maΣprofʃ.ru

Высшая математика – просто и доступно!

Математическая статистика. Начало

Практикум для начинающих

Настоящая книга *позволит вам в короткие сроки освоить азы математической статистики и научиться решать наиболее распространённые задачи стандартного вузовского курса.* ***Материал предназначен для*** *студентов-заочников и других читателей, которые хотят быстро освоить практику.*

Автор: Александр Емелин

Оглавление

1. Основы математической статистики

Есть правда, есть большая правда, а есть статистика на Матпрофи.ру!

1.1. Понятие предмета

Математическая статистика следует «вторым эшелоном» за теорией вероятностей, и это не случайность, а логическое продолжение. Отличие состоит в том, что теорвер даёт **теоретическую оценку** *случайным событиям*, а статистика работает с практическими или как говорят, ***эмпирическими*** *данными*, которые берутся непосредственно «из жизни».

Что изучает матстат?

Если кратко, то **математическая статистика изучает методы сбора и обработки *статистической информации* для получения научных и практических выводов.**

Статистическая – это та, которую можно выразить числами. Эта информация появляется в результате исследования **массовых** *(обычно)* **явлений**, которые носят случайный характер. Она может быть изначально числовой (например, длина чего-либо) или иметь качественную первооснову – «оцифровке» поддаётся даже доброта котиков.

Немедленный пример. Что главное орудие физика? Секундомер:

Пример 1

Студент Константин выполняет лабораторную работу по определению коэффициента вязкости жидкости методом Стокса.

…спокойствие, тут будет всего несколько чисел :)

Экспериментальная часть этой работы состоит в том, что в высокий цилиндрический сосуд с жидкостью сбрасывается достаточно маленький и тяжёлый шарик, после чего замеряется время его погружения.

Время погружения шарика зависит от множества случайных факторов: прямоты рук экспериментатора, погрешности измерения времени, хаотичного движения молекул жидкости и т.д., вплоть до влияния Луны. Поэтому эксперимент целесообразно провести 5-10 раз (как оно обычно и требуется).

Предположим, что в результате 5 опытов получены следующие результаты:

$t_1 = 6{,}9, \quad t_2 = 6{,}7, \quad t_3 = 7, \quad t_4 = 7{,}2, \quad t_5 = 6{,}8$ *(в секундах)*

Что произошло? Студент Костя собрал статистические данные. Они *эмпирические* (взяты непосредственно из опытов), носят случайный характер *(см. выше)*. И массовый. Ведь все однокурсники только и занимаются тем, что бросают в сосуды шарики, да и мало ли на Земле похожих шариков, которые тонут в похожей жидкости.

Ну а мы потихоньку погружаемся в терминологию:

Полученные экспериментальные значения называются ***вариантами***, а их совокупность – ***вариационным рядом***. Почему так? Потому что полученные значения *варьируются* под воздействием случайных факторов.

Справка: ***варианта*** *(существительное женского рода) – в статистике означает отдельно взятое эмпирическое значение.*

Далее. Далее Константин должен обработать полученные данные. Во-первых, посмотреть, а нет ли среди них *варианты*, которая сильно отличается от всех остальных? Наличие такого значения сигнализирует о том, что соответствующий опыт проведён неудачно и его следует исключить из рассмотрения.

Нет, все значения достаточно близкИ друг к другу, и теперь напрашивается вычислить среднюю величину – разделить сумму значений на их $n=5$ количество:

$$\bar{t}=\frac{\sum\limits_{i=1}^{5}t_i}{n}=\frac{t_1+t_2+t_3+t_4+t_5}{n}=\frac{6{,}9+6{,}7+7+7{,}2+6{,}8}{5}=\frac{34{,}6}{5}=6{,}92 \text{ секунды.}$$

Это значение называют ***простой средней*** или, как многие знают, ***средним арифметическим***. Его стандартно **обозначают** с чёрточкой наверху.

Справка на всякий случай: *математический значок* $\sum$ *означает суммирование, а переменная* i *играет роль «счётчика»; в данном случае* i *изменяется от 1 до 5.*

Если грызут сомнения на счёт точности, то лучше не полениться и провести 10 опытов, что, кстати, удобнее в плане вычислений (на 10 делить проще). И, разумеется, полученный результат будет надёжнее, чем в 1-м случае.

Всё. Статические данные обработаны, осталось сделать выводы. А именно, с помощью значения $\bar{t}$ вычислить коэффициент вязкости жидкости и ещё там вроде что-то, желающие могут найти эту лабу в Сети.

…Возможно, у вас возник вопрос, а почему я выбрал такой пример? Это немногое, что мне запомнилось из институтского курса физики :)

Пример 2

Студенческая группа сдала коллоквиум по матанализу со следующими результатами:

x_i	2	3	4	5
N_i	5	10	7	3

Требуется определить среднюю успеваемость группы

Сбором статистических данных здесь занимался преподаватель, и обратите внимание на их характер: они эмпирические, массовые *(громко, конечно, сказано, но таки массовые)* и отчасти случайные. Кому-то повезло с вопросом, кому-то нет, кто-то что-то вспомнил / забыл, списал, выпил, прогулял и так далее…, прямо какое-то броуновское движение студентов ☺

Как нетрудно понять, роль *вариант* x_i здесь играют полученные оценки, а N_i – это соответствующие ***частоты*** – количество студентов, которые получили ту или иную оценку. Подсчитаем общую численность группы:

$N = \sum_{i=1}^{4} N_i = N_1 + N_2 + N_3 + N_4 = 5 + 10 + 7 + 3 = 25$ человек и привыкаем к терминам: исследуемое множество называют ***статистической совокупностью***, а его численность – ***объёмом*** *совокупности*.

Теперь обратим внимание на следующую вещь: двоечников и отличников у нас мало, а нормальных студентов :) много. И возникает вопрос: как вычислить «справедливую» среднюю оценку по всей совокупности? Решение напрашивается – с помощью так называемой ***средневзвешенной средней***:

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i N_i}{\sum_{i=1}^{4} N_i} = \frac{x_1 N_1 + x_2 N_2 + x_3 N_3 + x_4 N_4}{N} = \frac{2 \cdot 5 + 3 \cdot 10 + 4 \cdot 7 + 5 \cdot 3}{25} =$$

$$= \frac{10 + 30 + 28 + 15}{25} = \frac{83}{25} = 3{,}32$$ – средняя успеваемость по группе.

…да, суровые у меня сегодня примеры :) **Давайте проанализируем их принципиальные отличия:**

1) В первом примере проводится статистическое исследование *количественной* величины (времени), а во втором «оцифровывается» и анализируется *качественный* признак (успеваемость).

2) В первой случае исследуемая величина *непрерывна*, и, строго говоря, все полученные значения **различны** (отличаются хоть какими-то миллисекундами). Во втором случае варианты *дискретны*, т.е. представляют собой отдельно взятые изолированные значения. Следует заметить, что они не обязаны быть целыми, так, например, можно ввести в рассмотрение оценки 2,5; 3,5 и 4,5. И у дискретной величины, как правило, есть *неоднократно* встречающиеся (одинаковые) варианты.

Ставлю важный подзаголовок и продолжаю:

1.2. Генеральная и выборочная совокупность

3) В первом примере речь идёт о ***выборке*** значений. Что это значит? Это значит, что шарик можно сбрасывать в воду гораздо бОльшее и теоретически вообще бесконечное количество раз. Таким образом, проведённые 5 опытов – есть, по сути, *выборка*, которую называют ***выборочной совокупностью***. При этом соответствующее среднее значение принято называть ***выборочной средней***.

Второй пример (с успеваемостью) отличен тем, что в нём исследуется ВСЯ совокупность, и поэтому её называют ***генеральной совокупностью***, а соответствующее среднее значение – ***генеральной средней***. Но такая ситуация редкость. Редко когда удаётся исследовать всю совокупность.

И сейчас мы подошли к краеугольному камню матстата:

1.3. Основной метод математической статистики

Задача

Федор пошёл на базу исследовать помидоры. Требуется определить среднюю массу помидора и среднюю долю первосортных помидоров.

Разбираемся в ситуации. Очевидно, что на базе находится очень и очень много помидоров, обозначим их общее количество через N. Это *генеральная совокупность объёма* N. Для того чтобы решить задачу, можно взвесить каждый овощ: $x_1, x_2, x_3, ..., x_N$ (в граммах, например) и вычислить *генеральную среднюю*:

$\overline{x}_{Г} = \dfrac{x_1 + x_2 + x_3 + ... + x_N}{N}$ – среднюю массу помидора.

Но это долго и трудоёмко, даже если Феде будут помогать все его однокурсники.

Поэтому для оценки параметров генеральной совокупности целесообразно использовать **выборочный метод**.

Его суть состоит в том, что из генеральной совокупности достаточно выбрать n объектов, которые хорошо характеризуют всю совокупность.

Это «хорошо» называют ***представительностью*** или, как говорят буржуи, ***репрезентативностью*** выборки.

Проговорим сиё модное слово вслух: ре-пре-зен-та-тив-ность.

…Молодцы! А то некоторым студентам из года в год слышится «презервативы» ☺ Радует, однако, что это не плохое слово :)

Что нужно для того, чтобы обеспечить репрезентативность?

Во-первых, выборка должна быть достаточно велика, помидоров так 500-1000 точно, что уже вполне по силам даже одному Феде.

Замечание: *в дальнейшем мы сформулируем более строгие статистические критерии на счёт оптимального объёма выборки.*

Во-вторых, отбор следует осуществлять *равномерно* – из каждого ящика.

В-третьих, отбор должен быть *случайным*. Для этого используются разные приёмы, самый простой из них – выбор «вслепую» из случайно выбранного места ящика, обязательно с разной глубины (а то мало ли что поставщик мог там спрятать).

…Да-да! Я буду обучать вас реальной статистике :)

И, в-четвёртых (а может быть, и, в-первых), есть и другие факторы. В частности, **важно знать**, а *однородна* ли генеральная совокупность? Так, если помидоры поступили от разных поставщиков, то каждую партию полезно исследовать по отдельности (сделать несколько независимых выборок).

Итак, пусть Фёдор по всем правилам выбрал n помидоров, и теперь дело за малым – взвесить каждый овощ: x_1, x_2, x_3, ..., x_n (граммы) и вычислить *выборочную среднюю*:

$\overline{x}_в = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$ – среднюю массу помидора в выборке.

При этом очевидно, что **чем больше** *объем n выборочной совокупности*, **тем полученное значение $\overline{x}_в$ будет точнее приближать** *генеральную среднюю* $\overline{x}_Г$.

Но фишка состоит в том, что если начать увеличивать выборку в два, три и бОльшее количество раз, то будут получаться выборочные средние, которые мало отличаются от уже рассчитанного значения $\overline{x}_в$. Вы спрОсите, как это установлено? Эмпирически, в результате огромного количества проведённых исследований. А затем данный факт был подтверждён и теоретически.

Таким образом, **нет никакого практического смысла тратить силы, время, нервы и деньги на исследование бОльшей выборки и тем более, всей генеральной совокупности**.

Вот оно как – в статистике есть и прямая экономическая выгода!

И технический момент, **обратите внимание на используемые буквы – они стандартны**. Вместо «иксов» иногда используют «игреки», а вместо «эн» – F и f. Иные буквы применяйте, только если их любит ваш преподаватель или они в вашей методичке.

Вторая часть задачи. Оценим вместе с Федей ***долю*** первосортных помидоров на базе. Для этого, разумеется, не нужно заново «шерстить» всю генеральную совокупность, анализируется та же самая выборка.

В отличие от первого пункта, здесь исследуется уже *качественный* признак, для которого, тем не менее, можно сформулировать чёткие критерии. Пусть первосортный помидор – это ~~чёрный, лысый~~ красный, спелый, без видимых дефектов, массой выше среднего. Совершенно понятно, что генеральная совокупность содержит K таких помидоров, и существует точное значение:

$\omega_Г = \frac{K}{N}$ – ***генеральная доля*** первосортных помидоров.

Однако по причине трудозатрат и нецелесообразности полного исследования, достаточно подсчитать количество k таких овощей в выборке и вычислить:

$\omega_в = \frac{k}{n}$ – ***выборочную долю***, которая будет близкА к истинному значению $\omega_Г$. Но это только, напомню, при условии грамотно организованной и проведённой выборки.

Доля, как вы догадываетесь, может принимать значения от 0 до 1, и нередко её домножают на 100, чтобы выразить этот показатель в процентах.

Готово.

Константин, Фёдор, спасибо за участие, а остальные, как в том анекдоте, поедут «на картошку» :) В качестве разминки предлагаю вам задачу с тремя пунктами различного уровня сложности:

Пример 3

а) Урожайность картофеля по трём областям за ** год составила 147, 145, 155 ц/га. Требуется вычислить среднюю урожайность.

Метрическая справка: *1 центнер = 100 кг, 1 тонна = 1000 кг;*
1 гектар (га) = 10000 квадратных метров;
показатель ц/га означает, сколько центнеров собрано в среднем с 1 гектара.

Вариация чуть сложнее:

б) Известны следующие данные по трём областям:

Область	Общая посевная площадь, тыс. га	Урожайность, ц/га
А	139,80	147
Б	102,34	145
В	63,29	155

…Вы думаете, тут исследована вся генеральная совокупность? Нет, эти циферки нарисовали чиновники для отчёта! – привыкайте к настоящей статистике:)))

Требуется вычислить среднюю урожайность.

И третий пункт, творческий:

в) Вычислить среднюю урожайность по следующим данным:

Область	Валовой сбор картофеля, тыс. тонн	Урожайность, ц/га
А	2055	147
Б	1484	145
В	981	155

«Валовой» – это значит, всего собрано по области.

ДУМАЕМ, ВНИКАЕМ и РАССУЖДАЕМ – принцип здесь точно такой же, как и при решении задач по теории вероятностей. И **не забываем приписывать к результатам размерность!** *(секунды, граммы и т.д., а в данном случае – ц/га)*. За сию небрежность вас накажут не только на физике ;)

Решения с пояснениями в конце книги.

В заключение вводной главы систематизируем самое важное:

Математическая статистика – это наука, изучающая методы сбора и обработки статистической информации для получения научных и практических выводов.

Основным методом математической статистики является **выборочный метод**, его суть состоит в исследовании представительной *выборочной совокупности* – для достоверной оценки совокупности *генеральной*. Данный метод экономит временнЫе, трудовые и материальные затраты, поскольку исследование всей совокупности зачастую затруднено либо невозможно.

Иногда матстат считают разделом математики. И это тоже правда! ☺ Желаю успехов в дальнейшем освоении курса! Вперёд без страха и сомнений:

2. Вариационные ряды

В самом широком смысле *вариационный ряд* – это множество значений статистической совокупности *(множество вариант x_i)*. Примеры нам встретились буквально на первых же страницах. Занёс преподаватель в журнал 25 оценок – получился вариационный ряд. Взвесил Фёдор помидоры и записал 500 чисел. Это вариационный ряд.

Статистические данные, полученные непосредственно в результате сбора информации, называют ***первичными***, и их очевидное неудобство состоит в хаотичности. Глаза разбегаются. Поэтому числа неплохо бы обработать: расположить по возрастанию, разбить на группы и тому подобное.

Одним из основных методов обработки данных является их ***группировка*** – разбиение стат. совокупности на группы по определённому признаку (или признакам). Учитель должен подсчитать количество «двоек», «троек», «четвёрок» и «пятёрок». Фёдор тоже должен выделить группы овощей по некому принципу. …Прикольно получилось:D

И в узком смысле под ***вариационным рядом*** понимают упорядоченную по возрастанию (как правило) статистическую совокупность, разбитую на группы. При этом все вариационные ряды можно разделить на два вида:

2.1. Дискретный вариационный ряд

Пусть количественная величина X может принимать лишь отдельные изолированные значения $x_1, x_2, x_3, ..., x_k$ (т.е. множество значений *дискретно* (прерывно)).

Дискретный вариационный ряд *(ДВР)* **– это упорядоченное по возрастанию дискретное множество** *вариант* $x_1, x_2, x_3, ..., x_k$ **и соответствующие им** *частоты* **либо *относительные частоты*.**

Частоты выборочной совокупности **обозначают** через $n_1, n_2, n_3, ..., n_k$, а частоты генеральной совокупности – через $N_1, N_2, N_3, ..., N_k$. Как вариант, вместо *частот* используют *относительные частоты*, **они рассчитывается по формулам:**

– для выборочной совокупности: $w_1 = \frac{n_1}{n}$, …, где $n = n_1 + n_2 + n_3 + ... + n_k$ – *объем* выборки,

– и буквы совокупности генеральной: $W_1 = \frac{N_1}{N}, ...$, где $N = N_1 + N_2 + N_3 + ... + N_k$ – её *объем*.

Легко видеть, что **сумма всех относительных частот совокупности равна единице**:

$$w_1 + w_2 + w_3 + ... + w_k = 1, \qquad W_1 + W_2 + W_3 + ... + W_k = 1.$$

Всегда проверяйте этот факт! Если в сумме не единица, то либо вы ошиблись, либо в условии задачи опечатка и решить её невозможно (корректным образом).

