

Наиболее популярные задачи регрессии — прогнозирование цен различных активов, спроса, объема продаж, доходов, лимитов и т. д.

В задачах классификации с обучением, зная набор значений факторных признаков $x^{(1)}$, $x^{(1)}$, ..., $x^{(m)}$, нужно предсказывать, к какому классу Y относится соответствующий объект. Например, по оси абсцисс на графике «Классификация с обучением» на рис. 7.4.6 может быть отложен возраст клиента, а по оси ординат — размер годового дохода. Синие точки при этом соответствуют известным данным о «хороших» клиентах, красные точки — известным данным о «плохих» клиентах, а зеленая линия используется для классификации: если очередному клиенту соответствует точка правее и выше зеленой линии, его можно считать «хорошим», иначе — «плохим». Замечаем, что алгоритм иногда ошибается: часть синих точек находится левее и ниже разделяющей линии (это соответствует отказам в выдаче кредита «хорошим» клиентам), часть красных точек находится правее и выше линии (это — кредиты, которые выданы «плохим» клиентам).

Самые популярные задачи, в которых применяется классификация с обучением, — это кредитный скоринг, анализ оттока клиентов, анализ откликов на маркетинговые кампании, идентификация клиентов, присвоение рейтингов эмитентам, прогнозирование банкротств, различные прогнозирование поломок оборудования, определение языка текстовых документов, поиск похожих документов, анализ тональности текстов, распознавание рукописных символов и т. д.

Регрессию и классификацию с обучением называют *обучением с учителем* или *обучением на размеченных данных*: есть некоторый «учитель», который размечает данные, рассказывая машине, как правильно предсказывать в каждом конкретном случае.[1] В задачах регрессии в исходном наборе данных известно, какие значения результативного признака были у каждого объекта. В задачах классификации с обучением, заранее ясно, какие из известных объектов относятся к классу «0», а какие — к классу «1».

Обучение без учителя используется, когда невозможно получить размеченный набор данных. Например, если мы ищем мошеннические транзакции, никто не разметит нам заранее все их возможные типы. Если мы сегментируем рынок, мы не знаем заранее, на какие сегменты делится рынок, сколько таких сегментов.

В задачах кластерного анализа требуется, зная набор значений признаков $x^{(1)}$, $x^{(1)}$, ..., $x^{(m)}$ на достаточно большом количестве объектов, разделить эти объекты на группы так, чтобы внутри каждой группы объекты были в определенном смысле похожи друг на друга, а в разных группах —

[1] Поэтому есть еще одно название обучения с учителем: *обучение по прецедентам.*

достаточно сильно отличались. Например, по оси абсцисс на графике «Кластерный анализ» (рис. 7.4.6) может быть отложен возраст клиента, а по оси ординат — размер годового дохода. Точки при этом соответствуют известным данным, а зеленые овалы обозначают классы клиентов, для каждого из которых можно предложить свой набор банковских продуктов.

Среди популярных задач, решаемых с помощью кластерного анализа, можно отметить сегментирование рынка (по типам покупателей, по лояльности и т. д.), разведочный анализ структуры новых данных, объединение близких точек на картах, сжатие изображений и др.

В задачах поиска аномалий требуется, зная набор значений признаков $x^{(1)}$, $x^{(1)}$, ..., $x^{(m)}$, находить объекты, сильно отличающиеся от остальных. Например, по оси абсцисс на графике «Поиск аномалий» (рис. 7.4.6) может быть отложена сумма последней операции, а по оси ординат — сумма предпоследней операции. Точки при этом соответствуют известным данным, а зеленый овал отделяет «честные» транзакции от мошеннических.

Самые популярные задачи поиска аномалий — определение мошеннических транзакций, поиск нарушителей налогового законодательства, поиск операций, связанных с отмыванием денег и финансированием терроризма, отсев нежелательных сообщений электронной почты, обнаружение вторжений, обнаружение инсайдеров на биржах и др.

Одну и ту же задачу почти всегда можно решить с помощью разных алгоритмов, но при этом, конечно, от выбора алгоритма зависит и точность, и скорость работы модели машинного обучения. Многие из алгоритмов машинного обучения используются уже достаточно давно (например, линейная и логистическая регрессия), другие методы (например, нейронные сети и ансамбли) стали применяться лишь в последние годы, потому что раньше не хватало производительности и памяти компьютеров (рис. 7.4.7).

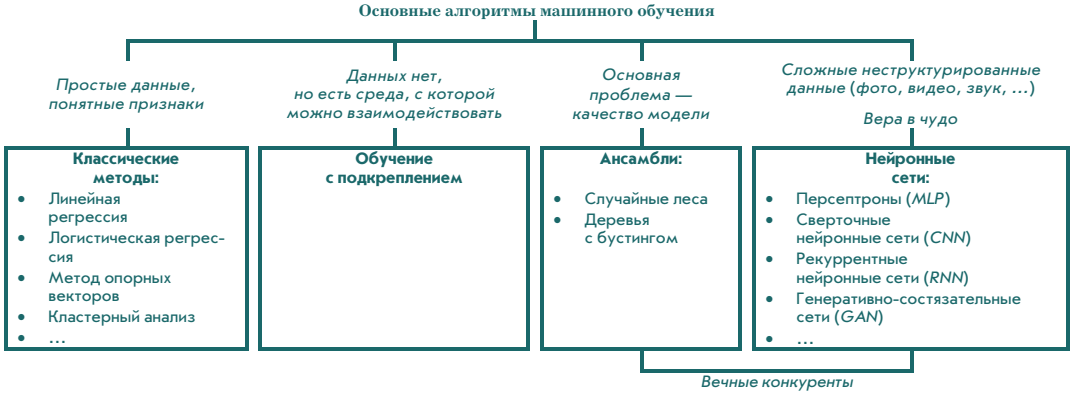


Рис. 7.4.7. *Основные алгоритмы машинного обучения*

Значительная часть систем машинного обучения использует классические алгоритмы: линейная и логистическая регрессия, метод опорных векторов, кластерный анализ. Эти алгоритмы относительно просты в разработке, они проверены практикой, а их результаты поддаются понятной для бизнеса интерпретации.

В том случае, когда точность предсказаний очень важна, как правило, удается добиться хороших результатов с помощью ансамблевых методов —

деревьев и лесов решений. Во многих крупных организациях очень популярны нейронные сети, с помощью которых стараются решать все возникающие задачи. На самом деле, сложность нейросетевых алгоритмов, трудоемкость их разработки и длительное время обучения оправданы в ситуациях анализа неструктурированных данных типа текстов на естественном языке, фото, видео и т. д. В случае простых и понятных признаков эффект от использования нейронных сетей не всегда сопоставим с затратами на их разработку и может проявиться лишь при использовании системы машинного обучения в очень больших масштабах. Некоторые нейронные сети, в которых (по мнению их разработчиков) достаточно много слоев, относят к системам глубокого (или «глубинного») обучения. При этом никакого общепринятого определения, какую сеть можно назвать глубокой, а какую нельзя, не существует.

В последнее время начали использоваться на практике системы обучения с подкреплением, когда модель обучается, взаимодействуя с некоторой средой. Обучение с подкреплением представляет собой частный случай обучения с учителем, только в качестве «учителя» выступает внешняя среда, выдающая на принятые решения *сигналы подкрепления*, позволяющие понять, насколько удачно было принято решение, и на основе этого скорректировать стратегию для достижения лучшего результата. Например, такая система может учиться торговать на финансовых рынках, получая положительное подкрепление в ситуациях, когда удается заработать, и отрицательное подкрепление, когда торговое решение приводит к потерям. Система *AlphaGo*, разработанная компанией *DeepMind* и выигравшая в 2016 г. у профессионала 9-го дана (высшего ранга) Л. Седоля, основана на использовании сверточных нейронных сетей и обучения с подкреплением: сеть постоянно обучалась, играя с одной из ранее обученных сетей, выбранной случайно.[1]

На рис. 7.4.8 и 7.4.9 представлены путеводители по сферам применения машинного обучения и по алгоритмам машинного обучения. Эти путеводители заведомо неполны, они включает лишь наиболее популярные методы и задачи.



Рис. 7.4.8. *Путеводитель по задачам машинного обучения*

[1] Ранее считалось, что компьютер никогда не сможет выиграть у человека в игру го, так как количество возможных ходов в этой игре настолько велико, что их перебор принципиально невозможен.

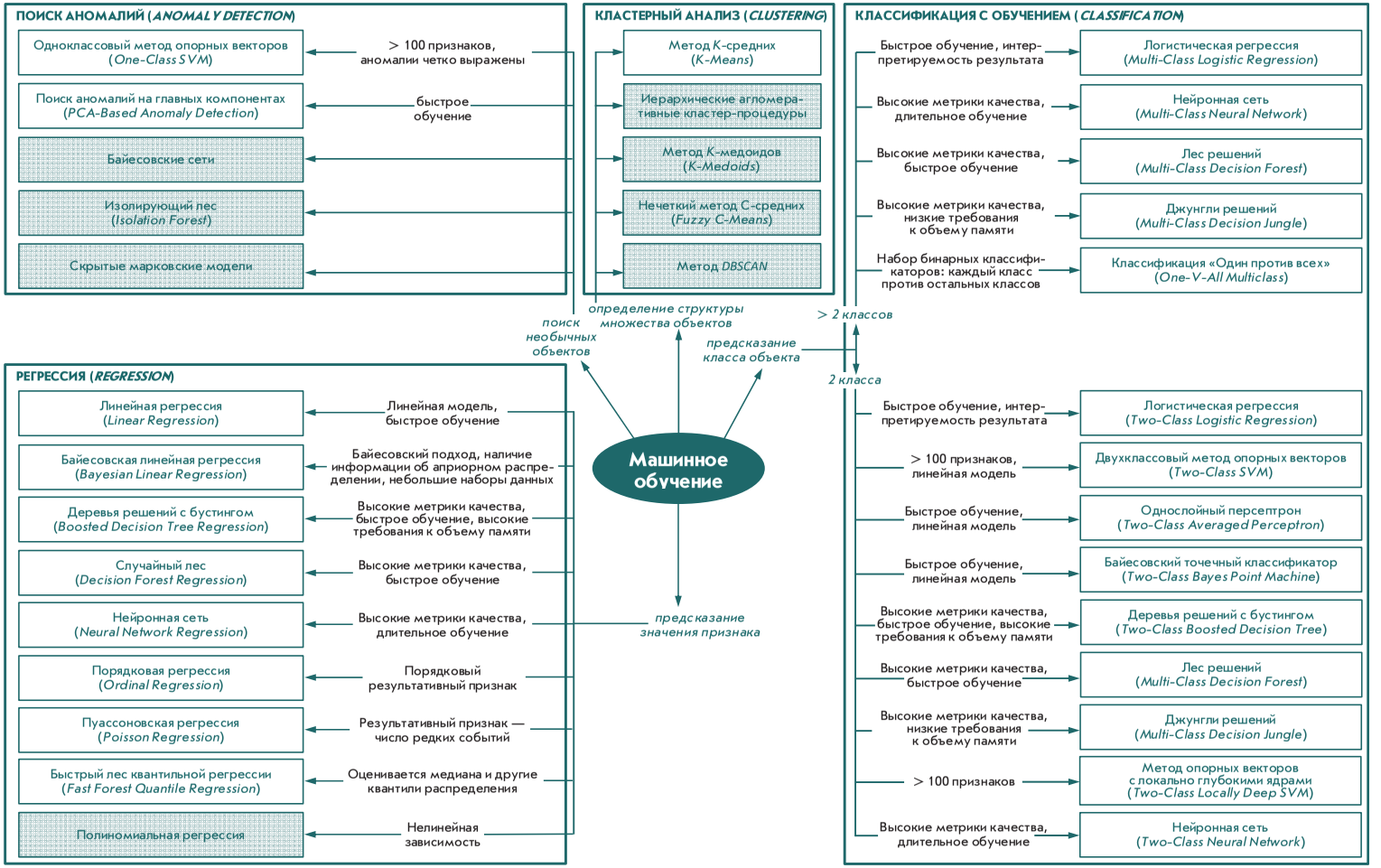


Рис. 7.4.9. *Путеводитель по алгоритмам машинного обучения*

В последние годы появились облачные платформы машинного обучения. Основным преимуществом таких систем является гибкая масштабируемость: выделение и высвобождение вычислительных ресурсов происходит мгновенно в соответствии с решаемыми задачами. Amazon Machine Learning, Google Machine Learning Engine и IBM Watson Machine Learning реализуют возможность запуска на своих облачных мощностях базовых алгоритмов регрессии и классификации с обучением.

При этом Google Machine Learning Engine предоставляет возможность запуска в облачной среде моделей TensorFlow, Amazon Machine Learning дает возможность выполнять алгоритмы не только из TensorFlow, но и из Caffe2 и Apache MXNet, а IBM Watson Machine Learning позволяет в интерактивной форме использовать алгоритмы Apache Spark и IBM SPSS.

Существует и российский продукт Loginom, позволяющий создавать без программирования модели машинного обучения. Loginom имеет очень продуманный и удобный интерфейс, при этом систему можно запускать в облачной среде, а можно развернуть локально и использовать без подключения к интернету — это актуально для тех организаций, которые опасаются по поводу безопасности передачи конфиденциальных наборов данных в облачные хранилища.

Однако по количеству реализованных алгоритмов и возможностям настройки и интеграции все эти облачные сервисы существенно уступают платформе Microsoft Azure Machine Learning Studio (сокращенно — Microsoft Azure ML Studio), позволяющей с помощью простого графического интерфейса строить модели машинного обучения с использованием множества стандартных алгоритмов классификации, регрессии, кластерного анализа и поиска аномалий, а также встраивать в эти модели собственный код на SQL, Python и R.

На путеводителе по машинному обучению, приведенному в этой книге на рис. 7.4.9, алгоритмы, реализованные в Microsoft Azure ML Studio, представлены на прямоугольниках с белым фоном.

В департаменте анализа данных, принятия решений и финансовых технологий Финуниверситета Microsoft Azure ML Studio используется не только для проектирования алгоритмов, но и для создания предсказательных облачных сервисов, которые могут получать на вход значения факторных признаков и в реальном времени выдавать в ответ результат предсказания.

В частности, в нашей системе интеллектуального мониторинга вовлеченности студентов на основе анализа видеопотоков из аудиторий для распознавания вовлеченности в реальном времени применяется веб-сервис, построенный в Microsoft Azure ML Studio.

Для работы в Microsoft Azure ML Studio нужна учетная запись *Microsoft*. Многие организации (в частности, многие университеты) предоставляют своим сотрудникам (студентам) адреса корпоративной электронной почты, автоматически являющиеся учетными записями *Microsoft* (у студентов Финансового университета это адрес электронной почты в домене @edu.fa.ru).

Если у Вас нет учетной записи *Microsoft*, то для ее получения необходимо пройти по адресу https://signup.live.com/ и заполнить соответствующую анкету (рис. 7.4.10). В качестве имени пользователя можно использовать любой имеющийся у Вас адрес электронной почты.

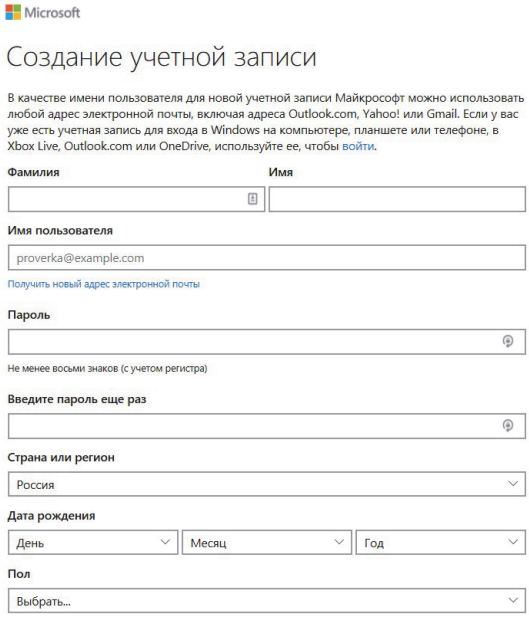


Рис. 7.4.10. *Анкета для получения учетной записи Microsoft*

Доступ к облачной службе Microsoft Azure ML Studio предоставляется по подписке, которая может быть приобретена с одним из двух уровней предоставления услуг:

- базовый уровень предоставляется бесплатно, при этом существуют ограничения: не более 100 модулей в каждом эксперименте; не более 1 ч вычислительного времени на каждый эксперимент; не более 10 Гбайт дискового пространства для хранения экспериментов и др.;
- стандартный уровень предоставляется на платной основе в рамках подписки Microsoft Azure (стоимость одного рабочего места по состоянию на лето 2018 г. составляет 624,38 руб./мес., дополнительно к этому оплачивается время экспериментов по тарифу 62,50 руб./ч.), при этом ограничения базового уровня снимаются.

Для выполнения заданий из этой книги вполне достаточно подписки на базовый уровень Microsoft Azure ML Studio, однако если у Вас есть подписка на стандартный уровень, то можно пользоваться ею.

Для получения подписки на базовый уровень Microsoft Azure ML Studio необходимо открыть сайт https://studio.azureml.net, затем нажать на кнопку «Sign up here»(рис. 7.4.11) и выбрать вариант «Free Workspace» (рис. 7.4.12).

Далее необходимо следовать инструкциям, при этом потребуется выполнить вход в учетную запись *Microsoft*.

Замечание. Подписка на базовый уровень Microsoft Azure ML Studio — это не то же самое, что пробная подписка на Microsoft Azure. В рамках подписки на базовый уровень Microsoft Azure ML Studio предоставляется бессрочный доступ к графической среде разработки моделей машинного обучения, подписка на Microsoft Azure при этом не требуется. В отличие от этого, в рамках пробной подписки на Microsoft Azure предоставляется доступ сроком на один месяц ко всем службам Microsoft Azure (в том числе, к инструменту Azure ML Studio).

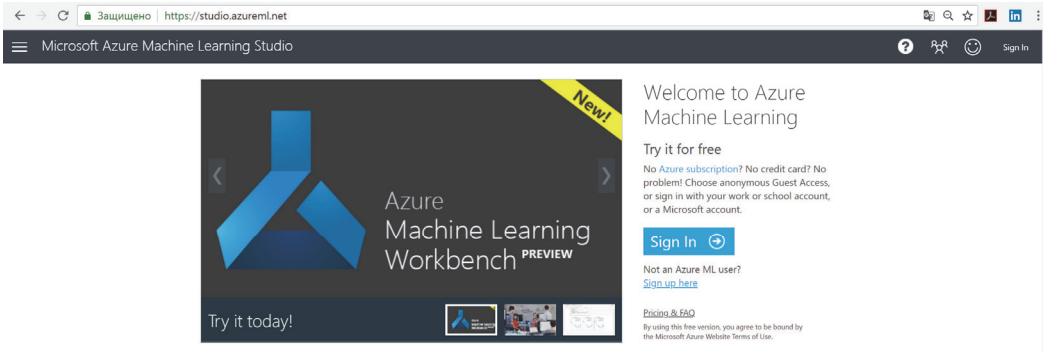


Рис. 7.4.11. *Страница приветствия Microsoft Azure ML Studio*

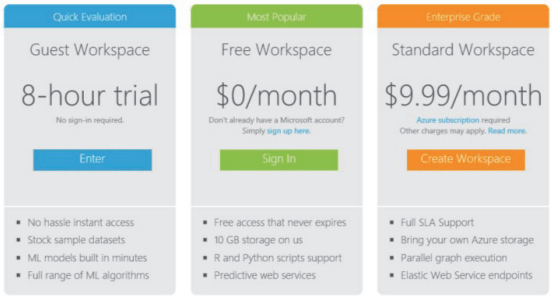


Рис. 7.4.12. *Тарифные планы Microsoft Azure ML Studio*

Задачи

445. Приведите примеры задач, для решения которых применимы алгоритмы: *а*) классификации с обучением; *б*) регрессии; *в*) кластерного анализа; *г*) поиска аномалий.

446. Объясните, в чем принципиальное отличие методов обучения с учителем и без учителя.

447. Получите подписку на базовый уровень *Microsoft Azure ML Studio*.

7.4.2. Предварительная обработка данных в *Microsoft Azure ML Studio*

Разработка моделей машинного обучения обычно проходит шесть стадий (рис. 7.4.13).[1]

На первой стадии формулируется **постановка задачи** с точки зрения бизнеса. При этом бизнес-цели соотносятся с возможностями, анализируются возможные риски и определяются способы и пути оценки качества будущих результатов.

Пример 7.4.1. Рассмотрим постановку задачи разработаки системы машинного обучения для автоматизации кредитного скоринга в банке.

Такая система должна использовать данные о заемщиках, которым ранее уже выдавались кредиты и о которых известно, кто из них погасил задолженность, а кто нет.

[1] Стрелки на рис. 7.4.13 показывают, что в большинстве проектов разработки систем машинного обучения приходится неоднократно возвращаться к предыдущим стадиям, а затем снова двигаться вперед.

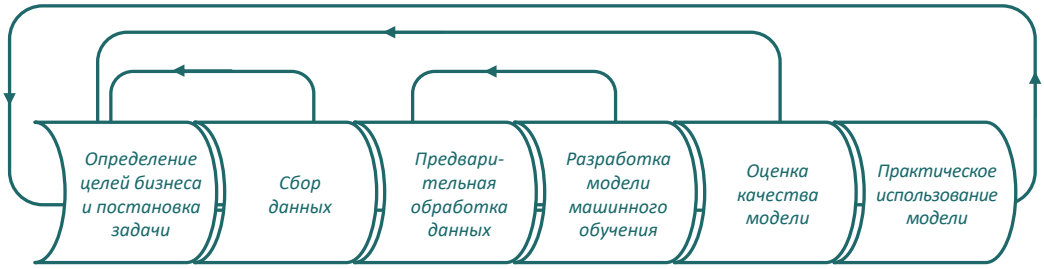


Рис. 7.4.13. *Список экспериментов в Microsoft Azure ML Studio*

Данные для обучения модели машинного обучения есть: в банковских информационных системах имеются данные о самих заемщиках (из заявлений о предоставлении кредита), а также данные о платежах, которыми заемщики погашали кредиты.

При практическом использовании системы кредитного скоринга существует два основных типа рисков:

- если отказать в выдаче кредита надежному заемщику, банк не получит процентный доход;
- если выдать кредит ненадежному заемщику, банк потеряет полную сумму кредита или значительную ее часть.

Поэтому качество разрабатываемой системы машинного обучения должно определяться как можно меньшим количеством ошибок первого рода, когда система называет «плохих» заемщиков «хорошими», и ошибок второго рода, когда «хороших» заемщиков система принимает за «плохих».

Особо отметим при этом, что для банка ущерб от ошибки первого рода существенно выше, чем от ошибки второго рода. ❑

Вторая стадия процесса разработки модели машинного обучения — это сбор данных. Основные задачи на этой стадии: понять, какие данные есть в наличии и каких не хватает, познакомиться с имеющимися данными как можно ближе, выявить проблемы с качеством данных, в том числе ошибки, пропущенные значения, повторяющиеся строки, попробовать отыскать сформулировать гипотезы о наличии скрытых закономерностей в данных.

После того, как данные собраны и проанализированы, происходит переход на третью стадию процесса разработки модели машинного обучения — стадию **предварительной обработки данных**, в ходе которой данные из разных источников собираются в единый набор, исправляются ошибки в данных, удаляются повторяющиеся строки, заполняются пропуски, а также с учетом бизнес-логики формируются синтетические признаки.

Microsoft Azure ML Studio позволяет создавать эксперименты, в которых можно управлять данными, создавать модели машинного обучения и визуализировать результаты. Проиллюстрируем практику сбора и предварительной обработки данных в среде Microsoft Azure ML Studio проиллюстрируем на примере решения задачи кредитного скоринга.

Пример 7.4.2. В файле «German Credit Cards UCI.csv» находится набор данных о клиентах некоторого банка, которые уже получали кредит, причем некоторые из них вернули кредит в срок, другие нарушили условия договора.

Цель разработки модели машинного обучения заключается в том, чтобы на основании известных данных о существующих клиентах научиться предсказывать, кто из новых клиентов, подающих заявления на получение кредита, сможет вернуть кредит в соответствии с условиями договора, а кто нет.

Пройдем по адресу https://studio.azureml.net и войдем в систему под учетной записью *Microsoft*, связанной с подпиской на Microsoft Azure ML Studio.

Если на экране появится окно приветствия, закроем его, нажав кнопку «×». Если затем появится страница, содержащая коллекцию примеров экспериментов (Microsoft Samples), закроем и ее, нажав кнопку «×».

После этого экран будет выглядеть, как на рис. 7.4.14.

Замечание. Эксперименты, создаваемые в Microsoft Azure ML Studio, и используемые наборы данных можно организовывать в проекты, чтобы, когда их станет много, в них было легче ориентироваться. Чтобы создать проект, можно нажать значок «Projects» и далее «Create Project».

Создадим новый эксперимент по машинному обучению. Для этого нажмем кнопку «+ NEW», которая расположена в левом нижнем углу среды Microsoft Azure ML Studio. Затем выберем категорию «Experiment» и из коллекции примеров экспериментов («Microsoft Samples») выберем пустой эксперимент — «Blank Experiment», который выглядит, как показано на рис. 7.4.15.

Изменим название эксперимента — вместо названия по умолчанию «Experiment created on <today's date>» напишем: «Кредитный скоринг».

Фоном в окне проектирования эксперимента служит подсказка по созданию моделей машинного обучения: начать нужно с того, что добавить в модель набор данных.

В файле «German Credit Cards UCI.csv» содержатся данные одного из немецких банков, собранные в начале 1990-х гг. профессором Гамбурского университета Г. Хофманном. Банк в этот момент создавал модель кредитного скоринга для принятия решений о выдаче кредитов иностранцам.

Если открыть файл «German Credit Cards UCI.csv» с помощью любого текстового редактора, мы увидим, что он содержит 1013 строк, в которых находятся значения признаков, разделенные запятыми, — обычно в машинном обучении удобно работать именно с такими csv-файлами (рис. 7.4.16).

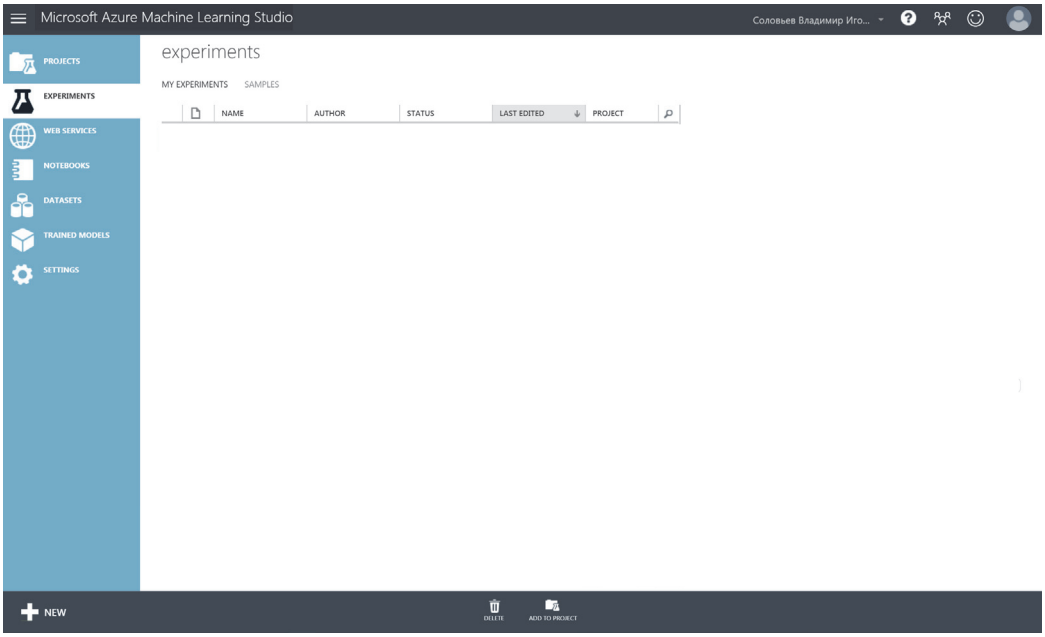


Рис. 7.4.14. *Список экспериментов в Microsoft Azure ML Studio*

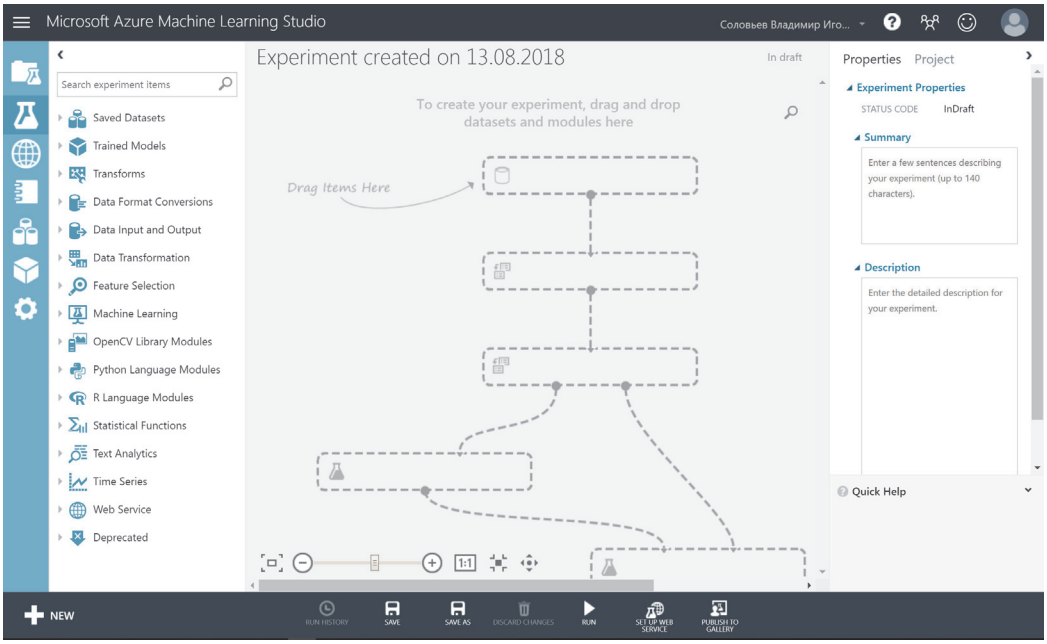


Рис. 7.4.15. *Пустой эксперимент*

```
Customer ID,Checking Account Status,Loan Duration (Months),Credit History,Purpose,Loan Amount,Savings Account Balance, ...  ,Loan Status
1122334,< 0 DM,6,critical account - other non-bank loans,radio/television,1169,unknown/none, ...                          , fully paid
6156361,0 - 200 DM,48,current loans paid,radio/television,5951,< 100 DM, ...                                            ,charged off
2051359,none,12,critical account - other non-bank loans,education,2096,< 100 DM, ...                                    ,fully paid
8740590,< 0 DM,42,current loans paid,furniture/equipment,7882,< 100 DM, ...                                              ,fully paid
... ... ... 
```

- Number of Dependents — количество иждивенцев;
- Telephone — наличие телефона;
- Foreign Worker Status — наличие статуса иностранного рабочего;
- Loan Status — статус кредита: погашен («fully paid») или не погашен («charged off»).

Далее каждая строка содержит разделенные запятыми значения признаков для конкретных клиентов — по одной строке на клиента.

Результативным признаком в данном случае является признак «Loan Status» — именно его необходимо научиться предсказывать. Сейчас в этом столбце находятся данные о том, кто из существующих клиентов вернул кредит, а кто нет.

Загрузим набор исходных данных в Microsoft Azure ML Studio. Для этого нажмем кнопку «+ NEW», расположенную в левом нижнем углу, а затем выберем категорию «DATASET» (рис. 7.4.17). Далее выберем загрузку набора данных их локального файла — «FROM LOCAL FILE», в появившемся диалоговом окне «Upload a new dataset» выберем файл «German Credit Cards UCI.csv» на локальном компьютере. Введем следующую информацию:

This is a new version of an existing dataset: Очистим выделение;
Enter a name for the new dataset: German Credit Cards UCI;
Select a type for the new dataset: Generic CSV file with a header (.csv);
Provide an optional description: Данные о кредитах.

После того, как мы нажмем кнопку «✓» (рис. 7.4.18), набор данных начнет загружаться. Дождемся завершения загрузки, затем нажмем кнопку «OK ✓» в строке состояния (в нижней части Microsoft Azure ML Studio).

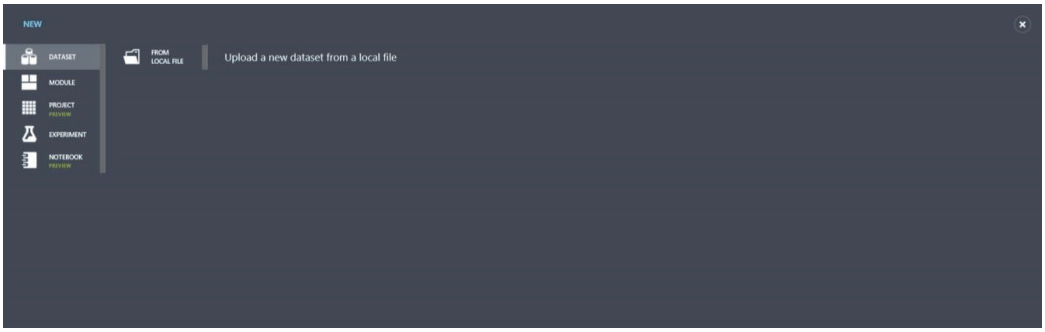


Рис. 7.4.17 *Загрузка нового набора данных*

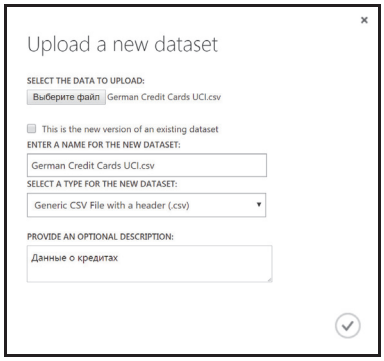


Рис. 7.4.18. *Окно ввода параметров нового набора данных*

На панели модулей эксперимента (в левой части экрана) развернем наборы данных «Saved Datasets > My Datasets», убедимся, что набор данных «German Credit Cards UCI» загружен, и перенесем его мышкой на рабочую область. Эксперимент при этом будет выглядеть, как на рис. 7.4.19.

Сохраним эксперимент, нажав кнопку «SAVE» в нижней части экрана.

Замечание. В Microsoft Azure ML Studio реализовано автоматическое сохранение экспериментов, но на всякий случай рекомендуется периодически сохранять их и вручную.

Нажмем левой кнопкой мыши на выходной порт модуля набора данных «German Credit Cards UCI» и выберем пункт меню «Visualize», чтобы просмотреть набор данных (рис. 7.4.20). На экране появится окно с соответствующей визуализацией (рис. 7.4.21).

Замечание. *Выходные порты* — это маленькие кружки в нижней части модуля, а *входные порты* — это маленькие кружки в нижней части модуля.

Убедимся, что это тот набор данных, который мы загружали (сравнив его с рис. 7.4.16). Заметим, что в анализируемом наборе данных 1014 объектов — строк (rows) и 22 признака — столбца (columns).

Нажмем левой кнопкой мыши на заголовок первого столбца («Customer ID»), в правой части экрана появятся некоторые данные описательной статистики этого признака (рис. 7.4.22). Замечаем, что хотя в наборе данных 1014 строк, признак «Customer ID» — идентификатор клиента — содержит всего 1000 уникальных значений (Unique Values). Значит, информация о некоторых клиентах в наборе данных повторяется. Пропущенных значений (Missing Values) этот признак не имеет.