За примером ДВР далеко ходить не будем, студенческая группа объёма $N = 25$:

x_i	2	3	4	5
N_i	5	10	7	3

где *варианты* x_i – это упорядоченные по возрастанию оценки по матанализу (отдельные изолированные значения), а *частоты* N_i – это количество студентов, получивших ту или иную оценку. Упорядоченный вариационный ряд также называют ***статистическим распределением совокупности***. Ибо вот так вот распределены оценки.

Для разминки найдём относительные частоты:

$W_1 = ...$

и непременно проконтролируем, что:
$W_1 + W_2 + W_3 + W_4 = 0{,}2 + 0{,}4 + 0{,}28 + 0{,}12 = 1.$

Все вычисления обычно проводят на калькуляторе либо в Экселе (MS Excel), а результаты заносят в таблицу, при этом **в статистике данные чаще располагают не в строках, а в столбцах:**

x_i	N_i	W_i
2	5	0,2
3	10	0,4
4	7	0,28
5	3	0,12
$\sum$	25	1

Такое расположение обусловлено тем, что количество *вариант* может быть достаточно велико, и они просто не вместятся в стандартную строку. Не редкость, когда их 10-20, а бывает и 100-200. Но мы осилим любое количество!

Откуда берутся дискретные вариационные ряды? Они появляются в результате исследования *дискретной* характеристики статистической совокупности, причём, *варианты* ряда не отличаются большим разнообразием. Например, оценки (коих не так много) в примере выше. И чтобы составить вариационный ряд, ещё нужно потрудиться:

Пример 4

По результатам выборочного исследования рабочих цеха были установлены их квалификационные разряды: 4, 5, 6, 4, 4, 2, 3, 5, 4, 4, 5, 2, 3, 3, 4, 5, 5, 2, 3, 6, 5, 4, 6, 4, 3.

Требуется: составить *вариационный ряд* и построить ***полигон частот***; найти *относительные частоты* и построить ***эмпирическую функцию распределения***.

Решение: в условии прямо сказано о том, что перед нами выборка из *генеральной совокупности* (всех рабочих цеха), и первое, что логично сделать – подсчитать её *объем*, т.е. количество рабочих. Здесь это легко сделать устно, циферок в условии: $n = 25$.

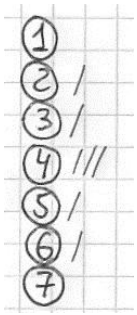
Квалификационные разряды – есть величина *дискретная* и этих разрядов немного. Поэтому нам предстоит составить **дискретный** вариационный ряд *(обратите внимание, что в условии ничего не сказано о характере ряда)*.

Как составить дискретный вариационный ряд?

Если у вас под рукой нет программ, то вручную. При этом оптимальным может быть следующий алгоритм: сначала окидываем взглядом все числа и определяем среди них минимальное (примерно) и максимальное (примерно). В данном случае ориентировочный диапазон – от 1 до 7. Записываем их в столбец на черновике и обводим в кружочки. Далее начинаем вычёркивать карандашом числа из исходного списка:

…

и делать засечки около соответствующих кружков:



После того, как все числа будут вычеркнуты, подсчитываем количество засечек в каждой строке:

…

Обязательно проверяем, получается ли у нас в сумме объём выборки: $3+5+8+6+3=25=n$ – отлично!

Заносим найденные значения в таблицу на чистовик:

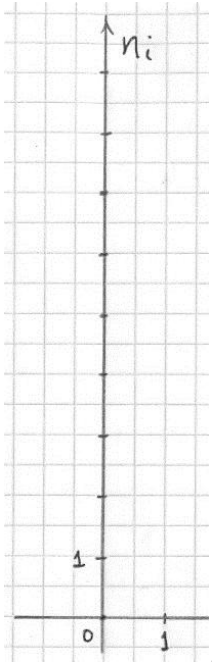
x_i	n_i
2	3
3	5
4	8
5	6
6	3
$\sum$	25

…ну что же, вполне и вполне логично – рабочих средней квалификации много, а учеников и мастеров – мало. Полученные результаты позволяются достаточно точно судить об уровне квалификации всего цеха (если, конечно, выборка представительна)

Дискретный вариационный ряд можно изобразить графически:

➢ Полигон распределения

Полигон частот **– это** *ломаная***, соединяющая соседние точки** …:



…

Кстати, с помощью полигона можно не только изобразить, но ещё и **однозначно задать дискретный вариационный ряд** (вместо таблицы с вариантами и частотами).

Теперь программный способ решения:

Задание

Самостоятельно решить данную задачу в Экселе (прямо в открывшемся файле).

Решаем! – все исходные данные с пошаговыми инструкциями прилагаются.

После чего переходим ко второй части задачи, в которой требуется найти относительные частоты и построить *эмпирическую функцию распределения*.

Относительные частоты рассчитываем по формуле $w_i = \frac{n_i}{n}$ – для этого каждую частоту n_i делим на *объём выборки* $n = 25$ и результаты заносим в дополнительный столбец, далее я перехожу к электронной версии оформления:

x_i	n_i	w_i
2	3	0,12
3	5	0,2
4	8	0,32
5	6	0,24
6	3	0,12
$\sum$	25	1

Обязательно проверяем, что сумма всех относительных частот равна единице: $0{,}12 + 0{,}2 + 0{,}32 + 0{,}24 + 0{,}12 = 1$, ОК.

Иногда требуется построить ***полигон относительных частот***. Как вы правильно догадались – это *ломаная*, соединяющая соседние точки (x_i, w_i). Но такое задание больше характерно для интервального вариационного ряда, до которого мы доберёмся в самом близком будущем.

А теперь посмотрим на *относительные частоты* w_i **и задумаемся**: на что они похожи? …Правильно, на вероятности. Так, например, можно сказать, что $w_3 = 0{,}32$ – есть *примерная* вероятность того, что наугад выбранный рабочий цеха будет иметь 4-й разряд. «Примерная» – по той причине, что перед нами выборка. А вот учесть ВСЕХ рабочих цеха (всю *генеральную совокупность*), то рассчитанные относительные частоты W_1, W_2, W_3, W_4, W_5 – в точности и есть эти вероятности.

Полигон относительных частот – это статистический аналог **многоугольника распределения** из теории вероятностей. Следует заметить, что он уже не задаёт вариационный ряд, так как относительные частоты $w_1, w_2, w_3, ..., w_k$ (сами по себе) ничего не говорят нам о частотах $n_1, n_2, n_3, ..., n_k$ и объеме совокупности.

Но не полигоном единым жив дискретный вариационный ряд, существует и другой подход к его заданию и изображению:

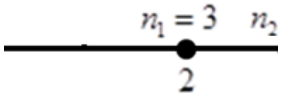
➢ Эмпирическая функция распределения

Это статистический аналог **функции распределения** из теорвера. Данная функция определяется, как отношение:

$F^*(x) = \frac{n_x}{n}$, где n_x – количество *вариант* СТРОГО МЕНЬШИХ, чем x,

при этом «икс» «пробегает» все значения от «минус» до «плюс» бесконечности.

Построим ***эмпирическую функцию распределения*** $F^*(x)$ для нашей задачи. Чтобы было нагляднее, отложу *варианты* x_i и их количество n_i на числовой оси:



…

На интервале $x \in (-\infty; 2)$: $F^*(x) = 0$ – по той причине, что левее ЛЮБОЙ точки этого интервала вариант x_i нет. Кроме того, функция равна нулю ещё и в точке $F^*(2) = 0$. Почему? Потому, что значение $F^*(2)$ определяет количество *вариант* (см. определение), которые СТРОГО меньше двух, а это количество равно нулю.

На промежутке $x \in (2; 3]$: $F^*(x) = \frac{n_x}{n} = \frac{n_1}{n} = w_1 = 0{,}12$ – и опять обратите внимание, что значение $F^*(3) = 0{,}12$ не учитывает рабочих 3-го разряда, т.к. речь идёт о вариантах, которые СТРОГО меньше трёх (по определению).

На промежутке $x \in (3; 4]$: $F^*(x) = \frac{n_x}{n} = ...$ – и далее процесс продолжается по принципу накопления частот:

– если $4 < x \le 5$, то $F^*(x) = \dfrac{n_1 + n_2 + n_3}{n} = w_1 + w_2 + w_3 = 0{,}12 + 0{,}2 + 0{,}32 = 0{,}64$;

– если $5 < x \le 6$, то $F^*(x) = w_1 + w_2 + w_3 + w_4 = 0{,}12 + 0{,}2 + 0{,}32 + 0{,}24 = 0{,}88$;

– и, наконец, если $6 < x < +\infty$, то $F^*(x) = w_1 + w_2 + w_3 + w_4 + w_5 = 1$ – и в самом деле, для ЛЮБОГО «икс» из интервала $(6; +\infty)$ ВСЕ частоты n_i расположены СТРОГО левее этого значения «икс» (см. чертёж выше).

Накопленные относительные частоты $w_н$ удобно заносить в отдельный столбец таблицы, при этом алгоритм вычислений очень прост: сначала сносим слева частоту w_1 *(красная стрелка)*, и каждое следующее значение $w_н$ получаем как сумму предыдущего и относительной частоты из текущего левого столбца *(зелёные обозначения)*:

x_i	n_i	
2	3	0
3	5	0
4	8	0
5	6	0
6	3	0
$\sum$	25	

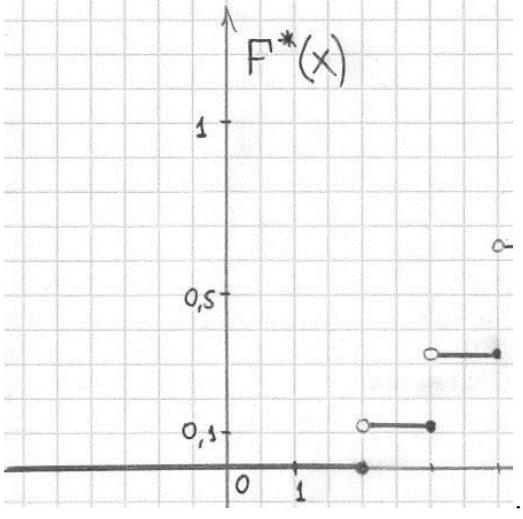
…

Вот ещё, кстати, один довод за вертикальную ориентацию данных – справа по надобности можно приписывать дополнительные столбцы.

Построенную функцию принято записывать в *кусочном* виде:

$$F^*(x) = \begin{cases} 0, & \text{если} \quad x \le 2 \\ 0{,}12, & \text{если} \quad 2 < x \le 3 \\ 0{,}32, & \text{если} \quad 3 < x \le 4 \\ 0{,}64, & \text{если} \quad 4 < x \le 5 \\ 0{,}88, & \text{если} \quad 5 < x \le 6 \\ 1, & \text{если} \quad x > 6 \end{cases}$$

а её график представляет собой ступенчатую фигуру:



…

Эмпирическая функция распределения *не убывает* и принимает значения лишь из промежутка $0 \le F^*(x) \le 1$, и если у вас вдруг получится что-то не так, то ищите ошибку.

Теперь смотрим видео, о том, **как построить эту функцию в Экселе** *(Ютуб)*.

И, конечно, вспомним основной метод математической статистики. *Эмпирическая функция распределения* $F^*(x)$ строится по выборке и приближает ***теоретическую функцию распределения*** $F(x)$. Легко догадаться, что последняя появляется в результате исследования всей *генеральной совокупности*, но если рабочих в цехе ещё пересчитать можно, то звёзды на небе – уже вряд ли. Вот поэтому и важнА функция *эмпирическая*, и ещё важнее, чтобы выборка была *репрезентативна*, дабы приближение было хорошим.

Миниатюрное задание для закрепления материала:

Пример 5

Дано статистическое распределение совокупности:

x_i	-2	1,5	5	7
n_i	12	8	20	10

Составить эмпирическую функцию распределения, выполнить чертёж

Решаем самостоятельно – все числа уже в Экселе! Свериться с образцом можно в конце книги. По поводу красоты чертежа сильно не запаривайтесь, главное, чтобы было правильно – этого обычно достаточно для зачёта.

2.2. Интервальный вариационный ряд

Предпосылкой построения ***интервального вариационного ряда*** (ИВР) является тот факт, что исследуемая величина X принимает слишком много различных значений x_i. Зачастую ИВР появляется в результате изучения *непрерывной* характеристики объектов. Типично – это время, масса, размеры и другие физические величины. Вспоминаем Константина, который замерял время на лабораторной работе и Фёдора, который взвешивал помидоры.

В таких ситуациях затруднительно либо невозможно применить тот же подход, что для дискретного ряда. Это связано с тем, что ВСЕ *варианты* x_i **различны** (во многих случаях). И даже если встречаются совпадающие значения, например, 50 грамм и 50 грамм, то связано это с округлением, а фактически значения всё равно отличаются хоть какими-то микрограммами.

Поэтому здесь используется **другой подход**, а именно определяется интервал, в пределах которого *варьируются* значения x_i, затем этот интервал делится на ***частичные интервалы*** (обычно равной длины h) и по каждому частичному интервалу подсчитываются *частоты* n_i *(либо* N_i*)* – количество *вариант*, которые в него попали. Если *варианта* попала на «стык» интервалов, то её относят к старшему интервалу.

Интервальный вариационный ряд *(ИВР) статистической совокупности* **– это упорядоченное множество смежных интервалов и соответствующие им** *частоты***, в сумме равные** *объёму* **совокупности.** Дабы не плодить лишних букв и индексов, я никак не обозначил эти интервалы. Придирчивый читатель, к слову, наверняка заметил, что через x_i я обозначаю как исходные варианты, так и значения сгруппированного ряда.

Следует отметить, что исследуемая характеристика не обязана быть непрерывной, и мы как раз начнём с такой задачи:

Пример 6

По результатам исследования цены некоторого товара в различных торговых точках города, получены следующие данные (в денежных единицах):

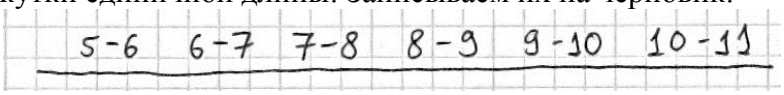
7,5	7,6	8,7
6,1	10,6	9,8
7	6	8,3
6	8,2	8,5
7,4	7,1	9,5
6,8	9,6	6,3
6,3	8,5	5,8
7,5	9,2	7,2
7	8	7,5
7,5	8	6,5

Составить вариационный ряд, построить ***гистограмму частот***, *гистограмму* и ***полигон относительных частот*** + бонус: ***эмпирическую функцию распределения***.

Решение: очевидно, что перед нами выборочная совокупность *объема* $n = 30$, и **вопрос номер один**: какой ряд составлять – дискретный или интервальный? Заметьте, что в вопросе задачи ничего не сказано о характере ряда. Строго говоря, цены дискретны и среди них даже есть одинаковые. Однако они могут быть округлены, да и разброс цен довольно велик. Поэтому здесь целесообразно провести интервальное разбиение.

Начнём с экстремальной ситуации, когда у вас под рукой нет Экселя или другого подходящего программного обеспечения. Только ручка, карандаш, тетрадь и калькулятор.

Тактика действий похожа на работу с дискретным вариационным рядом. Сначала окидываем взглядом предложенные числа и определяем примерный интервал, в который вписываются эти значения. «Навскидку» все значения заключены в пределах от 5 до 11. Далее делим этот интервал на удобные *подынтервалы*, в данном случае напрашиваются промежутки единичной длины. Записываем их на черновик:



Теперь начинаем вычёркивать числа из исходного списка и записываем их в соответствующие колонки нашей импровизированной таблицы:

…

После этого находим самое маленькое число в левой колонке *(минимальное значение)* и самое большое число – в правой *(максимальное значение)*. Тут даже ничего искать не пришлось, честное слово, не нарочно получилось:)

$x_{\min}=5{,}8,\quad x_{\max}=10{,}6$ ден. ед. – **не забываем указывать размерность!**

Вычислим ***размах вариации***:

$R=x_{\max}-x_{\min}=10{,}6-5{,}8=4{,}8$ ден. ед. – длина **общего интервала**, в пределах которого варьируется цена.

Теперь его нужно разбить на *частичные интервалы*. Сколько интервалов рассмотреть? По умолчанию на этот счёт существует ***формула Стерджеса***:

$k=1+3{,}322\lg n$, где $\lg n$ – десятичный логарифм* от объёма выборки и k – оптимальное количество интервалов, при этом результат округляют до ближайшего левого целого значения.

* *есть на любом более или менее приличном калькуляторе.*

В нашем случае получаем: $k=1+3{,}322\lg 30\approx 5{,}9\to 5$ интервалов.

Следует отметить, что *правило Стерджеса* носит рекомендательный, но не обязательный характер. Нередко в условии задачи прямо сказано, на какое количество интервалов следует проводить разбиение (на 4, 5, 6, 10 и т.д.), и тогда следует придерживаться именно этого указания.