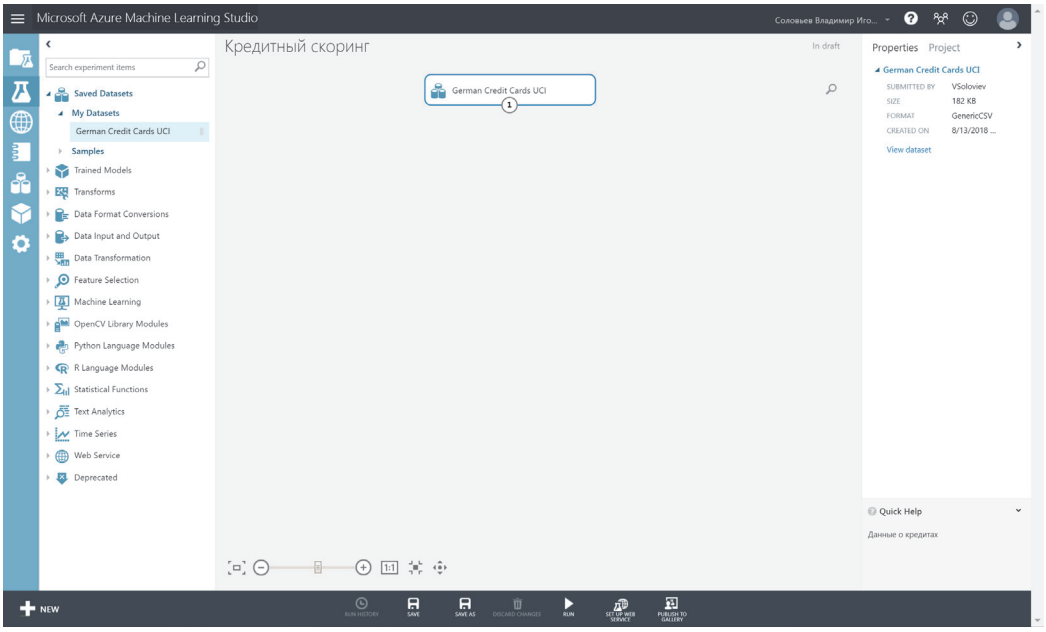


Рис. 7.4.19. *Эксперимент с добавленным набором данных*

Закроем окно визуализации набора данных, нажав кнопку «×» в правом верхнем углу этого окна.

Удалим повторяющиеся значения. Для этого на панели модулей раскроем меню «Data Transformation > Manipulation», выберем модуль «Remove

Duplicate Rows» и перенесем его мышкой в рабочую область Microsoft Azure ML Studio. Соединим выходной порт модуля набора данных «German Credit Cards UCI» со входным портом модуля «Remove Duplicate Rows». Эксперимент примет вид, представленный на рис. 7.4.23.

Замечание. Вместо того, чтобы искать модули по древовидному меню, для быстрого доступа к ним можно воспользоваться строкой поиска, которая находится над меню («Search experiment items»): в нее можно ввести название нужного модуля.

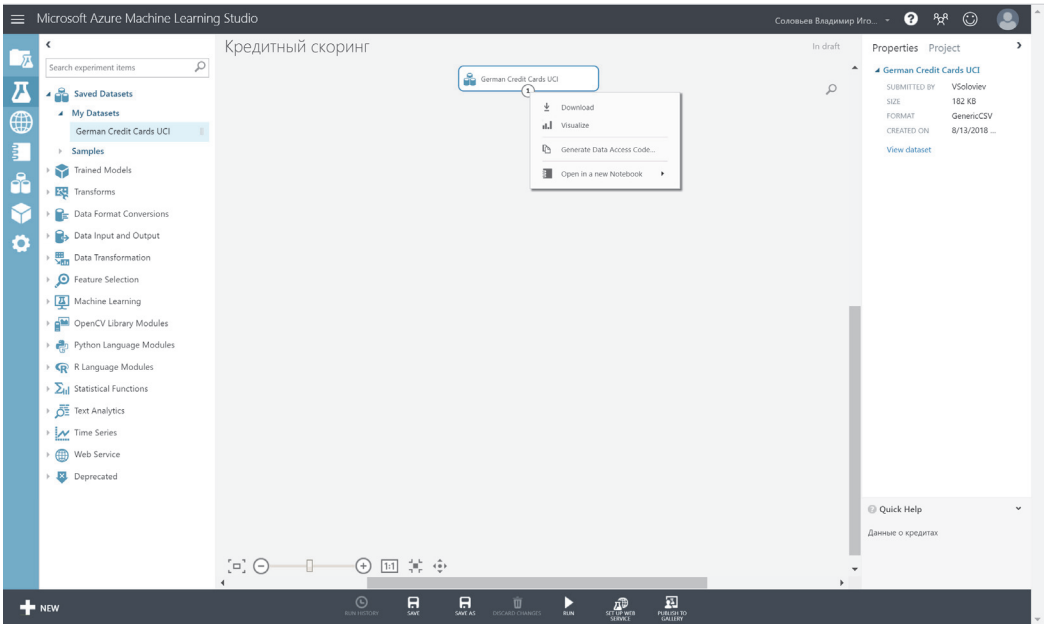


Рис. 7.4.20. *Меню нижнего порта набора данных*

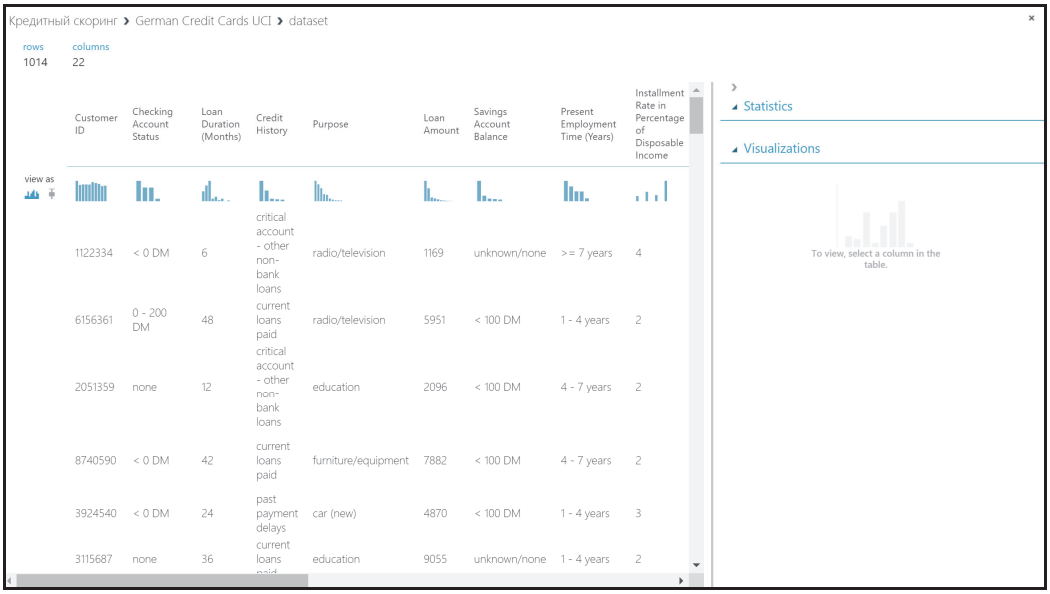


Рис. 7.4.21. *Визуализация набора данных*

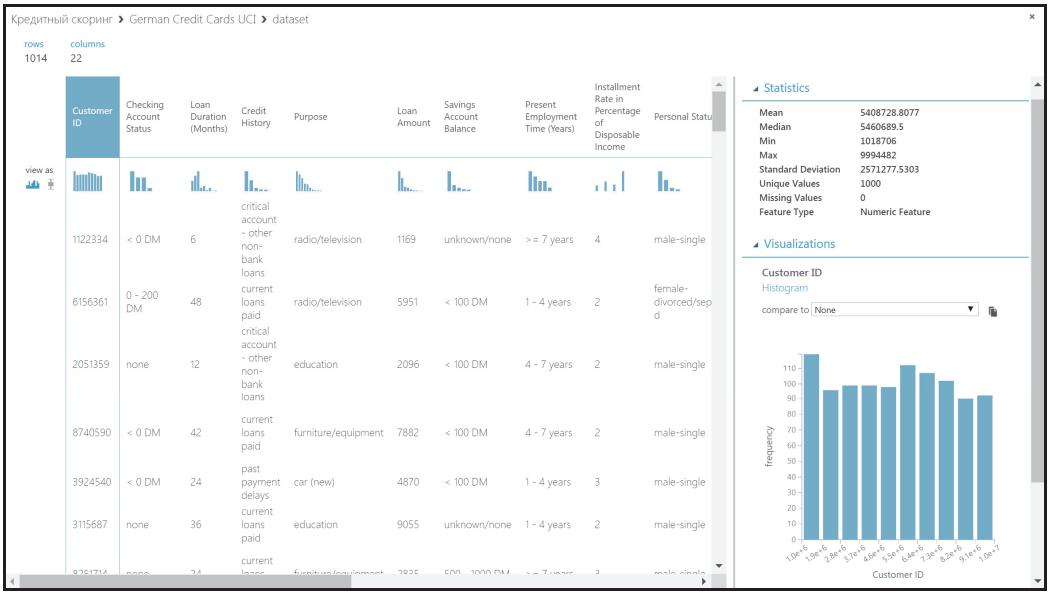


Рис. 7.4.22. *Визуализация описательной статистики признака «Customer ID»*

Замечаем подсказку — восклицательный знак в красном кружке («❗»). Если подвести к нему мышку, появится пояснение «Value required» (необходимо ввести значения параметров модуля).

Выберем модуль «Remove Duplicate Rows» и на панели параметров модуля («Properties»), которая находится в правой части экрана, нажмем кнопку «Launch column selector», чтобы выбрать признаки, повторяющиеся значения в которых идентифицируют повторяющиеся строки (рис. 7.4.23).

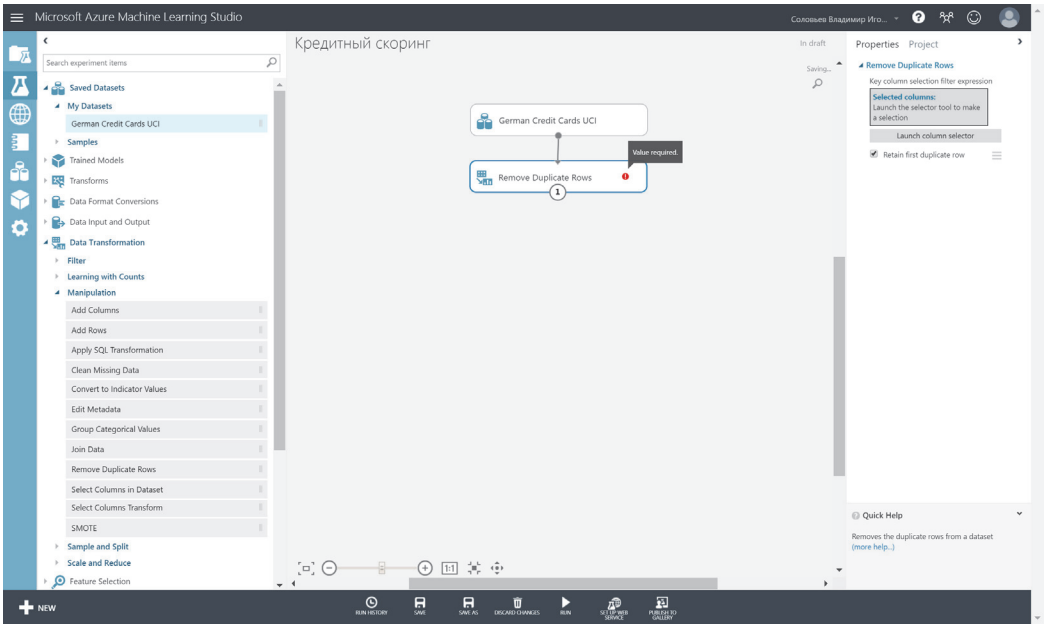


Рис. 7.4.23. *Эксперимент с добавленным модулем удаления повторяющихся строк*

В появившемся окне выбора столбцов (рис. 7.4.24) с помощью стрелки («>») перенесем признак «Customer ID» из списка доступных столбцов («AVAILABLE COLUMNS») в список выбранных столбцов («SELECTED COLUMNS») и нажмем кнопку «✓».

Запустим эксперимент, нажав кнопку «▶ RUN» в нижней части экрана. После того, как он закончит выполняться (в модуль удаления повторяющихся строк будет отмечен зеленой галочкой — «✔»), визуализируем результаты, нажав левой кнопкой мыши на выходной порт модуля «Remove Duplicate Rows». В появившемся окне (рис. 7.4.25) замечаем, что теперь в наборе данных осталось 1000 строк — повторяющиеся строки удалены.

Проанализируем описательную статистику признаков, поочередно нажимая левой кнопкой мыши на заголовки столбцов, и зафиксируем некоторые наблюдения.

Гистограмма распределения признака «Checking Account Status» показывает, что 394 клиента из 1000 не имеют текущего счета, 274 клиента имеют на текущем счете отрицательные остатки, 269 клиентов имеют остатки от 0 до 200 DM и у 63 клиентов остатки на текущих счетах превышают 200 DM (рис. 7.4.26). Пропущенных значений (Missing Values) данный признак не имеет.

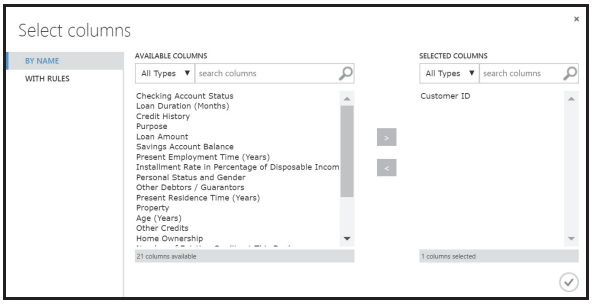


Рис. 7.4.24. *Окно выбора столбцов*

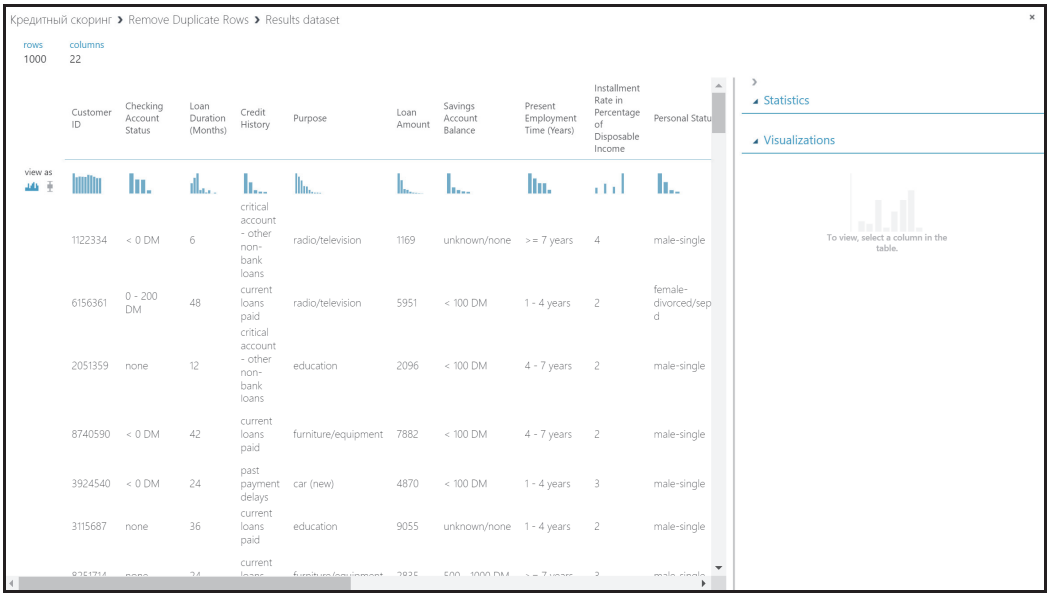


Рис. 7.4.25. *Визуализация результата удаления повторяющихся строк*

Гистограмма распределения признака «Loan Duration (Months)» (рис. 7.4.27) показывает, что минимальный срок кредита в анализируемом наборе данных составляет 4 мес. («Min»), а максимальный срок — 72 мес. («Max»). Выборочное среднее («Mean») для срока кредита составляет 20,903 мес., медиана («Median») срока кредита равна 18 мес. (т. е. половина клиентов брали кредит на срок, не больший 18 мес., а половина — на срок, больший 18 мес.). Кроме того, отметим, что 770 клиентов (= 171 + 262 + 337) брали кредиты на срок до 24 мес., а оставшиеся 230 клиентов (= 1000 − 770) — на срок более 24 мес. — распределение признака «Loan Duration (Months)» скошено вправо. Пропущенных значений (Missing Values) данный признак не имеет.

Гистограмма распределения результативного признака «Loan Status» (рис. 7.4.28) демонстрирует, что 70% клиентов банка погасили свои кредиты, а проблемные кредиты были у 30%. Это — распространенное явление. Обычно хороших клиентов больше чем плохих. В зависимости от профиля риска банка и группы клиентов «плохие» заемщики составляют долю от 5—7 до 30% от общего числа клиентов, поскольку банк старается выдавать кредиты тем, кто их вернет.

То, что хороших заемщиков больше, конечно, хорошо. Но из-за такой несбалансированности классов модель машинного обучения будет при обучении «видеть» много примеров «хороших» заемщиков и мало — «плохих», поэтому она «научится» распознавать «хороших» заемщиков лучше, чем «плохих», и, возможно, будет слишком часто рекомендовать выдавать кредиты ненадежным клиентам, считая их «хорошими». В связи с этим, возможно, следует каким-то образом сбалансировать классы.

Не закрывая окно визуализации описательной статистики признака «Loan Status», визуализируем его возможную зависимость от других признаков. Для этого в выпадающем списке признаков для сравнения («compare to») выберем признак «Loan Duration (Months)».

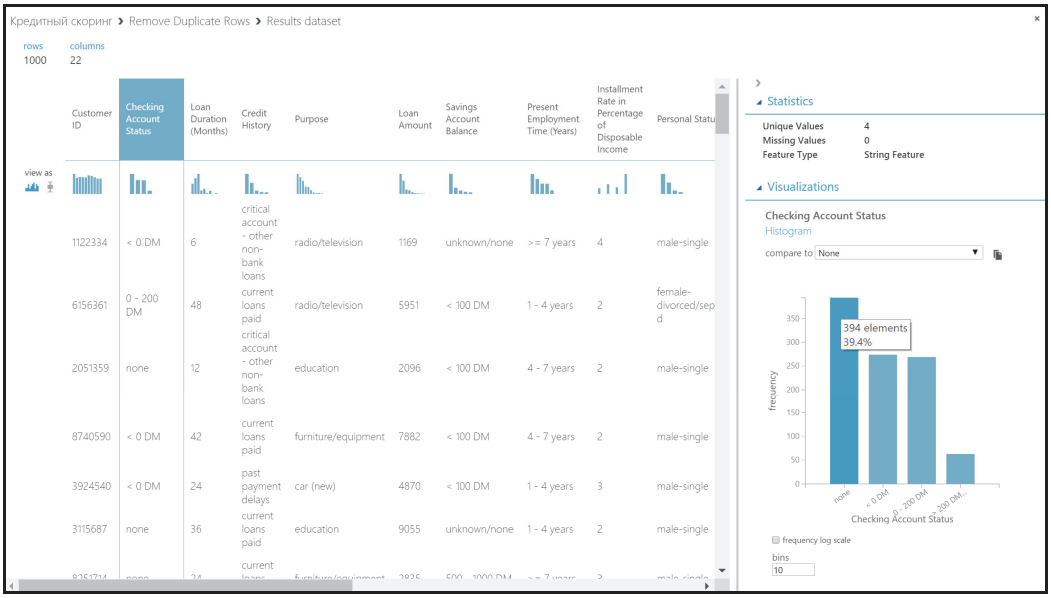


Рис. 7.4.26. *Визуализация описательной статистики признака «Checking Account Status»*

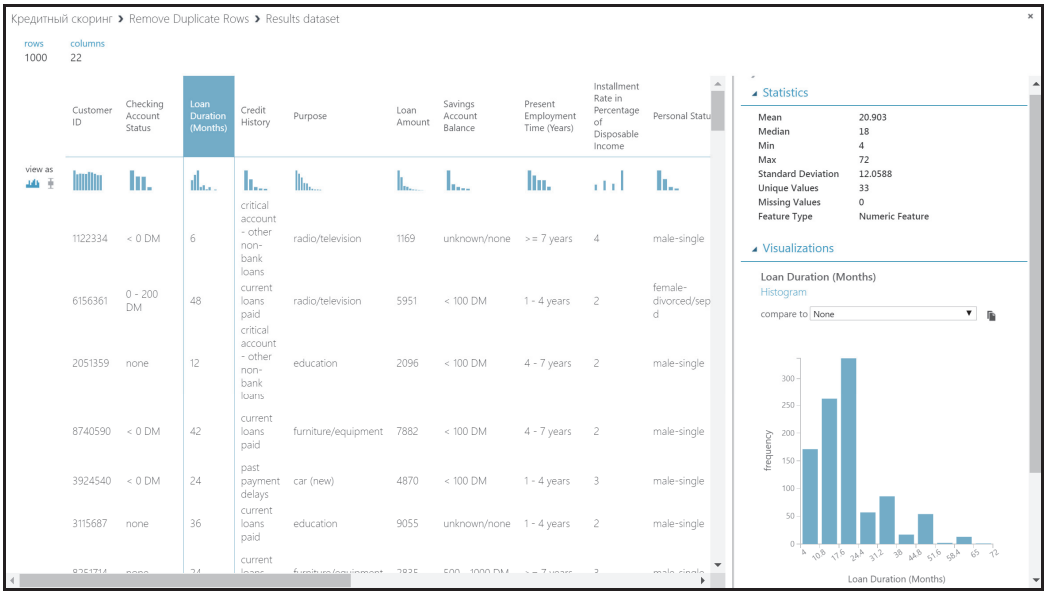


Рис. 7.4.27. *Визуализация описательной статистики признака «Loan Duration (Months)»*



Рис. 7.4.28. *Визуализация описательной статистики признака «Loan Status»*

Сравнение диаграмм размаха срока кредита для двух групп заемщиков (рис. 7.4.29) дает возможность заключить, что «хорошие» клиенты берут кредиты на более короткие сроки, чем «плохие» заемщики.

Визуализировав аналогичным образом связь признака «Loan Status» с признаками «Age», «Loan Amount», «Number of Existing Credits at This Bank», «Credit History», «Property» и «Foreign Worker Status» (рис. 7.4.30), можем заметить, что «хорошие» заемщики обычно старше «плохих», берут кредиты меньших размеров и имеют в собственности недвижимость.



Рис. 7.4.29. *Визуализация связи признаков «Loan Status» и «Loan Duration (Months)»*

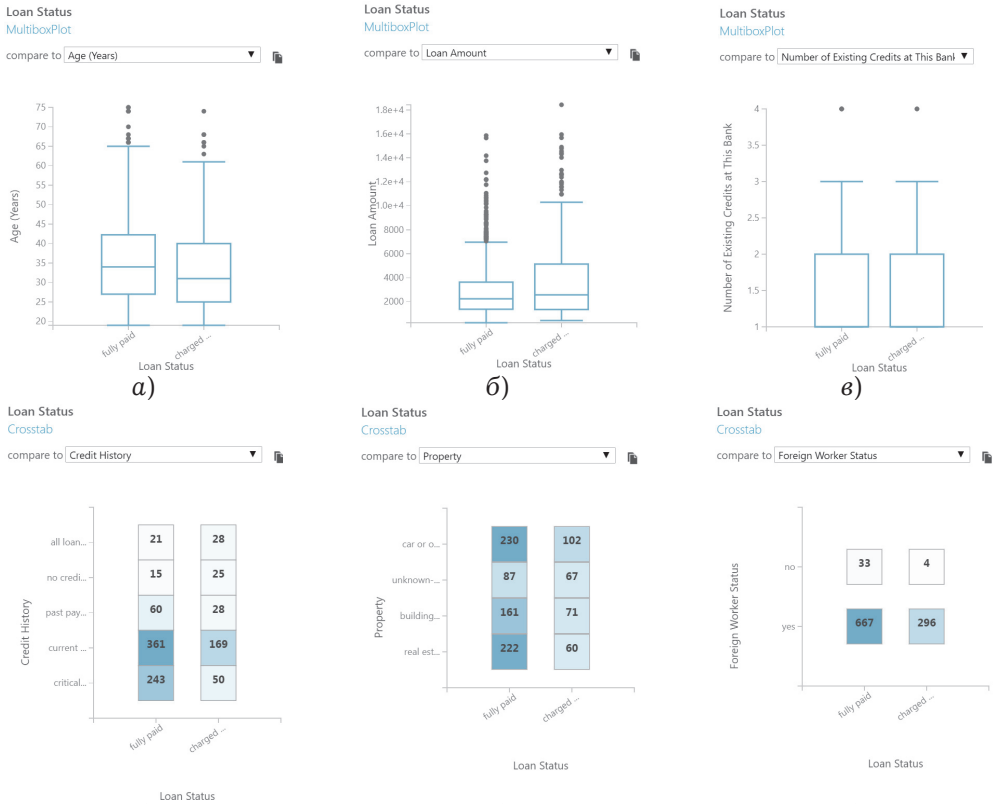


Рис. 7.4.30. *Визуализация связи признака «Loan Status» с признаками «Age» (а); «Loan Amount» (б); «Number of Existing Credits at This Bank» (в); «Credit History» (г); «Property» (д); «Foreign Worker Status» (е)*

Интересно, что наличие критических проблем чаще проявляется у «хороших» заемщиков — это, скорее всего, объясняется тем, что «плохим» заемщикам, имевшим такие критические проблемы, было отказано в выдаче кредитов. При этом количество кредитов в данном банке у «хороших» и «плохих» заемщиков примерно одинаково. Также не заметна связь надежности заемщика с фактором наличия статуса иностранного рабочего.

Укажем, какой признак является результативным, и какие признаки являются факторными. Для этого раскроем левое меню «Data Transformation > Manipulation», выберем модуль «Edit Metadata» и перенесем его мышкой в рабочую область Microsoft Azure ML Studio. Соединим выходной порт модуля «Remove Duplicate Rows» со входным портом модуля «Edit Metadata».

Выделив модуль «Edit Metadata», на панели параметров модуля («Properties») нажмем кнопку «Launch column selector» и выберем столбец «Loan Status». Введем следующие параметры:

Data Type: Unchanged;
Categorical: Unchanged;
Fields: Label (*Label* — результативный признак).

Эксперимент при этом будет выглядеть, как на рис. 7.4.31.

Добавим еще один модуль «Edit Metadata», на панели его параметров нажмем кнопку «Launch column selector» и выберем все столбцы, кроме «Customer ID» и «Loan Status». Также введем следующие параметры:

Data Type: Unchanged;
Categorical: Unchanged;
Fields: Features (*Feature* — факторный признак).

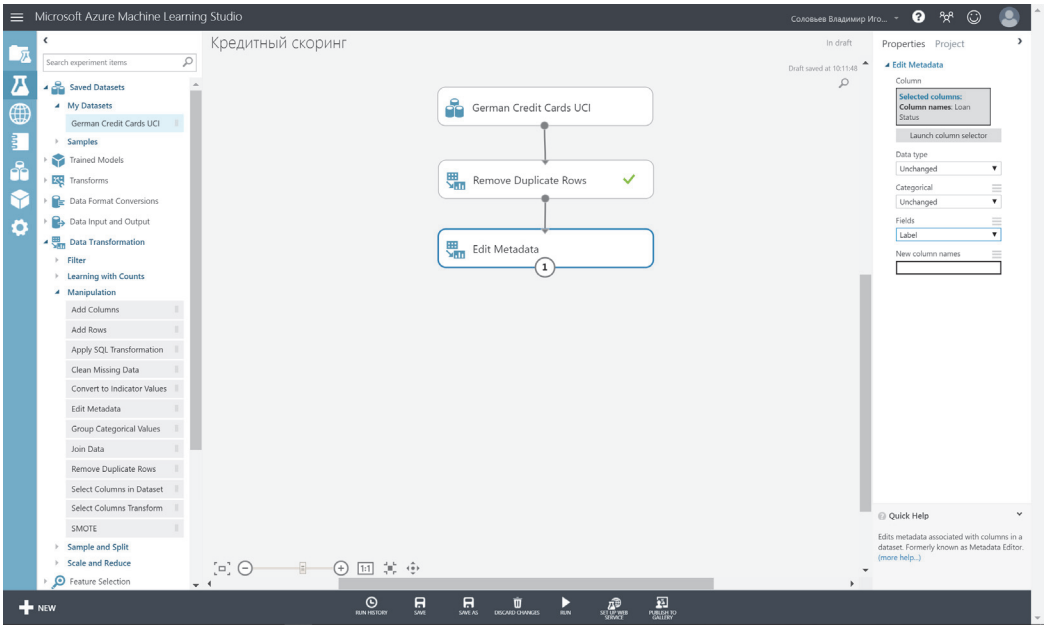


Рис. 7.4.31. *Эксперимент с добавленным модулем указания результативного признака*

Чтобы указать, что признак «Customer ID» не является ни результативным, ни факторным, а используется в качестве идентификатора, добавим еще один модуль «Edit Metadata», на панели его параметров выберем столбец «Customer ID» и введем следующие параметры:

Data Type: Unchanged;
Categorical: Unchanged;
Fields: Clear Feature.

Теперь нужно указать, что признаки, значения которых представляют собой текстовые строки, являются категориальными. Добавим еще один модуль «Edit Metadata». Чтобы выбрать все строковые признаки, на панели параметров модуля нажмем кнопку «Launch column selector» и обратим внимание, что в появившемся диалоговом окне «Select columns» можно выбирать столбцы по одному, указывая их названия («BY NAME»), а можно выбирать их группами в соответствии с определенными правилами («WITH RULES»). Выберем вариант «WITH RULES» и создадим правило, в соответствии с которым вначале нет ни одного столбца («NO COLUMNS»), а затем добавляются все строковые столбцы: Include — column type — String (рис. 7.4.32). Нажмем кнопку «✓».

Теперь введем следующие параметры:

Data Type: Unchanged;
Categorical: Make categorical;
Fields: Unchanged.

Эксперимент при этом будет выглядеть, как на рис. 7.4.33.



Рис. 7.4.32. *Выбор всех строковых признаков*

Итак, мы подготовили исходные данные к тому, чтобы на их основе разработать модель машинного обучения, которая будет предсказывать, какие из клиентов банка являются «хорошими» с позиции кредитного риска, а какие — «плохими». ❑

Замечание 1. В примере 7.4.2 набор исходных был загружен в Microsoft Azure ML Studio из локального csv-файла. Вместо этого можно использовать модуль «Import Data» на панели модулей он находится в категории «Data Input and Output»). Модуль «Import Data» позволяет автоматически импортировать наборы данных из различных внешних источников, включая внешние сайты, облачные и внутренние базы данных и др.

Замечание 2. В примере 7.4.2 ни один из признаков не содержал пропущенных значений. Это нетипично, обычно реальные наборы данных содержат много пропусков. Большинство алгоритмов машинного обучения не может работать со строками, в которых значения части признаков неизвестны, поэтому все пропущенные значения обязательно требуется каким-либо образом обработать. Для этого в Microsoft Azure ML Studio имеется модуль «Clean Missing Data» (на панели модулей он находится в категории «Data Input and Output»). Этот модуль можно настроить различными способами:

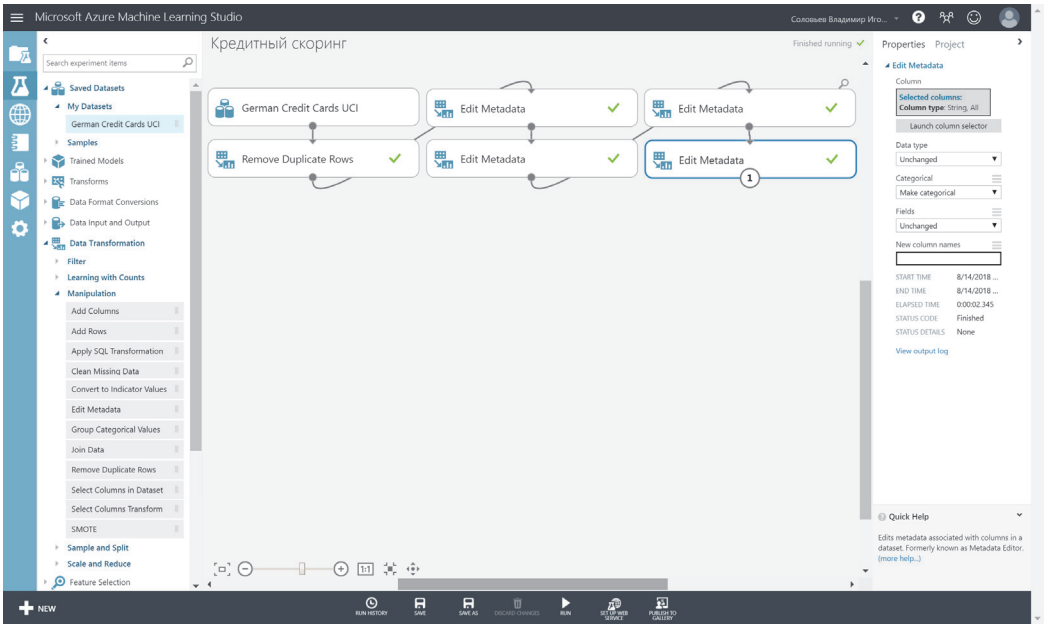


Рис. 7.4.33. *Эксперимент с отредактированными метаданными*

- можно удалить строки с пропущенными значениями;
- можно удалить признаки с пропущенными значениями.
- можно заменить пропущенные значения некоторыми фиктивными (например, для категориальных признаков ввести специальную категорию «Значение неизвестно», а для числовых значений вместо пропущенных значений подставить какое-либо стандартное число, например, 0);
- для числовых признаков можно вместо пропущенных значений подставить выборочное среднее, медиану или моду (при этом следует иметь в виду, что среднее очень чувствительно к выбросам, а мода у числовых признаков редко является информативным показателем, так как даже в больших наборах данных есть много близких значений, но мало одинаковых);
- у категориальных признаков выборочное среднее и медиана не существуют, но для заполнения пропусков можно использовать моду — наиболее часто встречающуюся категорию;
- можно спрогнозировать пропущенные значения на основании значений других признаков, используя метод многомерной оценки цепными уравнениями[1] (*Multivariate Imputation using Chained Equations*, *MICE*) или вероятностный метод главных компонент[2] (*Probabilictic Principal Component Analysis*, *Probabilistic PCA*).

Какой из этих способов использовать, зависит от конкретной задачи (в машинном обучении есть одно правило: «Никаких правил!»). Обычно экспериментируют с разными способами обработки пропущенных значений, и для разных признаков выбирают разные способы.

Пример 7.4.3. Файл «Video Games.csv» содержит набор данных о компьютерных играх, выпущенных с 1980 по 2017 г. Выбраны самые популярные игры — мировые продажи каждой из них 100 тыс. экземпляров.

Для 11 703 компьютерных игр известны значения следующих признаков:

[1] Метод многомерной оценки цепными уравнениями предложен в 1977 г. профессором Гарвардского университета Д.Б. Рубиным: *Rubin D.B.* Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys // Journal of the American Statistical Association. — 1977. — V. 72. — № 359. — P. 538—543.

[2] Вероятностный метод главных компонент предложен в 1999 г. сотрудниками *Microsoft Research* М.Е. Типпингом и К.М. Бишопом *Tipping M.E., Bishop C.M.* Probabilistic Principal Component Analysis // Journal of the Royal Statistical Society, Series. — 1999. — V. 61. — Part 3. — P. 611—622.

- Name — название игры;
- Platform — платформа (PC; DS; GBA;PS; PS2; PS3; PSP; Wii; XB или X360);
- Year_of_Release — год выпуска;
- Genre — жанр (Action; Adventure; Fighting; Platform; Puzzle; Racing; Role-Playing; Shooter; Simulation; Sports или Strategy; Misc означает все остальные жанры);
- Publisher — издатель;
- NA_Sales — объемы продаж игры в Северной Америке (в млн долл. США);
- EU_Sales — объемы продаж игры в Европе(в млн долл. США);
- JP_Sales — объемы продаж игры в Японии (в млн долл. США);
- Other_Sales — объемы продаж игры в других регионах (в млн долл. США);
- Critic_Score — сумма оценок, данных игре штатными критиками сайта *Metacritic*;[1]
- Critic_Count — количество штатных критиков, давших оценки игре;
- User_Score — сумма оценок, данных игре игры пользователями сайта *Metacritic*;
- User_Count — количество пользователей, давших оценки;
- Developer — разработчик;
- Rating — рейтинг игры (AO — только для взрослых; 18+ M — только для лиц старше 18 лет; 17+ — только для лиц старше 17 лет; E10+ для всех старше 10 лет; E — для всех; EC — для детей младшего возраста; RP — рейтинг не присвоен).

Требуется построить модель, предсказывающую объем продаж игры в Японии, зная ее объемы продаж в других странах. Япония — достаточно специфичная страна с точки зрения индустрии развлечений. К тому же государство активно защищает своих производителей, что не способствует импорту.

Перед издателями компьютерных игр всегда стоит вопрос: стоит ли игру переводить, подгонять под местный менталитет и платить пошлины для входа на японский рынок. Чтобы это понять, следует оценить, на какой кусок пирога можно вообще в Японии рассчитывать.

Один из вариантов — делать это по истории продаж в других регионах. Если игра еще в разработке, за историю продаж можно считать прогнозы: Европа и Америка не такие непредсказуемые рынки, как Япония.