Длины *частичных интервалов* могут быть различны, но в большинстве случаев использует ***равноинтервальную группировку***:

$h=\dfrac{x_{\max}-x_{\min}}{k}=\dfrac{4{,}8}{5}=0{,}96\approx 1$ – длина частичного интервала. В принципе, здесь можно было не округлять и использовать длину 0,96, но удобнее, ясен день, 1.

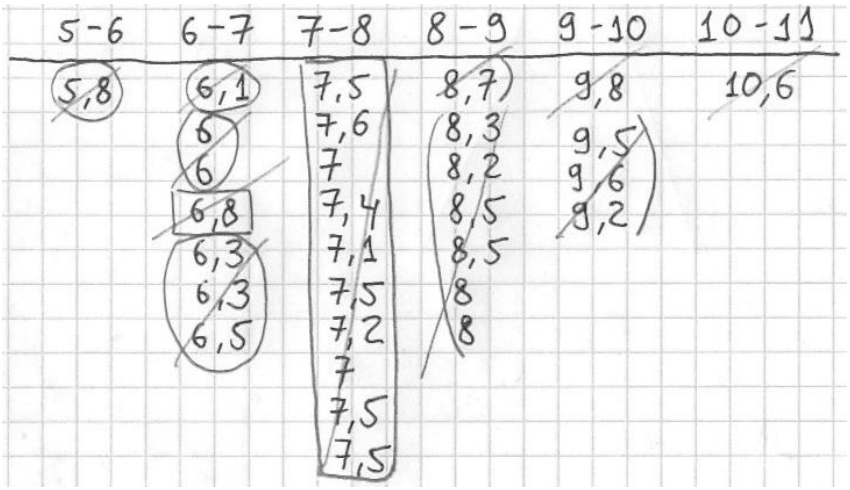
И коль скоро мы прибавили 0,04, то по пяти частичным интервалам получается «перебор»: $0{,}04\cdot 5=0{,}2$. Посему от самой малой варианты $x_{\min}=5{,}8$ отмеряем влево 0,1 влево *(половину «перебора»)* и к значению 5,7 начинаем прибавлять по $h=1$, получая тем самым частичные интервалы. При этом сразу рассчитываем их середины x_i (например, $x_1=\dfrac{5{,}7+6{,}7}{2}=6{,}2$) – они требуются почти во всех тематических задачах:

интервалы	
5,7 - 6,7	
6,7 - 7,7	
7,7 - 8,7	
8,7 - 9,7	
9,7 - 10,7	

…

– убеждаемся в том, что самая большая варианта $x_{\max}=10{,}6$ вписалась в последний частичный интервал и отстоит от его правого конца на 0,1.

Далее подсчитываем *частоты* по каждому интервалу. Для этого в черновой таблице обводим значения, попавшие в тот или иной интервал, подсчитываем их количество и вычёркиваем:



Так, значения из 1-го интервала я обвёл овалами (7 штук) и вычеркнул, значения из 2-го интервала – прямоугольниками (11 штук) и вычеркнул и так далее. Варианта $x = 8{,}7$ попала на «стык» интервалов и, согласно озвученному выше правилу, её следует отнести к последующему интервалу $(8{,}7;\ 9{,}7)$.

В результате получаем *интервальный вариационный ряд*:

интервалы:
5,7 – 6,7
6,7 – 7,7
7,7 – 8,7
8,7 – 9,7
9,7 – 10,7
суммы:

…

при этом **обязательно** убеждаемся в том, что ничего не потеряно:

$\sum_{i=1}^{k} n_i = 7 + 11 + 6 + 4 + 2 = 30 = n$, ОК.

…Да, кстати, все ли представили свой любимый товар, чтобы было интереснее разбирать это длинное решение? ☺

Точно также как и в дискретном случае, интервальный вариационный ряд можно (и нужно) изобразить графически. И здесь у нас весьма большое разнообразие. Но сначала добавим в таблицу дополнительные столбцы и продолжим расчёты:

По каждому интервалу рассчитываем (не тушуемся): ***плотность частот*** $\frac{n_i}{h}$,

относительные частоты $w_i = \frac{n_i}{n}$ (округляем их до 2 знаков после запятой), а также

плотность относительных частот $\frac{w_i}{h}$. Поскольку длина *частичного интервала* $h = 1$,

то вычисления заметно упрощаются:

интервалы	x
5,7 – 6,7	6,
6,7 – 7,7	7,
7,7 – 8,7	8,
8,7 – 9,7	9,
9,7 – 10,7	10,
суммы:	

…

Если интервалы имеют разные длины h_i, то при нахождении *плотностей* каждую частоту нужно разделить на длину **своего** интервала: $\frac{n_i}{h_i}$, $\frac{w_i}{h_i}$. Но у нас группировка *равноинтервальная*, да не абы какая, а с единичным *частичным интервалом*. Дело за чертежами. Один за другим:

➢ Гистограмма частот

– это фигура, состоящая из прямоугольников, ширина которых равна длинам *частичных интервалов*, а высота – соответствующим ***плотностям*** *частот*:

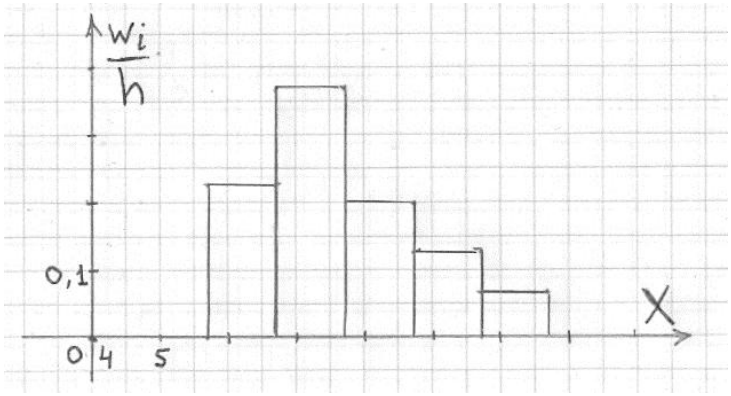
при этом вполне допустимо использовать нестандартную шкалу по оси абсцисс, в данном случае я начал нумерацию с четырёх.

Площадь *гистограммы частот* в точности равна объёму совокупности: …. В нашем случае $h = 1$ и *плотности* $\frac{n_i}{h}$ совпали с самими *частотами* n_i, таким образом:

$$1 \cdot \sum \frac{n_i}{1} = \sum n_i = n$$

➢ Гистограмма относительных частот

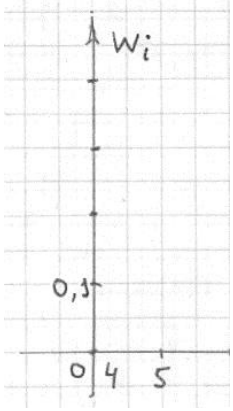
– это фигура, состоящая из прямоугольников, ширина которых равна длинам *частичных интервалов*, а высота – соответствующим ***плотностям*** *относительных частот*:



Площадь такой гистограммы равна единице: …, и это статистический аналог **функции плотности распределения** непрерывной случайной величины.

Построенный чертёж даёт наглядное и весьма точное представление о распределении цен на ботинки по всей генеральной совокупности. Но это при условии, что выборка представительна.

И для ИВР чаще всего требуется построить гистограмму именно относительных частот. А вместе с ней нередко и *полигон* таковых частот. Без проблем, *полигон относительных частот* **– это ломаная, соединяющая соседние точки …, где** x_i **– середины интервалов**:



По сути, **здесь мы приблизили** интервальный **ряд** дискретным, выбрав в качестве *вариант* x_i середины интервалов. **Это важнейший принцип и метод**, который неоднократно встретится нам в будущем.

Большим достоинством приведённого решения является тот факт, что многие вычисления здесь устные, а если вы помните, как делить «столбиком», то можно обойтись даже без калькулятора. Вот она где притаилась, смерть Терминатора :) ;)

Автоматизируем решение в Экселе *(видео на Ютуб)*.

И бонус:

➢ Эмпирическая функция распределения для ИВР

Она определяется точно так же, как в дискретном случае:

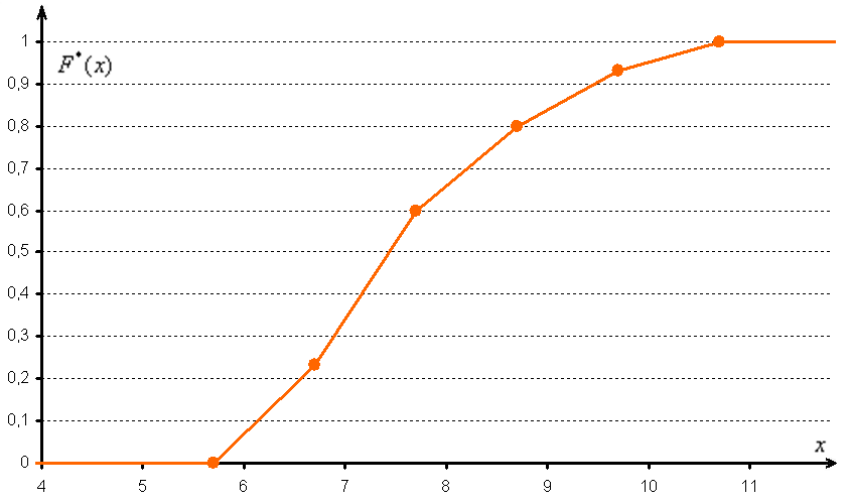
$F^*(x)=\frac{n_x}{n}$, где n_x – количество вариант СТРОГО МЕНЬШИХ, чем «икс», который «пробегает» все значения от «минус» до «плюс» бесконечности.

Но вот построить её для интервального ряда намного проще. Находим накопленные относительные частоты:

Интервалы		w_i
5,7	6,7	0,23
6,7	7,7	0,37
7,7	8,7	0,20
8,7	9,7	0,13
9,7	10,7	0,07
Суммы:		1

…

И строим *кусочно-ломаную* линию, с промежуточными точками $(x_{прав}, w_{нак})$, где $x_{прав}$ – правые концы интервалов, а $w_{нак}$ – относительная частота, которая успела накопиться на всех «пройденных» интервалах:



При этом $F^*(x)=0$ если $x\le 5{,}7$ и $F^*(x)=1$ если $x>10{,}7$.

Напоминаю, что данная функция *не убывает*, принимает значения из промежутка $0\le F^*(x)\le 1$ и, кроме того, для ИВР она ещё и *непрерывна*.

Эмпирическая функция является аналогом **функции распределения НСВ** и *приближает* теоретическую функцию $F(x)$, которую теоретически, а иногда и практически можно построить по всей генеральной совокупности.

Помимо перечисленных графиков, *вариационные ряды* также можно представить с помощью ***кумуляты*** и ***огивы*** *частот* либо *относительных частот*, но в классическом учебном курсе эта дичь редкая, и поэтому я не буду останавливаться на ней этой книге. Скажу только, что у вас вряд ли возникнут проблемы с их построением в случае такой необходимости.

Теперь что касаемо *объёма* выборки. Хорошо, если в вашей задаче всего лишь 20-30-50 *вариант*, но что делать, если их 100-200 и больше? В моей практике встречались десятки таких задач, и ручной подсчёт здесь уже не торт. Никаких проблем:

Как быстро составить ИВР при большом объёме выборки? *(Ютуб)*

Но не всё так сурово. Во многих задачах вам будет дан готовый вариационный ряд:

Пример 7

Выборочная проверка партии чая, поступившего в торговую сеть, дала следующие результаты:

Вес, грамм, x	47-49	49-50	50-51	51-53
Количество пачек, n_i	20	50	20	10

Требуется построить гистограмму и полигон относительных частот, эмпирическую функцию распределения

Проверяем свои навыки работы в Экселе! (исходные числа и краткая инструкция прилагается) И на всякий случай краткое решение для сверки есть в конце книги.

Иногда встречаются ИВР с открытыми крайними интервалами, например:

Суточный пробег автомобиля, км	Число автомобилей
до 160	12
от 160 до 180	36
от 180 до 200	28
свыше 200	24
Итого	**100**

В таких случаях интервалы «закрывают». Обычно поступают так: сначала смотрим на средние интервалы и выясняем длину *частичного интервала*: $h = 20$ км. И для дальнейшего решения можно считать, что крайние интервалы имеют такую же длину: от 140 до 160 и от 200 до 220 км. Соответственно, середины интервалов: 150 и 210 км.

И **самое важное по главе**, **обязательно прочитайте**, тут немного:)

Вариационный ряд – это множество значений *статистической совокупности*.

В узком смысле под ним понимают упорядоченные по возрастанию варианты, разбитые на группы, при этом возможны два случая:

Дискретный вариационный ряд *(ДВР)* – это упорядоченное дискретное множество *вариант* $x_1, x_2, x_3, ..., x_k$ и соответствующие им *частоты*.

Интервальный вариационный ряд *(ИВР)* – это упорядоченное множество смежных интервалов, разбивающее варианты на группы, и соответствующие им частоты.

ДВР чаще изображают с помощью *полигона частот*, а ИВР – *гистограммой относительных частот*. В обоих случаях определена *функция распределения*.

Упорядоченный вариационный ряд также называют *статистическим распределением совокупности*. Попросту говоря, он показывает, как распределены числа.

ДВР и ИВР появляются в результате группировки *первичных данных*, для более удобного их представления – в целях дальнейшего исследования.

3. Основные показатели статистической совокупности

Итак, в нашем распоряжении есть *первичные* статистические данные, собранные непосильным трудом. Чаще всего это выборка *объёма* n. Но бывает и генеральная совокупность *объёма* N. Кстати, сам по себе ***объём*** **– это элементарный и важнейший показатель статистической совокупности.**

Что дальше? Дальше нам нужно **исследовать эту статистическую совокупность и сделать выводы**. Во многих случаях целесообразно составить дискретный либо интервальный вариационный ряд**,** в чём мы только что потренировались**.** Но вообще, это не обязательная опция. Так, если чисел 5-10, то чего тут составлять?! И даже если *вариант* много, то они поддаются обработке и в *несгруппированном* (исходном) виде. Не особо удобно, конечно, однако… прочь лирику!

Так или иначе, для статистической совокупности можно рассчитать её ключевые показатели, среди которых выделяют две большие группы:

3.1. Показатели центральной тенденции

Простейший пример такого показателя нам уже встречался – это *среднее арифметическое* значение. Но *средней* дело не ограничивается, впрочем, обо всём по порядку:

➢ Генеральная и выборочная средняя

Пусть исследуется некоторая *генеральная совокупность* объёма N, а именно её числовая характеристика X, не важно, *дискретная* или *непрерывная*.

Генеральной средней называют *среднее арифметическое* всех значений этой совокупности:

$$\overline{x}_{Г} = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{\sum\limits_{i=1}^{N} x_i}{N}$$

Если среди чисел x_i есть одинаковые *(что характерно для дискретного ряда)*, то формулу можно записать в более компактном виде:

…, где:
варианта x_1 повторяется N_1 раз;
варианта x_2 – N_2 раз;
варианта x_3 – N_3 раз;
…
варианта x_K – N_K раз.

Живой пример вычисления *генеральной средней* встретился в Примере 2, но чтобы не занудничать, я даже не буду напоминать его содержание. Далее:

Как мы помним, обработка всей генеральной совокупности часто затруднена либо невозможна, и поэтому из неё организуют представительную выборку *объема* n, и на основании исследования этой выборки делают вывод обо всей совокупности.

Выборочной средней называется *среднее арифметическое* всех значений выборки:

$$\overline{x}_{в} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

и при наличии одинаковых *вариант* формула запишется компактнее:

…– как сумма произведений вариант x_i на соответствующие *частоты* n_i, делённая на объём совокупности n.

Выборочная средняя $\overline{x}_{в}$ позволяет достаточно точно оценить истинное значение $\overline{x}_{Г}$, при этом, чем больше выборка, тем точнее будет эта оценка.

Практику начнём с дискретного вариационного ряда и знакомого условия:

Пример 8

По результатам выборочного исследования $n = 25$ рабочих цеха были установлены их квалификационные разряды: 4, 5, 6, 4, 4, 2, 3, 5, 4, 4, 5, 2, 3, 3, 4, 5, 5, 2, 3, 6, 5, 4, 6, 4, 3.

Это числа из Примера 4, но теперь нам требуется: вычислить *выборочную среднюю*, и, не отходя от станка, найти ***моду*** и ***медиану***.

Как **решать** задачу? Если нам даны *первичные данные* (конкретные варианты x_i), то их можно тупо просуммировать и разделить результат на объём выборки:

$$\overline{x}_{в} = \frac{4+5+6+...+4+3}{25} = \frac{101}{25} = 4{,}04 \approx 4$$ – средний квалификационный разряд рабочих цеха.

Но здесь удобнее составить вариационный ряд:

x_i	n_i
2	3
3	5
4	8
5	6
6	3
$\sum$	25

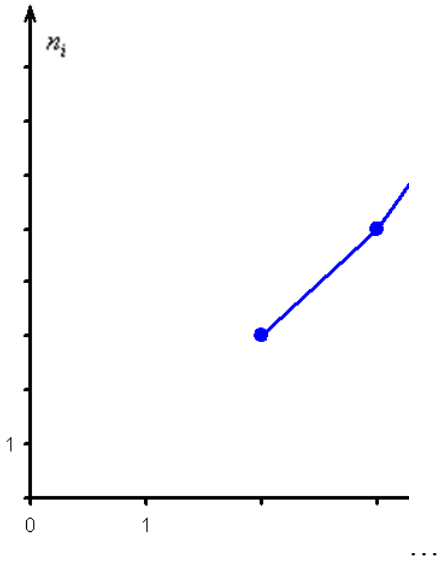
и использовать «цивилизованную» формулу:

$$\overline{x}_{в} = \frac{\sum\limits_{i=1}^{k} x_i n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + x_5 n_5}{n} = \frac{2 \cdot 3 + 3 \cdot 5 + 4 \cdot 8 + 5 \cdot 6 + 6 \cdot 3}{25} =$$

$$= \frac{6 + 15 + 32 + 30 + 18}{25} = \frac{101}{25} = 4{,}04$$

➢ Мода

Мода M_0 *дискретного вариационного ряда* **– это** *варианта* **с максимальной частотой**. Этот показатель определён как для выборочной, так и для генеральной совокупности. В нашей задаче $M_0 = x_3 = 4$. Моду легко отыскать по таблице и ещё легче на полигоне частот – это *абсцисса* самой высокой точки:



…

Иногда таковых значений несколько (с одинаковой максимальной частотой), и тогда модой считают каждое из них.

Если все или почти все *варианты* различны (что характерно для интервального ряда), то модальное значение определяется несколько другим способом.

➢ Медиана

Медиана m_e **вариационного ряда*** **– это значение, которая делит его на две равные части** (по количеству вариант).

* *не важно, дискретного или интервального, генеральной совокупности или выборочной.*

Медиану можно отыскать несколькими способами. Если даны *первичные данные*, то сортируем их по возрастанию либо убыванию и находим середину ранжированного ряда: $m_e = x_{13} = 4$. Почему именно 13-я варианта? Потому что перед ней находится 12 чисел и после неё тоже 12 чисел, таким образом, значение $x_{13} = 4$ разделило ряд на две равные части, а значит, является *медианой*. Этот номер можно найти аналитически:

– если совокупность содержит **нечётное** количество вариант (наш случай), то делим её объём пополам: $\frac{n}{2} = \frac{25}{2} = 12{,}5$ и округляем полученное значение в бОльшую сторону: 13 – получая тем самым номер искомой варианты;

– если совокупность содержит **чётное** количество вариант, например 20, то делаем то же самое: $\frac{20}{2} = 10$ и медианное значение рассчитываем как *среднее арифметическое* 10-й и следующей варианты: $m_e = \frac{x_{10} + x_{11}}{2}$.

Подчёркиваю, что изложенная выше инструкция работает для упорядоченного (по возрастанию либо убыванию) ряда. Но есть и более быстрый путь, где ничего не нужно сортировать. Это использование стандартной функции Экселя:

– забиваем в любую свободную ячейку =МЕДИАНА(, выделяем мышью все варианты, закрываем скобку) и жмём Enter. Попробуйте самостоятельно. Этот способ удобен, когда вам дано много чисел.

Следует отметить, что в Экселе существуют и отдельные функции для вычисления *средней* (=СРЗНАЧ), *моды* (=МОДА) и ещё много чего, но я против использования этих функций в учебном курсе, за исключением случаев, где это действительно целесообразно. …Почему против? Потому что они не помогают понять суть показателей. Так, *среднюю* гораздо вразумительнее рассчитывать следующим образом:

=СУММ(выделяем мышью диапазон) / объем совокупности. Вычисления рекомендую опробовать лично *(ссылка выше)*.

Ситуация вторая. Когда составлен либо изначально дан готовый дискретный ряд. Тут можно поступить «по любительски» – начать отсчитывать примерно равное количество вариант по краям ряда:

x_i	n_i	
2	3	8
3	5	
4	8	
5	6	9
6	3	

после чего мысленно либо на черновике их отбрасывать, в данном случае отбросим по 8 штук сверху и снизу:

4	8
5	~~6~~ 1

откуда становится ясно, что медианное значение: $m_e = 4$

Или в более солидном стиле, находим относительные накопленные частоты:

x_i	n_i	
2	3	
3	5	
4	8	
5	6	[
6	3	
$\sum$	25	

…

и ту варианту, у которой $w_н$ «перевалила» за отметку 0,5 (50% упорядоченной совокупности). Для варианты $x_2 = 3$ успело накопиться $w_н = 0{,}32$ (32% совокупности), а вот для $x_3 = 4$ – уже $w_н = 0{,}64$ (64%). Таким образом, отметка в 50% пройдена именно здесь, и, стало быть, $m_e = x_3 = 4$.

Запишем красивый **ответ**: $\bar{x}_в = 4{,}04, \quad M_0 = 4, \quad m_e = 4$

И тут возникает следующий **закономерный вопрос**: а зачем вообще нужна *мода* с *медианой*? – ведь есть *средняя*. А дело в том, что в ряде случаев среднее значение неудовлетворительно характеризует центральную тенденцию совокупности:

Пример 9

Известны результаты продаж пиджаков в универмаге города:

x_i	1	2	3	4	5
f_i	225	32	82	145	16

где x_i – количество пуговиц на пиджаке, f_i – число продаж.

Обратите внимание, что в условии задачи ничего не сказано о том, *генеральная* ли это совокупность или *выборочная*, и **в подобной ситуации я не рекомендую ничего додумывать** – *среднюю* просто **обозначаем** через $\bar{x}$, без подстрочного индекса.

Задание: вычислить среднюю. В экселевском файле уже забиты исходные данные и приведена краткая инструкция. Если под пальцами нет Экселя, считаем на калькуляторе. Ну а если вам не нравятся пиджаки, то представьте какие-нибудь шляпки с цветочками:)

….Есть? Какие мысли на счёт полученного значения $\bar{x}$? …С такой статистикой магазин разорится.

Ещё хуже в этом смысле ситуация с медианой – продолжаем решать задачу в Экселе либо в тетради. Особо зоркие читатели, медиану углядят устно.

И, конечно, важнейший показатель здесь мода: $M_0 = 1$. Потому что такая мода :) Более того, в прикладных исследованиях рассматривают несколько модальных значений, в частности, ещё одной модой можно считать варианту $x_4 = 4$. Но это уже «попсовая» статистика, которую я не буду развивать в настоящем курсе.

Теперь надеваем пиджаки / шляпы и возвращаемся на фабрику, где бухгалтер Петрова вычислила генеральную среднюю заработную плату рабочих: $\bar{x}_Г = 1000$ денежных единиц. Здесь мы плавно переходим к интервальному ряду, который целесообразно составлять для «денежных» показателей.

Что будет, если к совокупности добавить руководящий персонал и директора Петрова? Средняя зарплата немного увеличится: $\bar{x}_Г = 1100$, и это уже будет несколько искажённая картина.

А вот если сюда добавить олигарха Петровского, то полученная *средняя* $\bar{x}_Г = 5000$ вообще вызовет широкое возмущение общественности.

Поэтому если в статистической совокупности есть «аномальные» отклонения в ту или иную сторону, то в качестве оценки центрального значения как нельзя лучше подходит медиана, которая в нашем условном примере будет равна, скажем, $m_e = 1050$. Ниже этой планки зарабатывает ровно половина сотрудников фабрики и выше – другая половина, включая Петрова и Петровского. …Главное только, чтобы они наняли правильного статистика :)

➢ Как вычислить среднюю, моду и медиану интервального ряда?

Начнём опять с ситуации, когда нам даны *первичные* статические данные:

Пример 10

По результатам выборочного исследования цен на ботинки в магазинах города получены следующие данные (ден. ед.):

7,5	7,6	8,7
6,1	10,6	9,8
7	6	8,3
6	8,2	8,5
7,4	7,1	9,5
6,8	9,6	6,3
6,3	8,5	5,8
7,5	9,2	7,2
7	8	7,5
7,5	8	6,5

– это в точности числа из Примера 6. Но теперь нам нужно найти среднюю, моду и медиану.

Решение: чтобы найти среднюю по первичным данным, нужно просуммировать все варианты и разделить полученный результат на объём совокупности:

$$\bar{x}_в = \frac{7,5+7,6+8,7+...+7,5+8+6,5}{30} = \frac{230}{30} \approx 7,67 \text{ ден. ед.}$$

Эти подсчёты, кстати, займут не так много времени и при использовании оффлайн калькулятора. Но если есть Эксель, то, конечно, забиваем в любую свободную ячейку: =СУММ(, выделяем мышкой все числа, закрываем скобку), ставим знак деления /, вводим число 30 и жмём Enter. Готово.

Что касается моды, то её оценка по исходным данным, становится непригодна. Хоть мы и видим среди чисел одинаковые, но среди них запросто может найтись так 5-6-7 вариант с одинаковой максимальной частотой, например, частотой 2. Поэтому модальное значение рассчитывается по сформированному интервальному ряду *(см. ниже)*.

Чего не скажешь о медиане: забиваем в Эксель =МЕДИАНА(, выделяем мышью все числа, закрываем скобку) и жмём Enter: $m_e = 7,5$. Причём, здесь даже ничего не нужно сортировать.

Но в Примере 6 я проводил сортировку совокупности по возрастанию (вспоминаем и сортируем), и это хорошая возможность повторить формальный алгоритм отыскания медианы. Делим объём выборки пополам:

$\frac{n}{2} = \frac{30}{2} = 15$, и поскольку она состоит из чётного количества вариант, то медиана равна среднему арифметическому 15-й и 16-й варианты упорядоченного (!) вариационного ряда:

…ден. ед.

Ситуация вторая. Когда даны не первичные данные, а готовый интервальный ряд (что в учебных задачах бывает чаще).

Продолжаем анализировать этот же пример с ботинками, где по исходным данным был составлен ИВР. Для вычисления *средней* потребуются середины x_i интервалов:

Интервалы		x
5,7	6,7	6,
6,7	7,7	7,
7,7	8,7	8,
8,7	9,7	9,
9,7	10,7	10
Суммы:		

…

– чтобы воспользоваться знакомой формулой дискретного случая:

$$\bar{x}_в = \frac{\sum\limits_{i=1}^{k} x_i n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + x_5 n_5}{n} = \frac{6{,}2 \cdot 7 + 7{,}2 \cdot 11 + 8{,}2 \cdot 6 + 9{,}2 \cdot 4 + 10{,}2 \cdot 2}{30} =$$

$$= \frac{43{,}4 + 79{,}2 + 49{,}2 + 36{,}8 + 20{,}4}{30} = \frac{229}{30} \approx 7{,}63$$ – и это отличный результат!

Расхождение с более точным значением ($\approx 7{,}67$), вычисленным по первичным данным, составило всего 0,04!

Здесь мы использовали упомянутый ранее приём – **приблизили интервальный ряд дискретным, и это приближение оказалось весьма эффективным**. Впрочем, с современными программами не составляет особого труда вычислить точное значение даже по очень большому массиву первичных данных. Если они нам известны ;)

С другими центральными показателями всё занятнее.

Чтобы найти моду, нужно найти ***модальный интервал*** *(с максимальной частотой)* – в нашей задаче это интервал $(6{,}7;\ 7{,}7)$ с частотой 11, и воспользоваться следующей **страшненькой формулой**:

…, где:

$x_0 = 6{,}7$ – нижняя граница модального интервала;

$h = 7{,}7 - 6{,}7 = 1$ – длина модального интервала;

$n_M = 11$ – частота модального интервала;

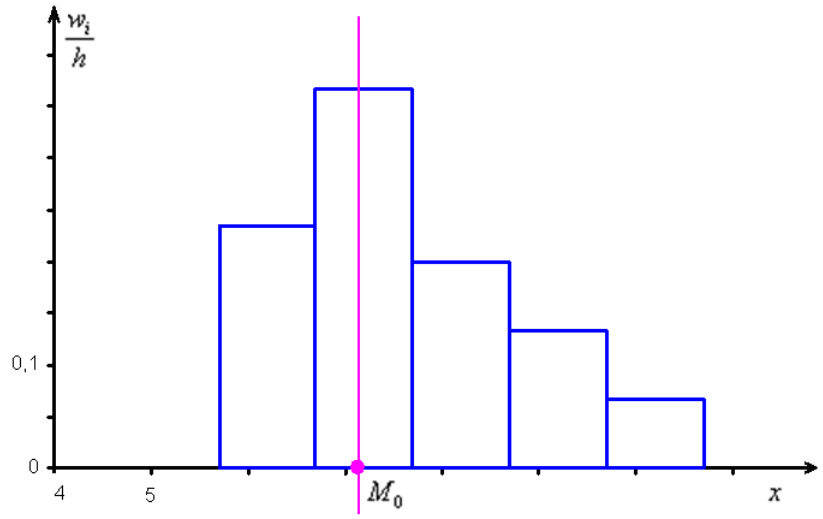
$n_{M-1} = 7$ – частота предыдущего интервала;

$n_{M+1} = 6$ – частота следующего интервала.

Таким образом:

$$M_0 = 6{,}7 + \frac{11-7}{(11-7)+(11-6)} \cdot 1 = 6{,}7 + \frac{4}{4+5} = 6{,}7 + \frac{4}{9} \approx 6{,}7 + 0{,}44 = 7{,}14$$ ден. ед. – как видите, «модная» цена на ботинки заметно отличается от среднего арифметического значения $\bar{x}_в \approx 7{,}67$.

Не вдаваясь в геометрию формулы, просто приведу гистограмму относительных частот и отмечу M_0:



откуда хорошо видно, что *мода* смещена относительно центра *модального интервала* в сторону левого интервала с бОльшей частотой. По той причине, что дешёвых ботинок больше. И, возможно, они тоже вполне себе модные.

Справочно остановлюсь на редких случаях:

– если модальный интервал крайний, то $n_{M-1}=0$ *либо* $n_{M+1}=0$*;*

– если обнаружатся два смежных модальных интервала, например, $(6{,}7;\,7{,}7)$ *и* $(7{,}7;\,8{,}7)$*, то рассматриваем модальный интервал* $(6{,}7;\,8{,}7)$*, при этом близлежащие интервалы (слева и справа) по возможности тоже укрупняем в два раза;*

– если между модальными интервалами есть расстояние, то применяем формулу к каждому интервалу, получая тем самым две или бОльшее количество мод.

Вот такой вот депеш мод :)

И медиана. Она рассчитывается **чуть по менее страшной формуле**. Для её применения нужно найти ***медианный интервал*** – это интервал, содержащий варианту (либо 2 варианты), которая делит вариационный ряд на две равные части.

Выше я рассказал, как определить медиану, ориентируясь на *относительные накопленные частоты* $w_н$, здесь же сподручнее рассчитать «обычные» ***накопленные частоты*** $n_н$. Вычислительный алгоритм такой же – первое значение сносим слева *(красная стрелка)*, а каждое следующее получается как сумма предыдущего с текущей частотой из левого столбца *(зелёные обозначения в качестве примера)*:

Интервалы		x_i	n_i
5,7	6,7	6,2	7
6,7	7,7	7,2	1
7,7	8,7	8,2	6
8,7	9,7	9,2	4
9,7	10,7	10,2	2
Суммы:			3

…

Всем понятен смысл чисел в правом столбце? – это количество вариант, которые успели «накопится» на всех «пройденных» интервалах, включая текущий.

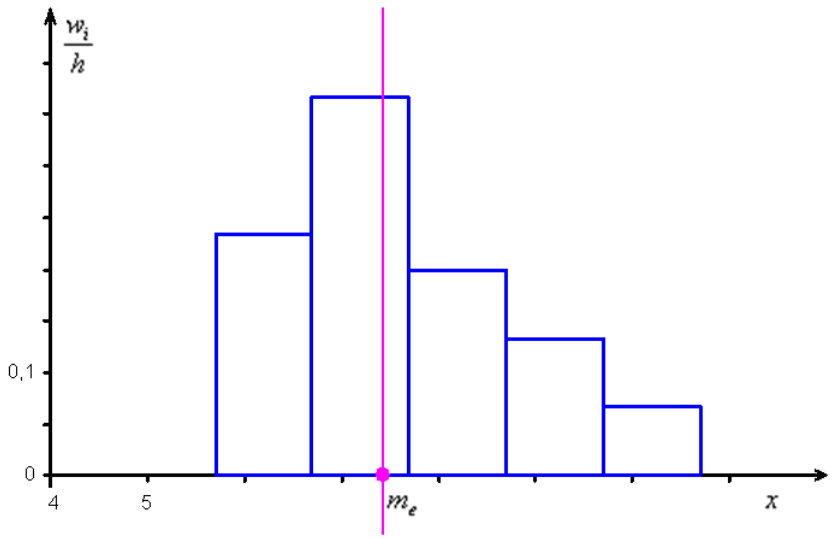
Поскольку у нас чётное количество вариант (30 штук), то медианным будет тот интервал, который содержит $\frac{30}{2}=15$-ю и 16-ю варианту. И ориентируясь по накопленным частотам, легко прийти к выводу, что эти варианты содержатся в интервале $(6,7;\ 7,7)$.