В качестве показателя качества целесообразнее всего выбрать является среднюю абсолютную ошибку предсказания (Mean Absolute Error).

Загрузим набор исходных данных в Microsoft Azure ML Studio («+ NEW» — «DATASET» — «FROM LOCAL FILE» — «Video Games.csv»).

Создадим новый эксперимент по машинному обучению («+ NEW» — «Experiment» — «Blank Experiment»).

Когда набор данных «Video Games» загрузится, он появится на панели модулей эксперимента в категории «Saved Datasets > My Datasets». Добавим его на рабочую область эксперимента.

Предварительную обработку данных начнем с удаления повторяющихся строк — тех, в которых одновременно совпадают названия, год выпуска и издатель.

Добавим модуль «Remove Duplicate Rows» (на панели модулей он находится в категории «Data Transformation > Manipulation») и соединим выходной порт модуля набора данных «Video Games» со входным портом модуля «Remove Duplicate Rows».

[1] *Metacritic* (http://metacritic.com) — англоязычный сайт-агрегатор, собирающий отзывы о музыкальных альбомах, компьютерных играх, фильмах, телевизионных шоу и т. п.

На панели параметров модуля нажмем кнопку «Launch column selector» и выберем признаки «Name», «Platform», «Year_of_Release» и «Publisher» (рис. 7.4.34).

Визуализировав выходной порт модуля набора данных «Video Games», видим, что исходный набор данных содержит 11 703 строки.

В результатах визуализации выходного порта модуля «Remove Duplicate Rows» (рис. 7.4.35) уже 11 702 строки — одна из игр была внесена в набор данных дважды.

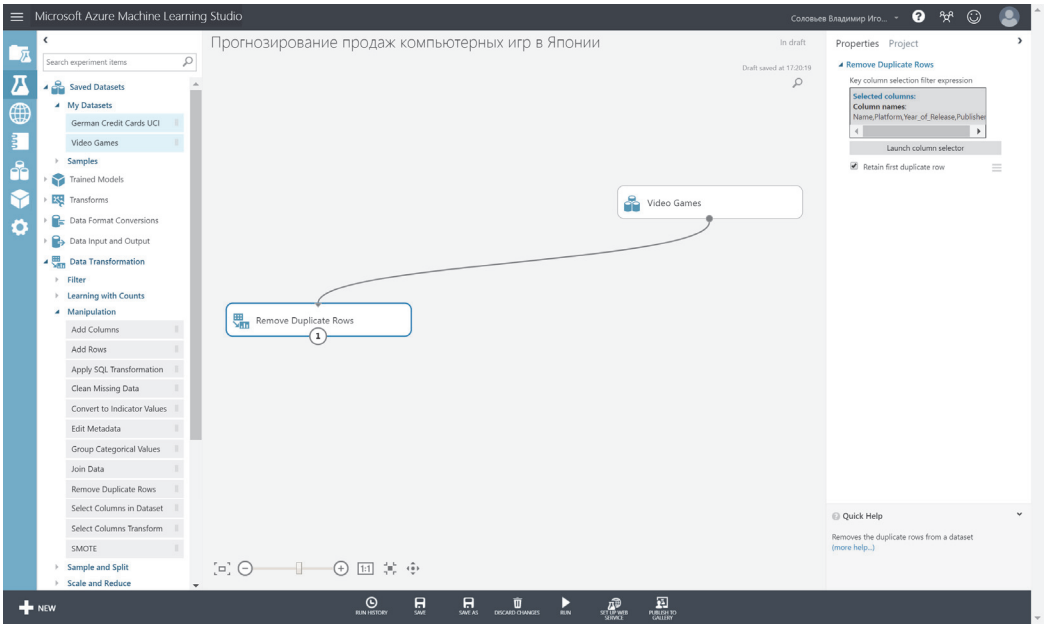


Рис. 7.4.34. *Эксперимент с добавленным модулем удаления повторяющихся строк*

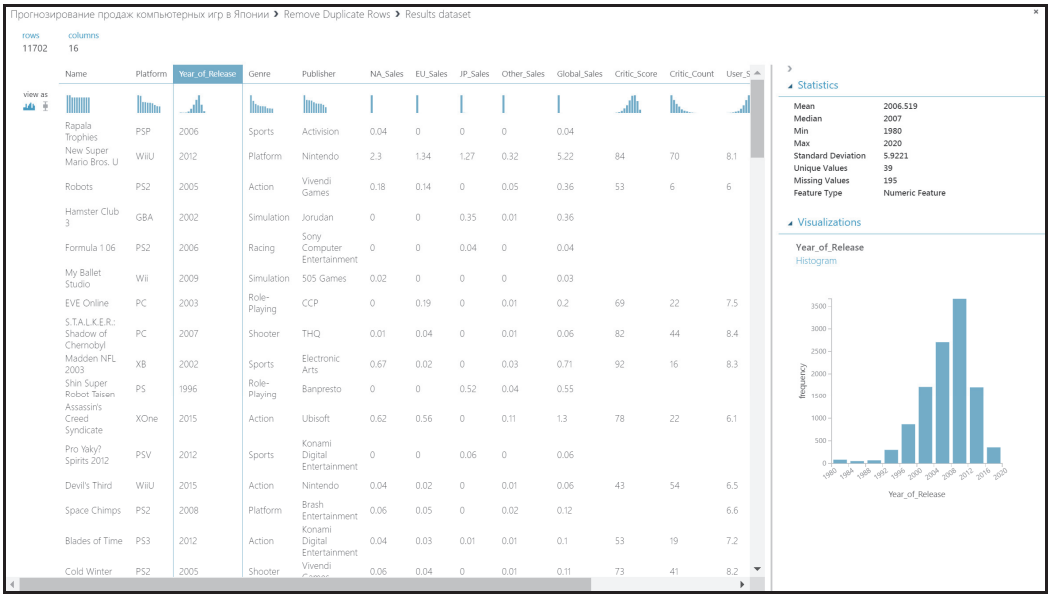


Рис. 7.4.35. *Визуализация результата удаления повторяющихся строк*

Проанализируем описательную статистику признаков, поочередно нажимая левой кнопкой мыши на заголовки столбцов, и зафиксируем некоторые наблюдения.

Замечаем, что признаки «Name» и «Genre» содержат по одному пропущенному значению, признак «Year_of_Release» — 195 пропущенных значений, признаки «Critic_Score» и «Critic_Count» — по 6027, признаки «User_Score» и «User_Count» — по 6384, признак «Developer» — 4639, признак «Rating» — 4746 пропущенных значений. В остальных признаках пропусков нет.

Также замечаем, что максимальное значение признака «Year_of_Release» равно 2020. Поскольку набор данных собран в 2017 г., по-видимому, значения года выпуска, большие 2017, являются выбросами, и их необходимо будет каким-то образом обработать.

Распределения признаков «Year_of_Release», «Critic_Score» и «User_Score» слегка скошены влево (рис. 7.4.36, *а*—*в*), распределение признака «Critic_Count» несильно скошено вправо рис. 7.4.36, *г*), распределение признака «User_Count» скошено вправо гораздо сильнее (рис. 7.4.36, *д*), а распределения объемов продаж — «NA_Sales», «EU_Sales», «JP_Sales» и «Other_Sales» — имеют очень значительный скос вправо (рис. 7.4.36, *е*—*и*).

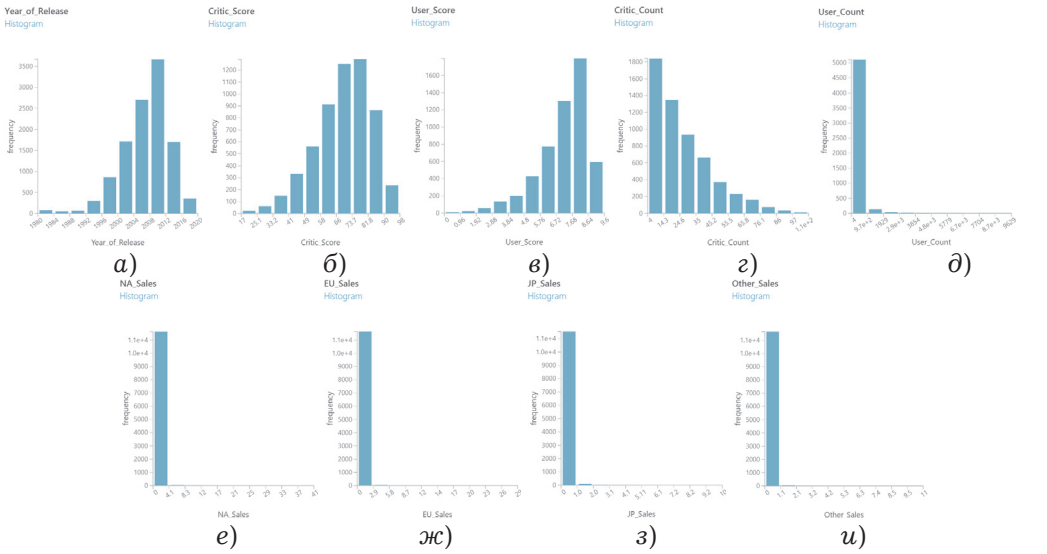


Рис. 7.4.36. *Гистограммы скошенных признаков*

Заменим единственное пропущенные значения признаков «Name» и «Genre» значением «Unknown». Для этого добавим в эксперимент модуль «Clean Missing Data» (на панели модулей он находится в категории «Data Transformation > Manipulation»). На панели параметров модуля нажмем кнопку «Launch column selector» и выберем признаки «Name» и «Genre». Также введем следующие параметры:

Minimum missing value ratio: 0;
Maximum missing value ratio: 1;[1]
Cleaning mode: Custom substitution value;
Replacement value: Unknown;
Generate missing value indicator column: установим выделение.[1]

[1] Параметры Minimum missing value ratio и Maximum missing value ratio можно использовать, чтобы указать минимальную и максимальную долю пропущенных значений, при которых выполняется операция их обработки.

Эксперимент при этом будет выглядеть, как на рис. 7.4.37. Аналогичным образом заменим пропущенные значения признаков «Critic_Score», «Critic_Count», «User_Score» и «User_Count» нулями (рис. 7.4.38). Но поставить нули вместо пропущенных значений года выпуска нецелесообразна, гораздо лучше заполнить пропуски в столбце «Year_of_Release» медианным значением этого признака (рис. 7.4.39).

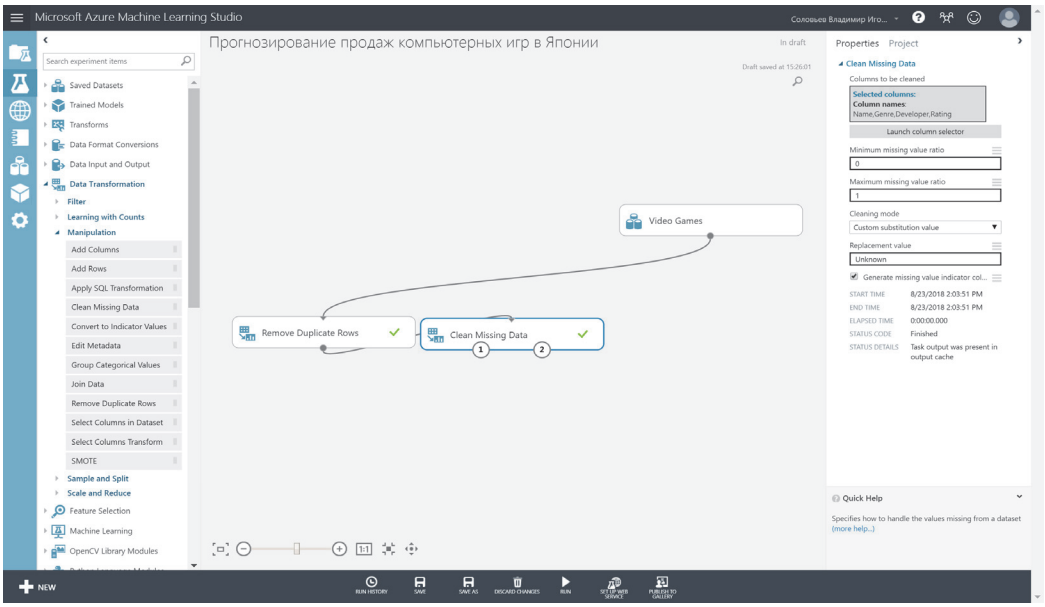


Рис. 7.4.37. *Обработка пропущенных значений признаков «Name», «Genre», «Developer» и «Rating»*

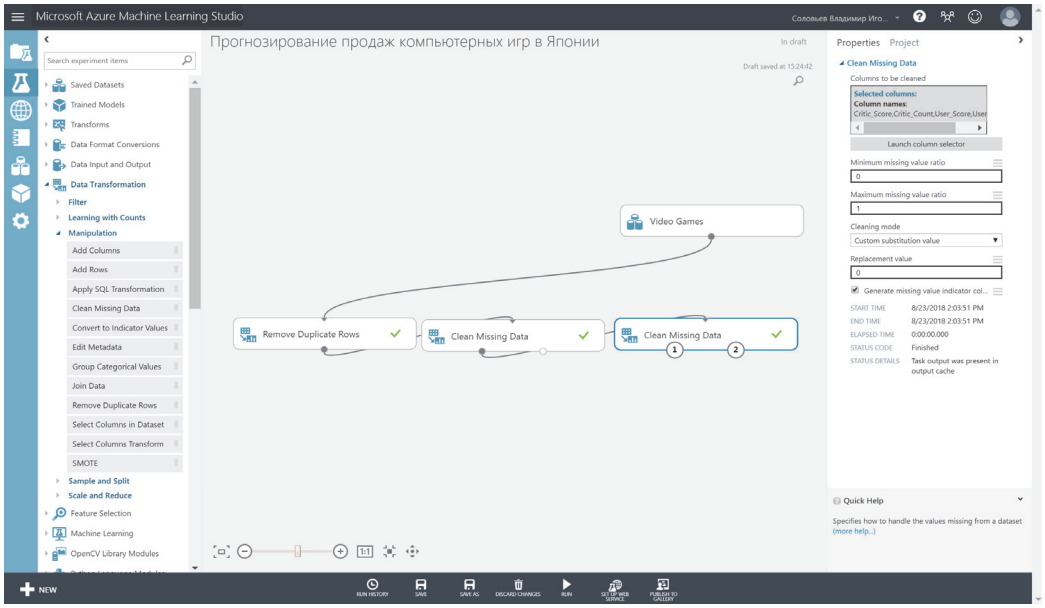


Рис. 7.4.38. *Обработка пропущенных значений признаков «Critic_Score», «Critic_Count», «User_Score» и «User_Count»*

[1] Лучше всегда генерировать дополнительный индикатор — признак, равный единице, если значения в данной строке были пропущены, и нулю в противном случае.

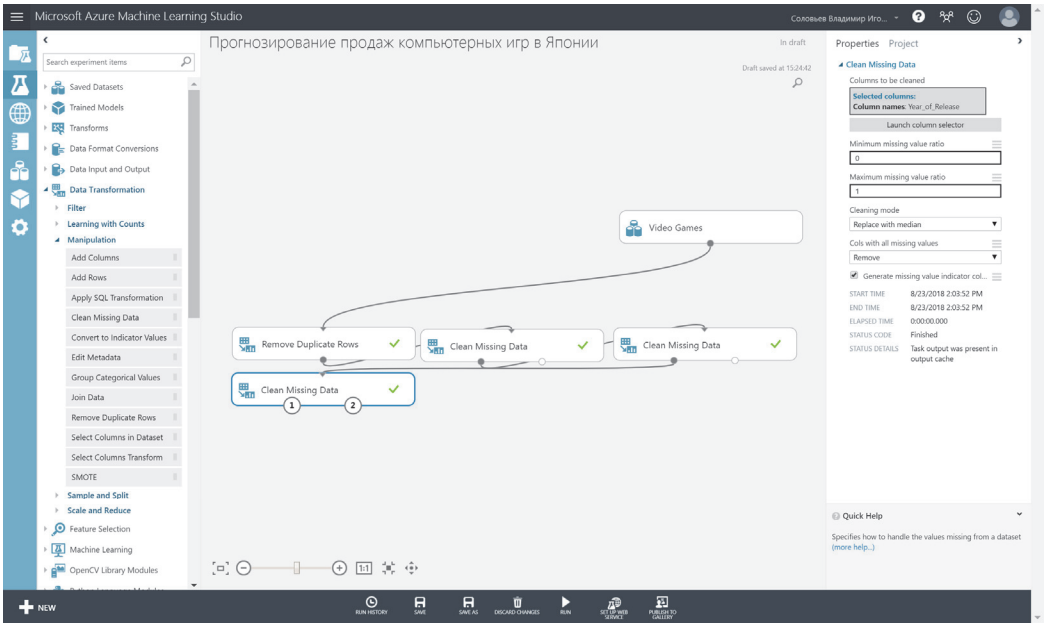


Рис. 7.4.39. *Обработка пропущенных значений признака «Year_of_Release»*

Удалим выбросы значений признака «Year_of_Release» (игры, у которых значение данного признака больше 2017). Для этого сначала преобразуем выбросы в пропущенные значения, а затем удалим строки с пропущенными значениями.

Добавим в эксперимент модуль «Clip Values» (на панели модулей он находится в категории «Data Transformation > Scale and Reduce»).

Нажав на панели параметров модуля кнопку «Launch column selector», выберем признак «Year_of_Release», затем укажем следующие параметры:

Set of thresholds: ClipPeaks;
Upper threshold: Constant;
Constant value for upper threshold: 2017;
Upper substitute value: Missing;
Generate missing value indicator column: снимем выделение.[1]

Эксперимент примет вид, как на рис. 7.4.40.

Теперь можно удалить строки, в которых пропущены значения признака «Year_of_Release», т. е. те строки, в которых значение года выпуска в исходном наборе данных было больше 2017 (рис. 7.4.41).

Чтобы завершить этап предварительной обработки данных, укажем, какой признак является результативным, и какие признаки являются факторными. Добавим в эксперимент по очереди четыре модуля «Edit Metadata» (на панели модулей их можно найти в категории левое меню «Data Transformation > Manipulation»).

На панели параметров первого модуля «Edit Metadata» с помощью кнопки «Launch column selector» выберем столбец «JP_Sales» и изменим значение параметра «Fields» на «Label» (остальные параметры оставим по умолчанию: «Unchanged»). У второго модуля «Edit Metadata» на панели параметров выберем все столбцы, кро-

[1] Ранее отмечалось, что генерировать дополнительный индикатор всегда полезно. В данном случае этого делать не следует, потому что в обучающей выборке мы просто удалим все игры, «выпущенные» после 2017 г., а прогнозировать мы можем объемы продаж и для игр, выпущенных в будущем, поэтому в контрольной выборке мы такие игры удалять не будем.

ме «JP_Sales», и изменим значение параметра «Fields» на «Features» (остальные параметры оставим по умолчанию). У третьего модуля «Edit Metadata» выберем все строковые столбцы («Launch column selector» > «WITH RULES» > «NO COLUMNS» > «Include — column type — String») и изменим значение параметра «Categorical» на «Make categorical» (остальные параметры — по умолчанию). У четвертого модуля выберем признаки «Name», «Publisher», «Developer» и изменим значение параметра «Fields» на «Clear feature» (остальные параметры оставим по умолчанию: «Unchanged»). Эксперимент при этом будет выглядеть, как на рис. 7.4.42.

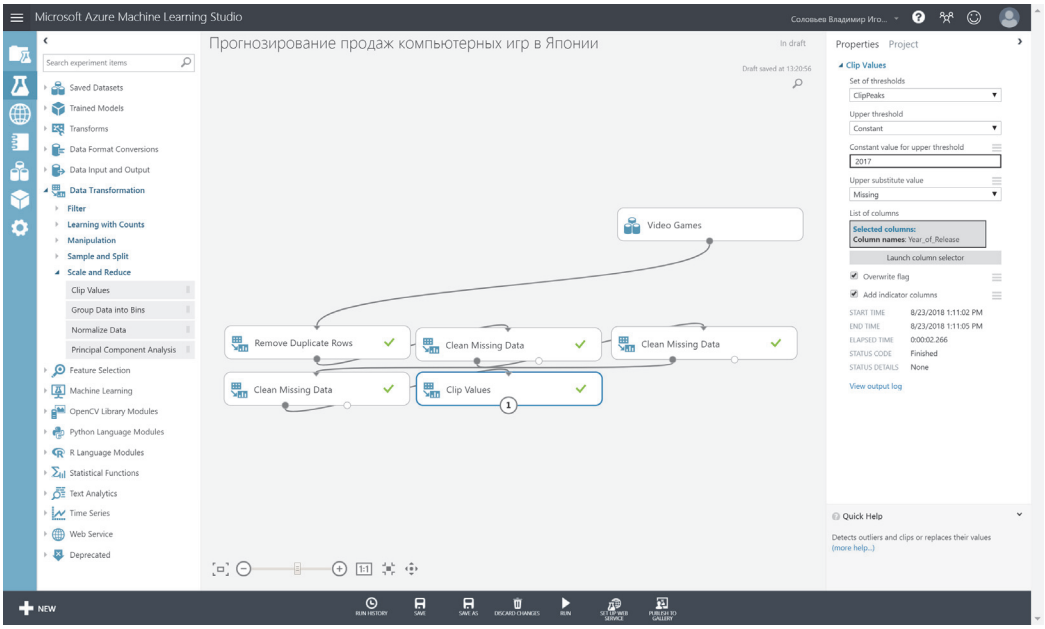


Рис. 7.4.40. *Очистка выбросов признака «Year_of_Release»*

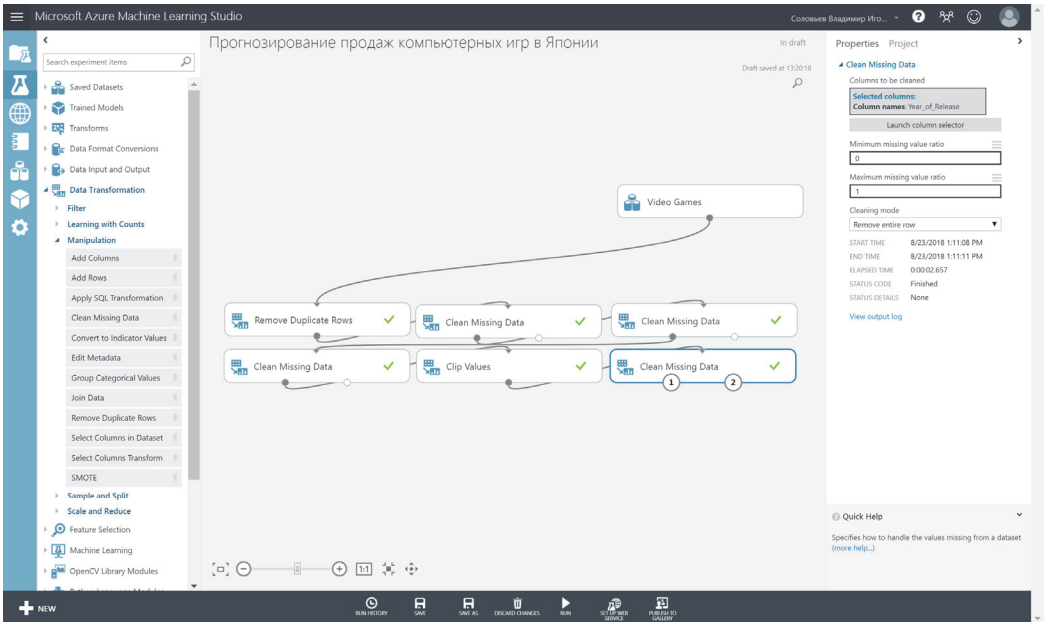


Рис. 7.4.41. *Удаление строк, имевших выбросы признака «Year_of_Release»*

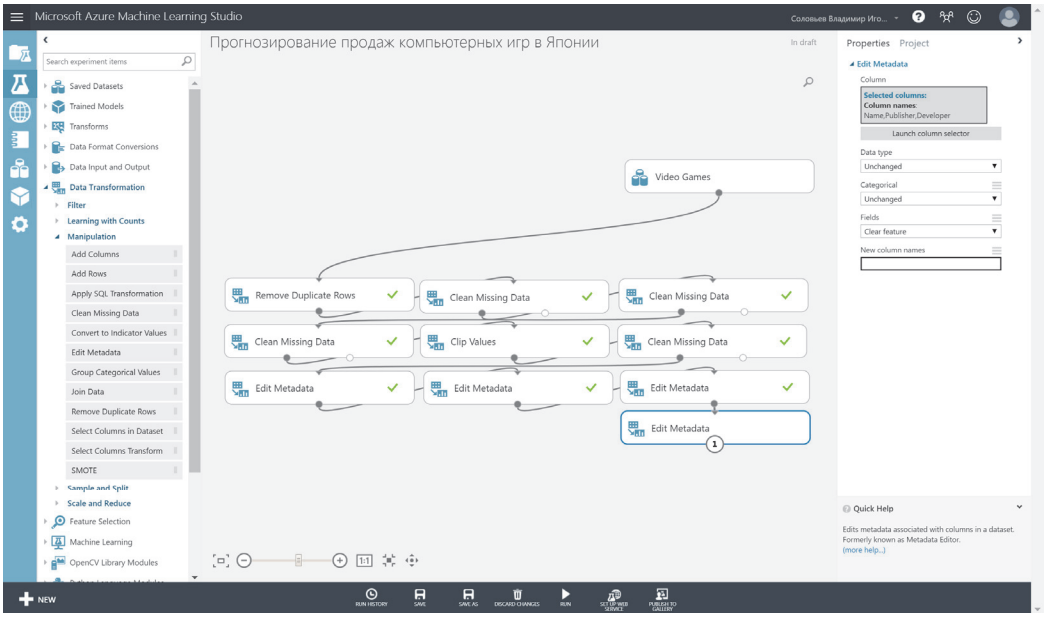


Рис. 7.4.42. *Эксперимент с добавленными модулями указания факторных, результативных и категориальных признаков*

Итак, мы провели предварительную обработку набора данных «Video Games»: удалили повторяющиеся элементы, заполнили пропущенные значения всех признаков. Кроме того, мы удалили строки с выбросами, соответствующие играм, выпущенным позже, чем был собран набор данных. Также мы отметили, что признаки, соответствующие объемам продаж, имеют значительный скос. Пока мы ничего с этим не делали, но впоследствии мы построим две модели машинного обучения. Первая будет прогнозировать объемы продаж в Японии на основании исходных, скошенных, признаков. Вторая модель будет использовать объемы продаж, подвергнутые преобразованию $\ln(1 + x)$.

Возможно, это позволит улучшить качество прогнозирования. Так ли это, увидим позже. ❑

Задачи

448. Повторите самостоятельно все действия по созданию эксперимента в условиях примера 7.4.2.

449. Повторите самостоятельно все действия по созданию эксперимента в условиях примера 7.4.3.

450. Объясните, почему номер договора в примере 7.4.2 и название игры в примере 7.4.2 нельзя использовать в качестве факторных признаков.

7.4.3. Разработка моделей машинного обучения и оценка их качества

Перед тем, как разрабатывать модель машинного обучения, нужно определить, как будет оцениваться ее качество. Чтобы понимать, насколько хорошо модель машинного обучения предсказывает неизвестное, можно воспользоваться **разделением исходного набора данных на обучающую (тренировочную) и контрольную (тестовую) выборки** — два независимых непересекающихся подмножества

элементов, выбираемых случайным образом. *Обучающая выборка* (первое подмножество, обычно составляющее 60—80% элементов исходного набора данных) используется для обучения модели, *контрольная выборка* (оставшиеся 20—40%) — для оценки предсказательной способности модели.

При обучении студентов сначала им предлагают возможность порешать задачи с известными ответами, а затем, чтобы понять, насколько хорошо усвоен материал, студентам дают контрольную работу, задачи из которой они никогда не видели. Точно так же модели машинного обучения вначале можно дать возможность натренироваться на обучающей выборке, а затем проверить, насколько хорошо она «научилась», предлагая предсказать значения результативного признака для объектов из тестовой выборки. Сравнив, насколько результаты предсказаний машины на тестовой выборке близки к истинным значениям результативного признака, можно сделать вывод о качестве модели.

Замечание 1. Выбор соотношения между размерами обучающей и контрольной выборок — это компромисс: чем больше примеров в обучающей выборке, тем лучше модель обучится, а чем больше примеров в контрольной выборке, тем лучше мы сможем понять, насколько хорошо модель обучилась.

Замечание 2. Разделение выборки на обучающую и контрольную — это самый простой метод, позволяющий оценить прогностическую способность моделей машинного обучения. Однако это не единственный метод. Обычно более эффективно использовать кросс-валидацию, при которой исходный набор делится на k частей, а затем обучение и оценка проходят k раз. Вначале алгоритм обучается на данных из 1-й, 2-й, ..., $(k-1)$-й подвыборок, а оценивается на k-й подвыборке, затем обучение происходит на подвыборках с номерами 1, 3, 4, ... $k-1$, а оценка качества на подвыборке с номером 2, и т. д.[1]

Пример 7.4.4. Продолжим рассмотрение набора данных «German Credit Cards UCI» из примера 7.4.2. Проведем разделение подготовленного набора данных на обучающую и контрольную выборку. Для этого на панели модулей выберем модуль разбиения набора данных («Split Data») — он находится в категории «Data Transformation > Sample and Split») — и перенесем его в рабочую область. Соединим выход («Results dataset») последнего из модулей «Edit Metadata» со входом («Dataset») модуля «Split Data».

Выделим модуль «Split Data» и на панели свойств («Properties») введем следующую информацию:

Splitting mode: Split Rows;
Fraction of rows in the first output dataset: 0.7;
Randomized split: установим выделение;
Random seed: 123;
Stratified split: False.

Теперь эксперимент будет выглядеть, как на рис. 7.4.43.

Замечание. Мы указали, что в левый выходной порт модуля «Split Data» необходимо случайным образом выбрать 70% строк из входного набора данных (а в правый порт — 30%). При этом на самом деле выбор строк происходит с помощью генератора псевдослучайных чисел, и параметр «Random seed» («зерно») отвечает за инициализацию этого генератора. Если оставить данный параметр равным нулю (по умолчанию), то результаты при каждом запуске алгоритма будут разными,[2] если же установить зерно равным какому-либо ненулевому значению, то «случайные» числа каждый раз будут одними и теми же. Использование ненулевого зерна позволяет обеспечить повторяемость результатов.

---

[1] Подробнее см. *Соловьев В.И.* Прикладные задачи машинного обучения в экономике и финансах: Практический курс с применением технологий *Microsoft Azure ML* и соревнованиями на платформе *Kaggle*. — М.: КНОРУС, 2018.

[2] При нулевом значении параметра «Random seed» инициализация датчика псевдослучайных чисел привязывается к системному времени.

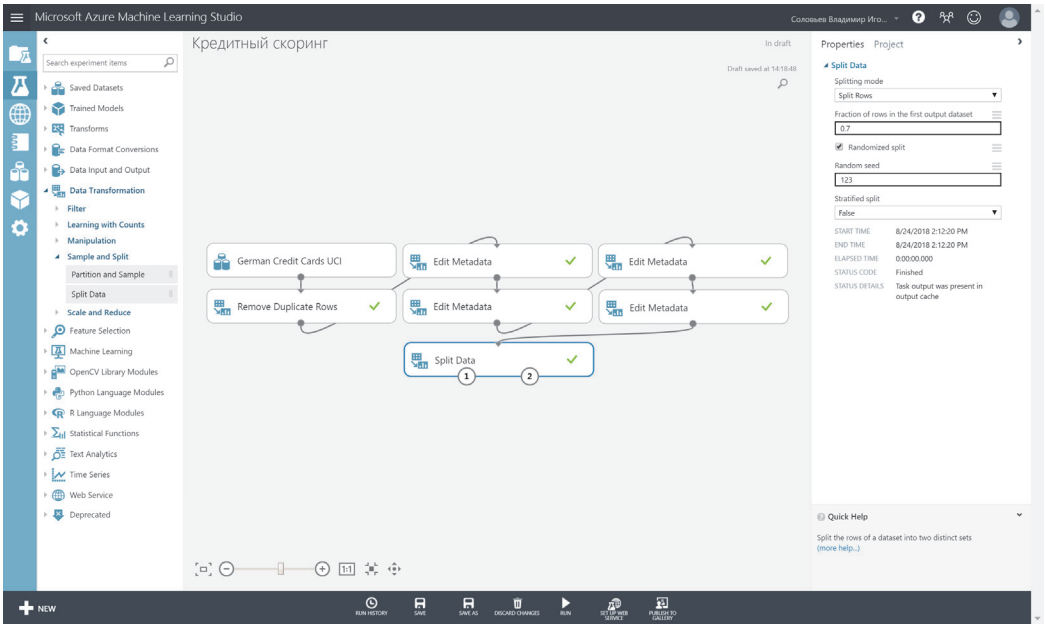


Рис. 7.4.43. *Эксперимент с добавленным модулем разбиения данных*

Теперь перейдем **к выбору алгоритма машинного обучения** и непосредственно к обучению модели.

Мы занимаемся машинным обучением, поэтому, очевидно, нам потребуются инструменты Microsoft Azure ML Studio, объединенные в категорию «Machine Learning». В этой категории есть четыре подкатегории:

- «Evaluate»;
- «Initialize Model»;
- «Score»;
- «Train».

Даже впервые строя модель машинного обучения, понятно, что начинать нужно с инициализации («Initialize Model»), т. е. с выбора алгоритма. Нам нужно научиться предсказывать, кто из клиентов является «хорошим», а кто является «плохим». Это — задача классификации наблюдений.

В Microsoft Azure ML Studio встроено 9 алгоритмов двухклассовой («Two-Class») классификации и 5 алгоритмов многоклассовой («Multi-Class») классификации (см. рис. 7.4.9). У нас классов два: «хорошие» и «плохие» заемщики.

На панели модулей раскроем категорию «Machine Learning > Initialize Model > Classification» и познакомимся с набором алгоритмов двухклассовой («Two-Class») и многоклассовой («Multi-Class») классификации, встроенных в Microsoft Azure ML Studio.

Воспользуемся алгоритмом логистической регрессии, для этого выберем модуль «Two-Class Logistic Regression» и перенесем его в рабочую область левее модуля «Split Data».

Замечание. Размещение модулей на экране никак не влияет на результат, а служит только для обеспечения наглядности рабочего процесса эксперимента. Модулей обычно достаточно много (большинство из них, как правило, обеспечивают предварительную обработку данных), и нужно стараться разместить модули так, чтобы они занимали как можно меньше места, но при этом чтобы рабочий процесс был максимально понятен.

Выделим модуль «Two-Class Logistic Regression» и на панели свойств («Properties») оставим значения параметров модели логистической регрессии, установленные по умолчанию:

Create trainer mode: Single Parameter;
Optimization tolerance: 1E–07;
L1 regularization weight: 1;
L2 regularization weight: 1;
Memory size for L-BFGS: 20;
Random seed: 123;
Allow unknown categorical levels: установим выделение.

В результате эксперимент будет выглядеть, как на рис. 7.4.44.

эксперимента выберем модуль обучения модели («Train Model») — он находится в категории «Machine Learning > Train» — и перенесем его в рабочую область ниже всех других модулей.

Соединим выход («Untrained Model») модуля «Two-Class Logistic Regression» с левым входным портом («Untrained model») модуля «Train Model», а левый выход («Results dataset1») модуля «Split Data» с правым входным портом («Dataset») модуля «Train Model».

На панели свойств («Properties») модуля «Train Model» нажмем кнопку «Launch column selector» и выберем признак «Loan Status» в качестве результативного признака («Label column»).

Сохраним эксперимент (он выглядит, как на рис. 7.4.45), нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN».

Замечание. Когда мы ведем стрелку из выходного порта какого-либо модуля, Microsoft Azure ML Studio подсвечивает модули, в которых можно эту стрелку закончить, зеленым пунктиром. При этом порты, в которые можно привести стрелку, визуализируются зелеными точками, а порты, в которые данная стрелка входить не может, — красными точками (рис. 7.4.46).

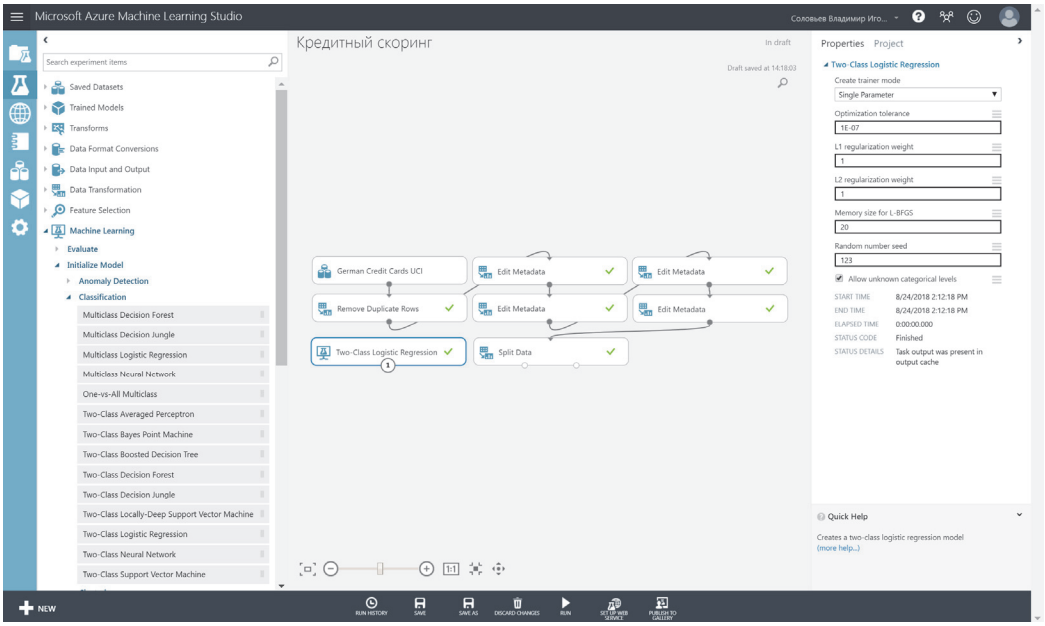


Рис. 7.4.44. *Эксперимент с добавленным модулем инициализации модели*

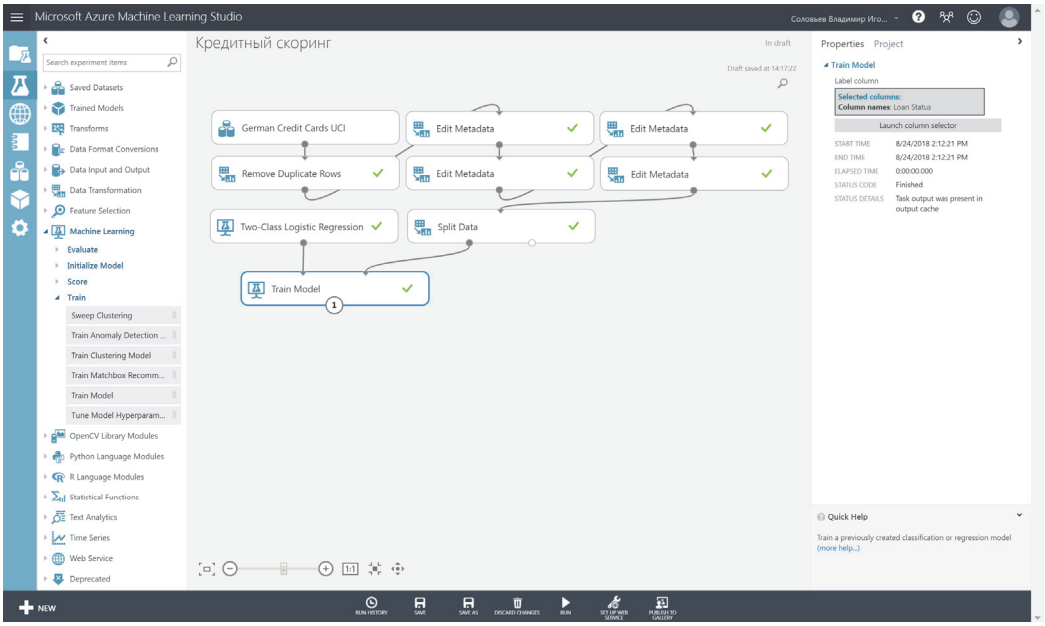


Рис. 7.4.45. *Эксперимент с добавленным модулем обучения модели*

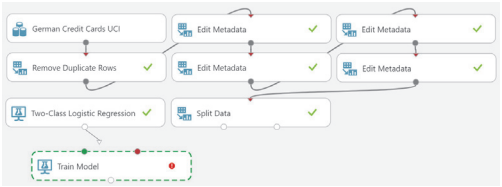


Рис. 7.4.46. *Подсветка допустимых и недопустимых модулей и входных портов*

Чтобы **обучить модель на тренировочной выборке**, на панели модулей Если визуализировать выход модуля «Train Model» (рис. 7.4.47), то мы увидим вначале параметры алгоритма (Settings), а затем коэффициенты модели логистической регрессии (Feature Weights), отсортированные по убыванию абсолютных значений (коэффициент «Bias» — это свободный член).

Замечаем, что отсутствие текущего счета, а также наличие созаемщиков и гарантов ведут к увеличению вероятности возврата кредита (т. е. вероятности того, что заемщик является «хорошим»), так как коэффициенты при этих признаках положительны, а срок кредита, размер кредита, доля платежа по кредиту в располагаемом доходе и отрицательный баланс текущего счета имеют обратную связь с вероятностью возврата кредита (соответствующие коэффициенты модели отрицательны).

Перейдем к **вычислению прогнозов на контрольной выборке**.

На панели модулей выберем модуль вычисления прогнозов «Score Model» (он находится в категории «Machine Learning > Score») и перенесем его в рабочую область ниже всех других.

Соединим выход («Trained model») модуля «Train Model» с левым входом («Trained model») модуля «Score Model», а правый выход («Results dataset2») модуля «Split Data» — с правым входом («Dataset») модуля «Score Model». После этого эксперимент будет выглядеть так, как показано на рис. 7.4.48.



Рис. 7.4.47. *Визуализация результатов обучения модели*

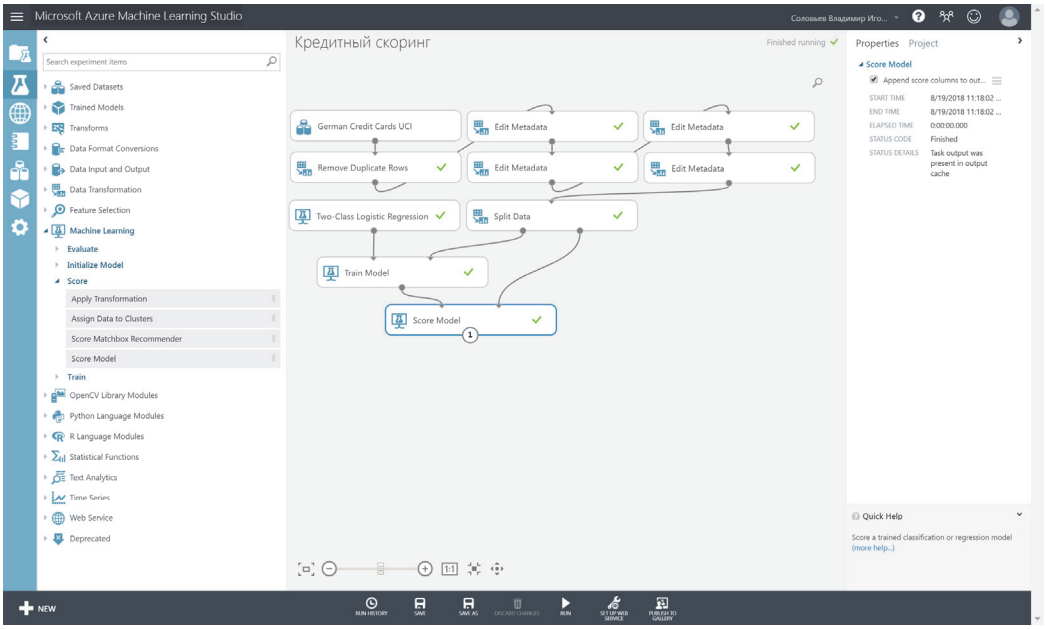


Рис. 7.4.48. *Эксперимент с добавленным модулем вычисления прогнозов*

Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN». Когда выполнение эксперимента закончится, визуализируем выход («Scored dataset») модуля «Score Model», щелкнув по выходному порту левой кнопкой мыши и выбрав пункт меню «Visualize» (рис. 7.4.49).

Обратим внимание на новые столбцы:

- «Scored Labels» — это прогноз статуса возврата кредита;
- «Scored Probabilities» — это прогнозируемая вероятность того, что клиент является «хорошим».

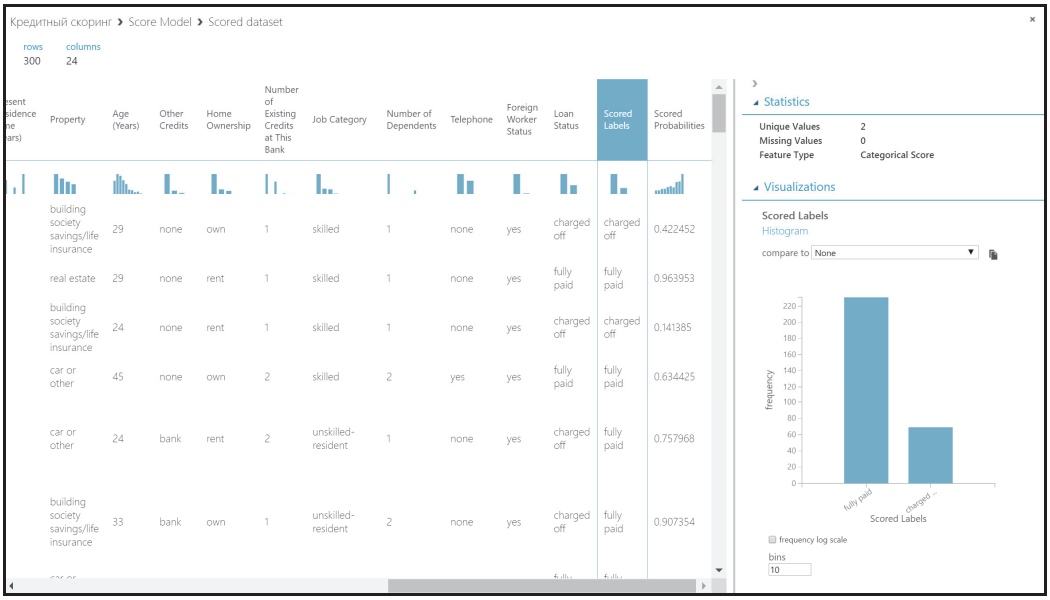


Рис. 7.4.49. *Визуализация прогнозов на контрольной выборке*

Перейдем к **оценке качества модели машинного обучения.**

Вспомним, что вычисления произведены на контрольной выборке, для каждого элемента которой известно, к какому классу («хороших» или «плохих» заемщиков) он относится, и сравним спрогнозированный статус клиента («Scored Labels») и его реальный статус («Loan Status»).

Замечание. Истинный результат и прогноз результата — это, конечно, не одно и то же. Естественно, модель делает прогнозы результативного признака для элементов контрольной выборки, основываясь только на значениях факторных признаков — истинных значений результативного признака она «не видит» (рис. 7.4.50).

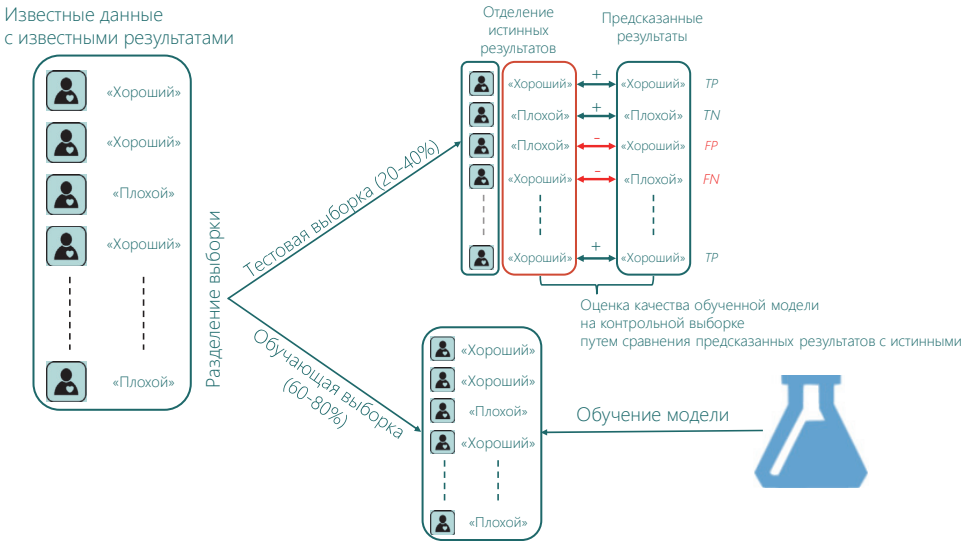


Рис. 7.4.50. *Оценка качества модели машинного обучения*

Видно, что, например, в первых четырех строках прогноз совпадает с истинным значением, а в пятой и шестой строках — не совпадает. Если на 300 элементах тестового набора данных у нас всего две ошибки, наверное, модель следует признать очень хорошей. Если же на весь тестовый набор у нас всего четыре верных прогноза, то такую модель хорошей признать нельзя.

Чтобы провести сравнение истинных значений признака «Loan Status» и прогнозов, сделанных нашей моделью, на панели модулей выберем модуль оценки качества модели («Evaluate Model») — он находится в категории «Machine Learning > Evaluate» — и перенесем его в нижнюю часть рабочей области.

Соединим выход («Scored dataset») модуля «Score Model» с левым входом («Scored dataset») модуля «Evaluate Model». После этого эксперимент должен выглядеть так, как на рис. 7.4.51.

Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN». Когда выполнение эксперимента закончится, визуализируем выход («Evaluation results») модуля «Evaluate Model» (рис. 7.4.52).

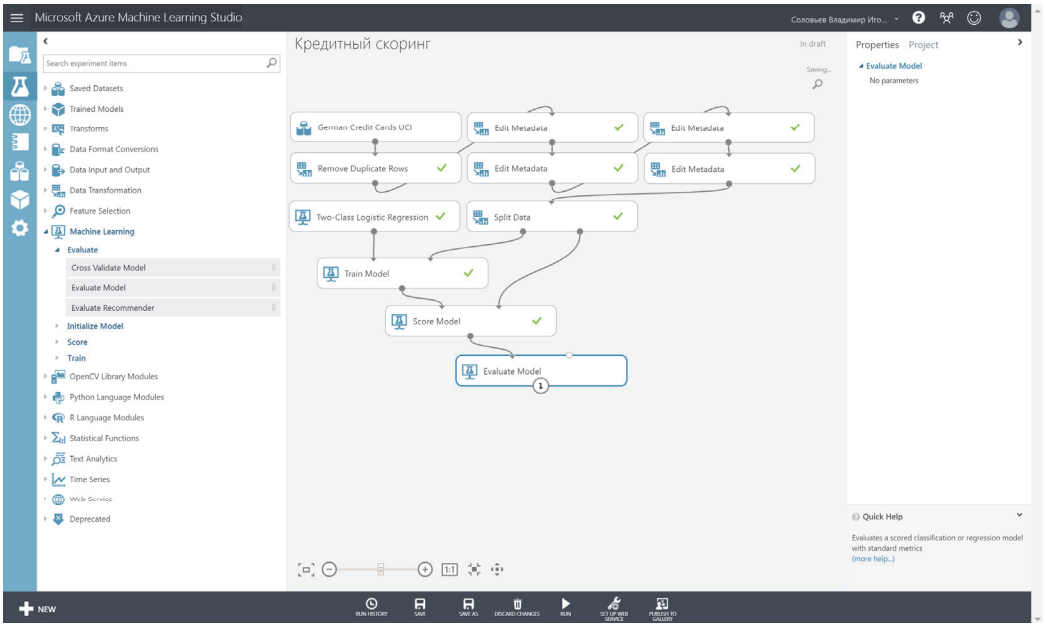
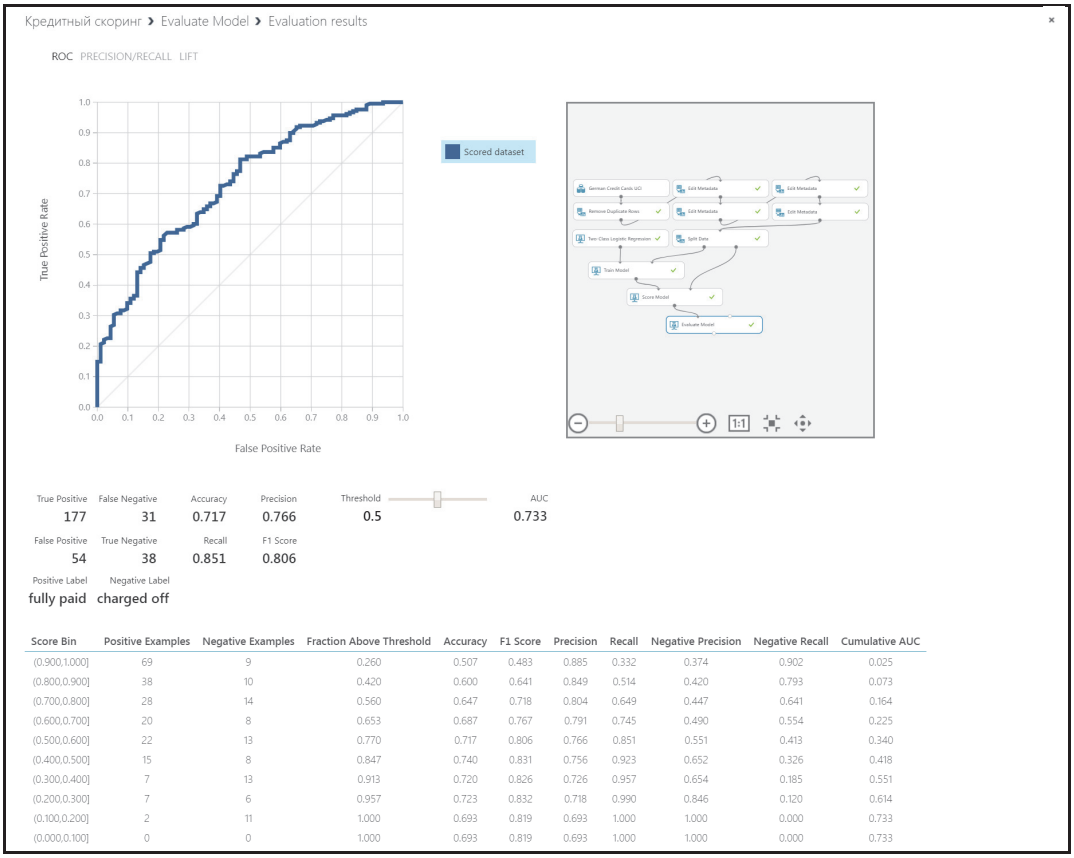


Рис. 7.4.51. *Эксперимент с добавленным модулем оценки качества модели*

Проанализируем результаты оценки качества построенной модели машинного обучения:

- *True Positive* (*TP*) = 177 — количество правильно распознанных «хороших» заемщиков;
- *True Negative* (*TN*) = 38 — количество правильно распознанных «плохих» заемщиков;
- *False Positive* (*FP*) = 54 — количество ошибок первого рода, т. е. «плохих» заемщиков, неправильно названных «хорошими»; в случае использования данной модели на практике ошибка первого рода означает, что кредит будет выдан «плохому» заемщику, чем больше *FP*, тем больше окажутся потери банка в результате невозвратов кредитов;



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	69	9	0.260	0.507	0.483	0.885	0.332	0.374	0.902	0.025
(0.800,0.900]	38	10	0.420	0.600	0.641	0.849	0.514	0.420	0.793	0.073
(0.700,0.800]	28	14	0.560	0.647	0.718	0.804	0.649	0.447	0.641	0.164
(0.600,0.700]	20	8	0.653	0.687	0.767	0.791	0.745	0.490	0.554	0.225
(0.500,0.600]	22	13	0.770	0.717	0.806	0.766	0.851	0.551	0.413	0.340
(0.400,0.500]	15	8	0.847	0.740	0.831	0.756	0.923	0.652	0.326	0.418
(0.300,0.400]	7	13	0.913	0.720	0.826	0.726	0.957	0.654	0.185	0.551
(0.200,0.300]	7	6	0.957	0.723	0.832	0.718	0.990	0.846	0.120	0.614
(0.100,0.200]	2	11	1.000	0.693	0.819	0.693	1.000	1.000	0.000	0.733
(0.000,0.100]	0	0	1.000	0.693	0.819	0.693	1.000	1.000	0.000	0.733

Рис. 7.4.52. *Визуализация оценки качества модели машинного обучения*

- *False Negative* (*FN*) = 31 — количество ошибок второго рода, т. е. «хороших» заемщиков, неправильно названных «плохими»; в случае использования данной модели на практике ошибка второго рода означает, что добросовестному заемщику будет отказано в кредите, чем больше *FN*, тем больше банк теряет процентного дохода в результате отказов «хорошим» заемщикам;
- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{177+38}{177+54+38+31} = 0{,}717$ — *доля правильных ответов*;
- $Precision = \frac{TP}{TP+FP} = \frac{177}{177+54} = 0{,}766$ — *мера точности*, характеризующая, сколько полученных в результате использования модели классификации положительных ответов являются правильными;
- $Recall = \frac{TP}{TP+FN} = \frac{177}{177+31} = 0{,}851$ — *мера полноты*, характеризующая способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых (ложноположительные ответы не влияют на эту метрику);
- $F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} = \frac{2 \cdot 0{,}766 \cdot 0{,}851}{0{,}766+0{,}851} = 0{,}806$ — среднее гармоническое точности и полноты;
- *Кривая ROC* (*Receiver Operating Characteristic*) отображает изменение отношения доли верно классифицированных «хороших» клиентов в их общем количестве $\left(TPR = Recall = \frac{TP}{TP+FN}\right)$ к доле «плохих» клиентов, неправильно от-

несенных к «хорошим», $\left(FPR = \frac{FP}{TN+FP}\right)$ при варьировании порога решающего правила (Threshold);

- *AUC* (*Area Under ROC Curve*) = 0,881 — площадь под кривой *ROC*; чем выше показатель *AUC*, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (это значение соответствует случайному гаданию на основе подбрасывания монеты), значение менее 0,5 свидетельствует о том, что классификатор действует с точностью до наоборот: если положительные назвать отрицательными и наоборот, классификатор будет работать лучше.

Судя по доле правильных ответов (*Accuracy*), построенная модель машинного обучения, очевидно, работает не идеально, но значительно лучше подбрасывания монеты, дающего 50% правильных ответов.

С другой стороны, если не пользоваться возможностями машинного обучения, а просто выдавать кредиты всем, и «хорошим», и «плохим» заемщикам, то правильных ответов будет 70% (в анализируемом наборе данных 700 клиентов из 1000 являются надежными).

Поэтому построенная модель машинного обучения требует улучшения, чтобы, прежде всего, уменьшить число «плохих» клиентов, неправильно относимых моделью к классу «хороших» (*FP*).[1] Это эквивалентно увеличению показателя точности классификации. Но при этом нужно, чтобы и число «хороших» заемщиков, которым выдаются кредиты, не снизилось, т. е. необходимо, чтобы и показатель полноты классификации остался достаточно большим.

Понятия точности $\left(Presicion = \frac{TP}{TP+FP}\right)$ и полноты $\left(Recall = \frac{TP}{TP+FN}\right)$ проиллюстрированы на рис. 7.4.53 с помощью диаграммы Венна — Эйлера.

На этой диаграмме объекты, которые на самом деле относятся к положительному классу, находятся в левой половине — они отмечены белыми кружками. Объекты, которые на самом деле относятся к отрицательному классу, находятся в правой половине — они отмечены черными кружками.

В круге находятся объекты, которые были названы «положительными» моделью машинного обучения.

Очевидно, в левой части круга находятся правильно распознанные представители положительного класса (*TP*), а в правой части круга — ошибки первого рода, т. е. представители отрицательного класса, ошибочно названные «положительными» (*FP*).

Точность (*Precision*) — это доля объектов, объединенных левым зеленым полукругом, среди всех объектов из круга.

Полнота (*Recall*) представляет собой отношение числа объектов из левого зеленого полукруга к общему числу объектов, являющихся на самом деле положительными и объединенных левым прямоугольником.

Посмотрим, что произойдет при изменении порога решающего правила (Threshold). По умолчанию как «положительные» классифицируются объекты, у которых вероятность отнесения к положительному классу (значение признака «Scored Probabilities») больше порога, равного 0,5.

[1] Ведь наша модель более половины «плохих» клиентов называет «хорошими»!



Рис. 7.4.53. *Точность и полнота классификации* (*Precision и Recall*)

Попробуем изменить порог отсечения. Для этого в окне визуализации оценки качества модели машинного обучения (рис. 7.4.52) можно подвигать мышью ползунок «Threshold». Метрики качества модели при различных значениях порога представлены на рис. рис. 7.4.54.

Замечаем, что, при увеличении порога с 0,5 (рис. 7.4.54, *а*) до 0,53 (рис. 7.4.54, *б*) доля правильных ответов, полнота и *F1* уменьшаются, а точность незначительно (на 0,001) увеличивается.

Увеличение порога отсечения до 0,56 приводит к значениям доли правильных ответов и точности, большим, чем при пороге, равном 0,5 (рис. 7.4.54, *в*). При этом полнота и *F1* уменьшаются незначительно. Заметим, что при переходе от порога, равного 0,5, к порогу, равному 0,56, число ошибок первого рода уменьшается на 11 (с 54 до 43), но при этом число ошибок второго рода увеличивается на 9 (с 31 до 40).

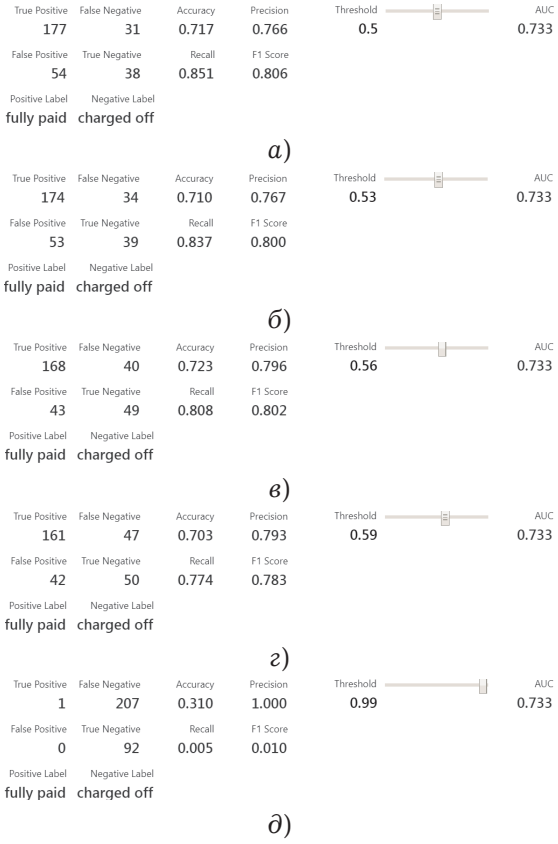


Рис. 7.4.54. *Метрики качества модели при значениях порога (Threshold), равных: а) 0, 5; б) 0, 53; в) 0, 56; г) 0, 59; д) 0, 99*

Таким образом, можно заключить, что в качестве порога отсечения 0,56 подходит лучше, чем 0,5.

Если увеличить порог до 0,59, то и доля правильных ответов, и точности, и полнота уменьшатся (рис. 7.4.54, *г*).

Если же задать порог на уровне 0,99, то мы получим 100%-ную точность, но при этом доля правильных ответов упадет до 0,31, а полнота — до 0,005 (рис. 7.4.54, *д*).

Теперь самостоятельно рассчитаем в Microsoft Excel все метрики, которые вычисляются в модуле «Evaluate Model».

Чтобы сохранить результаты прогнозов из облака *Microsoft* на свой компьютер, на панели модулей выберем модуль «Convert to CSV», который служит для преобразования наборов данных Microsoft Azure ML Studio в формат значений, разделенных запятыми (csv). Этот модуль находится в категории «Data Format Conversions». Разместим этот модуль левее модуля «Evaluate Model» и соединим выход («Scored dataset») модуля «Score Model» со входом («Dataset») модуля «Data Format Conversions».

Сохраним эксперимент, нажав кнопку «SAVE», запустим его, нажав кнопку «▶ RUN», а когда он закончит выполняться, сохраним результат на свой компьютер, нажав левой кнопкой на выходной порт модуля «Data Format Conversions» и выбрав пункт меню «Download» (рис. 7.4.55).

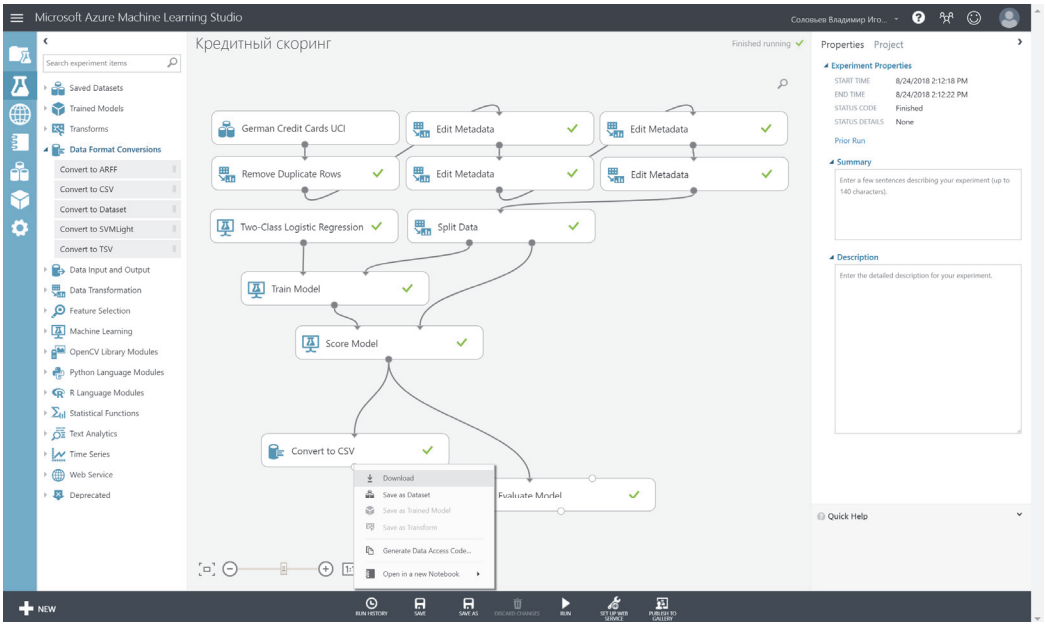


Рис. 7.4.55. *Сохранение результатов прогнозов на локальный компьютер*

Теперь откроем Microsoft Excel, создадим пустую книгу, на вкладке «Данные» нажмем кнопку «Получить данные», далее выберем пункты меню «Из файла > Из текстового/CSV-файла» и откроем сохраненные из Microsoft Azure ML Studio результаты прогнозов.

В мастере загрузки нажмем кнопку «Загрузить», и на рабочем листе появится таблица с исходным набором данных «German Credit Cards UCI», к которому добавлен столбец «Scored Labels» (содержащий прогнозы статуса возврата кредита) и «Scored Probabilities» (прогнозируемый вероятности возврата кредита). Эти данные сохранены на рабочем листе «Прогнозы (лог. регрессия)» файла «German Credit ROC.xlsx».

Скопируем столбцы «Loan Status», «Scored Labels» и «Scored Probabilities» на новый рабочий лист (в файле «German Credit ROC.xlsx» этот рабочий лист называется «Метрики качества»).

Поэкспериментируем с порогом отсечения на этом листе и посмотрим, как в зависимости от него меняется количество правильно распознанных «хороших» и «плохих» заемщиков (*TP* и *TN*), а также количество ошибок первого и второго рода (*FP* и *FN*).

Столбец E назовем «Threshold». Введем в ячейки E2:E12 следующие варианты значений порогов отсечения: 1; 0,9; 0,8; 0,7; ...; 0,1; 0.

Чтобы вычислить в ячейке F2, сколько будет правильно распознанных «хороших» клиентов, если считать «хорошими» всех, у кого прогноз вероятности возврата кредита больше значения ячейки E2, введем в F2 формулу

=СЧЁТЕСЛИМН(A:A; "=fully paid"; C:C; ">" & E2).

В результате в ячейке F2 будет рассчитано число клиентов из контрольной выборки, которые на самом деле вернули кредит (в столбце A стоит «fully paid»), и при этом прогноз вероятности возврата кредита этими клиентами из

столбца C больше порога, который задан в ячейке E2. Это не что иное, как число верно распознанных «хороших» клиентов — *TP*.

Аналогичным образом вычислим в ячейках G2, H2 и I2 соответственно *TN*, *FP* и *FN*:

TN =СЧЁТЕСЛИМН(A:A; "charged off"; C:C; "<=" & E2);

FP =СЧЁТЕСЛИМН(A:A; "charged off"; C:C; ">" & E2);

FN =СЧЁТЕСЛИМН(A:A; "fully paid"; C:C; "<="&E2) —

см. рис. 7.4.56.

Выделим ячейки F2:I2 и дважды щелкнем мышкой по правому нижнему углу выделенного диапазона — формулы автоматически распространятся на все ячейки, для которых указаны значения порога в столбце E. Результаты представлены на рис. 7.4.56.

Теперь легко вычислить для каждого варианта порога отсечения следующие характеристики (см. рис. рис. 7.4.57, *а*):

- общее число «хороших» и «плохих» клиентов: $P = TP + FN$; $N = TN + FP$ (естественно, значения *P* и *N* будут одними и теми же при всех значениях порога);
- долю «плохих» клиентов, неправильно отнесенных к «хорошим»: $FPR = \frac{FP}{TN+FP}$;
- долю верно классифицированных «хороших» клиентов в их общем количестве: $TPR = \frac{TP}{TP+FN}$;
- долю правильных ответов $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$, точность $Precision = \frac{TP}{TP+FP}$, полноту $Recall = \frac{TP}{TP+FN}$ (совпадающую с долей верно классифицированных «хороших» клиентов в их общем количестве, уже вычисленной как показатель *TPR*) и $F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$.

Поскольку с точки зрения алгоритма абсолютно не важно, какой из классов условно называть «положительным», для отрицательного класса можно вычислить аналоги точности и полноты: $Negative\ Precision = \frac{TN}{TN+FN}$ и $Negative\ Recall = \frac{TN}{TN+FP}$.

Замечание. В некоторых ячейках появилось сообщение «#ДЕЛ/0!». Это совершенно естественно. Например, при пороге отсечения, равном 1, ни один элемент не будет назван положительным, поэтому $TP = FP = 0$, и возникает неопределенность вида $0/0$: $Precision = TP(TP + FP) = 0 / (0 + 0)$. Таким образом, при пороге, равном единице, точность не определена. Также является неопределенным значение показателя *Negative Precision* при пороге, равном 0 и 0,1.

График зависимости *TPR* от *FPR* (рис. 7.4.58, *а*) — это приближенный график кривой *ROC*. Приближенный, так как мы строили его, изменяя порог отсечения с шагом, равным 0,1.

Кривая *ROC*, построенная с помощью Microsoft Azure ML Studio, более сложная, так как построена при шаге изменения порога отсечения, равном 0,001.

Мы можем построить в Microsoft Excel точно такую же такую кривую, как в Microsoft Azure ML Studio. Сделаем в файле «German Credit ROC.xlsx» копию рабочего листа «Метрики качества» и назовем ее «ROC». На рабочем листе «ROC» введем в ячейки E2:E4 числа 0; 0,001; 0,002, затем выделим диапазон E2:E4 и протянем его (за правый нижний угол) вниз до строки 2002.