Формула медианы:

…, где:

$n=30$ – объём статистической совокупности;
$x_0=6,7$ – нижняя граница медианного интервала;
$h=7,7-6,7=1$ – длина медианного интервала;
$n_m=11$ – частота медианного интервала;
$n^{н}_{m-1}=7$ – накопленная частота **предыдущего** интервала.

Таким образом:
$m_e=6,7+\frac{0,5\cdot 30-7}{11}\cdot 1=6,7+\frac{15-7}{11}=6,7+0,73\approx 7,43$ ден. ед. – заметим, что медианное значение, в отличие от моды, оказалось смещено правее, т.к. по правую руку находится значительное количество вариант:



Справочно особые случаи:

– если медианным является крайний левый интервал, то $n^{н}_{m-1}=0$;

– если вариационный ряд содержит чётное количество вариант и две средние варианты попали в разные интервалы, то объединяем эти интервалы, и по возможности удваиваем предыдущий интервал.

Ответ: $\bar{x}_в\approx 7,67,\quad M_0\approx 7,14,\quad m_e\approx 7,43$ ден. ед.

По сравнению с предыдущей задачей $(\bar{x}_в=4,04,\quad M_0=4,\quad m_e=4)$, центральные показатели оказались заметно отличны друг от друга. Это говорит об ***асимметрии («скошенности»)*** распределения цен, что хорошо видно по гистограмме и совершенно логично – ботинок низкого и среднего ценового сегмента много, а премиального – мало.

Задание для тренировки:

Пример 11

Для изучения затрат времени на изготовление одной детали рабочими завода проведена выборка, в результате которой получено следующее статистическое распределение:

Затраты на одну деталь, мин.	Число деталей, шт.
до 20	10
от 20 до 24	20
от 24 до 28	50
от 28 до 32	15
свыше 32	5
Итого	**100**

…да, тот самый завод Петровского :) Найти среднюю, моду и медиану.

Решаем эту задачу в Экселе – все числа и инструкции уже там. Если нет Экселя, считаем на калькуляторе, что в данном случае может оказаться даже удобнее. Образец решения, как обычно, в конце книги. Это, кстати, уже каноничная «интервальная» задача, в которой исследуется *непрерывная* величина – время.

Что ещё можно сказать по теме?

Несмотря на разнообразия рассмотренных показателей, их всё равно бывает не достаточно. Существуют крайне неоднородные совокупности, у которых варианты «кучкуются» во многих местах, и по этой причине *средняя*, *мода* и *медиана* плохо характеризуют положение дел.

В таких случаях вариационный ряд дробят с помощью ***квартилей***, ***децилей***, а в ~~упоротых~~ специализированных исследованиях – и с помощью ***перцентилей***.

Квартили упорядоченного вариационного ряда – это *варианты* Q_1, Q_2, Q_3, которые делят его на 4 равные (по количеству вариант) части. Из чего автоматически следует, что 2-я квартиль – есть в точности медиана: $Q_2 = m_e$.

В тяжёлых случаях проводится разбиение на 10 частей – **децилями** $D_1, D_2, D_3, ..., D_9$ – это варианты, который делят упорядоченный вариационный ряд на 10 равных (по количеству вариант) частей.

И в очень тяжелых случаях в ход пускается 99 **перцентилей** $P_1, P_2, P_3, ..., P_{99}$.

После разбиения вариационного ряда каждый участок исследуется по отдельности – рассчитываются локальные средние и другие показатели.

В учебном курсе квартили, децили, перцентили встречаются редко, и посему я оставляю этот материал (их нахождение) для самостоятельного изучения.

Ну а сейчас мы переходим к изучению второй группы статистических показателей:

3.2. Показатели вариации

Они показывают, КАК варьируются *статистические данные*, а именно – насколько далеко «разбросаны» *варианты* относительно средних значений, да и просто друг от друга.

➢ Размах вариации

Он уже встречался. **Это разность между максимальным и минимальным значением *статической совокупности***:

$R = x_{max} - x_{min}$, при этом не имеет значения, *генеральная* ли нам дана совокупность или *выборочная*, сгруппированы ли данные или нет.

Очевидно, что все варианты x_i исследуемой совокупности (той или иной) заключены в промежутке $[x_{min}; x_{max}]$, а размах R – есть не что иное, как его длина.

Такой вот простой и понятный показатель. Но, несмотря на его элементарность, рассмотрим технику вычисления, и, конечно, это отличный повод размяться:

Пример 12

Дана статистическая совокупность: 15, 17, 13, 10, 21, 17, 23, 9, 14, 19. Найти размах вариации

Решить задачу можно несколькими способами.

Способ первый, суровый (продолжаю вас готовить к борьбе с киборгами :)) Это когда под рукой нет вычислительной техники. Или когда она есть, но вы сами понимаете, как важно «прокачать» свои человеческие способности.

Если чисел не так много (наш случай), то максимальное и минимальное значения виднЫ устно: $x_{min} = 9$, $x_{max} = 23$ и размах равен: $R = x_{max} - x_{min} = 23 - 9 = 14$ единиц.

Если чисел больше (20-30 и даже больше), то надёжен следующий алгоритм:

1) Ищем минимальное значение. Сначала самым маленьким будет первое число: 15. Второе число (17) больше, и поэтому его пропускаем. Третье число (13) меньше, чем 15, и теперь 13 – самое малое число. И так далее, пока не закончится список.

2) Ищем максимальное значение. Сначала самым большим будет первое число: 15. Второе число (17) больше и теперь оно становится самым большим. И так далее – до конца списка.

Способ второй, более быстрый (обычно). Использование программного обеспечения, при этом числа можно просто отсортировать (по возрастанию либо убыванию) или использовать специальные функции:

Задание: найти минимальное и минимальное значения в Экселе – данные уже там, данные вас ждут! …Отлично, молодцы! Запишем **ответ** $R = 14$ ед. и с нетерпением перелистнём страницу:

О смысле и важности *показателей вариации* я рассказывал ещё в курсе теорвера. Рассмотрим двух студентов, каждый из которых *в среднем* учится на 3,5 балла. Но есть один нюанс. Один стабильно получает тройки-четвёрки, а другой – то пятёрки, то двойки. И поэтому важно знать не только средние значения, но и *меру рассеяния* оценок относительно средней величины. Чем она меньше – тем стабильнее учится студент.

Эту меру можно оценить следующим образом: из каждой оценки x_i *(пусть их будет* n *штук)* вычитаем среднее значение $\bar{x}$.

Величина $x_i - \bar{x}$ **называется *отклонением*** *(значения* x_i*)* ***от средней*****.**

Теперь эти *отклонения* нужно просуммировать, но тут появляется проблема: среди разностей $x_i - \bar{x}$ есть как положительные, так и отрицательные, и при их суммировании будет происходить взаимоуничтожение отклонений. Более того, итоговая сумма равна нулю: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, и мы не получаем желаемого результата.

Вопрос можно решить с помощью **модуля**, который уничтожает минусы: $\sum_{i=1}^{n}|x_i - \bar{x}|$, после чего осталось разделить сумму на о*бъём совокупности* n и получить

➢ **Среднее линейное отклонение**

$\bar{l} = \dfrac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$ **– есть среднее арифметическое абсолютных отклонений всех значений статистической совокупности от *средней*.** Это формула для несгруппированных статистических данных.

Если же в нашем распоряжении есть сформированный дискретный либо интервальный вариационный ряд, то формула будет такой:

…, где x_i *– варианты* (для дискретного ряда) либо середины частичных интервалов (для интервального ряда), а n_i – соответствующие частоты.

Напоминаю, что маленькая буква n обычно используется для *выборочной* совокупности, а большая – для *генеральной*: N – объём ген. совокупности, N_i – частоты.

Пример 13

В результате 10 независимых измерений некоторой величины, выполненных с одинаковой точностью, получены опытные данные, которые представлены в таблице

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7,1	6,3	6,2	5,8	7,7	6,8	6,7	5,9	5,7	5,1

Требуется вычислить среднее линейное отклонение.

Решение: очевидно, что перед нами первичные данные и выборочная совокупность *(теоретически измерений можно провести бесконечно много)*. На первом шаге вычислим выборочную среднюю:

$$\bar{x}_в = \frac{x_1 + x_2 + x_3 + ... + x_{10}}{n} = \frac{7,1 + 6,3 + 6,2 + ... + 5,1}{10} = \frac{63,3}{10} = 6,33$$

Теперь находим модули *отклонений от средней*:

$$|x_1 - \bar{x}_в| = |7,1 - 6,33| = |0,77| = 0,77$$

$$|x_2 - \bar{x}_в| = |6,3 - 6,33| = |-0,03| = 0,03$$

…

и так далее до:

$$|x_{10} - \bar{x}_в| = |5,1 - 6,33| = |-1,23| = 1,23$$

Вычисления удобно проводить на калькуляторе или в Экселе *(видео ниже)*, а результаты заносить в таблицу:

x_i	$\lvert x_i - \bar{x}_в \rvert$
7,1	0,77
6,3	0,03
6,2	0,13
5,8	0,53
7,7	1,37
6,8	0,47
6,7	0,37
5,9	0,43
5,7	0,63
5,1	1,23
63,3	5,96

На завершающем этапе рассчитываем сумму модулей:

$\sum_{i=1}^{n} |x_i - \bar{x}_в| = \sum_{i=1}^{10} |x_i - \bar{x}_в| = 0,77 + 0,03 + 0,13 + ... + 1,23 = 5,96$ и среднее линейное отклонение:

$\bar{l} = ... = \frac{5,96}{10} = 0,596 \approx 0,6$ ед. – оно означает, что измеренные значения x_i *в среднем* отличаются от $\bar{x}_в = 6,33$ примерно на 0,6 ед.

Ответ: $\bar{l} = 0,596$

Среднее линейное отклонение – это хорошо, но помимо него, для оценки рассеяния *вариант* относительно *средней* существует более совершенный и распространённый подход. Он состоит в том, чтобы использовать не модули, а возведение отклонений в квадрат: $(x_i - \bar{x})^2$ *(для ликвидации возможных «минусов»)*.

В результате получается:

➢ Генеральная и выборочная дисперсия

Дисперсия с латыни так и переводится – рассеяние.

…Не сломать бы язык :) …так: ***выборочная дисперсия*** **– это *среднее арифметическое* квадратов отклонений всех *вариант* выборки от её *средней***:

$D_в = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x}_в)^2}{n}$ – для несгруппированных данных, и:

$D_в = \dfrac{\sum\limits_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot n_i}{n}$ – для сформированного вариационного ряда, где x_i – кратные (одинаковые по значению) *варианты* в дискретном случае либо середины *частичных интервалов* – в интервальном, и n_i – соответствующие частоты.

Ещё раз не спеша и ОСМЫСЛЕННО прочитайте определение и выполните Задание: сформулировать и записать (на бумагу!) определение ***генеральной дисперсии*** и соответствующие формулы. Свериться можно в конце книги.

Вычислим дисперсию по данным Примера 13. Здесь вместо модулей нужно рассчитать квадраты отклонений:

$(x_1 - \bar{x}_в)^2 = (7{,}1 - 6{,}33)^2 = 0{,}77^2 = 0{,}5929$

$(x_2 - \bar{x}_в)^2 = (6{,}3 - 6{,}33)^2 = (-0{,}03)^2 = 0{,}0009$

…

$(x_{10} - \bar{x}_в)^2 = (5{,}1 - 6{,}33)^2 = (-1{,}23)^2 = 1{,}5129$

Заполняем табличку:

x_i	$(x_i - \bar{x}_в)^2$
7,1	0,5929
6,3	0,0009
6,2	0,0169
5,8	0,2809
7,7	1,8769
6,8	0,2209
6,7	0,1369
5,9	0,1849
5,7	0,3969
5,1	1,5129
63,3	5,221

и порядок: $D_в = \ldots$ квадратных (!) единиц – коль скоро, мы возводили в квадрат. И, чтобы вернуться в размерность задачи, из дисперсии следует извлечь квадратный корень. Но мы не будем торопить события, лучше посмотрим, **как выполнять вычисления в Экселе** *(Ютуб)*.

Ответ: $D_в = 0{,}5221\,ед.^2$

Разобранная только что задача часто встречается в лабораторных работах по физике (да и не только) – когда некоторая величина замеряется раз десять и затем рассчитывается среднее значение.

А теперь представьте, что вся ваша группа выполняет лабу по физике, и каждый провёл по 10 испытаний в схожих условиях. Очевидно, что у всех получились *несколько разные* выборочные значения $\bar{x}_в$, но все они *без какой-либо закономерности* (в общем случае) будут варьироваться вокруг истинного значения показателя $\bar{x}_Г$ (роль генеральной средней может играть некий теоретический эталон). Это свойство (отсутствие закономерности) называется ***несмещённостью*** *оценки* генеральной средней, и справедливо оно, как мы увидим ниже, не для всех показателей.

Теперь пару ласковых об отклонениях. В чём их смысл? Всё просто: у кого эти показатели ниже, тот качественнее проводит опыты *(плавнее выполняет действия, точнее снимает показания с приборов, засекает время и т.п.)*. В идеале эти отклонения равны нулю, но это только в идеале – сам эмпиризм ситуации порождает генеральное линейное отклонение $\bar{l}_Г$ и генеральную дисперсию $D_Г$, которые обусловлены человеческим фактором, погрешностью приборов и так далее – вплоть до магнитных бурь.

В случае с полученными линейными отклонениями $\bar{l}$ – всё то же самое, они будут безо всякой закономерности варьироваться вокруг генерального значения $\bar{l}_Г$. Но вот с дисперсией это не так. Полученные значения выборочной дисперсии $D_в$ будут давать *систематически* заниженную оценку генеральной дисперсии $D_Г$. И поэтому выборочную дисперсию следует «поправить» **по формуле**:

… – желающие могут найти обоснование этого факта и этой формулы в специализированной литературе по математической статистике.

Показатель s^2 так и называется –

➢ **Исправленная выборочная дисперсия**

и вот она уже является ***несмещённой*** *оценкой* генеральной дисперсии.

Таким образом, каждый студент должен поправить свою дисперсию, в частности, для данных Примера 13: $s^2 = \frac{n}{n-1} \cdot D_в = \frac{10}{9} \cdot 0{,}5221 \approx 0{,}58$

Следует отметить, что в больших выборках (от 30 вариант) этой поправкой можно пренебречь, так как при $n \to \infty$ дробь $\frac{n}{n-1}$ стремится к единице и $D_в \to s^2$.

И иногда дисперсию лучше вовсе не поправлять. Так, в разобранном примере от нас требовалось просто вычислить выборочную дисперсию и всё. Поэтому в ответе записываем $D_в$. Но вот если дисперсия будет «участвовать» в дальнейших действиях, то, конечно, приводим её к виду s^2. Более того, встречаются задачи, где вообще не понятно – выборочная ли дана совокупность или генеральная, и тогда разумно проявить аккуратность, используя обозначения без подстрочных индексов, в частности, $\bar{x}$ и D.

Теперь случай, когда дан готовый вариационный ряд. У меня опять есть подходящая советская задача про телефонную станцию, но я скорректирую условие в соответствии с современными реалиями:

Пример 14

В результате выборочного исследования звонков, статистик МТС получил следующие данные (за некоторый временной промежуток):

Длительность соединения, мин., t_i	Количество звонков, n_i
0	8
1	18
2	11
3	7
4	4
5	2

…У ОпСоСов, как известно, своя статистика – с округлением до ближайшей целой минуты :), впрочем, это тоже устареет…, как метко заметил современник, дети дружно играли во дворе – каждый в своём смартфоне.

Найти размах вариации, среднее линейное отклонение и выборочную дисперсию. Дать несмещённую оценку генеральной дисперсии и пояснить, что это означает.

Решить данную задачу в Экселе *(данные и гайд уже там)* либо на бумаге с помощью калькулятора. Краткое решение и ответ в конце книги.

Теперь вернёмся к технике вычисления дисперсии. Выше мы её рассчитывали по определению: $D_в = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x}_в)^2}{n}$ – для несгруппированных данных и $D_в = \dfrac{\sum\limits_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot n_i}{n}$ – для *дискретного* либо *интервального вариационного ряда*. Это для выборки. Если же речь идёт о генеральной совокупности, то используем обозначения $D_Г$, $\bar{x}_Г$, N и N_i.

Расчёт дисперсии по определению прост и реально используется на практике, но **существует ещё более простой и удобный способ** –

➢ **Вычисление дисперсии по формуле**

Эта формула выводится непосредственно из определения:

… – дисперсия равна разности *средней арифметической* квадратов всех *вариант* статистической совокупности и квадрата *средней* самих этих вариант.

ОСМЫСЛЕННО повторяем ВСЛУХ и вникаем!

… Карл украл у Клары кораллы, а Клара украла у Карла кларнет!

Если что-то не очень понятно, то сейчас всё станет на свои места:

Для несгруппированных *вариант* $x_1, x_2, x_3, ..., x_n$ выборочной совокупности формула детализируется следующим образом:

$$D_в = \frac{\sum\limits_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum\limits_{i=1}^{n} x_i}{n} \right)^2$$

и для готового вариационного ряда – так:

…, где x_i – кратные (одинаковые) варианты дискретного ряда либо середины интервалов интервального ряда, а n_i – соответствующие частоты.

Для генеральной дисперсии $D_Г$ формулы те же, только с прописными буквами N, N_i, K. Часто используют просто значок суммирования $\sum$ – без переменной-счётчика, поскольку в контексте той или иной задачи и так понятно, что суммируется.

И начнём мы со знакомой подопытной задачи:

Пример 15

В результате 10 независимых измерений получены следующие данные:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7,1	6,3	6,2	5,8	7,7	6,8	6,7	5,9	5,7	5,1

В Примере 13 мы нашли дисперсию по определению: $D_в = 0{,}5221\, ед.^2$, таким образом, ответ известен заранее, и это всегда круто. Всегда, когда он правильный.

Решение: используем формулу $D_в = \overline{x^2} - (\overline{x}_в)^2 = \frac{\sum\limits_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum\limits_{i=1}^{n} x_i}{n} \right)^2$.

Для её применения нужно найти выборочную среднюю, повторим действие:

$$\overline{x}_в = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \frac{\sum\limits_{i=1}^{10} x_i}{10} = \frac{x_1 + x_2 + x_3 + ... + x_{10}}{n} = \frac{7{,}1 + 6{,}3 + 6{,}2 + ... + 5{,}1}{10} = \frac{63{,}3}{10} = 6{,}33,$$

вычислить квадраты всех вариант:

$7{,}1^2 = 50{,}41$

$6{,}3^2 = 39{,}69$

$6{,}2^2 = 38{,}44$

...

$5{,}1^2 = 26{,}01$

и их сумму: $\sum\limits_{i=1}^{n} x_i^2 = 50{,}41 + 39{,}69 + 38{,}44 + ... + 26{,}01 = 405{,}91$

Результаты вычислений удобно заносить в таблицу:

x_i	x_i^2
7,1	50,41
6,3	39,69
6,2	38,44
5,8	33,64
7,7	59,29
6,8	46,24
6,7	44,89
5,9	34,81
5,7	32,49
5,1	26,01
63,3	405,91

Осталось применить формулу:

$D_в = \frac{\sum\limits_{i=1}^{n} x_i^2}{n} - (\overline{x}_в)^2 = \frac{405,91}{10} - 6,33^2 = 40,591 - 40,0689 = 0,5221$, что мы и хотели увидеть – результат, естественно, совпал с полученным ранее по определению.

Ответ: $D_в = 0,5221$

Теперь случай сформированного вариационного ряда. В Примере 14 мы потренировались на дискретном ряде, и сейчас очередь интервального:

Пример 16

С целью изучения вкладов в Сбербанке города проведено выборочное исследование, в результате которого получены следующие данные:

Размер вклада, д.е.	Число вкладчиков
до 400	32
от 400 до 600	56
от 600 до 800	120
от 800 до 1000	104
свыше 1000	88
Итого:	**400**

Вычислить *выборочную дисперсию* и ***среднее квадратическое отклонение***, оценить соответствующие показатели генеральной совокупности.

Автор задачи заботливо подсчитал объем выборки $n = 400$, но не «закрыл» крайние интервалы. Такая вещь уже встречалась, и **решение** мы начинаем с этого закрытия:

поскольку длины внутренних интервалов составляют $h = 200$ д. е., то логично рассмотреть такую же длину и по краям, то бишь, интервалы от 200 до 400 и от 1000 до 1200 денежных единиц.

…Возможно, у вас возник вопрос, а как быть, если даны интервалы разной длины? В этом случае можно принять за «эталон» среднюю длину известных интервалов.

Для расчёта числовых характеристик перейдём к дискретному вариационному ряду, выбрав в качестве *вариант* x_i середины интервалов, которые здесь видны устно:

Интервалы	Середины, x_i	Частоты, n_i
200-400	300	32
400-600	500	56
600-800	700	120
800-1000	900	104
1000-1200	1100	88
Суммы:		**400**

В тяжёлых случаях, напоминаю, суммируем концы интервалов и делим их пополам, например: $x_1 = \frac{200+400}{2} = \frac{600}{2} = 300$.

Кроме того, варианты целесообразно уменьшить в 1000 раз, поскольку в ходе дальнейших вычислений будут получаться гигантские числа. С современной техникой, это, конечно, не проблема, но смотреться будет некрасиво.

Сначала вычислим выборочную среднюю. Этот алгоритм уже обкатан: находим произведения $x_i n_i$, их сумму:

Интервалы	x_i	n_i	x_i
200-400	0,3	32	9
400-600	0,5	56	2
600-800	0,7	120	8
800-1000	0,9	104	93
1000-1200	1,1	88	96
Суммы:		400	31

…

и по соответствующей формуле:

$\bar{x}_в = \frac{\sum x_i n_i}{n} = \frac{312}{400} = 0{,}78$ тыс. д. е. (или 780 д. е.) – средний размер вклада.

Примечание: *далее для компактной записи я буду использовать просто значок* $\sum$ *– без переменной-«счётчика».*

Теперь дисперсия. Её никто не запрещает рассчитать по определению $D_в = \frac{\sum (x_i - \bar{x}_в)^2 \cdot n_i}{n}$, но заметьте, насколько легче формула … – для её применения всего-то лишь нужно рассчитать произведения $x_i^2 n_i$ и их сумму $\sum x_i^2 n_i$ *(правый столбец таблицы)*. Несмотря на то, что многие читатели уже освоили технику вычислений в Экселе, я записал **ещё один ролик** (Ютуб):

Итак, по формуле вычисления дисперсии, получаем:

$D_в = ... = \frac{266{,}4}{400} - 0{,}78^2 = 0{,}666 - 0{,}6084 = 0{,}0576$ тыс. д. е. в квадрате (так как по определению, дисперсия – есть величина квадратичная).

И, чтобы вернуться в размерность задачи, из дисперсии следует извлечь квадратный корень:

$\sqrt{D_в} = \sqrt{0{,}0576} = 0{,}24$ тыс. д. е. или 240 денежных единиц. Полученный показатель называется

➢ **Среднее квадратическое отклонение.**

Или ***среднеквадратическое отклонение***. Или ***стандартное отклонение***. Это синонимы. Оно **обозначается** греческой буквой «сигма», и коль скоро у нас выборочная совокупность, то добавляем соответствующий подстрочный индекс:

$\sigma_в = 0{,}24$ – выборочное среднее квадратическое отклонение.

Чем меньше стандартное отклонение (и дисперсия), тем меньше *вариация* – тем бОльшее количество вариант находится вблизи выборочной средней. Но у нас, как нетрудно «прикинуть на глазок», разброс довольно-таки велик – значительное количество вкладов расположено далековато от среднего значения $\bar{x}_в = 0{,}78$, и поэтому стандартное отклонение $\sigma_в$ получилось немалым.

Следующая часть задачи состоит в том, чтобы корректно **оценить** *генеральную дисперсию* $D_Г$ и *генеральное среднее квадратическое отклонение* $\sigma_Г$.

Не так давно я рассказал о том, что выборочная дисперсия представляет собой **смещённую** оценку генеральной дисперсии. Это означает, что если мы будем проводить неоднократные выборки из той же генеральной совокупности, то полученные значения $D_в$ будут *систематически занижено* оценивать $D_Г$. Обращаю ваше внимание, что это не значит, что $D_в$ будет всегда меньше, чем $D_Г$.

И поэтому выборочную дисперсию, как намекает условие, нужно поправить:

$s^2 = ... \approx 0{,}057744$ – *исправленная выборочная дисперсия*

и, соответственно:

$s = \sqrt{s^2} \approx \sqrt{0{,}057744} \approx 0{,}2403$ или 240,30 денежных единиц – ***исправленное среднее квадратическое отклонение***.

s^2 и s – это уже *несмещённые* оценки генеральной дисперсии $D_Г$ и генерального стандартного отклонения $\sigma_Г$ соответственно.

Ввиду большого объёма выборки (100 вариант) этой поправкой можно пренебречь, но мы всё же не будем «разбрасываться» 30 «копейками».

Ответ: $D_в = 0{,}0576$, $\sigma_в = 0{,}24$; в качестве оценки соответствующих генеральных показателей принимаем $s^2 \approx 0{,}057744$ и $s \approx 0{,}2403$.

Рассмотренные выше показатели (размах вариации, среднее линейное отклонение, дисперсия, стандартное отклонение) входят в группу ***абсолютных*** *показателей вариации*, которые обладают рядом неудобств.

Так, если в прорешанной задаче не уменьшать варианты в 1000 раз, то дисперсия получится в миллион раз больше! Да-да, не $D_в = 0{,}0576$, а $D_в = 57600$. И возникает естественное желание привести результаты к некому единому стандарту.

Для этого существуют показатели ***относительные***, и самый известный из них –

➢ **Коэффициент вариации**

– это отношение *стандартного отклонения* к *средней*, выраженное в процентах:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\%$$

И вот теперь совершенно без разницы, в д. е. мы считали:

…

или в тысячах д. е.:

$$V = \frac{0{,}24}{0{,}78} \cdot 100\% \approx 30{,}77\%$$

Примечание: *на практике часто считают именно через* $\sigma_в$, *но для оценки коэффициента вариации всей генеральной совокупности, конечно же, корректнее использовать исправленное стандартное отклонение* s.

В статистике существует следующий **эмпирический ориентир:**

– если коэффициент вариации составляет примерно 30% и меньше, то статистическая совокупность считается ***однородной***. Это означает, что большинство *вариант* находится недалеко от *средней*, и найденное значение $\bar{x}$ хорошо характеризует центральную тенденцию совокупности.

– если коэффициент существенно больше 30%, то совокупность ***неоднородна***, то есть, значительное количество *вариант* находятся далеко от $\bar{x}$, и *средняя* плохо характеризует типичную варианту. В таких случаях целесообразно рассмотреть квартили, децили, а иногда и перцентили, которые делят вариационный ряд на части, и для каждого участка рассчитать свои показатели. Но это уже немного дебри статистики.

Другое преимущество относительных показателей – это возможность сравнивать **разнородные** статистические совокупности. Например, множество слонов и множество хомяков. Совершенно понятно, что дисперсия веса слонов по сравнению с *дисперсией* веса хомяков – будет просто конской, и их сопоставление не имеет смысла. Но вот анализ *коэффициентов вариации* веса вполне осмыслен, и может статься, что у слонов он составляет 10%, а у хомячков 40% *(пример, конечно, условный)*. Это говорит о сбалансированном питании и размеренной жизни слонов :) А вот хомяки, то носятся с голодухи по полям, то отъедаются и спят в норах, и поэтому среди них есть много худощавых и много упитанных особей :)

Помимо *коэффициента вариации*, существуют и другие относительные показатели, но в реальных студенческих работах они почти не встречаются, и поэтому я не буду их рассматривать в рамках данного курса. Лучше порешаем задачки, первая – на отработку терминов и формул, вторая – творческая:

Пример 17

а) Стандартное отклонение выборочной совокупности равно 5, а средний квадрат её вариант – 250. Найти выборочную среднюю.

б) Определите среднее квадратическое отклонение, если известно, что средняя равна 260, а коэффициент вариации составляет 30%.

Пример 18

Производство стальных труб на предприятии (тонн) в 1-м полугодии составило:

Январь	Февраль	Март	Апрель	Май	Июнь
263	284	310	296	288	251

Определить:
– среднемесячный объем производства;
– среднее квадратическое отклонение;
– коэффициент вариации.

Сделать краткие содержательные выводы. **– Да, это тоже типичный пункт статистической задачи! Даже не пункт – это цель статистического исследования.**

Обратите внимание, что здесь не понятно, выборочной ли считать эту совокупность или генеральной. В таких случаях лучше не заниматься домыслами, просто используем обозначения без подстрочных индексов. Все числа уже в Экселе – не ленимся, решаем!

Подведём итоги:

Статистическую совокупность характеризуют следующие основные показатели:

– ***объём***;

– ***показатели центральной тенденции*** – *мода* на экране, *медиана* в треугольнике, а *средние* – это температура по больнице и в палате.

– ***показатели вариации***, характеризующие разброс *вариант*.

Среди показателей вариации выделяют *абсолютные* и *относительные*. **К первой группе** относят *размах вариации*, *среднее линейное отклонение*, *дисперсию* и *среднее квадратическое отклонение*, а **ко второй** – *коэффициент вариации* и некоторые другие.

В малых выборках (30 и менее вариант) дисперсию *поправляют* – чтобы получить *несмещённую* оценку генеральной дисперсии.

В случае *неоднородности* совокупности *(коэффициент вариации 30% и более)*, её целесообразно разделить на части с помощью квартилей децилей либо перцентилей и исследовать эти части локально.

Всё ли вам понятно в этих нескольких строчках? Все ли термины?

Я вам приснюсь ;)

4. Статистические оценки параметров генеральной совокупности

Вспомним основной метод математической статистики. Он состоит в том, что для изучения *генеральной совокупности* объёма N из неё производится *выборка* объёма n, которая хорошо характеризует всю совокупность (свойство *представительности*). И на основании исследования этой *выборочной совокупности* мы с некоторой достоверностью можем оценить генеральные характеристики. Само собой, чем выше достоверность – тем лучше, тем качественнее исследование. Этому вопросу и посвящена данная глава.

Чаще всего требуется выявить закон распределения генеральной совокупности *(о чём пойдёт речь позже)* и **оценить** его важнейшие числовые параметры, такие как генеральная средняя $\overline{x}_{Г}$, генеральная дисперсия $D_{Г}$ и стандартное отклонение $\sqrt{D_{Г}}=\sigma_{Г}$.

4.1. Точечные оценки

Очевидно, что для оценки этих параметров нужно вычислить соответствующие выборочные значения. Так, выборочная средняя $\overline{x}_{в}$ позволяет нам оценить генеральную среднюю $\overline{x}_{Г}$, причём, оценить её *точечно*. Почему точечно? Потому что $\overline{x}_{в}$ – это отдельно взятое, конкретное значение. Если из той же генеральной совокупности мы будем проводить многократные выборки, то в общем случае у нас будут получаться *различные* выборочные средние, и каждая из них представляет собой ***точечную оценку*** генерального значения $\overline{x}_{Г}$.

Аналогично, *точечной оценкой* генеральной дисперсии $D_{Г}$ является исправленная выборочная дисперсия s^2, и соответственно, стандартного отклонения $\sqrt{D_{Г}}=\sigma_{Г}$ – исправленное стандартное отклонение s.

4.2. Интервальная оценка и доверительный интервал

Недостаток точечных оценок состоит в том, что при небольшом *объёме* выборки (как оно часто бывает), мы можем получать выборочные значения, которые далеки от истины. И в этих случаях логично потребовать, чтобы выборочная характеристика $\theta_{в}$ *(средняя, дисперсия или какая-то другая)* отличалась от своего генерального значения $\theta_{Г}$ **не более** чем на некоторое положительное значение δ.

Справка: *θ – греческая буква «тета», δ – греческая буква «дельта», вместо «дельты» также используют ε («эпсилон»).*

Значение δ называется ***точностью оценки***, и озвученное выше требование можно записать с помощью модуля: $|\theta_{Г}-\theta_{в}|<\delta$

Но статистические методы не позволяют 100%-но утверждать, что рассчитанное значение $\theta_{в}$ будет удовлетворять этому неравенству – ведь в статистике всегда есть место случайности, когда мы можем «выиграть в лотерею» в плохом смысле этого слова. Таким образом, можно говорить лишь о **вероятности** γ *(«гамма»)*, с которой это неравенство осуществится: $P\left(|\theta_{Г}-\theta_{в}|<\delta\right)=\gamma$.

А теперь я **раскрою модуль**:

…

и сформулирую суть:

Интервал $(\theta_в - \delta; \theta_в + \delta)$ называется ***доверительным интервалом*** и представляет собой ***интервальную оценку*** генерального значения $\theta_Г$ по найденному выборочному значению $\theta_в$. Данный интервал с вероятностью γ «накрывает» истинное значение $\theta_Г$. Эта вероятность называется ***доверительной вероятностью*** или ***надёжностью*** интервальной оценки. Надёжность «гамма» часто задаётся наперёд, популярные варианты:
$\gamma = 0{,}95$, $\gamma = 0{,}99$, $\gamma = 0{,}999$.