	A	B	C	D	E	F	G	H	I
1	Loan Status	Scored Labels	Scored Probabilities		Threshold	TP	TN	FP	FN
2	charged off	charged off	0,4224524497985		1	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E2)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E2)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E2)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E2)
3	fully paid	fully paid	0,9639533162117		0,9	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E3)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E3)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E3)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E3)
4	charged off	charged off	0,1413848847150		0,8	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E4)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E4)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E4)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E4)
5	fully paid	fully paid	0,6344245672225		0,7	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E5)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E5)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E5)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E5)
6	charged off	fully paid	0,7579677104949		0,6	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E6)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E6)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E6)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E6)
7	charged off	fully paid	0,9073535799980		0,5	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E7)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E7)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E7)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E7)
8	fully paid	fully paid	0,8384476304054		0,4	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E8)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E8)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E8)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E8)
9	fully paid	fully paid	0,9754257798194		0,3	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E9)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E9)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E9)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E9)
10	charged off	fully paid	0,8199307322502		0,2	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E10)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E10)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E10)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E10)
11	fully paid	fully paid	0,5657136440277		0,1	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E11)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E11)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E11)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E11)
12	fully paid	fully paid	0,9560250043869		0	=СЧЁТЕСЛИМН(A:A;"=fully paid";C:C;">"&E12)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;"<="&E12)	=СЧЁТЕСЛИМН(A:A;"charged off";C:C;">"&E12)	=СЧЁТЕСЛИМН(A:A;"fully paid";C:C;"<="&E12)
13	...	...	...						
14	fully paid	fully paid	0,7167625427246						

а)

	A	B	C	D	E	F	G	H	I
1	Loan Status	Scored Labels	Scored Probabilities		Threshold	TP	TN	FP	FN
2	charged off	charged off	0,4224524497985		1	0	92	0	208
3	fully paid	fully paid	0,9639533162117		0,9	69	83	9	139
4	charged off	charged off	0,1413848847150		0,8	107	73	19	101
5	fully paid	fully paid	0,6344245672225		0,7	135	59	33	73
6	charged off	fully paid	0,7579677104949		0,6	155	51	41	53
7	charged off	fully paid	0,9073535799980		0,5	177	38	54	31
8	fully paid	fully paid	0,8384476304054		0,4	192	30	62	16
9	fully paid	fully paid	0,9754257798194		0,3	199	17	75	9
10	charged off	fully paid	0,8199307322502		0,2	206	11	81	2
11	fully paid	fully paid	0,5657136440277		0,1	208	0	92	0
12	fully paid	fully paid	0,9560250043869		0	208	0	92	0
13	...	...	...						
14	fully paid	fully paid	0,7167625427246						

б)

Рис. 7.4.56. *Расчет показателей TP, TN, FP, FN при различных значениях порога отсечения: формулы Microsoft Excel* (*а*); *результаты расчетов* (*б*)

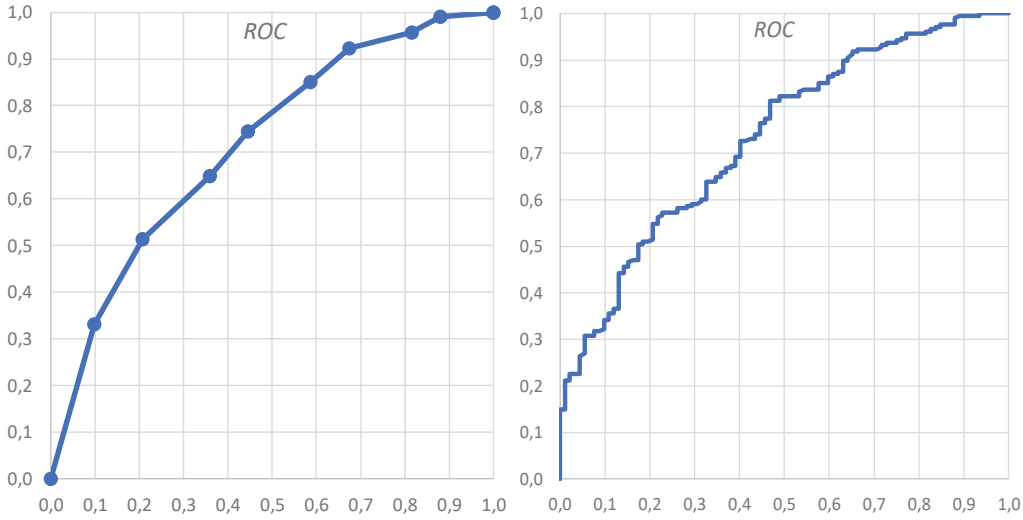
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Loan Status	Scored Labels	Scored Probabilities		Threshold	TP	TN	FP	FN	P = TP + FN	N = TN + FP	FPR	TPR	Accuracy	Precision	Recall	F1	Negative Precision	Negative Recal
2	charged off	charged off	0,4224524497985		1	0	92	0	208	208	92	0,0000000	0,0000000	0,3066667	#ДЕЛ/0!	0,0000000	#ДЕЛ/0!	0,3066667	1,0000000
3	fully paid	fully paid	0,9639533162117		0,9	69	83	9	139	208	92	0,0978261	0,3317308	0,5066667	0,8846154	0,3317308	0,4825175	0,3738739	0,9021739
4	charged off	charged off	0,1413848847150		0,8	107	73	19	101	208	92	0,2065217	0,5144231	0,6000000	0,8492063	0,5144231	0,6407186	0,4195402	0,7934783
5	fully paid	fully paid	0,6344245672225		0,7	135	59	33	73	208	92	0,3586957	0,6490385	0,6466667	0,8035714	0,6490385	0,7180851	0,4469697	0,6413043
6	charged off	fully paid	0,7579677104949		0,6	155	51	41	53	208	92	0,4456522	0,7451923	0,6866667	0,7908163	0,7451923	0,7673267	0,4903846	0,5543478
7	charged off	fully paid	0,9073535799980		0,5	177	38	54	31	208	92	0,5869565	0,8509615	0,7166667	0,7662338	0,8509615	0,8063781	0,5507246	0,4130435
8	fully paid	fully paid	0,8384476304054		0,4	192	30	62	16	208	92	0,6739130	0,9230769	0,7400000	0,7559055	0,9230769	0,8311688	0,6521739	0,3260870
9	fully paid	fully paid	0,9754257798194		0,3	199	17	75	9	208	92	0,8152174	0,9567308	0,7200000	0,7262774	0,9567308	0,8257261	0,6538462	0,1847826
10	charged off	fully paid	0,8199307322502		0,2	206	11	81	2	208	92	0,8804348	0,9903846	0,7233333	0,7177700	0,9903846	0,8323232	0,8461538	0,1195652
11	fully paid	fully paid	0,5657136440277		0,1	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
12	fully paid	fully paid	0,9560250043869		0	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
13	...	...	...																
14	fully paid	fully paid	0,7167625427246																

а)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Loan Status	Scored Labels	Scored Probabilities		Threshold	TP	TN	FP	FN	P = TP + FN	N = TN + FP	FPR	TPR	Accuracy	Precision	Recall	F1	Negative Precision	Negative Recal
2	charged off	charged off	0,4224524497985		0	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
3	fully paid	fully paid	0,9639533162117		0,005	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
4	charged off	charged off	0,1413848847150		0,01	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
5	fully paid	fully paid	0,6344245672225		0,015	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
6	charged off	fully paid	0,7579677104949		0,02	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
7	charged off	fully paid	0,9073535799980		0,025	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
8	fully paid	fully paid	0,8384476304054		0,03	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
9	fully paid	fully paid	0,9754257798194		0,035	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
10	charged off	fully paid	0,8199307322502		0,04	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
11	fully paid	fully paid	0,5657136440277		0,045	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
12	fully paid	fully paid	0,9560250043869		0,05	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
13	...	...	...		0,055	208	0	92	0	208	92	1,0000000	1,0000000	0,6933333	0,6933333	1,0000000	0,8188976	#ДЕЛ/0!	0,0000000
14	fully paid	fully paid	0,7167625427246		...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

б)

Рис. 7.4.57. *Расчет показателей P, N, FPR, TPR, Accuracy, Precision, Recall, F1, Negative Precision и Negative Recall:*
при изменении порога с шагом, равным 0, 1 (а) и при изменении порога с шагом, равным 0, 005 (б)



а) *б*)

Рис. 7.4.58. *Построение кривой ROC при шаге изменения порога, равном: 0, 1 (а); 0, 001 (б)*

В результате диапазон E2:E2002 будет заполнен числами, возрастающими от 0 до 1 с шагом 0,001. Теперь выделим диапазон F2:S12 и дважды щелкнем по правому нижнему углу — формулы распространятся на все строки до 2002-й. После этого график кривой *ROC* станет выглядеть, как на рис. 7.4.58, *б*, полностью совпадая с тем, который был построен с помощью Microsoft Azure ML Studio. На рис. 7.4.59 эти два графика наложены друг на друга.

Результаты расчетов сохранены в файле «German Credit ROC.xlsx» на рабочем листе «Метрики качества».

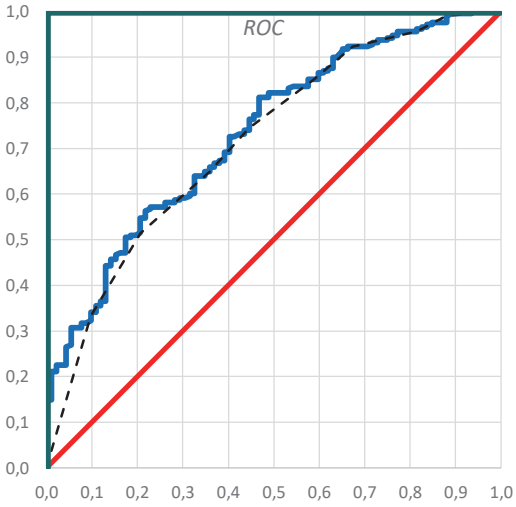


Рис. 7.4.59. *Сравнение кривых ROC*

Также на рис. 7.4.59 построены *ROC*-кривые для идеального классификатора, который никогда не ошибается, (зеленый «уголок») и для «самого плохого» классификатора, который всегда ошибается с 50%-ной вероятностью, (красная

линия). Площадь квадрата под зеленой *ROC*-кривой, очевидно, равна 1, а площадь треугольника под красной *ROC*-кривой равна 0,5.

Площадь под кривой *ROC* в нашем эксперименте составляет *AUC* = 0,733. Это значение было вычислено с помощью Microsoft Azure ML Studio (см. рис. 7.4.52). Предлагаем читателям убедиться, что это так, вычислив значение *AUC* с помощью Microsoft Excel в задаче **453**.

На рис. 7.4.60 представлен график зависимости точности от полноты, рассчитанный в результате работы модуля «Evaluate Model» (чтобы его увидеть, нужно в окне визуализации результатов оценки качества построенной модели машинного обучения нажать на вкладку «PRECISION/RECALL»). В задаче **454** предлагается построить этот график в Microsoft Excel.

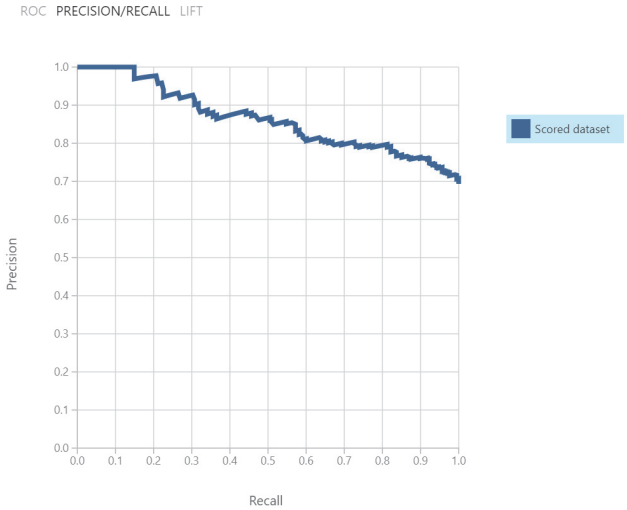


Рис. 7.4.60. *Кривая Precision/Recall*

Попробуем синтезировать новые признаки, чтобы улучшить качество предсказания. Перейдем от признаков «Loan Duration (Months)», «Loan Amount», «Age (Years)» и «Number of Existing Credits at This Bank» к их логарифмам.

Для этого добавим в эксперимент модуль «Apply Math Operation» (он находится в категории «Statistical Functions») и соединим левый выход модуля «Split Data» со входом модуля «Apply Math Operation». На панели параметров модуля «Apply Math Operation» нажмем кнопку «Launch column selector» и выберем столбцы «Loan Duration (Months)», «Loan Amount», «Age (Years)» и «Number of Existing Credits at This Bank». Далее укажем следующие параметры:

Category: Basic;
Basic math function: Ln;
Output mode: Inplace.[1]

Сделаем копию модуля «Apply Math Operation» и соединим с его входом правый выход модуля «Split Data» (рис. 7.4.61).

[1] По умолчанию режим вывода этого модуля — «ResultOnly» (вывести единственный столбец с результатом операции). Как правило, этот режим целесообразно изменить на «Append» (добавить новый столбец с результатом операции к существующему набору данных) или «Inplace» (добавить столбец с результатом операции вместо столбца, являвшегося первым аргументом операции).

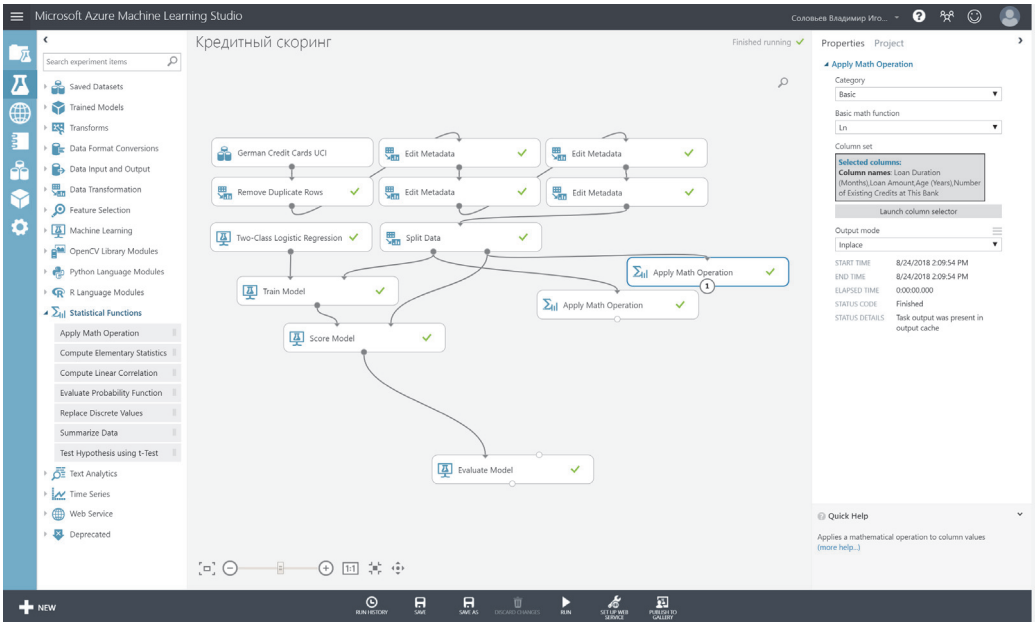


Рис. 7.4.61. *Логарифмирование признаков в обучающей и контрольной выборках*

Теперь скопируем модули «Two-class Logistic Regression», «Train Model» и «Score Model» (сначала выделим мышью один из модулей, затем нажмем клавишу «Ctrl» и, не отпуская ее, выделим остальные, после этого вызовем меню правой кнопкой мыши, и воспользуемся по очереди пунктами «Copy» и «Paste», как показано на рис. 7.4.62).

Соединим выход модуля «Apply Math Operation», используемого для преобразования обучающей выборки, с правым входом нового модуля «Train Model», а выход модуля «Apply Math Operation», используемого для преобразования контрольной выборки, с правым входом нового модуля «Score Model». Выход нового модуля «Score Model» соединим с правым входом модуля «Evaluate Model» (рис. 7.4.63).

Запустим эксперимент на выполнение, а затем визуализируем выход модуля «Evaluate Model» и сравним метрики качества (рис. 7.4.64): все показатели улучшились, но совсем незначительно, в пределах статистической погрешности.

Теперь попробуем улучшить качество предсказания, воспользовавшись другим алгоритмом машинного обучения, например, методом случайного леса.

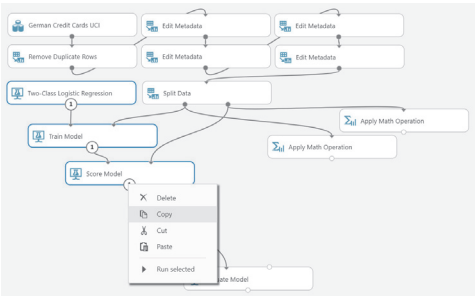


Рис. 7.4.62. *Копирование нескольких модулей*

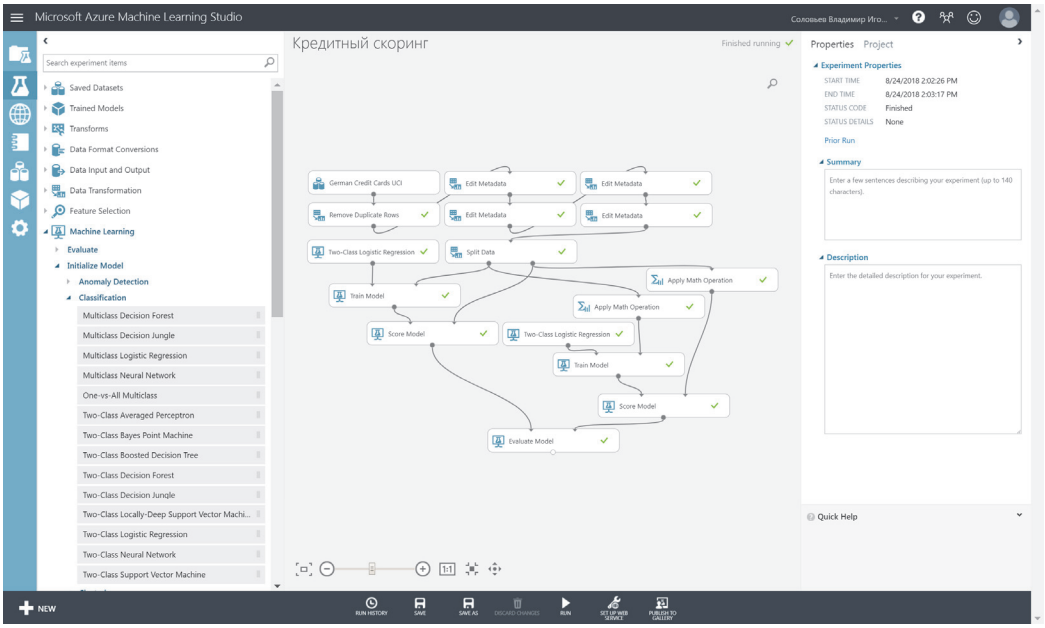


Рис. 7.4.63. *Эксперимент по сравнению качества прогнозирования двух моделей логистической регрессии*

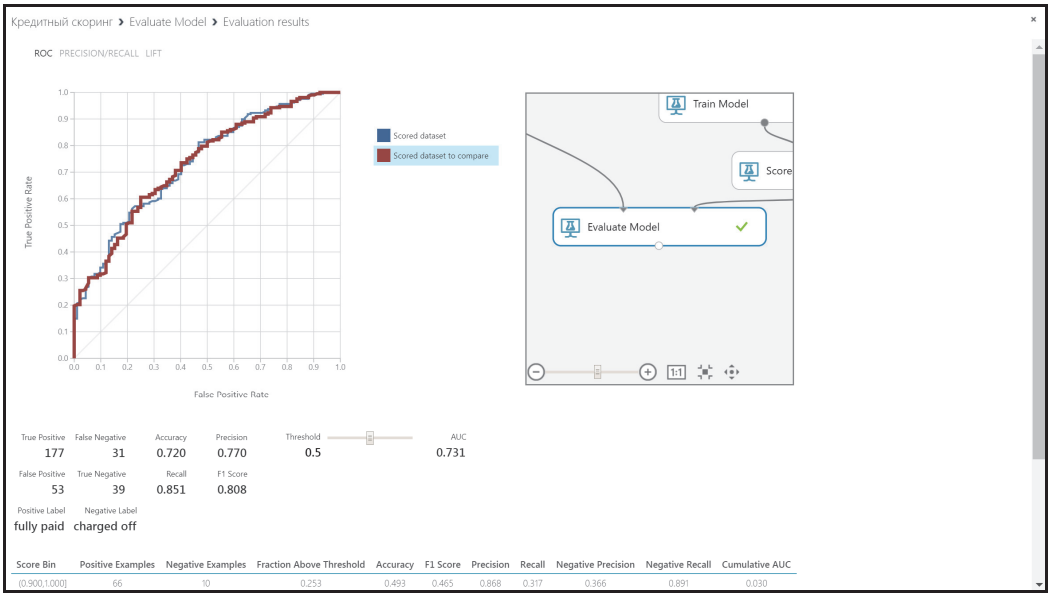


Рис. 7.4.64. *Сравнение результатов оценки качества прогнозирования двух моделей логистической регрессии*

Удалим второй модуль «Two-class Logistic Regression», и на его место поставим модуль «Two-Class Decision Forest» (его можно найти в категории «Machine Learning > Train»). На панели свойств модуля «Two-Class Decision Forest» оставим без изменения параметры, установленные по умолчанию:

Resampling method: Bagging;
Create trainer mode: Single Parameter;

Number of decision trees: 8;
Maximum depth of the decision tree: 32;
Number of random splits per node: 128;
Minimum number of samples per leaf node: 1;
Allow unknown values for categorical features: установим выделение.

Эксперимент примет вид, как на рис. 7.4.65.

Запустим его на выполнение, а затем визуализируем выход модуля «Evaluate Model» и сравним метрики качества (рис. 7.4.66).

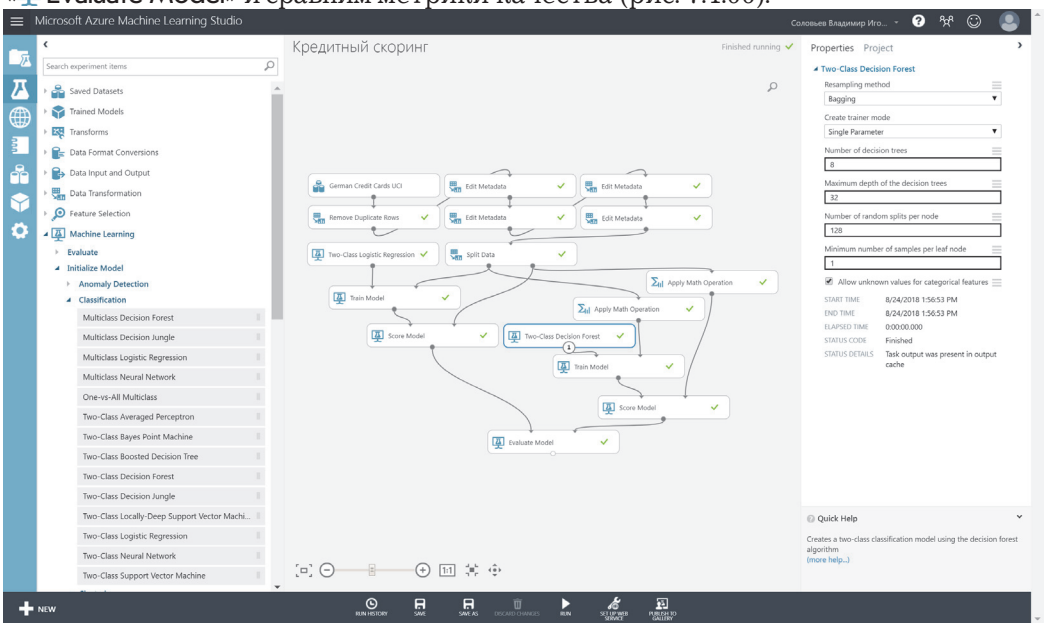


Рис. 7.4.65. *Эксперимент по сравнению качества прогнозирования модели логистической регрессии и модели случайного леса*

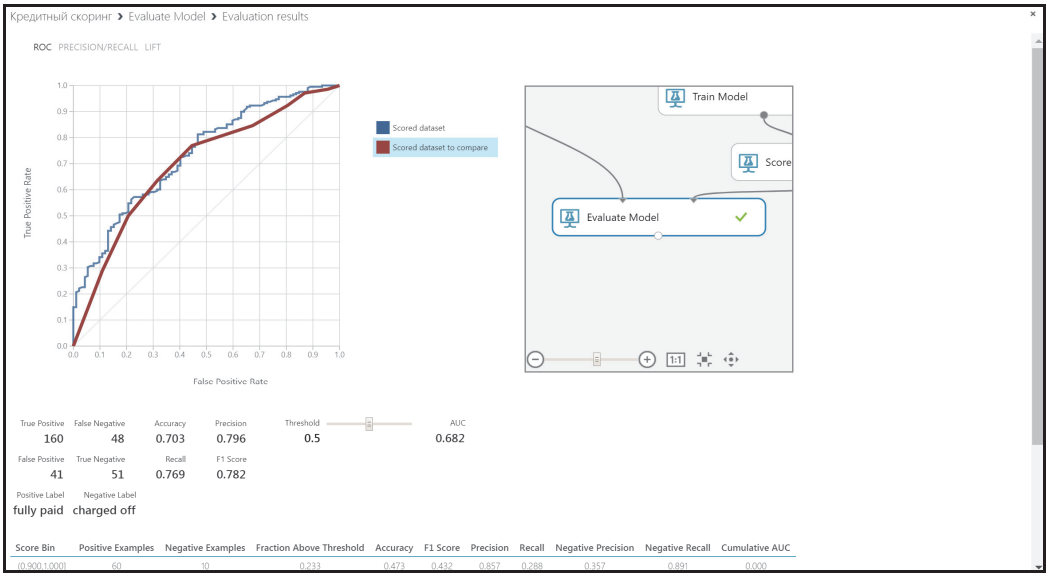


Рис. 7.4.66. *Сравнение результатов оценки качества прогнозирования модели логистической регрессии и модели случайного леса*

Видим, что все показатели стали существенно хуже.

Улучшение качества модели достигается путем варьирования параметров алгоритма.

Изменим параметры модели леса решений (на панели свойств модуля «Two-Class Decision Forest»):

Resampling method: Bagging;
Create trainer mode: Single Parameter;
Number of decision trees: 50;
Maximum depth of the decision tree: 32;
Number of random splits per node: 32;
Minimum number of samples per leaf node: 4;
Allow unknown values for categorical features: установим выделение.

Запустив эксперимент на выполнение, а затем визуализировав результаты оценки качества, замечаем, что результат стал заметно лучше, причем и по сравнению с первым вариантом случайного леса, и по сравнению с логистической регрессией (рис. 7.4.67).

Еще раз изменим параметры модели леса решений:

Resampling method: Bagging;
Create trainer mode: Single Parameter;
Number of decision trees: 50;
Maximum depth of the decision tree: 50;
Number of random splits per node: 16;
Minimum number of samples per leaf node: 4;
Allow unknown values for categorical features: установим выделение.

Запустим эксперимент на выполнение, а затем визуализируем оценку его качества — выход модуля «Evaluate Model» (рис. 7.4.68). Замечаем, что показатели качества стали еще лучше.

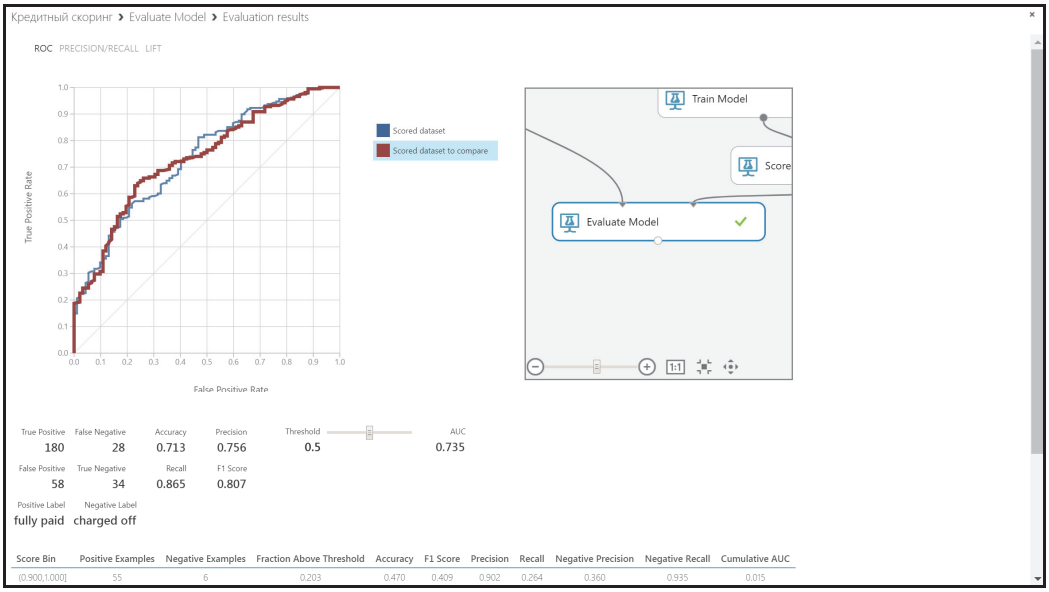


Рис. 7.4.67. *Сравнение результатов оценки после изменения параметров случайного леса*

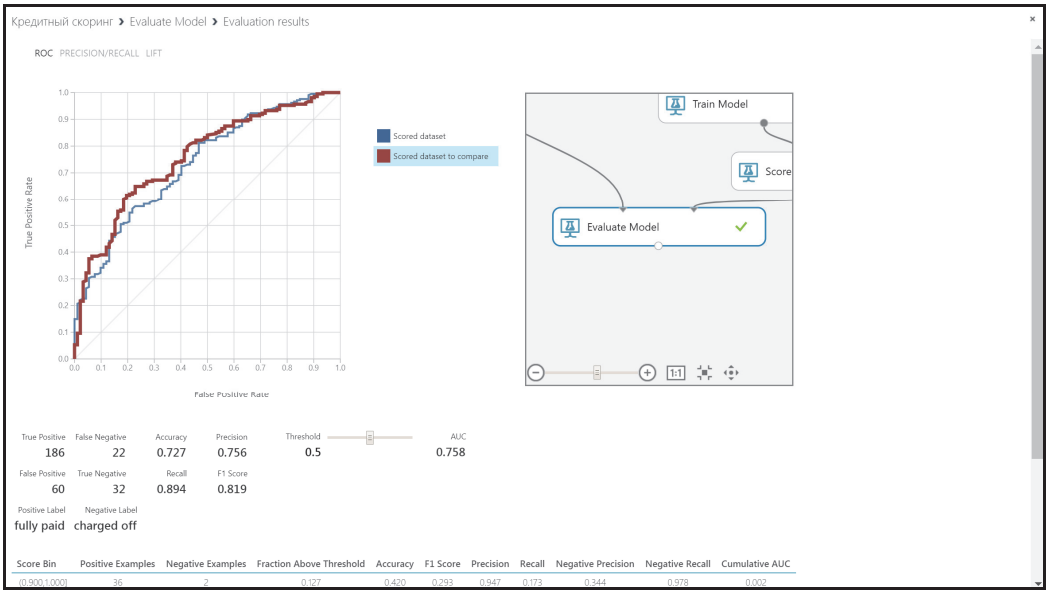


Рис. 7.4.68. *Сравнение результатов оценки после второго изменения параметров случайного леса*

Итак, мы попробовали для решения задачи кредитного скоринга синтезировать дополнительные признаки и использовать различные алгоритмы. Оказалось, что это позволяет как улучшить, так и ухудшить качество прогнозирования вероятности возврата кредита.

При этом во всех вариантах количество ненадежных клиентов, распознаваемых алгоритмом как «хороших» (*FP*), слишком велико.

Данный эксперимент можно скопировать себе — он опубликован по адресу https://is.gd/findata_creditscoring. ❑

Замечание. Мы подбирали параметры алгоритмов вручную, «с потолка». На практике параметры подбирают, используя полный или частичный перебор различных вариантов. В Microsoft Azure ML Studio для этого предназначен модуль «Tune Model Hyperparameters» из категории «Machine Learning > Train».

Пример 7.4.5. Теперь построим модель машинного обучения для прогнозирования объемов продаж компьютерных игр в Японии на основе набора данных «Video Games» из примера 7.4.3.

Разделим подготовленный набор данных на обучающую и контрольную выборку. Для этого на панели модулей выберем модуль «Split Data» («Data Transformation > Sample and Split») и перенесем его в рабочую область. Соединим выход набора данных «Video Games» модуля «Split Data», а левый выход модуля «Split Data» (который будет соответствовать обучающей выборке) — со входом модуля «Remove Duplicate Rows».

На панели свойств модуля «Split Data» введем следующую информацию:

Splitting mode: Split Rows;
Fraction of rows in the first output dataset: 0.7;
Randomized split: установим выделение;
Random seed: 123;
Stratified split: False.

Теперь эксперимент будет выглядеть, как на рис. 7.4.69.

С тестовой выборкой мы должны провести все те же преобразования, что и с обучающей выборкой, за исключением удаления строк.

Поэтому скопируем три модуля «Clean Missing Data» (рис. 7.4.70) и вставим их в правую часть рабочей области эксперимента (рис. 7.4.71).

Для прогнозирования воспользуемся алгоритмом линейной регрессии. Добавим в рабочую область эксперимента модуль «Linear Regression» («Machine Learning > Initializate Model > Regression») и на панели его свойств оставим значения параметров модели линейной регрессии, установленные по умолчанию:

Solution method: Ordinary Least Squares;
L2 regularization weight: 0,001;
Include intercept term: установим выделение;[1]
Random number seed: 123;
Allow unknown categorical levels: установим выделение.

В результате эксперимент будет выглядеть, как на рис. 7.4.72.

Добавим в эксперимент модуль обучения («Train Model» («Machine Learning > Train»), соединим выход модуля «Linear Regression» с левым входом модуля «Train Model», а выход последнего из модулей «Edit Metadata» с правым входом модуля «Train Model».

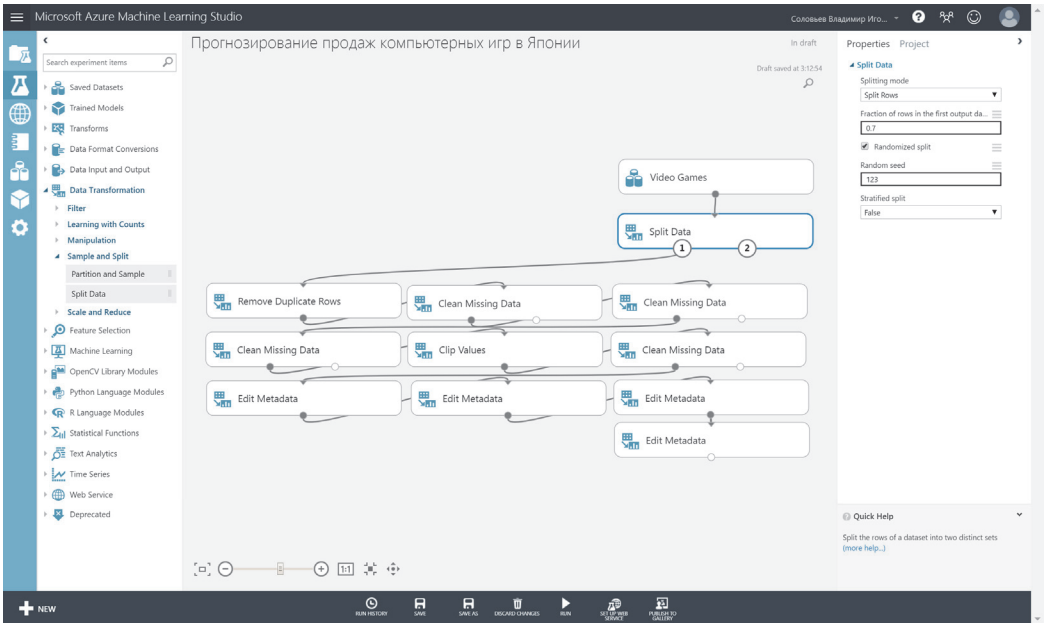


Рис. 7.4.69. *Разделение набора данных «Video Games»*

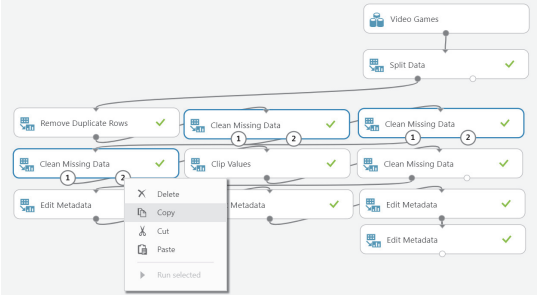


Рис. 7.4.70. *Копирование модулей обработки пропусков*

[1] Мы строим регрессионную модель со свободным членом.