Переходим к конкретике:

4.3. Оценка генеральной средней нормально распределенной совокупности

Если вы не знаете, что такое **нормальное распределение**, то это, конечно, большое упущение – обязательно ознакомьтесь с материалом по ссылке. И мы сразу разберём «заезженную» задачу, которую предлагают даже студентам-гуманитариям:

Пример 19

Известно, что генеральная совокупность распределена *нормально* со средним квадратическим отклонением $\sigma = 5$. Найти доверительный интервал для оценки математического ожидания a с надежностью 0,95, если выборочная средняя $\bar{x}_в = 24{,}15$, а объем выборки $n = 100$.

Прежде всего, **обращаю ваше внимание на принципиальный момент:** здесь

➢ **Известно стандартное отклонение генеральной совокупности.**

Дело в том, что в похожих задачах оно бывает и не известно, и тогда решение будет отличаться! Этот случай тоже будет. А сейчас **решение** таково, разбираемся в ситуации:

– из генеральной совокупности проведена выборка в $n = 100$ попугаев и по её результатам найдена выборочная средняя: $\bar{x}_в = 24{,}15$ (средний рост птицы).

Выборочная средняя – это точечная оценка неизвестной нам генеральной средней $\bar{x}_Г = a$. Как отмечалось выше, недостаток *точечной оценки* состоит в том, что она может оказаться далёкой от истины. И по условию, требуется найти интервал $(\bar{x}_в - \delta; \bar{x}_в + \delta)$, который с вероятностью $\gamma = 0{,}95$ **накроет** истинное значение $\bar{x}_Г = a$.

Именно так! Здесь некорректно говорить, что «истинное значение $\bar{x}_Г = a$ **попадёт** в этот интервал». Генеральная средняя – это конкретное (пусть и не известное нам) значение, и оно не может никуда «попасть». В разных выборках мы будем получать разные значения $\bar{x}_в$ и разные доверительные интервалы, которые могут лишь *накрыть* генеральную среднюю. А могут и не накрыть (некоторые из них).

Найдём *точность оценки*, она рассчитывается по формуле …, где t_γ – так называемый ***коэффициент доверия***. Этот коэффициент отыскивается из соотношения $2\Phi(t_\gamma)=\gamma$, где $\Phi(x)$ – ***функция Лапласа***.

По условию, $\gamma=0{,}95$, следовательно:

$$2\Phi(t_\gamma)=0{,}95 \quad \Rightarrow \quad \Phi(t_\gamma)=\frac{0{,}95}{2}=0{,}475$$

И по таблице значений функции Лапласа либо пользуясь приложенным к курсу расчётным макетом (пункт 1*), выясняем, что значению $\Phi(t_\gamma)=0{,}475$ соответствует аргумент $t_\gamma\approx 1{,}96$.

Таким образом, точность оценки:
…

и искомый доверительный интервал:
$(\bar{x}_в-\delta;\ \bar{x}_в+\delta)$
$(24{,}15-0{,}98;\ 24{,}15+0{,}98)$
$(23{,}17;\ 25{,}13)$

Этот интервал с вероятностью $\gamma=0{,}95$ *(надёжностью)* накрывает истинное генеральное значение $\bar{x}_Г=a$ среднего роста попугая. Но всё же остаётся 5%-ная вероятность того, что генеральная средняя окажется вне найденного интервала.

Ответ: $23{,}17<a<25{,}13$.

И тут возникает **светлая мысль** уменьшить этот интервал – чтобы получить более точную оценку. Что для этого можно сделать? Давайте посмотрим на формулу $\delta=\frac{t_\gamma\sigma}{\sqrt{n}}$.

Очевидно, что чем меньше стандартное отклонение (мера разброса значений), тем ***у***же доверительный интервал. Но это в отдельно взятой задаче ни на что не влияет – ведь нам известно конкретное значение σ и изменить его невозможно.

Поэтому для уменьшения «дельты» можно уменьшить *коэффициент доверия*, например, вместо $t_\gamma=1{,}96$ рассмотреть $t_\gamma=1$ и тогда $\delta=\frac{1\cdot 5}{\sqrt{100}}=0{,}5$, в результате чего доверительный интервал $(\bar{x}_в-\delta;\ \bar{x}_в+\delta)=(23{,}65;\ 24{,}65)$ – действительно стал в 2 раза короче. Но засада в том, что упала и *доверительная вероятность*:

пользуясь таблицей значений функции Лапласа либо расчётным макетом (пункт 1), находим: $\gamma=2\Phi(t_\gamma)=2\Phi(1)=2\cdot 0{,}3413=0{,}6826$ – то есть о том, что этот более узкий интервал накроет *генеральную среднюю*, мы теперь можем утверждать **лишь с вероятностью 68,26%**. Что, конечно, неудовлетворительно, для серьёзного статистического исследования.

Поэтому для уменьшения доверительного интервала (при том же значении γ) остаётся увеличивать объём выборки n. Что совершенно понятно и без формулы $\delta = \frac{t_\gamma \sigma}{\sqrt{n}}$, ведь чем больше объём выборки, тем точнее она характеризует генеральную совокупность (при прочих равных условиях). Об объёме выборки мы поговорим позже, ну а пока творческая задача для самостоятельного решения:

Пример 20

По результатам выборочного исследования $n = 100$ объектов найдена выборочная средняя $\overline{x}_в = 93$.

1) С какой вероятностью можно утверждать, что генеральная средняя отличается от найденного значения не более чем на 3, если известно, что генеральная совокупность распределения нормально с дисперсией 400?

2) Определить доверительный интервал, который с надежностью $\gamma = 0{,}99$ накроет истинное значение генеральной средней.

Образец в конце книги, таблица либо расчётный макет (пункты 1 и 1*) в помощь.

И тут, наверное, у вас назрели вопросы – **а откуда известно, что генеральная совокупность распределена нормально**, и тем более, **откуда известно её стандартное отклонение?**

Обычно эта информация известна из предыдущих исследований. Классический пример – измерительный прибор. Очевидно, что его случайные погрешности удовлетворяют условию **теоремы Ляпунова**, а значит, распределены нормально. Кроме того, производитель, как правило, тестирует прибор, и указывает в его паспорте стандартное отклонение *случайной погрешности*, которое можно принять за σ.

Но если установить нормальность распределения достаточно просто (в том числе статистическими методами), то с генеральным значением σ всё сложнее – зачастую вычислить его трудно или невозможно. В такой ситуации остаётся ориентироваться на исправленную выборочную дисперсию s^2 и решение несколько изменится.
Возвращаемся к нашей любимой задаче:

Пример 21

В результате 10 независимых измерений некоторой величины X, выполненных с одинаковой точностью, полученные опытные данные, которые представлены в таблице:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7,1	6,3	6,2	5,8	7,7	6,8	6,7	5,9	5,7	5,1

Предполагая, что результаты измерений подчинены нормальному закону распределения вероятностей, оценить истинное значение величины X при помощи доверительного интервала, покрывающего это значение с вероятностью 0,95.

Обратите внимание, что здесь речь идёт уже не о погрешностях прибора, а об измерениях, и помимо технических, велико влияние других, в частности, человеческого фактора, особенно, если вы используете махрово-аналоговый инструмент – что-нибудь вроде механического секундомера или линейки.

Решение следует начать с вычисления выборочных характеристик, и задача облегчается тем, что в Примере 13 они уже вычислены: $\bar{x}_в = 6{,}33$, $D_в = 0{,}5221$. По условию, требуется оценить генеральную совокупность (а именно, параметр $\bar{x}_Г = a$), и поэтому дисперсию нужно обязательно поправить:

… – *несмещённая* оценка неизвестной генеральной дисперсии σ^2. И нас будет интересовать *несмещённая* оценка генерального стандартного отклонения σ:

$s = \sqrt{s^2} \approx \sqrt{0{,}58} \approx 0{,}76$ – *исправленное среднее квадратическое отклонение*.

Теперь построим доверительный интервал для оценки истинного (генерального) значения a величины X.

➢ **Если генеральная дисперсия нормального распределения не известна**

то этот интервал строится по **похожей формуле**:

…, с той поправкой, что коэффициент доверия t_γ рассчитывается с помощью ***распределения Стьюдента***. Я не буду рассказывать об этом распределении и ограничусь технической стороной вопроса.

Значение t_γ можно найти с помощью *таблицы значений распределения Стьюдента*, в частности популярна таблица, специально адаптированная для данной задачи*. И, согласно таблице, доверительной вероятности $\gamma = 0{,}95$ и объёму выборки $n = 10$ соответствует коэффициент доверия: $t_\gamma = 2{,}2622$

* *в таблице, которую можно встретить чаще, приводятся значения для так называемого уровня значимости* $\alpha = 1 - \gamma$ *и для количества степеней свободы* $k = n - 1$.

Другой, более универсальный способ – воспользоваться Экселем, и чтобы далеко не ходить, я добавил этот функционал в расчётный макет: ищем *пункт 2б*, забиваем значения $\gamma = 0{,}95$, $k = n - 1 = 10 - 1 = 9$ и получаем «на выходе» $t_\gamma \approx 2{,}2622$.

Вычислим точность оценки:

…

Таким образом, искомый доверительный интервал:

$$\bar{x}_в - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x}_в + \frac{t_\gamma s}{\sqrt{n}}$$

$6{,}33 - 0{,}54 < a < 6{,}33 + 0{,}54$

$5{,}79 < a < 6{,}87$ – данный интервал с вероятностью $\gamma = 0{,}95$ накрывает истинное генеральное значение $\bar{x}_Г = a$ измеряемой величины X.

Ответ: $5{,}79 < a < 6{,}87$

Для самостоятельного решения:

Пример 22

На основании $n = 20$ испытаний установлено, что в среднем для изготовления ~~шавермы~~ полупроводникового диода требуется $\bar{x}_в = 76$ секунд, а исправленное среднее квадратическое отклонение составляет $s = 11$ секунд. Предположив, что время изготовления диода есть нормальная случайная величина, определить с надежностью $\gamma = 0{,}999$ доверительный интервал для оценки среднего времени изготовления диода

Краткое решение в конце книги, таблица или макет *(пункт 2б)* – в помощь.

Итак, что главное в разобранных задачах? Главное, обратить внимание, генеральное ли нам дано отклонение σ или исправленное выборочное s. От этого зависит, какую формулу нужно использовать, эту:

…, где $2\Phi(t_\gamma) = \gamma$,

или эту:

…, где t_γ отыскивается с помощью распределения Стьюдента.

При увеличении объёма выборки n, *распределение Стьюдента* стремится к *нормальному распределению*, и поэтому уже при $n > 30$ во 2-м случае допускается нахождение t_γ с помощью того же соотношения $2\Phi(t_\gamma) = \gamma$. Но я бы не рекомендовал так делать. Потому что если дано s, то предполагается, что решать нужно именно через «Стьюдента», и при наличии Экселя с этим никаких проблем – можно рассчитать любые значения, которые отсутствуют в таблицах.

Коварные авторы могут предложить «простое» выборочное отклонение $\sigma_в$, и тогда его следует поправить по формуле: …, которая следует из соотношения дисперсий: $s^2 = \frac{n}{n-1} \cdot D_в$. Иногда бывает предложена и дисперсия (та или иная). **Именно здесь нужно проявлять аккуратность**, сами же вычисления достаточно примитивны.

4.4. Оценка генеральной дисперсии нормально распределенной совокупности

Этот интервал можно построить несколькими способами, которые я постараюсь уместить буквально в пару страниц. Продолжаем решать ту же задачу об измерениях:

Пример 23

По $n = 10$ равноточным измерениям найдено исправленное среднее квадратическое отклонение $s = 0{,}76$. Предполагая, что результаты измерений распределены нормально, построить доверительный интервал для оценки истинного значения σ (генерального стандартного отклонения) с надёжностью $\gamma = 0{,}95$.

Обратите внимание, что для **решения** этой задачи нам не обязательно знать выборочную среднюю *(хотя в Примере 13 мы её нашли)*.

Способ первый. *Доверительный интервал* для оценки неизвестной дисперсии σ^2 нормальной генеральной совокупности определяется следующим образом (не пугаемся):

…, где χ^2 – ***распределение «хи-квадрат»*** (ещё один скелет в шкафу:)), а $\chi^2_{\alpha_1,k}$, $\chi^2_{\alpha_2,k}$ – критические значения, вычисленные для $\alpha_1=\frac{1-\gamma}{2}$, $\alpha_2=\frac{1+\gamma}{2}$ и $k=n-1$. Что это всё значит, я рассказывать тоже не буду ☺.

Данный интервал с вероятностью γ *(надёжностью)* накрывает истинное значение генеральной дисперсии σ^2. А если из всех частей неравенства извлечь корни, то получим соответствующий интервал для оценки генерального стандартного отклонения:

$$\frac{\sqrt{(n-1)s}}{\chi_{\alpha_1,k}}<\sigma<\frac{\sqrt{(n-1)s}}{\chi_{\alpha_2,k}}$$

Значения $n=10,\ s=0{,}76$ известны, осталось разобраться с нижним этажом.

Вычислим $\alpha_1=\frac{1-\gamma}{2}=\frac{1-0{,}95}{2}=0{,}025,\ \alpha_2=\frac{1+\gamma}{2}=\frac{1+0{,}95}{2}=0{,}975,\quad k=n-1=9$ и по таблице критических значений распределения χ^2 либо по макету *(пункт 3б)* найдём:

$$\chi^2_{\alpha_1,k}=\chi^2_{0{,}025;9}\approx 19$$

$$\chi^2_{\alpha_2,k}=\chi^2_{0{,}975;9}\approx 2{,}7$$

В результате:

$\frac{\sqrt{9}\cdot 0{,}76}{\sqrt{19}}<\sigma<\frac{\sqrt{9}\cdot 0{,}76}{\sqrt{2{,}7}}$ – не забываем извлечь корни из знаменателей!

$0{,}52<\sigma<1{,}39$ – таким образом, с вероятностью $\gamma=0{,}95$ можно утверждать, что данный интервал накроет генеральное стандартное отклонение σ.

Полученный интервал асимметричен относительно выборочного значения $s=0{,}76$, и его широкий диапазон объясним малым объёмом выборки – велика вероятность, что при 10 измерениях значение «эс» действительно далеко от истинного значения «сигма».

Способ второй, более простой. Он состоит в построении симметричного интервала по формуле:

… , где значение q отыскивается по соответствующей таблице.

Согласно таблице, *доверительной вероятности* $\gamma=0{,}95$ и объёму $n=10$ соответствует значение $q=0{,}65$, таким образом:

$$0{,}76\cdot(1-0{,}65)<\sigma<0{,}76\cdot(1+0{,}65)$$

$$0{,}266<\sigma<1{,}254$$

В результате мы получили примерно такой же широкий интервал. Для малых выборок может даже получиться $q>1$, в таких случаях принимают ещё более грубую интервальную оценку: $0<\sigma<s(1+q)$

Ответ: 1) $0{,}52<\sigma<1{,}39$, 2) $0{,}266<\sigma<1{,}254$.

Как и для *распределения Стьюдента*, при увеличении n *распределение хи-квадрат* стремится к *нормальному распределению*, и уже при $n > 30$ можно использовать приближенную формулу:

…, где *коэффициент доверия* t_γ, который определяется из знакомого лапласовского соотношения $2\Phi(t_\gamma) = \gamma$.

Иногда встречаются обратная задача – по известной точности оценки (фактически по известному интервалу) найти доверительную вероятность γ. Иногда требуется построить одностороннюю оценку. Но ввиду их исключительного «иногда», я передаю привет студентам Московского института статистики и продолжаю :) Точнее, предлагаю продолжить вам:

Пример 24

В результате обработки экспериментальных данных объёма $n = 100$ получены следующие выборочные характеристики: $\bar{x}_в = 18{,}52$, $\sigma_в \approx 2{,}3237$. В предположении о нормальном распределении генеральной совокупности, с надёжностью $\gamma = 0{,}9$ определить доверительные интервалы:

1) для оценки неизвестной генеральной средней $\bar{x}_Г$;

2) для оценки генерального среднего квадратического отклонения σ двумя способами – с помощью *распределения хи-квадрат*: $\frac{\sqrt{(n-1)}s}{\chi_{\alpha_1,k}} < \sigma < \frac{\sqrt{(n-1)}s}{\chi_{\alpha_2,k}}$ и приближённо, по формуле $\frac{\sqrt{2n}}{\sqrt{2n-3}+t_\gamma} \cdot s < \sigma < \frac{\sqrt{2n}}{\sqrt{2n-3}-t_\gamma} \cdot s$, где $2\Phi(t_\gamma) = \gamma$.

Заметьте, что здесь «плакал» лёгкий способ построения интервала $s(1-q) < \sigma < s(1+q)$, так как в стандартной таблице отсутствуют значения для $\gamma = 0{,}9$. А рассчитывать значения «ку» программным способом – геморрой ещё тот. Краткое решение и примерный образец оформления задачи в конце книги. Далее по курсу:

4.5. Повторная и бесповторная выборка

Что это означает? Слова говорят сами за себя: если случайно отбираемые объекты **не возвращаются** в *генеральную совокупность*, то это ***бесповторная выборка***. Если же выбранный объект возвращается обратно (перед выбором следующего), то это ***повторная выборка***, т.е. здесь один и тот же попугай может быть выбран неоднократно.