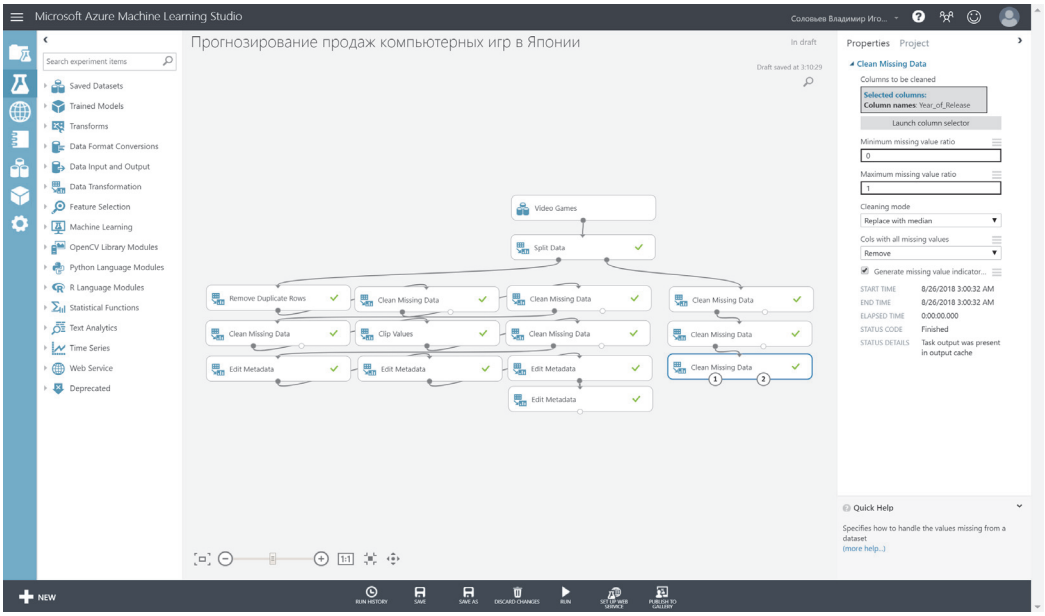


Рис. 7.4.71. *Обработка пропусков в контрольной выборке*

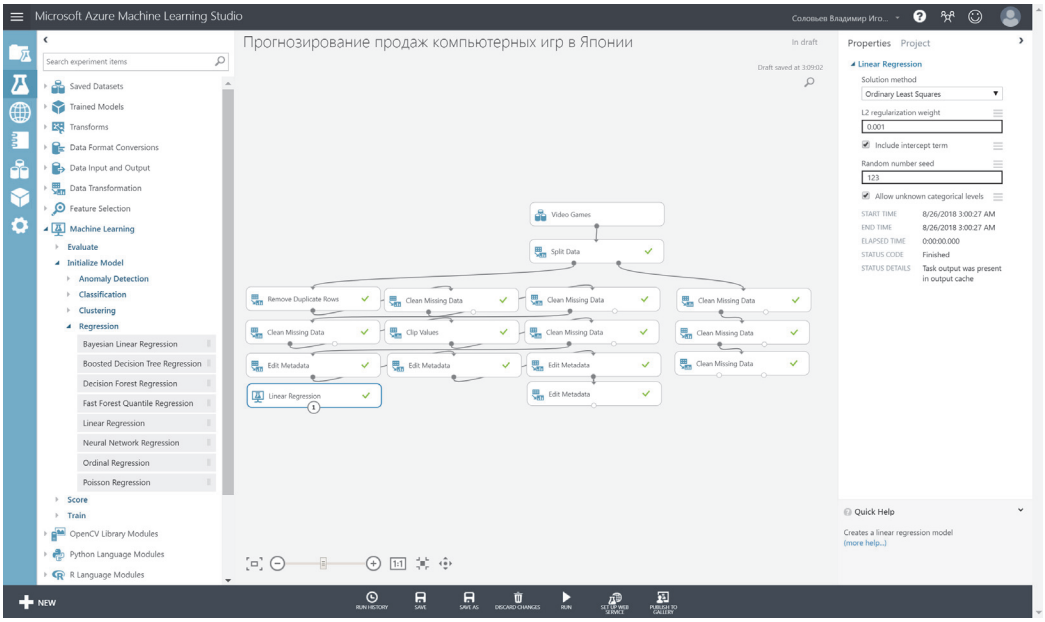


Рис. 7.4.72. *Добавление модуля инициализации модели линейной регрессии*

На панели свойств модуля «Train Model» нажмем кнопку «Launch column selector» и выберем признак «JP_Sales» в качестве результативного (рис. 7.4.73). Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN». Когда выполнение эксперимента завершится, визуализируем выход модуля «Train Model» (рис. 7.4.74) и увидим вначале параметры алгоритма (Settings), а затем коэффициенты модели логистической регрессии (Feature Weights), отсортированные по убыванию абсолютных значений (коэффициент «Bias» — это свободный член).

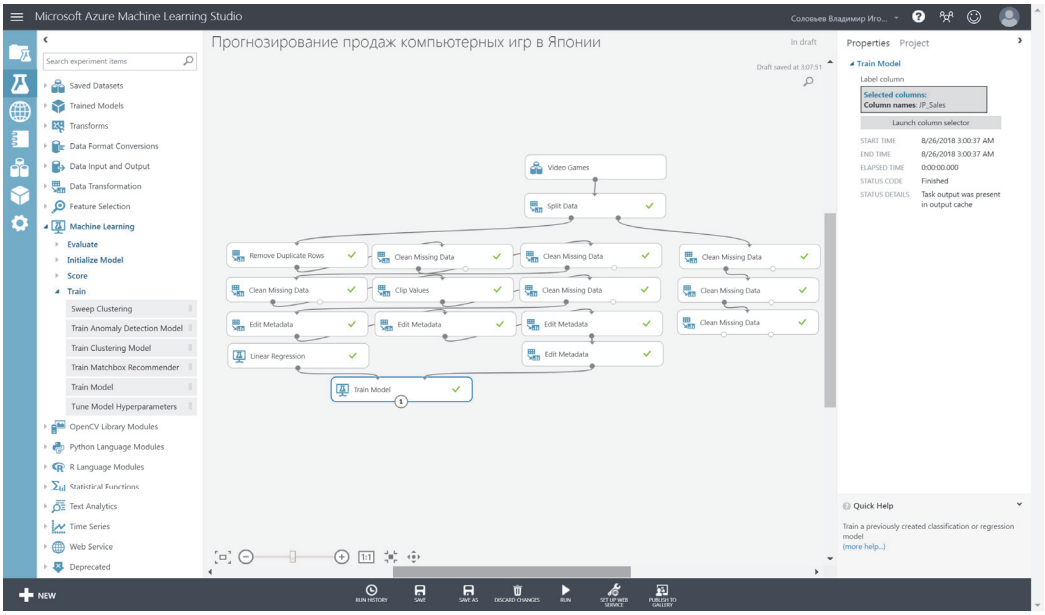


Рис. 7.4.73. *Добавление модуля обучения модели линейной регрессии*



Рис. 7.4.74. *Визуализация результатов обучения модели линейной регрессии*

Добавим в эксперимент модуль вычисления прогнозов «Score Model» (он находится в категории «Machine Learning > Score») и перенесем его в рабочую область ниже всех других. Соединим выход модуля «Train Model» с левым входом модуля «Score Model», а выход последнего из модулей «Clean Missing Data» — с правым входом модуля «Score Model». После этого эксперимент будет выглядеть так, как показано на рис. 7.4.75.

Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN». Когда выполнение эксперимента закончится, визуализируем выход модуля «Score Model», щелкнув по выходному порту левой кнопкой мыши и выбрав пункт меню «Visualize» (рис. 7.4.76).

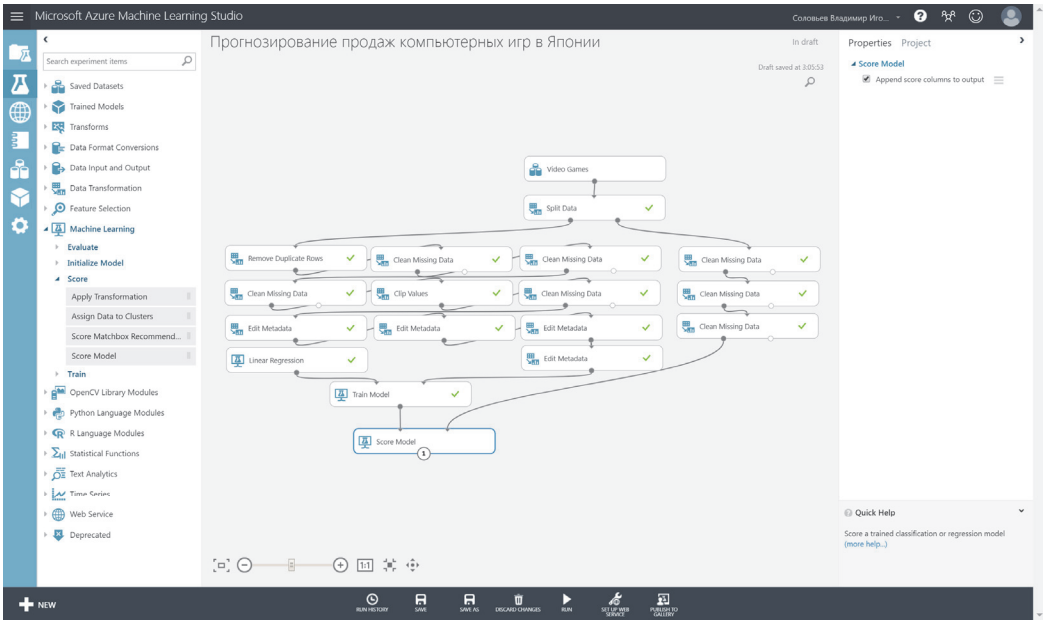


Рис. 7.4.75. *Добавление модуля вычисления прогнозов*

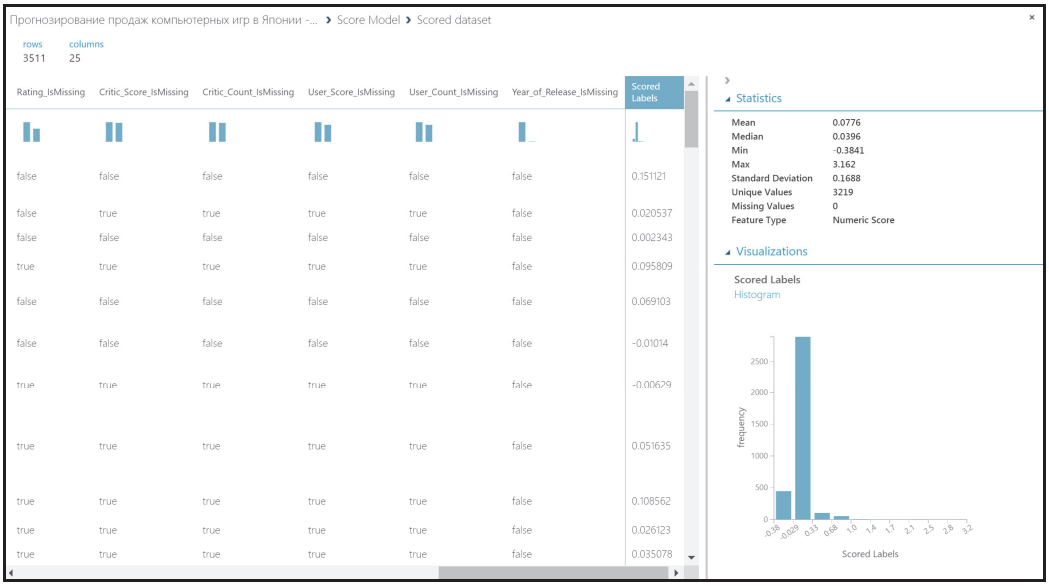


Рис. 7.4.76. *Визуализация результатов вычисления прогнозов на контрольной выборке*

Обратим внимание на новый (самый правый) столбец «Scored Labels» — это прогноз объема продаж компьютерной игры в Японии.

Прогноз результативного признака Y по модели линейной регрессии определяется как линейная функция $\hat{Y}_x = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + \cdots + a_m x^{(m)}$ (где $x^{(1)}$, $x^{(2)}$, $x^{(m)}$ — значения факторных признаков). При этом значение прогноза может быть как положительным, так и отрицательным. Но по содержательному смыслу задачи объемы продаж отрицательными быть не могут, поэтому если прогноз окажется отрицательным его целесообразно заменить на нулевой. Для

этого добавим в эксперимент модуль модулей «Clip Values» («Data Transformation > Scale and Reduce»), как показано на рис. 7.4.77.

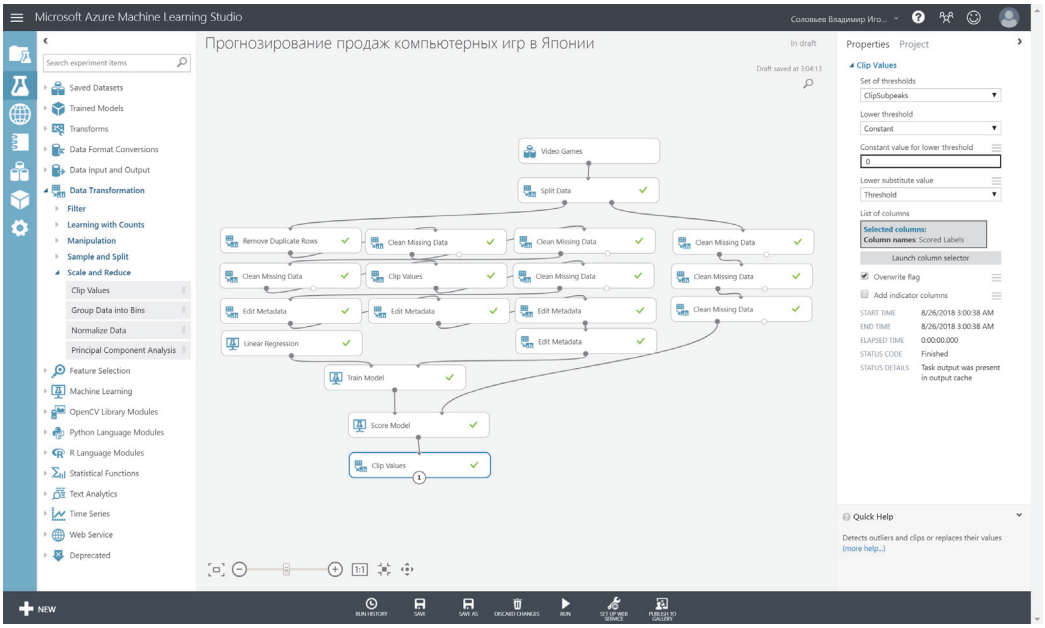


Рис. 7.4.77. *Добавление модуля «Clip Values»*

Соединим выход модуля «Score Model» с входом модуля «Clip Values». На панели свойств модуля «Clip Values» с помощью кнопки «Launch column selector» выберем признак «Scored Labels» и установим следующие значения параметров:

Set of thresholds: ClipSubpeaks;
Lower threshold: Constant;
Constant value for lower threshold: 0;
Lower substitute value: Threshold;
Overwrite flag: установим выделение;
Add indicator columns: очистим выделение.

Чтобы оценить качество прогнозирования, добавим в эксперимент модуль «Evaluate Model» («Machine Learning > Evaluate»). Соединим выход модуля «Clip Values» (выходящего из модуля «Score Model») с левым входом модуля «Evaluate Model» (рис. 7.4.78). Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN». Когда выполнение эксперимента закончится, визуализируем выход («Evaluation results») модуля «Evaluate Model» (рис. 7.4.79).

Наиболее часто качество моделей регрессии оценивается с помощью *средней абсолютной ошибки предсказания* (*Mean Absolute Error, MAE*) — среднего значения модулей отклонений реальных значений y_i от прогнозов $\hat{y}_i$:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n},$$

средней квадратической ошибки предсказания (*Mean Squared Error, MSE*) — среднего значения квадратов отклонений реальных значений y_i от прогнозов $\hat{y}_i$:

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

и *квадратного корня из средней квадратической ошибки предсказания* (*Root Mean Squared Error, RMSE*):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}.$$

Сравнение функции абсолютного значения и квадратичной функции (рис. 7.4.80) позволяет заключить, что средняя квадратическая ошибка предполагает значительно более сильные штрафы за большие отклонения, чем средняя абсолютная ошибка.

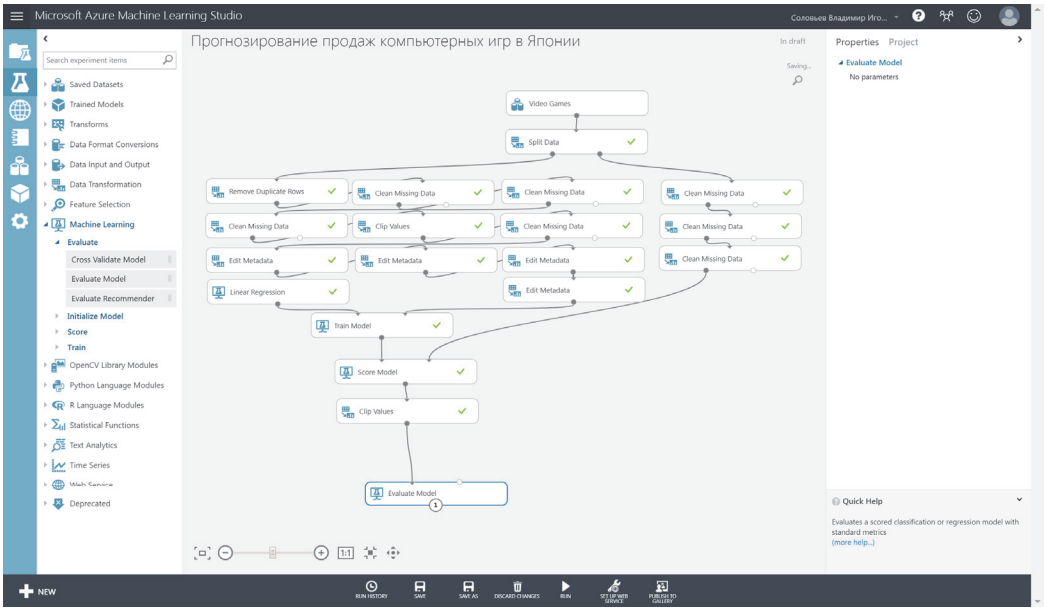


Рис. 7.4.78. *Добавление модуля оценки качества модели*

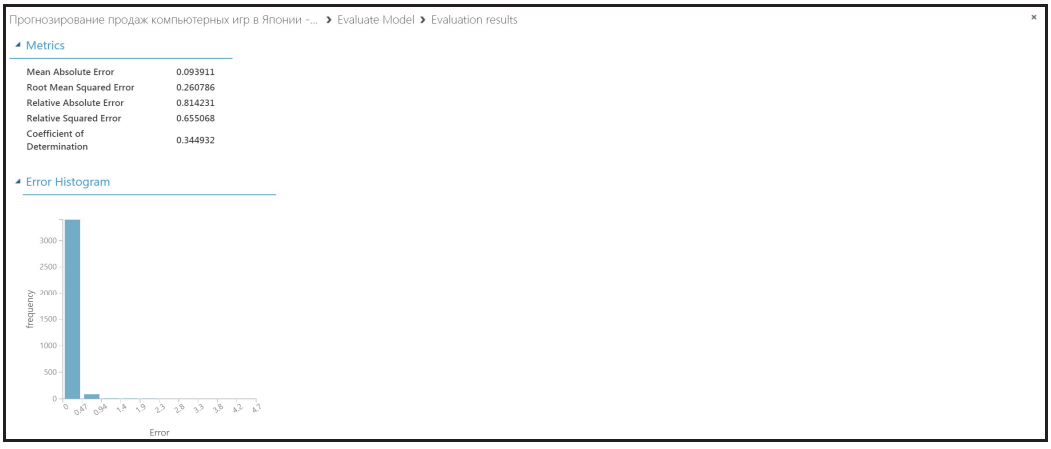


Рис. 7.4.79. *Визуализация результатов обучения модели линейной регрессии*

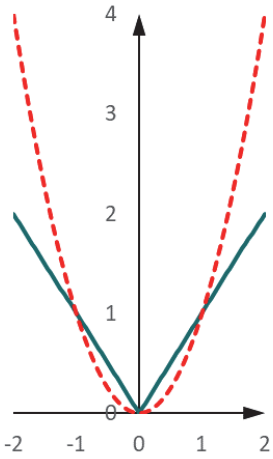


Рис. 7.4.80. *Модуль (сплошная зеленая линия) и квадрат (красный пунктир)*

Поэтому в качестве основной оценки качества модели *RMSE* используют в том случае, когда большие отклонения совершенно неприемлемы. В остальных случаях чаще всего используют *MAE*, поскольку среднее абсолютное значение отклонения интерпретируется гораздо понятнее, чем квадратный корень из среднего значения квадратов отклонений.

В рассматриваемом примере *MAE* = 0,093911 (в среднем прогноз отличается от истины примерно на 0,09 млн долл.), *RMSE* = 0,260786.

Как было отмечено в примере 7.4.3, все объемы продаж («NA_Sales», «EU_Sales», «JP_Sales», «Other_Sales») имеют скошенные распределения. Такие признаки, у которых имеется сильный скос, как правило, полезно заменить на их логарифмы, тогда их распределения будут более симметричными и более близкими к нормальному — для многих алгоритмов машинного обучения это важно.

Но поскольку указанные признаки содержат много нулевых значений, а логарифм нуля не определен, лучше воспользоваться преобразованием $\ln(1 + x)$.

Добавим в эксперимент модуль «Apply Math Operation» (он находится в категории «Statistical Functions») и соединим выход последнего из модулей «Edit Metadata» со входом модуля «Apply Math Operation». На панели параметров модуля «Apply Math Operation» нажмем кнопку «Launch column selector» и выберем столбцы «NA_Sales», «EU_Sales», «JP_Sales» и «Other_Sales». Далее укажем следующие параметры:

Category: Basic;
Basic math function: LnPlus1;
Output mode: Inplace.

Соответствующий эксперимент проиллюстрирован рис. 7.4.81.

Сделаем копию модуля «Apply Math Operation» и соединим с его входом выход последнего из модулей очистки контрольного набора данных — модуля «Clean Missing Data».

Естественно, логарифмировать признак «JP_Sales» в контрольной выборке не следует, ведь он будет использоваться для оценки качества прогнозирования.

Поэтому нажмем кнопку «Launch column selector» и оставим выбранными столбцы «NA_Sales», «EU_Sales» и «Other_Sales» (рис. 7.4.82).

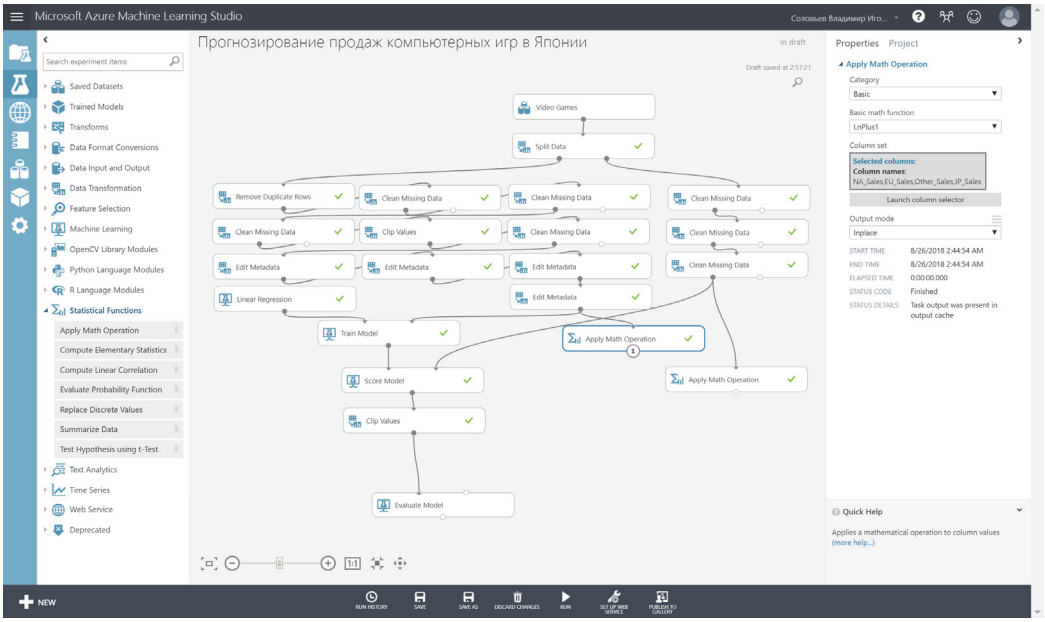


Рис. 7.4.81. *Логарифмирование объемов продаж в обучающей выборке*

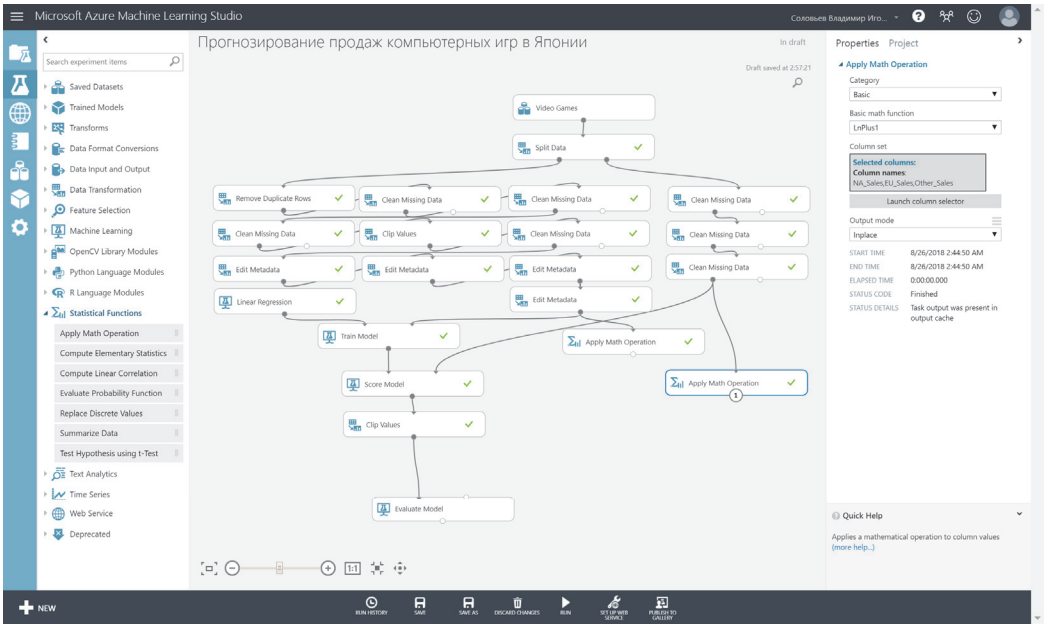


Рис. 7.4.82. *Логарифмирование объемов продаж в контрольной выборке*

Теперь скопируем модули «Linear Regression», «Train Model», «Score Model» и «Clip Values» (для этого можно мышкой выделить их сразу все, как показано на рис. 7.4.83, а затем воспользоваться меню, выпадающим по нажатию правой кнопки мышки).

Соединим выход модуля «Apply Math Operation», используемого для преобразования обучающей выборки, с правым входом модуля «Train Model», а выход модуля «Apply Math Operation», используемого для преобразования контрольной выборки, с правым входом модуля «Score Model».

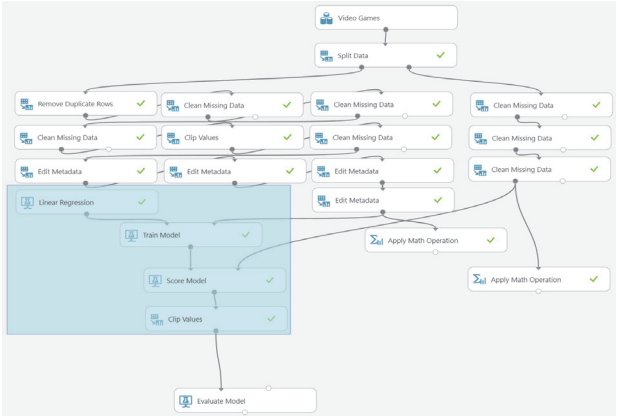


Рис. 7.4.83. *Выделение нескольких модулей*

Поскольку мы в обучающей выборке логарифмировали признак «JP_Sales», теперь нужно выполнить обратное преобразование ($e^y - 1$). Для этого добавим еще один модуль «Apply Math Operation» и соединим с его входом выход нового модуля «Score Model». На панели параметров модуля «Apply Math Operation» выберем единственный столбец «JP_Sales» и укажем следующие параметры:

Category: Basic;
Basic math function: ExpMinus1;
Output mode: Inplace.

Выход модуля «Apply Math Operation» соединим со входом вставленного недавно нового модуля «Clip Values», а выход модуля «Clip Values» — с правым входом модуля «Evaluate Model» (рис. 7.4.84).

Запустим эксперимент на выполнение, а затем проанализируем метрики качества (рис. 7.4.85). Отмечаем, что средняя абсолютная ошибка предсказания (*MAE*) уменьшилась (улучшилась), а квадратный корень из средней квадратичной ошибки (*RMSE*) предсказания увеличился (ухудшился).

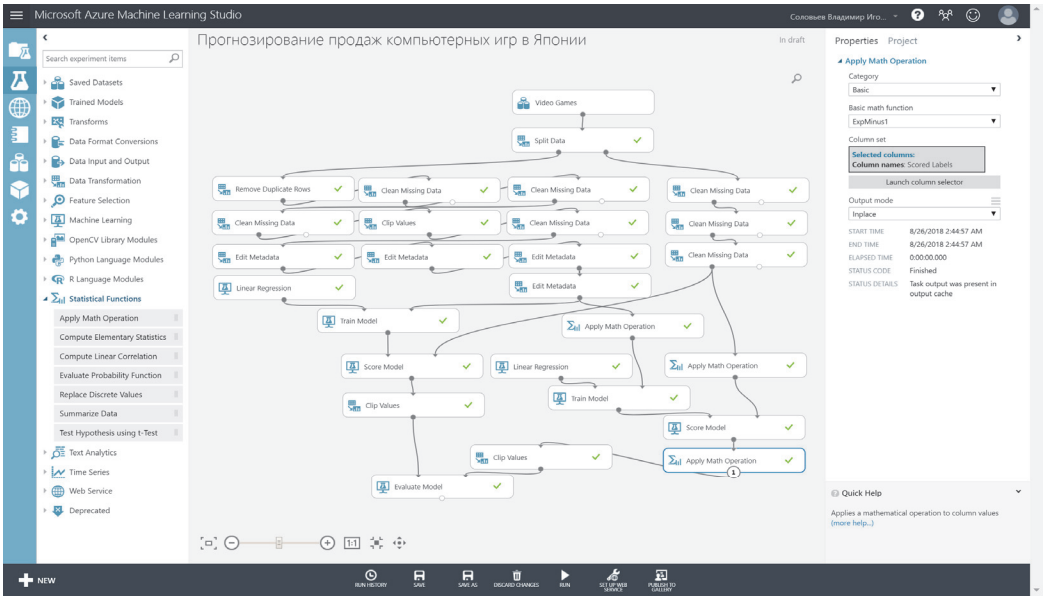


Рис. 7.4.84. *Обратное преобразование прогноза логарифма объема продаж*

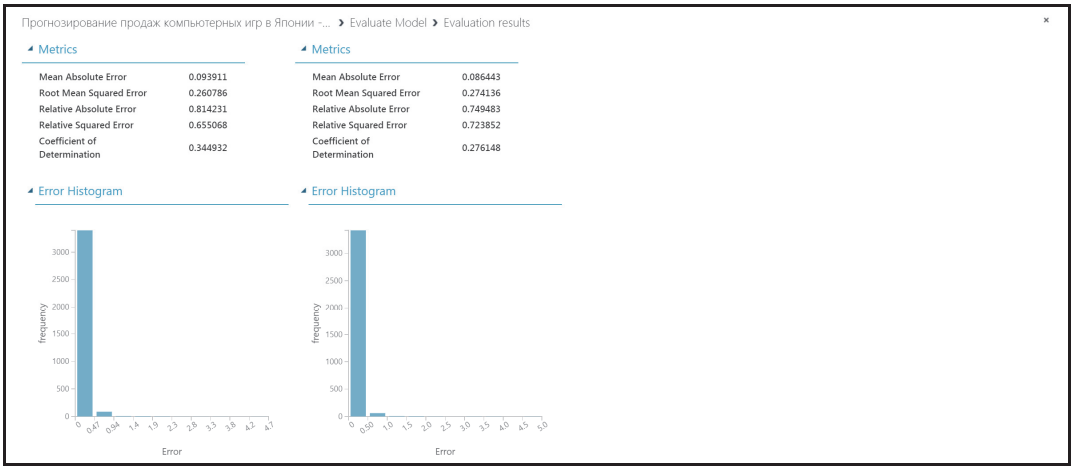


Рис. 7.4.85. *Сравнение результатов оценки двух моделей прогнозирования*

В заключение этого примера добавим к эксперименту краткое резюме — Summary и более подробное описание — Description (рис. 7.4.86).

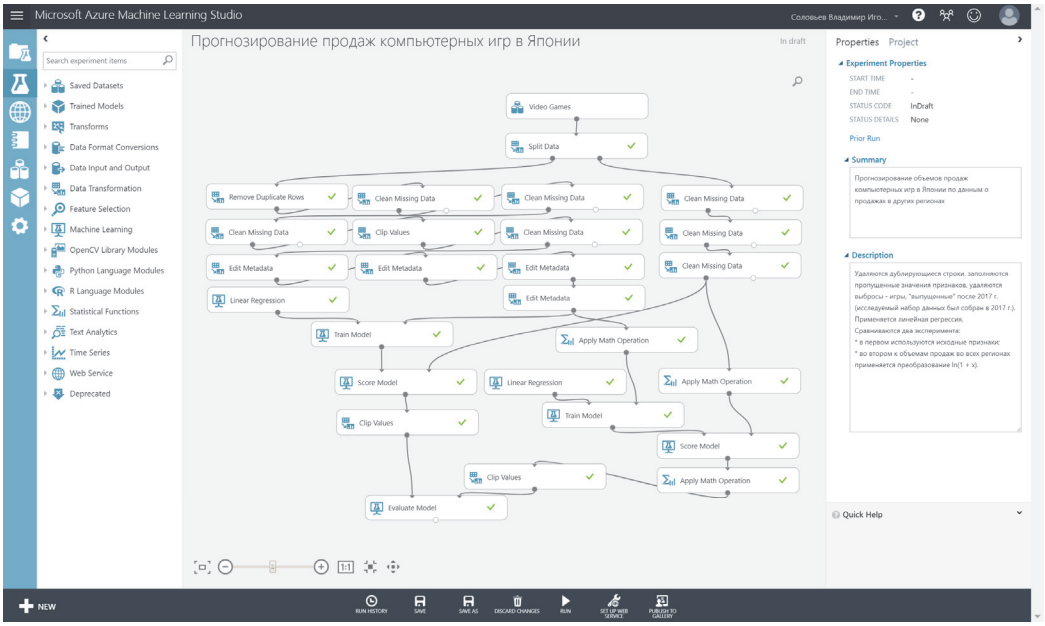


Рис. 7.4.86. *Добавленное резюме и описание эксперимента*

Данный эксперимент можно скопировать себе — он опубликован по адресу https://is.gd/findata_videogames. ❑

Замечание. Чтобы опубликовать эксперимент и иметь возможность поделиться им с другими, достаточно нажать кнопку «PUBLISH TO GALLERY».

На основе разработанной модели машинного обучения можно опубликовать веб-сервис, к которому дальше можно обращаться из различных приложений, с веб-сайтов или из таблиц Microsoft Excel.

Пример 7.4.6. Опубликуем веб-сервис, предсказывающий вероятность возврата кредита по анкетным данным заемщика и основанный на модели случайного леса, построенной в примере 7.4.4.

Для этого откроем эксперимент «Кредитный скоринг», запустим его, нажав кнопку «▶ RUN», затем выделим модуль «Train Model», соответствующий модели случайного леса,[1] нажмем кнопку «SET UP WEB SERVICE», далее выберем вариант «Predictive Web Service [Recommended]».

После этого будет создан предсказательный эксперимент, соответствующий исходному эксперименту (рис. 7.4.87).

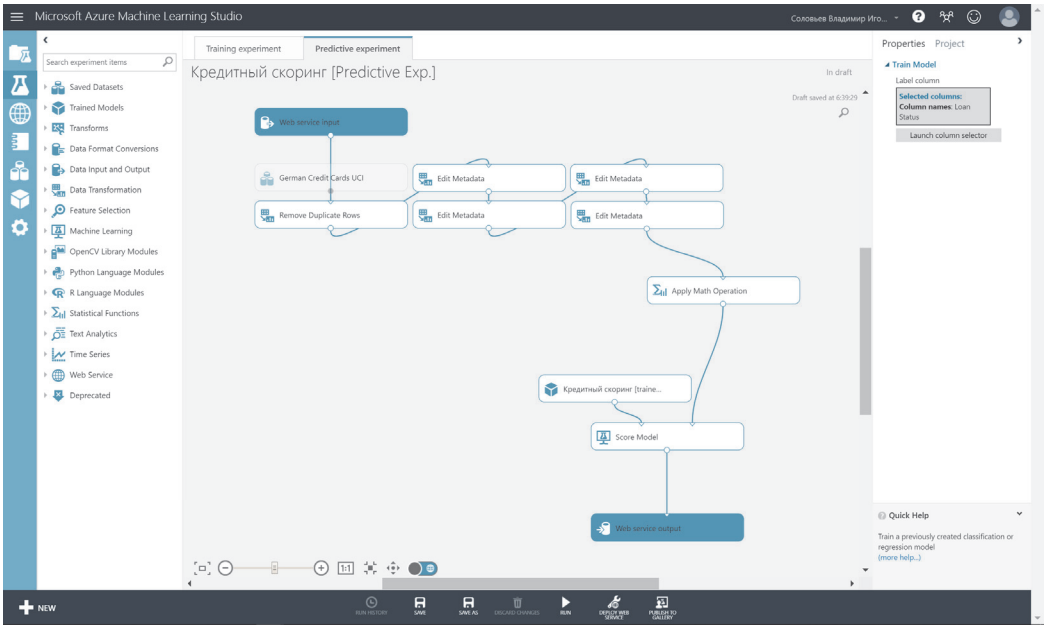


Рис. 7.4.87. *Предсказательный эксперимент*

На выходе модуля «Score Model» выдаются все исходные признаки, а также признаки «Scored Labels» и «Scored Probabilities» (для моделей классификации, как в рассматриваемом примере) или только «Scored Labels» (для моделей регрессии). При практическом использовании важен только ответ, т. е. только значение признаков «Scored Labels» и «Scored Probabilities».

Поэтому добавим в эксперимент модуль «Select Columns in Dataset» и выберем на панели его свойств два столбца: «Scored Labels» и «Scored Probabilities» (рис. 7.4.88).

Сохраним эксперимент, нажав кнопку «SAVE», и запустим его, нажав кнопку «▶ RUN».

Когда выполнение эксперимента закончится, нажмем кнопку «DEPLOY WEB SERVICE». Появится окно (рис. 7.4.89), в котором приводятся параметры опубликованного веб-сервиса. В частности, там есть ключ для доступа к веб-сервису из приложений и с сайтов (API key), возможность получить доступ к веб-сервису из таблиц Microsoft Excel, а кнопка «Test» дает возможность проверить работу веб-сервиса при помощи стандартной веб-формы (рис. 7.4.90).

В эту веб-форму можно ввести значения факторных признаков и получить предсказание вероятности возврата кредита заемщиком с такими значениями анкетных характеристик.

[1] Если в эксперименте несколько модулей «Train Model», нужно выбрать тот из них, на котором должен основываться публикуемый веб-сервис.

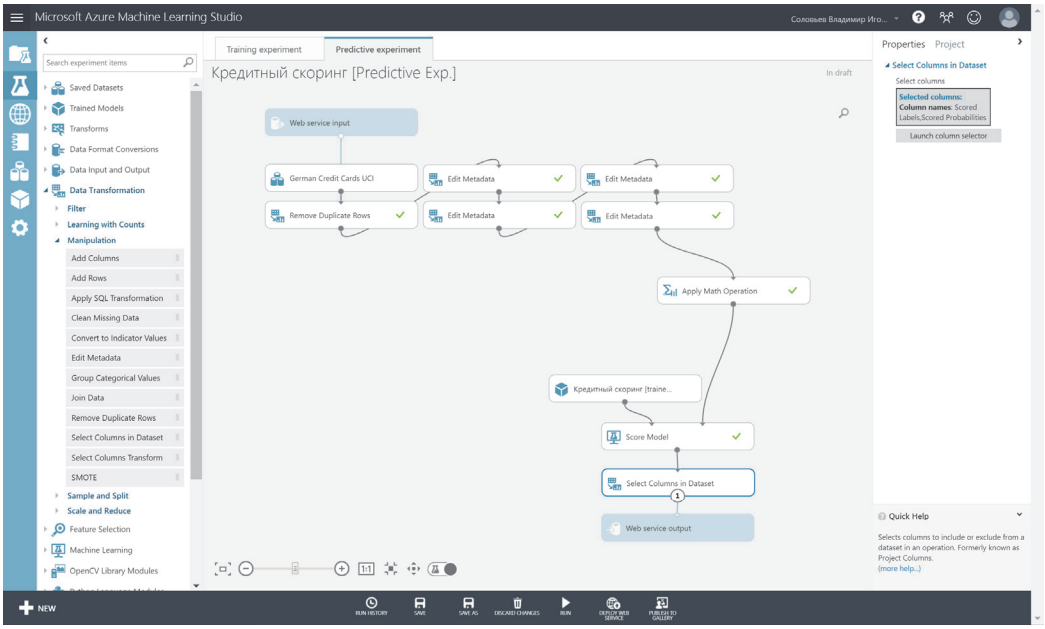


Рис. 7.4.88. *Предсказательный эксперимент, выдающий не полный набор признаков, а только предсказанный результат*

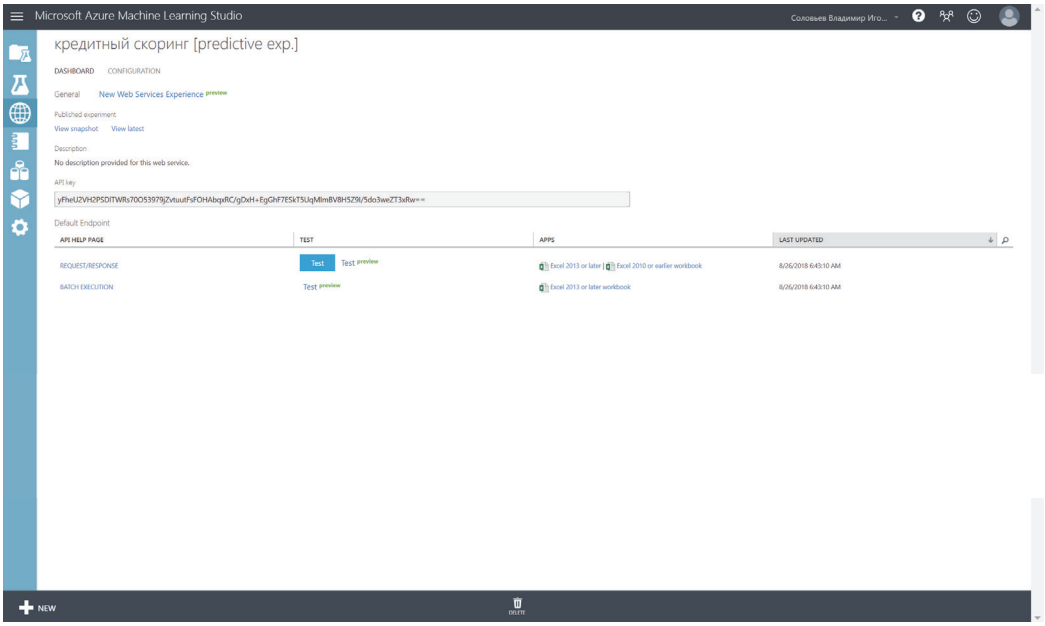


Рис. 7.4.89. *Параметры опубликованного веб-сервиса*

Кнопка «Sample Data > Enable» позволяет заполнить поля произвольными значениями, которые при желании можно скорректировать и нажать для получения прогноза кнопку «Test Request-Responce» (она расположена в самом низу).

В результате будет получен прогноз кредитного статуса заемщика, а также вероятности возврата кредита (рис. 7.4.91). ❑

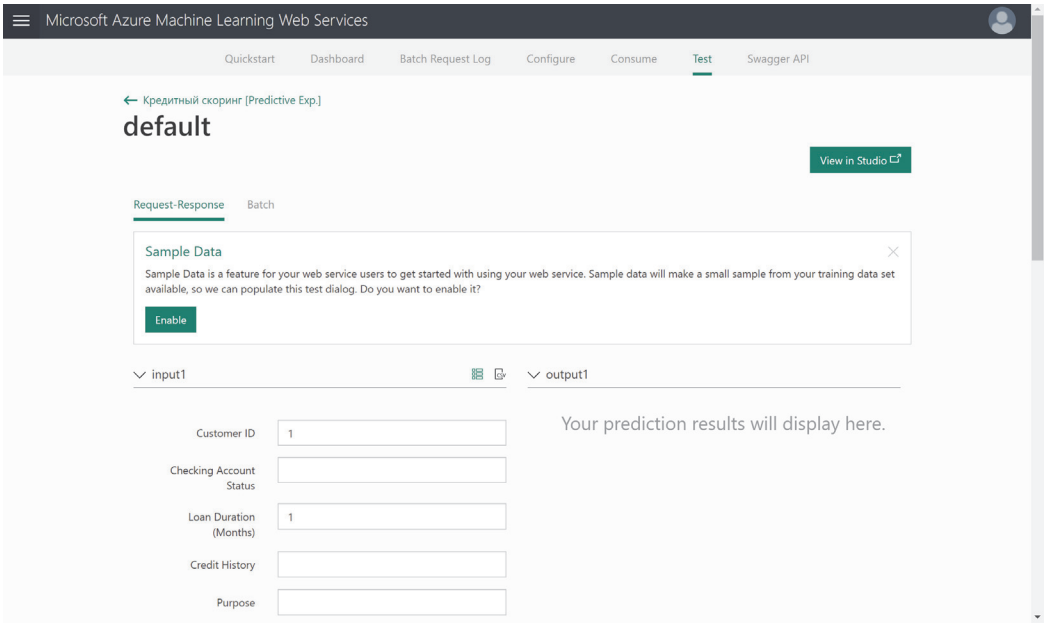


Рис. 7.4.90. *Стандартная веб-форма для проверки работы веб-сервиса*

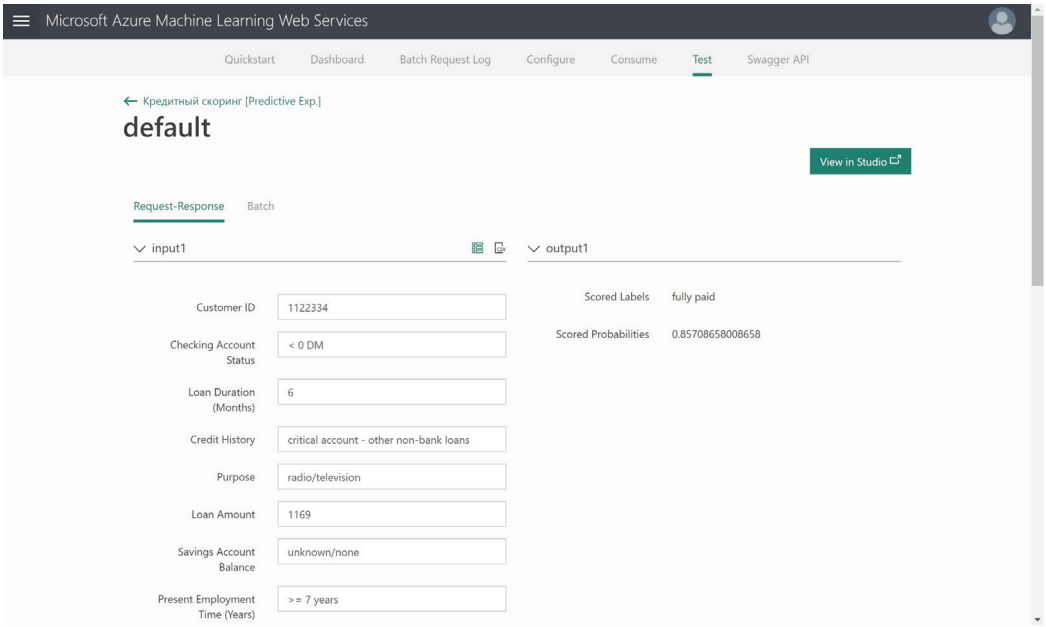


Рис. 7.4.91. *Результат работы веб-сервиса*

Задачи

451. Повторите самостоятельно все действия по созданию эксперимента в условиях примера 7.4.4.

452. Проанализируйте результаты обучения моделей в примере 7.4.4 и дайте ответ на вопрос, какие признаки в наибольшей степени влияют на вероятность возврата кредита

453. Вычислите с помощью Microsoft Excel значение показателя *AUC* по результатам примера 7.4.4.

454. Постройте с помощью Microsoft Excel кривую *Precision/Recall* по результатам примера 7.4.4.

455. Повторите самостоятельно все действия по созданию эксперимента в условиях примера 7.4.5.

456. Проанализируйте результаты обучения моделей в примере 7.4.4 и дайте ответ на вопрос, какие признаки в наибольшей степени влияют на объем продаж компьютерных игр в Японии.

457. В условиях примера 7.4.5 введите признак «Publisher» в состав факторных, предусмотрев, что в контрольной выборке он может не совпадать ни с одним из известных по обучающей выборке значений, а может быть неизвестен. Обучите модель линейной регрессии на новом наборе признаков. Сравните оценки качества моделей с результатами примера 7.4.5.

Заключение

Самые главные инвестиции — это инвестиции в человеческий капитал. Автор надеется, что изучение анализа данных стало серьезным вкладом в интеллектуальный капитал читателей.

Все больше интеллектуальных задач в ближайшее время можно будет передать машинам, все больше ценности можно будет извлечь с помощью описательной, предсказательной и предписывающей аналитики. И поэтому компетенции, связанные с применением искусственного интеллекта, обработки и интеллектуального анализа данных, машинного обучения в экономике и финансах станут еще более востребованными.

В седьмой главе мы очень коротко рассмотрели примеры алгоритмов и прикладных задач машинного обучения. Этот раздел искусственного интеллекта сейчас очень бурно развивается. Многие организации, не только ИТ-гиганты, с помощью интеллектуальных технологий начинают улучшать качество принятия решений и увеличивать масштаб задач. Сегодня уже стандарт, что решение об одобрении потребительского кредита банки принимают с помощью систем машинного обучения в течение нескольких минут, а не в течение нескольких дней, а то и недель, как раньше. Появляются новые сервисы, основанные на технологиях анализа данных и машинного обучения. Появляются новые модели бизнеса, основанные на технологиях искусственного интеллекта, открываются новые рынки.

Детальному изучению задач машинного обучения в экономике и финансах, их практической реализации в Microsoft Azure ML Studio, детальному разбору различных аспектов преобразования данных, разработки моделей машинного обучения, их тонкой настройки, проведения экспериментов и оценки качества результатов в Финансовом университете посвящен курс «Прикладные задачи машинного обучения».[1] При этом благодаря использованию инструмента Microsoft Azure ML Studio, позволяющего разрабатывать системы машинного обучения с помощью графического интерфейса, курс доступен всем людям, не имеющим специальной подготовки в области программирования. А модули, связанные с использованием языков Python, R, SQL, позволяют задействовать свой потенциал и тем, кто знаком с этими языками программирования.

Тем, кто интересуется созданием финтех-стартапов, основанных на интеллектуальных технологиях, рекомендуется к изучению курс «Финтех: Инфраструктура и модели бизнеса».[2] В этом курсе изучаются и информационные технологии, на основе которых сегодня трансформируется финансовый сектор,[3] и модели бизнеса в современном финтехе.

Тех же, кто хочет стать профессионалом в области анализа данных и машинного обучения, автор приглашает на магистерскую программу Финансового университета «Анализ данных и машинное обучение в экономике и финансах», руководителем которой он является, а также в аспирантуру департамента анализа данных, принятия решений и финансовых технологий Финуниверситета. Наши студенты не только глубоко погружаются в теорию искусственного интеллекта, но и участвуют в интересных проектах создания интеллектуальных систем для конкретных заказчиков.

---

[1] См. *Соловьев В.И.* Прикладные задачи машинного обучения в экономике и финансах: Практический курс с применением технологий *Microsoft Azure ML* и соревнованиями на платформе *Kaggle*. — М.: КНОРУС, 2018.

[2] См. *Соловьев В.И.* Финтех: Инфраструктура и модели бизнеса: Учебник. — М.: КНОРУС, 2018.

[3] В том числе технологии машинного обучения и технологии распределенных реестров.

Список использованной литературы

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика в задачах и упражнениях. — М.: ЮНИТИ-ДАНА, 2001.
2. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики: Т. 1. Теория вероятностей и прикладная статистика. — М.: ЮНИТИ-ДАНА, 2001.
3. *Беннинга Ш.* Финансовое моделирование с использованием Excel. — М.: Вильямс, 2017.
4. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983.
5. *Браилов А.В., Глебов В.И., Криволапов С.Я., Рябов П.Е.* Теория вероятностей и математическая статистика. — Ижевск: ИКИ, 2016.
6. *Бабешко Л.О., Бич М.Г., Орлова И.В.* Эконометрика и эконометрическое моделирование. — М.: Вузовский учебник : ИНФРА-М, 2018.
7. *Бывшев В.А.* Эконометрика. — М.: Финансы и статистика, 2008.
8. *Вадзинский Р.* Статистические вычисления в среде Excel. — СПб.: Питер, 2008.
9. *Ван дер Варден Б.Л.* Математическая статистика. — М.: Издательство иностранной литературы, 1960.
10. *Ватутин В.А., Ивченко Г.И., Медведев Ю.И., Чистяков В.П.* Теория вероятностей и математическая статистика в задачах. — М.: URSS, 2015.
11. *Вентцель Е.С.* Теория вероятностей. — М.: Юстиция, 2018.
12. *Вентцель Е.С., Овчаров Л.А.* Задачи и упражнения по теории вероятностей. — М.: Высшая школа, 2006.
13. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и ее инженерные приложения. — М.: Юстиция, 2018.
14. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. — М.: Большая Российская энциклопедия, 1999.
15. *Винстон У.* Бизнес-моделирование и анализ данных: Решение актуальных задач с помощью Microsoft Excel. — СПб.: Питер, 2018.
16. *Вуколов Э.А., Ефимов А.В., Земсков В.Н., Поспелов А.С.* Сборник задач по математике для втузов. Ч. 4. Теория вероятностей. Математическая статистика. — М.: Физматлит, 2004.
17. *Галочкин В.Т.* Эконометрика. — М.: Юрайт, 2017.
18. *Геворкян П.С., Потемкин А.В., Эйсымонт И.М.* Теория вероятностей и математическая статистика. — М.: Физматлит, 2016.
19. *Гихман И.И., Скороход А.В., Ядренко М.И.* Теория вероятностей и математическая статистика. — Киев: Вища школа, 1979.
20. *Гнеденко Б.В.* Беседы о математической статистике. – М.: URSS, 2017.
21. *Гнеденко Б.В.* Курс теории вероятностей. — М.: URSS, 2015.
22. *Гнеденко Б.В.* Очерк по истории теории вероятностей. — М.: URSS, 2013.
23. *Дороговцев А.Я., Сильвестров Д.С., Скороход А.В., Ядренко М.И.* Теория вероятностей: Сборник задач. — Киев: Вища школа, 1980.
24. *Зубков А.М., Севастьянов Б.А., Чистяков В.П.* Сборник задач по теории вероятностей. — СПб.: Лань, 2009.
25. *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику: Статистика знает всё. — М.: URSS, 2017.
26. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. — М.: URSS, 2018.
27. *Ивченко Г.И., Медведев Ю.И., Чистяков В.П.* Математическая статистика в задачах. — М.: URSS, 2015.
28. *Калинина В.Н.* Теория вероятностей и математическая статистика: Компьютерно-ориентированный курс. — М.: Юрайт, 2018.
29. *Калинина В.Н., Соловьев В.И.* Анализ данных: Компьютерный практикум. — М.: КНОРУС, 2017.
30. *Карандаев И.С., Малыхин В.И., Соловьев В.И.* Прикладная математика. — М.: ИНФРА-М, 2002.
31. *Кендалл М., Стюарт А.* Статистические выводы и связи. — М.: Наука, 1973.
32. *Кендалл М., Стюарт А.* Теория распределений. — М.: Наука, 1966.
33. *Колемаев В.А.* Эконометрика. — М.: ИНФРА-М, 2010.
34. *Колемаев В.А., Калинина В.Н.* Теория вероятностей и математическая статистика. — М.: КНОРУС, 2017.
35. *Колемаев В.А., Калинина В.Н., Соловьев В.И.* Математическая статистика в примерах и задачах. — М.: ГУУ, 2001.

36. *Колемаев В.А., Калинина В.Н., Соловьев В.И., Малыхин В.И., Курочкин А.П.* Теория вероятностей в примерах и задачах. — М.: ГУУ, 2001.
37. *Колемаев В.А., Соловьев В.И., Карандаев И.С., Малыхин В.И., Гатауллин Т.М., Ершов А.Т., Прохоров Ю.Г., Юнисов Х.Х.* Методы оптимальных решений: Практикум. — М.: КНОРУС, 2017.
38. *Колесников А.Н.* Теория вероятностей в финансах и страховании. — М.: Анкил, 2008.
39. *Колмогоров А.Н.* Основные понятия теории вероятностей. — М.: URSS, 2018.
40. *Королев В.Ю.* Теория вероятностей и математическая статистика. — М.: Проспект, 2008.
41. *Коршунов Д.А., Фосс С.Г., Эйсымонт И.М.* Сборник задач и упражнений по теории вероятностей. — СПб: Лань, 2004.
42. *Коршунов Д.А., Чернова Н.И.* Сборник задач и упражнений по математической статистике. — Новосибирск: Издательство Института математики. — 2004.
43. *Крамер Г.* Математические методы статистики. — М.: Мир, 1975.
44. *Лагутин М.Б.* Наглядная математическая статистика. — М.: Бином. Лаборатория знаний, 2018.
45. *Левин Д.М., Стефан Д., Кребиль Т.С., Беренсон М.Л.* Статистика для менеджеров с использованием Microsoft Excel. — М.: Вильямс, 2005.
46. *Лоэв М.* Теория вероятностей. — М.: Издательство иностранной литературы, 1962.
47. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика: Начальный курс. — М.: Дело, 2007.
48. *Малыхин В.И.* Высшая математика. — М.: ИНФРА—М, 2017.
49. *Мешалкин Л.Д.* Сборник задач по теории вероятностей. — М.: Издательство Московского университета, 1963.
50. *Миркин Б.Г.* Введение в анализ данных. — М.: Юрайт, 2017.
51. *Мхитарян В.С., Архипова М.Ю., Дуброва Т.А., Миронкина Ю.Н., Сиротин В.П.* Анализ данных. — М.: Юрайт, 2017.
52. *Прохоров А.В., Ушаков В.Г., Ушаков Н.Г.* Задачи по теории вероятностей: Основные понятия. Предельные теоремы. Случайные процессы. — М.: КДУ, 2009.
53. *Реньи А.* Трилогия о математике. — М.: Мир, 1980.
54. *Репин О.А., Суханова Е.И., Ширяева Л.К.* Задачи всероссийских студенческих олимпиад по теории вероятностей и математической статистике. — СПб.: Лань, 2011.
55. *Розанов Ю.А.* Теория вероятностей, математическая статистика и случайные процессы. — М.: Наука, 1989.
56. *Ротарь В.И.* Теория вероятностей. — М.: Высшая школа, 1992.
57. *Севастьянов Б.А.* Вероятностные модели. — М.: Наука, 1992.
58. *Севастьянов Б.А.* Курс теории вероятностей и математической статистики. — Ижевск.: Институт компьютерных исследований, 2004.
59. *Секей Г.* Парадоксы в теории вероятностей и математической статистике. — Ижевск.: Институт компьютерных исследований, 2004.
60. *Сигел Э.* Практическая бизнес-статистика. — М.: Вильямс, 2016.
61. *Соловьев В.И.* Прикладные задачи машинного обучения в экономике и финансах: Практический курс с применением технологий Microsoft Azure ML и соревнованиями на платформе Kaggle. — М.: КНОРУС, 2018.
62. *Соловьев В.И.* Финансовая математика. — М.: КНОРУС, 2016.
63. *Соловьев В.И.* Финтех: Инфраструктура и модели бизнеса. — М.: КНОРУС, 2018.
64. *Тутубалин В.Н.* Теория вероятностей. — М.: Academia, 2008.
65. *Уилкс С.* Математическая статистика. — М.: Наука, 1967.
66. *Феллер В.* Введение в теорию вероятностей и ее приложения: В 2-х тт. — М.: URSS, 2010.
67. *Форман Дж.* Много цифр: Анализ больших данных при помощи Excel. — М.: Альпина Паблишер, 2016.
68. *Чистяков В.П.* Курс теории вероятностей. — М.: URSS, 2015.
69. *Шеффе Г.* Дисперсионный анализ. — М.: Наука, 1980.
70. *Ширяев А.Н.* Вероятностно-статистические методы в теории принятия решений. — М.: МЦНМО, 2017.
71. *Ширяев А.Н.* Вероятность — 1. — М.: МЦНМО, 2017.
72. *Ширяев А.Н.* Вероятность — 2. — М.: МЦНМО, 2017.
73. *Winston W., Munson M.* Essential Statistics for Data Analysis using Excel: Массовый открытый онлайн-курс / Microsoft. — https://www.edx.org/course/essential-statistics-data-analysis-using-microsoft-dat222x-2

Приложения

Приложение 1. Материалы для контроля знаний

Расчетно-графическое задание

Каждый студент выполняет индивидуальный вариант расчетно-графического задания.

Начинать выполнение задания следует после того, как будет освоен материал п. 1.1.2.

Защиту заданий в форме беседы студента с преподавателем целесообразно проводить после того, как будет изучен п. 6.2.6.

1. В соответствии с вариантом, соответствующим номеру журнала в группе, соберите недельные данные о ценах закрытия и объемах торгов по трем акциям, обращающимся на Московской бирже, с 01.01.2010 г. по сегодняшний день. Названия акций в соответствии с вариантами представлены в табл. П.1.1.

Таблица П.1.1

Варианты акций для расчетно-графического задания

Вариант	Тикер 1	Тикер 2	Тикер 3	Вариант	Тикер 1	Тикер 2	Тикер 3
1	iИСКЧ ао	АВТОВАЗ ао	Акрон	36	МРСК СК	МРСК Центр	МРСК ЦП
2	АВТОВАЗ ап	Аптеки36и6	Арсагера	37	МРСК Ур	МРСК ЦП	МРСКВол
3	Акрон	Арсагера	АшинскийМЗ	38	МРСК Центр	МРСКВол	МРСКСиб
4	iИСКЧ ао	АшинскийМЗ	Аэрофлот	39	МРСК ЦП	МРСКСиб	МРСКЮга ао
5	Аптеки36и6	Аэрофлот	Белон ао	40	МРСКВол	МРСКЮга ао	НКНХ ао
6	АшинскийМЗ	Белон ао	БСП ао	41	МРСКСиб	НКНХ ао	НКНХ ап
7	Аэрофлот	БСП ао	Возрожд-ао	42	МРСКЮга ао	НКНХ ап	НЛМК ао
8	Аэрофлот	Белон ао	Возрожд-ао	43	НКНХ ао	НЛМК ао	НМТП ао
9	БСП ао	Возрожд-п	ВТБ ао	44	НКНХ ап	НМТП ао	Новатэк ао
10	Возрожд-ао	ВТБ ао	Газпрнефть	45	НЛМК ао	Новатэк ао	ОГК-2 ао
11	Возрожд-п	Газпрнефть	ГАЗПРОМ ао	46	НМТП ао	ОГК-2 ао	ОПИН ао
12	ВТБ ао	ГАЗПРОМ ао	ГМКНорНик	47	Новатэк ао	ОПИН ао	ПИК ао
13	Газпрнефть	ГМКНорНик	ДагСб ао	48	ОГК-2 ао	ПИК ао	Распадская
14	ГАЗПРОМ ао	ДагСб ао	ДВМП ао	49	ОПИН ао	Распадская	Росбанк ао
15	ГМКНорНик	ДВМП ао	Дорогбж ао	50	ПИК ао	Росбанк ао	РОСИНТЕРао
16	ДагСб ао	Дорогбж ао	ДЭК ао	51	Распадская	РОСИНТЕРао	Роснефть
17	ДВМП ао	ДЭК ао	ИРКУТ-3	52	Росбанк ао	Роснефть	РусГидро
18	Дорогбж ао	ИРКУТ-3	ИркЭнерго	53	РОСИНТЕРао	РусГидро	Сбербанк
19	ДЭК ао	ИркЭнерго	КАМАЗ	54	Роснефть	Сбербанк	Сбербанк-п
20	ИРКУТ-3	КАМАЗ	Кубанэнр	55	РусГидро	Сбербанк-п	СевСт-ао
21	ИркЭнерго	Кубанэнр	Лензол. ап	56	Сбербанк	СевСт-ао	Система ао
22	КАМАЗ	Лензол. ап	Лензолото	57	Сбербанк-п	Система ао	СОЛЛЕРС
23	Кубанэнр	Лензолото	Ленэнерго	58	СевСт-ао	СОЛЛЕРС	Сургнфгз
24	Лензол. ап	Ленэнерго	Ленэнерг-п	59	Система ао	Сургнфгз	Сургнфгз-п
25	Лензолото	Ленэнерг-п	ЛСР ао	60	СОЛЛЕРС	Сургнфгз-п	Татнфт 3ао
26	Ленэнерго	ЛСР ао	ЛУКОЙЛ	61	Сургнфгз	Татнфт 3ао	Татнфт 3ап
27	Ленэнерг-п	ЛУКОЙЛ	М.видео	62	Сургнфгз-п	Татнфт 3ап	Таттел. ао
28	ЛСР ао	М.видео	Магнит ао	63	Татнфт 3ао	Таттел. ао	ТГК-1
29	ЛУКОЙЛ	Магнит ао	Мечел ао	64	Татнфт 3ап	ТГК-1	ТГК-14
30	М.видео	Мечел ао	ММК	65	Таттел. ао	ТГК-14	ТГК-2
31	Магнит ао	ММК	МОЭСК	66	ТГК-1	ТГК-2	ТМК ао
32	Мечел ао	МОЭСК	МРСК СЗ	67	ТГК-14	ТМК ао	Транснф ап
33	ММК	МРСК СЗ	МРСК СК	68	ТГК-2	Транснф ап	Уркалий-ао
34	МОЭСК	МРСК СК	МРСК Ур	69	ТМК ао	Уркалий-ао	ФСК ЕЭС ао
35	МРСК СЗ	МРСК Ур	МРСК Центр	70	Транснф ап	ФСК ЕЭС ао	ЧеркизГ-ао

2. Рассчитайте для каждой акции недельные доходности и натуральные логарифмы объемов торгов.

3. Постройте гистограммы всех показателей и опишите их: являются ли они симметричными или скошенными, похожи ли они на «колокол» нормального распределения.

4. Проведите условное форматирование данных об акциях.

5. Постройте графики зависимости цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов от времени. Прослеживаются ли какие-нибудь зависимости данных признаков от времени?

6. Постройте диаграммы рассеяния для каждой пары признаков. Между какими признаками можно предположить наличие зависимости?

7. С помощью программы **«Описательная статистика»** определите основные статистические показатели для данных об акциях.

8. Постройте диаграммы размаха по данным о ценах, объемов торгов и доходностей акций. Есть ли в данных выбросы? Чем можно их объяснить? Проанализируйте выбросы отдельно, исключив их из общего набора данных.

9. Постройте интервальные вариационные ряды, гистограммы, полигоны частот и графики выборочных функций распределения для каждого из признаков: цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов.

10. В предположении нормального закона распределения цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов замените для каждого из этих признаков параметры нормального закона a и σ их несмещенными состоятельными оценками, рассчитайте значения функции плотности нормального закона в серединах интервалов и функции распределения в правых концах интервалов. Для каждого из анализируемых признаков постройте на одном рисунке полигон частот и график теоретической функции плотности распределения, а на другом — кумуляту и график теоретической функции распределения.

11. Постройте 95%-ные доверительные интервалы для математических ожиданий и средних квадратических отклонений цен, объемов торгов, логарифмов объемов торгов и доходностей акций.

12. Постройте 95%-ные доверительные интервалы для прогнозов на следующую неделю значений цен, объемов торгов, логарифмов объемов торгов и доходностей.

13. Для каждой акции рассчитайте оценку 95%-ной суммы под риском на ближайшую неделю для пакета акций, приобретаемого сегодня за 100 000 руб.

14. Для каждой акции проверьте на 5%-ном и 1%-ном уровне значимости гипотезу о том, что математическое ожидание недельной доходности равно нулю при альтернативной гипотезе о том, что оно больше нуля. В каждом случае вычислите также наблюдаемый уровень значимости.

15. Для каждой акции проверьте на 5%-ном уровне значимости гипотезы о нормальном законе распределения каждого из признаков: цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов.

16. Подготовьте и представьте в электронном и печатном виде отчет в формате Microsoft Word со следующей структурой.

Титульный лист
Содержание
Введение
Краткое описание эмитентов трех акций. Краткое описание результатов анализа, полученных в ходе выполнения расчетно-графического задания.

1. **Структура данных о торгах по акциям пяти эмитентов**
Описание структуры исходных данных о ценах закрытия, объемах торгов, недельных доходностях и натуральных логарифмах объемов торгов по пяти акциям. Обсуждение гистограмм, графиков зависимости показателей от времени, диаграмм размаха, выбросов, описательной статистики (для трех акций). Обсуждение диаграмм рассеяния (для трех пар акций).
2. **Интервальные оценки и прогнозы**
Обсуждение интервальных оценок математических ожиданий и средних квадратических отклонений цен, объемов торгов, логарифмов объемов торгов и доходностей акций, а также интервальных прогнозов значений цен, объемов торгов, логарифмов объемов торгов и доходностей акций на следующую неделю (для трех акций).
3. **Сумма под риском (VaR).**
Обсуждение результатов оценки 95%-ных сумм под риском на ближайшую неделю для пакетов акций (для трех акций).
4. **Доходность**
Обсуждение результатов проверки гипотез о том, что математическое ожидание недельной доходности равно нулю (для трех акций).
5. **Возможность применимости нормального закона распределения**
Сравнение полигонов частот с графиками теоретической функции плотности распределения. Сравнение кумулят с графиками теоретических функций распределения. Обсуждение результатов проверки гипотез о нормальном законе распределения каждого из признаков: цен закрытия, объемов торгов, доходностей и логарифмов объемов торгов (для трех акций).

Заключение
Общие выводы о структуре и динамике анализируемых показателей.
Приложение.
Данные о ценах закрытия, объемах торгов, недельных доходностях и натуральных логарифмах объемов торгов по пяти акциям (с условным форматированием).

Примеры экзаменационных заданий

Если дисциплина «Анализ данных» преподается в течение двух семестров, то в каждом семестре предусмотрен экзамен, проводимый в форме компьютерного тестирования.

Первый семестр

Каждый студент получает индивидуальный вариант задания. Максимальная возможная оценка — 60 баллов.

1. Студент пользуется тремя электронными библиотеками, комплектование которых осуществляется независимо друг от друга. Нужная ему книга может быть в данных библиотеках с вероятностями 0,29; 0,85 и 0,42 соответственно. Найдите вероятность того, что студент найдет нужную ему книгу, обратившись наугад в одну из этих библиотек. Введите ответ (3 балла).

2. Монета подбрасывается до тех пор, пока герб не выпадет 6 раз. Найдите вероятность того, что будет произведено 12 бросаний. Введите ответ (4 балла).

3. Дискретные случайные величины $X_1, X_2, \ldots, X_8$ распределены по одному и тому же закону:

X	-1	0	1
p	0,3	0,4	0,3

Найдите математическое ожидание $\mathbf{E}\left(X_1^2 + X_2^2 + \cdots X_8^2\right)$. Введите ответ (9 баллов).

4. Известна функция распределения случайной величины X:

$$F_X(x) = \begin{cases} 0, & x \leqslant -1, \\ \dfrac{-x^2 + 2x + 3}{4}, & -1 < x \leqslant 1, \\ 1, & x \geqslant 1. \end{cases}$$

Дайте ответы на следующие вопросы:

а) введите значение математического ожидания $\mathbf{E}(X)$ (3 балла);

б) введите значение дисперсии $\mathbf{Var}(X)$ (4 балла);

в) введите значение среднего квадратического отклонения σ_X (1 балл);

г) введите значение вероятности $\mathbf{P}\{0 < X < 2\}$ (3 балла).

5. Вероятность того, что стоимость портфеля инвестора в течение месяца вырастет на 10%, равна 0,6, вероятность того, что стоимость портфеля останется без изменений, равна 0,3, а вероятность того, что стоимость портфеля уменьшится в два раза, равна 0,1. Дайте ответы на следующие вопросы:

а) введите значение математического ожидания стоимости портфеля через три месяца (4 балла);

б) введите значение дисперсии стоимости портфеля через три месяца (6 баллов).

6. Случайная величина X распределена по нормальному закону с параметрами $a = 10$ и $\sigma = 3$. Дайте ответы на следующие вопросы:

а) введите значение вероятности $\mathbf{P}\{X = a\}$ (1 балл);

б) введите значение вероятности $\mathbf{P}\{X > a\}$ (1 балл);

в) введите значение вероятности $\mathbf{P}\{a - \sigma < X < a + \sigma\}$ (3 балла);

г) введите значение вероятности $\mathbf{P}\{a - 3\sigma < X < a + 3\sigma\}$ (3 балла);

д) введите значение вероятности $\mathbf{P}\{a - 6\sigma < X < a + 6\sigma\}$ (3 балла).

7. Случайная величина X распределена по логарифмически нормальному закону с параметрами $a = 10$ и $\sigma = 3$. Дайте ответы на следующие вопросы:

а) введите значение математического ожидания $\mathbf{E}(X)$ (3 балла);

б) введите значение дисперсии $\mathbf{Var}(X)$ (3 балла);

в) введите значение вероятности $\mathbf{P}\{a - \sigma < X < a + \sigma\}$ (3 балла);

Второй семестр

Каждый студент получает индивидуальный вариант задания. Максимальная возможная оценка — 60 баллов.

1. Известны следующие данные о доходности некоторой акции:

−0,210213; −0,116919; −0,020412; −0,141634; −0,035086; −0,003390; −0,233073; −0,159131;
−0,081404; 0,001276; −0,056943; 0,201194; 0,038972; −0,159985; 0,182039; 0,034253;
−0,033045; 0,058164; 0,247887; 0,178342; 0,087341; 0,011220; 0,046311; 0,142699;
0,029185; −0,048088; −0,022337; 0,003468; 0,254932; 0,084951; −0,087226; −0,013889;
−0,201966; 0,040950; 0,069952; 0,195232; 0,093880; −0,039968; 0,007320; 0,133994;
0,098998; −0,189503; −0,097347; 0,156260; −0,006332; −0,032823; −0,077507; −0,211080;
−0,152849; 0,150149; −0,030200; 0,156738; −0,090925; 0,073917; −0,011122; 0,021327;
0,028718; 0,132425; −0,006059; 0,014349; −0,033004; 0,108691; 0,232725; 0,473893;
0,008754; −0,179219; 0,162969; 0,093659; 0,215817; −0,066566; −0,127094; 0,175602;
0,003316; −0,033984; 0,261705; 0,141647; 0,096423; 0,118615; 0,064315; 0,045676;
0,027359; 0,163649; 0,001875; 0,131921; −0,121201; −0,161890; 0,044747; 0,174921;
0,837495; 0,119782; −0,117599; 0,158810; 0,164757; 0,186553; −0,039511; 0,209467;
0,097181; −0,284898; 0,047745; 0,088314; 0,138735; 0,068131; −0,215986; −0,290508;
−0,236615.

Сохраните этот набор данных в файл в формате csv в виде одной строки или одного столбца, откройте этот файл в Microsoft Excel и проведите дальнейший анализ, чтобы ответить на следующие вопросы.

1.1. Каков объем у данной выборки (1 балл)?

1.2. Чему равно выборочное среднее доходности акции (1 балл)?

1.3. Чему равна исправленная выборочная дисперсия доходности акции (1 балл)?

1.4. Чему равна выборочная дисперсия доходности акции (1 балл)?

1.5. Чему равно стандартное отклонение доходности акции (1 балл)?

1.6. Чему равна выборочная медиана доходности акции, рассчитанная с помощью функции МЕДИАНА (1 балл)?

1.7. Какое значение представляет собой выброс (если выброс не один, то каков самый большой по модулю выброс, 1 балл)?

1.8. Чему равно выборочное среднее доходности акции, рассчитанное после удаления выброса (1 балл)?

1.9. Чему равна исправленная выборочная дисперсия доходности акции, рассчитанная после удаления выброса (1 балл)?

1.10. Чему равна выборочная дисперсия доходности акции, рассчитанная после удаления выброса (1 балл)?

1.11. Чему равно стандартное отклонение доходности акции, рассчитанное после удаления выброса (1 балл)?

1.12. Чему равна выборочная медиана доходности акции, рассчитанная после удаления выброса с помощью функции МЕДИАНА (1 балл)?

1.13. Чему равна нижняя граница 95%-ного доверительного интервала для математического ожидания доходности акции (2 балла)?

1.14. Чему равна верхняя граница 95%-ного доверительного интервала для математического ожидания доходности акции (2 балла)?

1.15. Чему равна нижняя граница 95%-ного доверительного интервала для генеральной дисперсии доходности акции (2 балла)?

1.16. Чему равна верхняя граница 95%-ного доверительного интервала для генеральной дисперсии доходности акции (2 балла)?

1.17. Чему равна нижняя граница 95%-ного доверительного интервала для генерального среднего квадратического отклонения доходности акции (2 балла)?

1.18. Чему равна верхняя граница 95%-ного доверительного интервала для генерального среднего квадратического отклонения доходности акции (2 балла)?

1.19. Чему равна нижняя граница 95%-ного доверительного интервала для прогноза доходности акции на следующий день (2 балла)?

1.20. Чему равна верхняя граница 95%-ного доверительного интервала для прогноза доходности акции на следующий день (2 балла)?

1.21. Чему равна сумма под риском (VaR) для пакета из данных акций стоимостью 1 000 000 руб., рассчитанная в рублях с 95%-ной надежностью (2 балла)?

1.22. Вы хотите быть уверены на уровне значимости 0,05, что по приобретаемым Вами акциям средняя доходность инвестиций превышает 1%. Чему равно наблюдаемое значение статистики соответствующего критерия (2 балла)?

1.23. Чему равен наблюдаемый уровень значимости (p-значение) соответствующего критерия (2 балла)?

1.24. Введите 0, если по результатам проверки гипотезы нет оснований отвергнуть гипотезу H_0, и 1, если по результатам проверки гипотезы есть основания отвергнуть H_0 и принять гипотезу H_1 (2 балла).

1.25. *P*-значение — это (введите номер правильного варианта ответа, 2 балла):

а) вероятность ошибиться, отвергнув гипотезу о том, что по приобретаемым Вами акциям средняя доходность инвестиций превышает 1%;

б) вероятность оказаться правым, отвергнув гипотезу о том, что по приобретаемым Вами акциям средняя доходность инвестиций равна 1%;

в) вероятность ошибиться, отвергнув гипотезу о том, что по приобретаемым Вами акциям средняя доходность инвестиций равна 1%;

г) вероятность оказаться правым, приняв гипотезу о том, что по приобретаемым Вами акциям средняя доходность инвестиций равна 1%.

2. В некоторых городах банк активно рекламировал новый кредитный продукт ($A^{(1)}$), в некоторых городах давал минимальное количество рекламы ($A^{(2)}$), а в некоторых городах вовсе не рекламировал ($A^{(3)}$). В таблице приведены сведения о числе кредитов, выданных в отделениях банка, расположенных в городах этих трех типов.

Номер отделения	$A^{(1)}$	$A^{(2)}$	$A^{(3)}$
1	100	84	0
2	50	54	40
3	10	0	0
4	0	10	10
5	70	0	0
6	0	0	10

Проверьте гипотезу H_0 об отсутствии влияния уровня рекламы на число выданных кредитов. Дайте ответы на следующие вопросы:

а) введите 0, если по результатам проверки гипотезы нет оснований отвергнуть гипотезу H_0, и 1, если по результатам проверки гипотезы есть основания отвергнуть H_0 и принять гипотезу H_1 (5 баллов);

б) введите значение коэффициента детерминации (6 баллов).

3. Среди заемщиков некоторого банка 50% составляют люди в возрасте от 21 до 30 лет, 30% — люди в возрасте от 31 до 50 лет, оставшиеся 20% — люди старше 50 лет. В одном из отделений банка в определенный месяц заявлениия о предоставлении кредита подали 26 человек в возрасте 21—30 лет, 15 человек в возрасте 31—50 лет и 9 человек старше 50 лет. Проверьте на 5%-ном уровне значимости гипотезу H_0 о том, что отклонение процентного состава заемщиков от среднего по банку объясняется исключительно случайными факторами. Дайте ответы на следующие вопросы:

а) введите 0, если по результатам проверки гипотезы нет оснований отвергнуть гипотезу H_0, и 1, если по результатам проверки гипотезы есть основания отвергнуть H_0 и принять гипотезу H_1 (5 баллов);

б) введите *p*-значение (6 баллов).

Приложение 2. Законы распределения случайных величин

Таблица П.2.1

Законы распределений дискретных случайных величин, часто встречающиеся на практике

Название закона распределения	Краткое обозначение закона	Механизм формирования случайной величины	Функция вероятности, математическое ожидание и дисперсия, квантили
закон распределения индикатора события A	$X = I_A$	$I_A = 1$, если событие A наступило, и $I_A = 0$ — в противном случае	$\mathbf{P}\{I_A = x\} = \begin{cases} 1-\mathbf{P}(A), & x = 0, \\ \mathbf{P}(A), & x = 1, \end{cases}$ $\mathbf{E}(X) = \mathbf{P}(A);\ \mathbf{Var}(X) = \mathbf{P}(A)[1-\mathbf{P}(A)]$
биномиальный	$X = \mathrm{Bin}(n;\ p)$	X — число успехов в n испытаниях Бернулли с вероятностью p успеха в единичном испытании	$\mathbf{P}\{X = x\} = C_n^x p^x (1-p)^{n-x} =$ $= \text{БИНОМ.РАСП}(\langle x\rangle;\ \langle n\rangle;\ \langle p\rangle;\ \text{ЛОЖЬ})$, $\mathbf{P}\{X \leqslant x\} =$ $= \text{БИНОМ.РАСП}(\langle x\rangle;\ \langle n\rangle;\ \langle p\rangle;\ \text{ИСТИНА})$, $\mathbf{P}\{k_1 \leqslant X \leqslant k_2\} =$ $= \text{БИНОМ.РАСП.ДИАП}(\langle n\rangle;\ \langle p\rangle;\ \langle k_1\rangle;\ \langle k_2\rangle)$, $\mathbf{E}(X) = np;\ \mathbf{Var}(X) = np(1-p)$, $x_\alpha = \text{БИНОМ.ОБР}(\langle n\rangle;\ \langle p\rangle;\ \langle \alpha\rangle)$
геометрический	$X = \mathrm{Geom}(p)$	X — число испытаний Бернулли, которые придется произвести до первого успеха (иногда — номер испытания, на котором впервые произошел успех)	$\mathbf{P}\{X = x\} = p(1-p)^{x-1} =$ $= \text{ОТРБИНОМ.РАСП}(\langle x-1\rangle;\ 1;\ \langle p\rangle)$, $\mathbf{E}(X) = 1/p,\ \mathbf{Var}(X) = (1-p)/p^2$
гипергеометрический	$X =$ $= \mathrm{HGeom}(L;\ K;\ l)$	X — число изделий первого сорта среди l изделий, отобранных случайным образом из партии, состоящей из L изделий, K из которых первого сорта, а остальные $(L-K)$ второго сорта	$\mathbf{P}\{X = x\} = \dfrac{C_K^x C_{L-K}^{l-x}}{C_L^l} =$ $= \text{ГИПЕРГЕОМ.РАСП}(\langle x\rangle;\langle l\rangle;\langle K\rangle;\langle L\rangle;\text{ЛОЖЬ})$, $\mathbf{P}\{X \leqslant x\} =$ $= \text{ГИПЕРГЕОМ.РАСП}(\langle x\rangle;\langle l\rangle;\langle K\rangle;\langle L\rangle;\text{ИСТИНА})$, $\mathbf{E}(X) = l\dfrac{K}{L},\ \mathbf{Var}(X) = l\dfrac{K}{L}\left(1-\dfrac{K}{L}\right)\dfrac{1-l/L}{1-1/L}$
Пуассона	$X = \mathrm{Pois}(\lambda)$	**1.** X — число успехов в n испытаниях Бернулли с вероятностью p успеха в единичном испытании, причем $n \to \infty$, $np \to \lambda = \text{const}$; на практике данным распределением пользуются в случае, когда n велико (несколько десятков или более), а $\lambda = np < 10$	$\mathbf{P}\{X = x\} = \dfrac{\lambda^x}{x!}\mathrm{e}^{-\lambda} =$ $= \text{ПУАССОН.РАСП}(\langle x\rangle;\ \langle\lambda\rangle;\ \text{ЛОЖЬ})$, $\mathbf{P}\{X \leqslant x\} =$ $= \text{ПУАССОН.РАСП}(\langle x\rangle;\ \langle\lambda\rangle;\ \text{ИСТИНА})$ $\mathbf{E}(X) = \mathbf{Var}(X) = \lambda$
	$X = \mathrm{Pois}(\mu t)$	**2.** X — число наступлений события простейшего потока с интенсивностью μ за время t	

Таблица П.2.2

Законы распределений непрерывных случайных величин, часто встречающиеся на практике

Название закона распределения	**Краткое обозначение закона**	**Механизм формирования случайной величины**	**Функция распределения, плотность распределения, математическое ожидание, дисперсия, квантили**
равномерный	$X = \mathrm{Unif}(a; b)$	X — случайная величина, принимающая значения только из некоторого отрезка $[a;b]$, причем с содержательной точки зрения все значения внутри этого отрезка одинаково возможны	$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & x\in[a;b],\\ 1, & x>b,\end{cases} f(x)=\begin{cases}0, & x<a.\\ \dfrac{1}{b-a}, & a<x<b,\\ 0, & x>b,\end{cases}$ $\mathbf{E}(X)=(a+b)/2,\ \mathbf{Var}(X)=(b-a)^2/12$
показательный (экспоненциальный)	$X = \mathrm{Exp}(\mu)$	X — интервал времени между двумя последовательными наступлениями события в простейшем потоке с интенсивностью μ	$F_X(x)=\begin{cases}0, & x<0,\\ 1-\mathrm{e}^{-\mu x}, & x\geqslant 0,\end{cases} f(x)=\begin{cases}0, & x<0,\\ \mu\mathrm{e}^{-\mu x}, & x>0,\end{cases}$ $f(x)=$ ЭКСП.РАСП(<x>; <μ>; ЛОЖЬ), $F(x)=$ ЭКСП.РАСП(<x>; <μ>; ИСТИНА), $\mathbf{E}(X)=1/\mu,\ \mathbf{Var}(X)=1/\mu^2$
нормальный	$X = \mathcal{N}(a; \sigma)$	$X = X_1 + X_2 + \cdots + X_n$ при $n\to\infty$, где $X_1, X_2, \ldots, X_n$ — независимые в совокупности одинаково распределенные случайные величины (согласно центральной предельной теореме, см. **§ 4.4**); на практике этим распределение можно пользоваться, когда воздействие каждой из величин $X_1, X_2, \ldots, X_n$ равномерно, незначительно и равновероятно по знаку, а число n достаточно велико ($n \geqslant 60$)	$F(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{x}\mathrm{e}^{-\frac{(t-a)^2}{2\sigma^2}}dt=$ $=$ НОРМ.РАСП(<x>; <a>; <σ>; ИСТИНА), $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}}=$ НОРМ.РАСП(<x>; <a>; <σ>; ЛОЖЬ), $\mathbf{E}(X)=a,\ \mathbf{Var}(X)=\sigma^2,$ $u_\alpha=$ НОРМ.ОБР(<α>; <a>; <σ>), для стандартного нормального закона $\mathcal{N}(0;1)$ используются функции НОРМ.СТ. РАСП и НОРМ.СТ.ОБР
логарифмически нормальный	$X = \mathcal{LN}(a; \sigma)$	$X = \ln \mathcal{N}(a; \sigma)$	$F(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\ln x}\mathrm{e}^{-\frac{(t-a)^2}{2\sigma^2}}dt=F_{\mathcal{N}(a;\sigma)}(\ln x)=$ $=$ ЛОГНОРМ.РАСП(<x>; <a>; <σ>; ИСТИНА), $f(x)=\begin{cases}0, & x<0,\\ \dfrac{1}{\sigma x\sqrt{2\pi}}\mathrm{e}^{-\frac{(\ln x-a)^2}{2\sigma^2}}, & x>0\end{cases}=$ $=\begin{cases}0, & x<0,\\ \dfrac{1}{x}f_{\mathcal{N}(a;\sigma)}(\ln x), & x>0\end{cases}=$ $=$ ЛОГНОРМ.РАСП(<x>; <a>; <σ>; ЛОЖЬ), $\mathbf{E}(X)=\mathrm{e}^{a+\sigma^2/2},\ \mathbf{Var}(X)=\mathrm{e}^{2a+\sigma^2}(\mathrm{e}^{\sigma^2}-1),$ $u_\alpha=$ ЛОГНОРМ.ОБР(<α>; <a>; <σ>)

Таблица П.2.3

Законы распределений случайных величин, важных в математической статистике

Название закона распределения	**Краткое обозначение закона**	**Механизм формирования случайной величины**	**Функция распределения, плотность распределения, математическое ожидание, дисперсия, квантили**
«Хи квадрат» с k степенями свободы	$X = \chi^2_k$	$\chi^2(n) = Z_1^2 + Z_2^2 + \cdots + Z_k^2$, где Z_1, Z_2, ..., Z_n — независимые в совокупности случайные величины, распределенные по стандартному нормальному закону $\mathcal{N}(0; 1)$	$F(x) =$ ХИ2.РАСП(<x>; <k>; ИСТИНА), $\mathbf{E}\left(\chi^2_k\right) = k$, $\mathbf{Var}\left(\chi^2_k\right) = 2k$, $\chi^2_{k;\alpha}$ = ХИ2.ОБР(<α>; <k>)
Стьюдента с k степенями свободы	$X = T_k$	$T_k = \dfrac{Z}{\sqrt{\chi^2_k / k}}$, где $Z = \mathcal{N}(0; 1)$ и χ^2_k — независимые случайные величины	$F(x) =$ СТЬЮДЕНТ.РАСП(<x>; <k>; ИСТИНА), $t_{k;\alpha}$ = СТЬЮДЕНТ.ОБР(<α>; <k>), $\overline{t}_{k;\alpha} = t_{k;1-\alpha/2}$ = СТЬЮДЕНТ.ОБР.2Х(<α>; <k>)
Фишера с k_1 и k_2 степенями свободы	$X = F_{k_1;k_2}$	$F_{k_1;k_2} = \dfrac{\chi^2_{k_1} / k_1}{\chi^2_{k_1} / k_2}$, где $\chi^2_{k_1}$ и $\chi^2_{k_2}$ — независимые случайные величины	$F(x) =$ F.РАСП(<x>; <k_1>; <k_2>; ИСТИНА), $f_{k_1;k_2;\alpha} =$ F.ОБР(<α>; <$k1$>; <$k2$>)

Приложение 3. Интервальные оценки

Таблица П.3.1

Двусторонние интервальные оценки числовых характеристик и параметров случайных величин

Параметр	Предположения	Статистика, используемая для построения интервальной оценки	Точечная оценка (максимального правдоподобия)	Интервальная оценка
математическое ожидание $a = \mathbf{E}(X)$	σ^2 известно	$\dfrac{(\overline{X}-a)\sqrt{n}}{\sigma} \sim \mathcal{N}(0;1)$	$\overline{X}$	в предположении бесконечного объема генеральной совокупности $\mathbf{P}\left\{\overline{X}-z_{(1+\gamma)/2}\dfrac{\sigma}{\sqrt{n}} < a < \overline{X}+z_{(1+\gamma)/2}\dfrac{\sigma}{\sqrt{n}}\right\}=\gamma,$ при конечном объеме генеральной совокупности $\mathbf{P}\left\{\overline{X}-z_{(1+\gamma)/2}\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}} < a < \overline{X}+z_{(1+\gamma)/2}\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}\right\}=\gamma,$ $z_{(1+\gamma)/2} =$ НОРМ.СТ.ОБР((1 + γ)/2)
	σ^2 не известно	$\dfrac{(\overline{X}-a)\sqrt{n}}{s} \sim T_{n-1}$	$\overline{X}$	в предположении бесконечного объема генеральной совокупности $\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}\dfrac{s}{\sqrt{n}} < a < \overline{X}+\overline{t}_{n-1;1-\gamma}\dfrac{s}{\sqrt{n}}\right\}=\gamma,$ при конечном объеме генеральной совокупности $\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}\dfrac{s}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}} < a < \overline{X}+\overline{t}_{n-1;1-\gamma}\dfrac{s}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}\right\}=\gamma,$ $\overline{t}_{n-1;1-\gamma} =$ СТЬЮДЕНТ.ОБР.2Х(1 − γ; n − 1)
вероятность p успеха в серии из n испытаний Бернулли	n порядка нескольких десятков или более, $m \neq 0; 1; n-1; n$	$\dfrac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim \mathcal{N}(0;1)$	$\hat{p}=\dfrac{m}{n}$	$\mathbf{P}\left\{\hat{p}-z_{(1+\gamma)/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p}+z_{(1+\gamma)/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right\}=\gamma,$ $z_{(1+\gamma)/2} =$ НОРМ.СТ.ОБР((1 + γ)/2)
	$m = 0; 1; n-1; n$			при $m=0$ $\mathbf{P}\left\{0 < p < 1-(1-\gamma)^{1/n}\right\}=\gamma,$ при $m=1$ $\mathbf{P}\left\{1-\left(1-\dfrac{1-\gamma}{2}\right)^{1/n} < p < 1-\left(\dfrac{1-\gamma}{2}\right)^{1/n}\right\}=\gamma,$ при $m=n-1$ $\mathbf{P}\left\{\left(\dfrac{1-\gamma}{2}\right)^{1/n} < p < \left(1-\dfrac{1-\gamma}{2}\right)^{1/n}\right\}=\gamma,$ при $m=n$ $\mathbf{P}\left\{(1-\gamma)^{1/n} < p < 1\right\}=\gamma$
дисперсия $\sigma^2 = \mathbf{Var}(X)$	a не известно	$\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$	s^2	$\mathbf{P}\left\{\dfrac{(n-1)s^2}{\chi^2_{n-1;(1+\gamma)/2}} < \sigma^2 < \dfrac{(n-1)s^2}{\chi^2_{n-1;(1-\gamma)/2}}\right\}=\gamma,$ $\chi^2_{n-1;(1+\gamma)/2} =$ ХИ2.ОБР((1 + γ)/2; n − 1), $\chi^2_{n-1;(1-\gamma)/2} =$ ХИ2.ОБР((1 − γ)/2; n − 1)
прогноз следующего элемента выборки X_{n+1}	σ^2 не известно	$\dfrac{(X_{n+1}-\overline{X})\sqrt{n}}{s\sqrt{n+1}} \sim T_{n-1}$	$\overline{X}$	$\mathbf{P}\left\{\overline{X}-\overline{t}_{n-1;1-\gamma}s\sqrt{1+\dfrac{1}{n}} < X_{n+1} < \overline{X}+\overline{t}_{n-1;1-\gamma}s\sqrt{1+\dfrac{1}{n}}\right\}=\gamma$

Таблица П.3.2

Односторонние интервальные оценки числовых характеристик и параметров случайных величин

Тип доверительного интервала	**Доверительный интервал**
нижняя граница математического ожидания	$\mathbf{P}\left\{\mathbf{E}(X) > \bar{X} - t_{n-1;\,\gamma}\dfrac{s}{\sqrt{n}}\right\} = \gamma$
верхняя граница математического ожидания	$\mathbf{P}\left\{\mathbf{E}(X) < \bar{X} - t_{n-1;\,1-\gamma}\dfrac{s}{\sqrt{n}}\right\} = \gamma$
верхняя граница дисперсии	$\mathbf{P}\left\{\sigma^2 < \dfrac{(n-1)s^2}{\chi^2_{n-1;\,1-\gamma}}\right\} = \gamma$
нижняя граница прогноза следующего элемента выборки X_{n+1}	$\mathbf{P}\left\{X_{n+1} > \bar{X} - t_{n-1;\,\gamma}s\sqrt{1+\dfrac{1}{n}}\right\} = \gamma$
верхняя граница прогноза следующего элемента выборки X_{n+1}	$\mathbf{P}\left\{X_{n+1} < \bar{X} + t_{n-1;\,\gamma}s\sqrt{1+\dfrac{1}{n}}\right\} = \gamma$

Приложение 4. Проверка гипотез

Критерии проверки гипотез о числовых

Проверяемая гипотеза H_0	Предположения	Статистика критерия	Распределение статистики при выполнении H_0	Альтернативная гипотеза H_1	Критическая область (область принятия гипотезы H_1)
$a = a_0$	σ^2 известно	$Z = \frac{(\overline{X} - a_0)\sqrt{n}}{\sigma}$	$\mathcal{N}(0;1)$	$a = a_1 > a_0$	$Z > z_{1-\alpha}$, $z_{1-\alpha}$ = = НОРМ.СТ.ОБР(<1 – α>)
				$a = a_1 < a_0$	$Z < -z_{1-\alpha}$, $-z_{1-\alpha} = z_\alpha$ = = НОРМ.СТ.ОБР(<α>)
				$a = a_1 \neq a_0$	$\lvert Z\rvert < z_{1-\alpha/2}$, $z_{1-\alpha/2}$ = = НОРМ.СТ.ОБР(<1 – α/2>)
	σ^2 неизвестно	$T_{n-1} = \frac{(\overline{X} - a_0)\sqrt{n}}{s}$	T_{n-1}	$a = a_1 > a_0$	$T_{n-1} > t_{n-1;1-\alpha}$, $t_{n-1;1-\alpha}$ = = СТЬЮДЕНТ.ОБР(<1 – α>; <n – 1>)
				$a = a_1 < a_0$	$T_{n-1} < -t_{n-1;1-\alpha}$, $-t_{n-1;1-\alpha} = t_{n-1;\alpha}$ = = СТЬЮДЕНТ.ОБР(<α>; <n – 1>)
				$a = a_1 \neq a_0$	$\lvert T_{n-1}\rvert > t_{n-1;1-\alpha/2}$, $t_{n-1;1-\alpha/2}$ = = СТЬЮДЕНТ.ОБР(<1 – α/2>; <n – 1>)
$\sigma^2 = b_0$ $(b_0 > 0)$	a неизвестно	$\chi^2_{n-1} = \frac{(n-1)s^2}{b_0}$	χ^2_{n-1}	$\sigma^2 = b_1 > b_0$	$\chi^2_{n-1} > \chi^2_{n-1;1-\alpha}$, $\chi^2_{n-1;1-\alpha}$ = = ХИ2.ОБР(<1 – α>; <n – 1>)
				$\sigma^2 = b_1 < b_0$	$\chi^2_{n-1} < \chi^2_{n-1;\alpha}$, $\chi^2_{n-1;\alpha}$ = = ХИ2.ОБР(<α>; <n – 1>)
				$\sigma^2 = b_1 \neq b_0$	$\chi^2_{n-1} \in \left(-\infty; \chi^2_{n-1;\alpha/2}\right) \cup \left(\chi^2_{n-1;1-\alpha/2}; +\infty\right)$, $\chi^2_{n-1;\alpha/2}$ = ХИ2.ОБР(<α/2>; <n – 1>), $\chi^2_{n-1;1-\alpha/2}$ = ХИ2.ОБР(<1 – α/2>; <n – 1>)
$p = p_0$	n порядка нескольких десятков или более, $np_0 > 10$, $n(1-p_0) > 10$	$Z = \frac{(\hat{p} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$	$\mathcal{N}(0;1)$	$p = p_1 > p_0$	$Z > z_{1-\alpha}$, $z_{1-\alpha}$ = = НОРМ.СТ.ОБР(<1 – α>)
				$p = p_1 < p_0$	$Z < -z_{1-\alpha}$, $-z_{1-\alpha} = z_\alpha$ = = НОРМ.СТ.ОБР(<α>)
				$p = p_1 \neq p_0$	$\lvert Z\rvert < z_{1-\alpha/2}$, $z_{1-\alpha/2}$ = = НОРМ.СТ.ОБР(<1 – α/2>)

Таблица П.4.1

значениях параметров нормального распределения

Наблюдаемый уровень значимости (*p*-значение)	**Мощность критерия** $1-\beta$
1 – НОРМ.СТ.РАСП(<Z>; ИСТИНА)	1 – НОРМ.СТ.РАСП(<$(a_0-a_1)\sqrt{n}/\sigma+z_{1-\alpha}$>; ИСТИНА)
НОРМ.СТ.РАСП(<Z>; ИСТИНА)	1 – НОРМ.СТ.РАСП(<$(a_1-a_0)\sqrt{n}/\sigma+z_{1-\alpha}$>; ИСТИНА)
2*НОРМ.СТ.РАСП(ABS(<Z>); ИСТИНА)	1 – НОРМ.СТ.РАСП(<$(a_0-a_1)\sqrt{n}/\sigma+z_{1-\alpha/2}$>; ИСТИНА)+ + НОРМ.СТ.РАСП(<$(a_0-a_1)\sqrt{n}/\sigma-z_{1-\alpha/2}$>; ИСТИНА)
1 – СТЬЮДЕНТ.РАСП(<T_{n-1}>; <$n-1$>; ИСТИНА)	1 – СТЬЮДЕНТ.РАСП(<$(a_0-a_1)\sqrt{n}/s+t_{n-1;1-\alpha}$>; <$n-1$>; ИСТИНА)
СТЬЮДЕНТ.РАСП(<T_{n-1}>; <$n-1$>; ИСТИНА)	1 – СТЬЮДЕНТ.РАСП(<$(a_1-a_0)\sqrt{n}/s+t_{n-1;1-\alpha}$>; <$n-1$>; ИСТИНА)
СТЬЮДЕНТ.РАСП.2Х(ABS(<T_{n-1}>; <$n-1$>)	1 – СТЬЮДЕНТ.РАСП(<$(a_0-a_1)\sqrt{n}/s+t_{n-1;1-\alpha/2}$>; <$n-1$>; ИСТИНА)+ + СТЬЮДЕНТ.РАСП(<$(a_0-a_1)\sqrt{n}/s-t_{n-1;1-\alpha/2}$>; <$n-1$>; ИСТИНА)
ХИ2.РАСП.ПХ(<χ^2_{n-1}>; <$n-1$>; ИСТИНА) = = 1 – ХИ2.РАСП(<χ^2_{n-1}>; <$n-1$>; ИСТИНА)	1 – ХИ2.РАСП(<$\chi^2_{n-1;1-\alpha}\,b_0/b_1$>; <$n-1$>; ИСТИНА)
ХИ2.РАСП(<χ^2_{n-1}>; <$n-1$>; ИСТИНА)	ХИ2.РАСП(<$\chi^2_{n-1;\alpha}\,b_0/b_1$>; <$n-1$>; ИСТИНА)
2*МИН(ХИ2.РАСП.ПХ(<χ^2_{n-1}>; <$n-1$>; ИСТИНА); ХИ2.РАСП.ПХ(<χ^2_{n-1}>; <$n-1$>; ИСТИНА))	1 + ХИ2.РАСП(<$\chi^2_{n-1;1-\alpha/2}\,b_0/b_1$>; <$n-1$>; ИСТИНА) – – ХИ2.РАСП(<$\chi^2_{n-1;\alpha/2}\,b_0/b_1$>; <$n-1$>; ИСТИНА)
1 – НОРМ.СТ.РАСП(<Z>; ИСТИНА)	НОРМ.СТ.РАСП(<$\frac{(p_1-p_0)\sqrt{n}}{\sqrt{p_1(1-p_1)}}-z_{1-\alpha}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}$>; ИСТИНА)
НОРМ.СТ.РАСП(<Z>; ИСТИНА)	НОРМ.СТ.РАСП(<$\frac{(p_0-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}-z_{1-\alpha}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}$>; ИСТИНА)
2*НОРМ.СТ.РАСП(ABS(<Z>); ИСТИНА)	НОРМ.СТ.РАСП(<$\frac{(p_1-p_0)\sqrt{n}}{\sqrt{p_1(1-p_1)}}-z_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}$>; ИСТИНА)+ + НОРМ.СТ.РАСП(<$\frac{(p_0-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}-z_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}$>; ИСТИНА)

Таблица П.4.2

Критерии проверки гипотез о сравнении параметров двух случайных величин, распределенных по нормальному закону

Проверяемая гипотеза H_0	**Предположения**	**Статистика критерия**	**Распределение статистики при выполнении H_0**	**Альтернативная гипотеза H_1**	**Критическая область (область принятия H_1)**
$a_1 = a_2$	σ_1^2, σ_2^2 известны	$Z = \dfrac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\mathcal{N}(0;1)$	$a_1 > a_2$	$Z > z_{1-\alpha}$
				$a_1 < a_2$	$Z < -z_{1-\alpha}$
				$a_1 \neq a_2$	$\lvert Z\rvert < z_{1-\alpha/2}$
	σ_1^2, σ_2^2 не известны, но равны	$T_{n_1+n_2-2} = \dfrac{\overline{X^{(1)}} - \overline{X^{(2)}}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, где $s = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	$T_{n_1+n_2-2}$	$a_1 > a_2$	$T_{n_1+n_2-2} < t_{n_1+n_2-2;1-}$
				$a_1 < a_2$	$T_{n_1+n_2-2} > -t_{n_1+n_2-2;}$
				$a_1 \neq a_2$	$\lvert T_{n_1+n_2-2}\rvert > t_{n_1+n_2-2;1-}$
$\sigma_1^2 = \sigma_2^2$	a_1, a_2 не известны	$F_{n_1-1;n_2-1} = \dfrac{s_1^2}{s_2^2}$ $\left(s_1^2 > s_2^2\right)$	$F_{n_1-1;n_2-1}$	$\sigma_1^2 > \sigma_2^2$	$F_{n_1-1;\,n_2-1} > f_{n_1-1;\,n_2-1;\,1-\alpha}$
				$\sigma_1^2 \neq \sigma_2^2$	$F_{n_1-1;\,n_2-1} > f_{n_1-1;\,n_2-1;\,1-\alpha/2}$
$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_\nu^2$	$n_i > 3$, $i = 1,2,\ldots,\nu$	$\chi^2_{\nu-1} = \dfrac{\sum\limits_{i=1}^{\nu}(n_i-1)\ln\dfrac{s^2}{s_i^2}}{1 + \dfrac{1}{3(\nu-1)}\left(\sum\limits_{i=1}^{\nu}\dfrac{1}{n_i-1} - \dfrac{1}{\sum\limits_{i=1}^{\nu}(n_i-1)}\right)}$, где $s^2 = \dfrac{\sum\limits_{i=1}^{\nu}(n_i-1)s_i^2}{\sum\limits_{i=1}^{\nu}(n_i-1)}$	$\chi^2_{\nu-1}$	—	$\chi^2_{\nu-1} > \chi^2_{\nu-1;\,\alpha}$ (*критерий Бартлетта*)
$p_1 = p_2$	n_1, n_2 порядка нескольких десятков или более	$Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, где $\hat{p} = \dfrac{m_1 + m_2}{n_1 + n_2}$	$\mathcal{N}(0;1)$	$p_1 > p_2$	$Z > z_{1-\alpha}$
				$p_1 < p_2$	$Z < -z_{1-\alpha}$
				$p_1 \neq p_2$	$\lvert Z\rvert < z_{1-\alpha/2}$

Об авторе



Владимир Игоревич Соловьев

Руководитель Департамента анализа данных, принятия решений и финансовых технологий Финансового университета.

Доктор экономических наук, профессор, лауреат премии имени профессора Б.Л. Овсиевича, лауреат премии *CEEMAN Champion Award* Ассоциации развития менеджмента *CEEMAN*, трижды лауреат премии «Проект года» Официального сообщества ИТ-директоров России *GlobalCIO*.

С 1990 г. реализовал множество проектов в области анализа финансовых рынков, моделирования умных производств, промышленного, финансового и образовательного интернета вещей, других задач искусственного интеллекта и обработки данных для компаний из финансового и реального секторов экономики.

С 1997 г. преподает учебные дисциплины в области анализа данных, машинного обучения, теории вероятностей, математической и прикладной статистики, методов оптимизации, теории игр и исследования операций, финансовой математики и математической экономики студентам математических и экономических направлений подготовки, слушателям программ *MBA* и курсов повышения квалификации.

Подготовил двух кандидатов наук, опубликовал более 100 научных статей и более 30 статей в профессиональных ИТ-журналах.

Автор 25 книг, в том числе семи учебников, опубликованных в издательстве «КНОРУС»:

- Анализ данных в экономике: Теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel (2018);
- Анализ данных: Практикум (2016);
- Методы оптимальных решений: Практикум (2016);.
- Методы принятия управленческих решений: Количественный подход (2016);
- Прикладные задачи машинного обучения в экономике и финансах: Практический курс с применением технологий Microsoft Azure ML и соревнованиями на платформе Kaggle (2018);
- Финансовая математика (2016);
- Финтех: Инфраструктура и модели бизнеса (2018).

Обладает профессиональными статусами в области наук о данных:

- Charter Member, Microsoft Professional Program for Data Science;
- Microsoft Certified Solutions Expert (Data Management and Analytics);
- Microsoft Certified Trainer.

Выпускник факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова и Московской школы управления СКОЛКОВО.