И те и другие примеры уже встречались ранее, но, конечно, нам привычнее и понятнее *бесповторный отбор*. Вспоминаем основной метод статистики и Фёдора с помидорами. Совершенно понятно, что после случайного выбора помидора нет никакого смысла возвращать его обратно в коробку, более того, в этом даже есть вредный смысл – ибо овощ может попасться снова, что ухудшит *репрезентативность* выборки. Или исследование успеваемости студентов ВУЗа. Однозначно и лучше бесповторный отбор. Другой пример, это телефонный опрос, давайте под праздник: *«Верите ли вы в Деда Мороза?»*, как вариант, анкетирование: *«да / нет / по праздникам»*. Здесь тоже вредно спрашивать каждого респондента дважды :), и поэтому опрос проводится без повторов.

Но вот в иных случаях это полезно, например, при статистическом исследовании прогулов в университете. Очевидно, что один и тот же студент может попасть в выборку неоднократно, и было бы неправильно не учитывать его повторные прогулы. Или количество обращений в поликлинику – то же самое, один тот же человек может обратиться несколько раз. Другой распространённый пример – многократное измерение некоторой величины. Теоретически генеральная совокупность бесконечна, и из неё исследователь «выбирает» несколько значений, которые могут повторяться, причём, не только теоретически, но и практически, по причине округления измерений.

Теперь в свете новой информации детализируем задачу о доверительном интервале генеральной средней. Детализация состоит в том, что **построение доверительного интервала зависит от того, бесповторная была проведена выборка или повторная**. Как и прежде, полагаем, что *генеральная совокупность* распределена *нормально*, либо её распределение близкО к таковому.

4.6. Оценка генеральной средней по повторной и бесповторной выборкам

Итак, вникаем: пусть из нормально распределенной *(или около того)* генеральной совокупности объёма N проведена выборка объёма n и по её результатам найдена выборочная средняя $\overline{x}_в$. Тогда доверительный интервал для оценки *генеральной средней* $\overline{x}_Г = a$ имеет вид:

$\overline{x}_в - \Delta < \overline{x}_Г < \overline{x}_в + \Delta$, где Δ *(«дельта» большая)* – точность оценки, которую также называют ***предельной ошибкой*** выборки.

Точность оценки рассчитывается как произведение $\Delta = t_\gamma \cdot \mu$ – *коэффициента доверия* t_γ на ***среднюю ошибку*** выборки μ *(«мю»)*.

Если известна дисперсия генеральной совокупности σ^2, то *коэффициент доверия* t_γ отыскивается из лапласовского соотношения $2\Phi(t_\gamma) = \gamma$, а *средняя ошибка* рассчитывается по формуле:

$\mu = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$ – для бесповторной выборки или … – для повторной.

Если же генеральная дисперсия не известна, то в качестве её приближения используют исправленную выборочную дисперсию s^2. В этом случае *коэффициент доверия* t_γ определяют с помощью *распределения Стьюдента*, а при $n \geq 30$ можно использовать соотношение $2\Phi(t_\gamma) = \gamma$. *Средняя* же *ошибка* рассчитывается по аналогичным формулам:

$\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$ – для бесповторной или … – для повторной выборки.

Напоминаю, что доверительная вероятность (надёжность) γ задаётся наперёд и показывает, с какой вероятностью построенный *доверительный интервал* $(\overline{x}_в - \Delta;\ \overline{x}_в + \Delta)$ накрывает истинное значение $\overline{x}_Г$.

С конспектом отмучились, теперь задачи :)

Модифицируем задание Примера 19, а именно уточним способ отбора попугаев:

Пример 25

Известно, что генеральная совокупность распределена *нормально* со средним квадратическим отклонением $\sigma = 5$. По результатам 4%-ной бесповторной выборки объёма $n = 100$, найдена выборочная средняя $\bar{x}_в = 24{,}15$ *(условно средний рост птицы)*.

1) Найти доверительный интервал для оценки генеральной средней $\bar{x}_Г = a$ с надежностью $\gamma = 0{,}95$.

2) Выборку какого объёма нужно организовать, чтобы уменьшить данный интервал в два раза?

Не решение даже, а целое **исследование** впереди, начинаем. Прежде всего, найдём объём генеральной совокупности:

$N = \frac{100\%}{4\%} \cdot n = 25 \cdot 100 = 2500$ попугаев, и на самом деле нам предстоит ответить на следующий вопрос: **а достаточно ли выборки объёма** $n = 100$**?** Или для качественного исследования роста попугаев нужно выбрать побольше птиц?

1) Доверительный интервал для оценки генеральной средней составим по формуле:

$\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$, где $\Delta = t_\gamma \cdot \mu$ – точность оценки. В задачах данного типа у коэффициента доверия часто опускают подстрочный индекс и пишут просто t, однако я не буду следовать мейнстриму, т. к. эта «кастрация» ухудшает понимание.

По условию, нам **известна генеральная дисперсия**, поэтому *коэффициент доверия* найдём из соотношения $2\Phi(t_\gamma) = \gamma \Rightarrow 2\Phi(t_\gamma) = 0{,}95 \Rightarrow \Phi(t_\gamma) = \frac{0{,}95}{2} = 0{,}475$. По таблице значений функции Лапласа либо на макете (пункт 1*) определяем, что этому значению функции соответствует аргумент $t_\gamma \approx 1{,}96$.

Поскольку выборка бесповторная, то *среднюю ошибку* рассчитаем по формуле:
…

Таким образом, точность оценки $\Delta = t_\gamma \cdot \mu \approx 1{,}96 \cdot 0{,}49 \approx 0{,}96$ и соответствующий доверительный интервал:

$\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$

$24{,}15 - 0{,}96 < \bar{x}_Г < 24{,}15 + 0{,}96$

$23{,}19 < \bar{x}_Г < 25{,}11$ – с вероятностью $\gamma = 0{,}95$ данный интервал накроет истинное значение генерального среднего роста $\bar{x}_Г = a$ попугая.

Теперь предположим, что нас не устраивает точность полученного результата. Хотелось бы уменьшить интервал. Или оставить его таким же, но повысить доверительную вероятность. Этим вопросам и посвящён следующий пункт решения:

2) Выясним, сколько попугаев нужно взять, чтобы уменьшить полученный интервал в два раза. Иными словами, была точность 0,96, а мы хотим $\Delta = \frac{0{,}96}{2} = 0{,}48$. **При условии сохранения доверительной вероятности** необходимый объём выборки можно рассчитать **по формуле** …, которая выводится из $\Delta = t_\gamma \cdot \mu = t_\gamma \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$.

А нашей задаче:

$$n \approx \frac{5^2 \cdot (1{,}96)^2 \cdot 2500}{5^2 \cdot (1{,}96)^2 + (0{,}48)^2 \cdot 2500} \approx 357{,}27$$ и **обязательно проверочка:**

$$\Delta = t_\gamma \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} = 1{,}96 \cdot \sqrt{\frac{5^2}{357{,}27}\left(1-\frac{357{,}27}{2500}\right)} \approx 1{,}96 \cdot 0{,}245 \approx 0{,}48$$, ч.т.п.

Таким образом, чтобы обеспечить точность $\Delta = 0{,}48$ при надёжности $\gamma = 0{,}95$ нужно провести выборку объёмом **не менее** 358 попугаев *(округлили в бОльшую сторону)*. В этом случае получится *доверительный интервал* в два раза короче:

$$\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$$

$$\bar{x}_в - 0{,}48 < \bar{x}_Г < \bar{x}_в + 0{,}48$$

И внимание! Здесь нельзя использовать значение $\bar{x}_в = 24{,}15$ предыдущего пункта! Почему? Потому что **в новой выборке** мы почти наверняка получим НОВУЮ выборочную среднюю. Вот её-то и нужно будет подставить.

Осталось прикинуть, а не много ли это – 358 попугаев? Объём выборки составит: $\frac{358}{2500} \cdot 100\% \approx 14{,}32\%$ от генеральной совокупности – ну, в принципе, сносно, хотя и многовато. Поэтому здесь можно использовать другой подход: оставить *точность оценки* $\Delta = 0{,}96$ прежней, но повысить доверительную вероятность до $\gamma = 0{,}99$. В этом случае нужно найти новый коэффициент доверия t_γ (из соотношения $2\Phi(t_\gamma) = \gamma$) и решить уравнение $\Delta = t_\gamma \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$, получив в качестве корня необходимый объём выборки n. Желающие могут выполнить этот пункт самостоятельно, в результате получается выборка в $n = 169$ попугаев или $\frac{169}{2500} \cdot 100\% = 6{,}76\%$ генеральной совокупности. Что лучше, конечно, ведь измерить линейкой 358 попугаев – задача хлопотная, они явно будут сопротивляться, а некоторые ещё и говорить нехорошие слова ☺.

Теперь распишем доверительный интервал $\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$ подробно:

$$\bar{x}_в - t_\gamma \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} < \bar{x}_Г < \bar{x}_в + t_\gamma \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$$

и ответим вот на какой **вопрос:** а что будет, если генеральная совокупность великА или даже бесконечна? В этом случае дробь $\frac{n}{N}$ близкА к нулю, и мы получаем интервал:

$\bar{x}_в - \frac{t_\gamma \sigma}{\sqrt{n}} < \bar{x}_Г < \bar{x}_в + \frac{t_\gamma \sigma}{\sqrt{n}}$, который фигурировал в Примере 19. То есть по умолчанию (когда не сказано, бесповторная выборка или нет), считают именно так.

Следует отметить, что полученный выше интервал соответствует повторной выборке со *средней ошибкой* …, таким образом, при слишком большом объёме N генеральной совокупности математическое различие между бесповторной и повторной выборкой стирается.

Пришло время запланировать собственное статистическое исследование:

Пример 26

В результате многократных независимых измерений некоторой физической величины X в прошлом достаточно точно определена генеральная дисперсия $\sigma^2 = 1{,}2$ ед.; при этом средняя величина склонна изменениям (от исследования к исследованию). Сколько измерений нужно осуществить, чтобы с вероятностью $\gamma = 0{,}9974$ заключить текущее истинное значение генеральной средней $\bar{x}_{\Gamma}$ в интервале длиной 0,5 ед.

И это как раз только что описанный случай: данную выборку можно считать бесповторной, при этом ген. совокупность теоретически бесконечна; либо повторной, так как округлённые результаты измерений могут повторяться.

Краткое решение в конце книги, числа можете выбрать по своему вкусу ☺. Но здесь есть одно «странное» значение $\gamma = 0{,}9974$. **Оно не случайно** и соответствует **правилу «трёх сигм»**, т. е., **практически достоверным** является тот факт, что построенный интервал накроет истинное значение $\bar{x}_{\Gamma}$.

Разумеется, на практике генеральная дисперсия чаще не известна, и поэтому за неимением лучшего, используют исправленную выборочную дисперсию:

Пример 27

С целью изучения урожайности подсолнечника в колхозах области проведено 5%-ное выборочное обследование 100 га посевов, отобранных в случайном порядке, в результате которого получены следующие данные:

Урожайность, ц/га	Посевная площадь, га
до 13	10
от 13 до 15	25
от 15 до 17	40
от 17 до 19	20
свыше 19	5
Итого	100

С вероятностью 0,9974 определить *предельную ошибку* выборки и возможные границы, в которых ожидается средняя урожайность подсолнечника в области.

Решение: в условии не указан тип отбора, но исходя из логики исследования, положим, что он *бесповторный*. Поскольку выборка 5%-ная, то объем генеральной совокупности (общая посевная площадь области) составляет:

$N = \frac{100\%}{5\%} \cdot n = 20 \cdot 100 = 2000$ гектаров – не знаю, насколько это реалистично, оставим этот вопрос на совести автора задачи.

По условию, требуется найти *предельную ошибку* выборки (точность оценки) $\Delta = t_\gamma \cdot \mu$, где t_γ – коэффициент доверия, соответствующий доверительной вероятности $\gamma = 0{,}9974$, и коль скоро выборка бесповторна и **генеральной дисперсии мы не знаем**, то *средняя ошибка* рассчитывается по формуле $\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$. Далее нужно составить интервал $\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$, который с вероятностью 99,74% (практически достоверно) накроет *генеральную среднюю* $\bar{x}_Г$ урожайность подсолнечника по области.

И если с коэффициентом «тэ гаммовое» трудностей никаких, то коэффициент «мю» здесь трудовой – по той причине, что нам не известна исправленная выборочная дисперсия s^2. Ну что же, хороший повод освежить пройденный материал. Смотрим на таблицу выше и приходим к выводу, что нам предложен интервальный вариационный ряд с открытыми крайними интервалами. Поскольку длина *частичного интервала* составляет $h = 2$ га, то вопрос закрываем так: 11-13 и 19-21 га.

Находим середины x_i интервалов (переходим к дискретному ряду), произведения $x_i n_i$, $x_i^2 n_i$ и их суммы:

Интервалы	x_i	n_i	
11-13	12	10	
13-15	14	25	
15-17	16	40	
17-19	18	20	
19-21	20	5	
Суммы:		100	

…

Вычислим выборочную среднюю: $\bar{x}_в = \frac{\sum x_i n_i}{n} = \frac{1570}{100} = 15{,}7$ центнеров с гектара.

Выборочную дисперсию вычислим по формуле:

$\sigma_в^2 = \frac{\sum x_i^2 n_i}{n} - (\bar{x}_в)^2 = \frac{25060}{100} - 15{,}7^2 = 250{,}6 - 246{,}49 = 4{,}11$ и этим частенько пренебрегают, но я призываю поправлять дисперсию:

…– мелочь, а приятно.

Теперь составляем доверительный интервал $\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$, где $\Delta = t_\gamma \cdot \mu$.

Найдём *коэффициент доверия* t_γ. Поскольку нам известна лишь исправленная выборочная дисперсия (а не генеральная), то правильнее использовать *распределение Стьюдента*. Но, к сожалению, в таблице нет значений для $\gamma = 0{,}9974$, но зато есть расчётный макет (пункт 2б). Для заданной надёжности и количества степеней свободы $k = n - 1 = 100 - 1 = 99$ получаем $t_\gamma \approx 3{,}0898$. Поскольку объём выборки $n > 30$, то можно использовать нормальное распределение, и тут получается конфетка:

$2\Phi(t_\gamma) = 0{,}9974 \Rightarrow \Phi(t_\gamma) = \frac{0{,}9974}{2} = 0{,}4987 \Rightarrow t_\gamma = 3$, какой способ выбрать – зависит от вашей методички, и я так подозреваю, второй ☺. Но сейчас выберем первый.

Вычислим *среднюю ошибку* бесповторной выборки:

…ц/га, таким образом, предельная ошибка составляет
$\Delta = t_\gamma \cdot \mu \approx 3{,}0898 \cdot 0{,}1986 \approx 0{,}6136$ ц/га, и искомый доверительный интервал:

$\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$

$15{,}7 - 0{,}6136 < \bar{x}_Г < 15{,}7 + 0{,}6136$

$15{,}0864 < \bar{x}_Г < 16{,}3136$ (ц/га) – границы, в которых ожидается средняя урожайность подсолнечника в области с вероятностью $\gamma = 0{,}9974$ (практически достоверно).

Ответ: $\Delta \approx 0{,}6136$ ц/га, $15{,}0864 < \bar{x}_Г < 16{,}3136$ (ц/га)

В рассмотренной задаче можно поставить вопросы, аналогичные Примеру 25, а именно попытаться улучшить исследование, в частности, уменьшить точность оценки Δ. В этом случае для определения необходимого объема выборки используется та же формула $n = \dfrac{s^2 t_\gamma^2 N}{s^2 t_\gamma^2 + \Delta^2 N}$, но она менее достоверна, поскольку в разных выборках мы будем получать разные значения s^2. Такие задачи, однако, встречаются, будьте готовы. Да, и аналогичная формула для повторной выборки: $\Delta = \dfrac{t_\gamma s}{\sqrt{n}} \Rightarrow \sqrt{n} = \dfrac{t_\gamma s}{\Delta} \Rightarrow n = \dfrac{t_\gamma^2 \cdot s^2}{\Delta^2}$.

Пример 28

По результатам 10%-ной бесповторной выборки объёма $n = 50$, найдены выборочная средняя $\bar{x}_в = 107{,}92$ и дисперсия $D_в = 84{,}51$.

а) Найти пределы, за которые с доверительной вероятностью 0,954 не выйдет среднее значение генеральной совокупности.

б) Найти эти пределы, если выборка повторная. Какой способ точнее?

Значение 0,954 обусловлено тем, что автор задачи пощадил студентов, в методичке используется функция Лапласа и получается целое значение t_γ. Решаем самостоятельно!

4.7. Оценка генеральной доли

Быстренько освежим в памяти, что такое *доля*. Вспоминаем N помидоров на базе, среди которых K первосортных. Тогда отношение $\omega_Г = \dfrac{K}{N}$ является *генеральной долей* первосортных помидоров. Однако исследовать все овощи затруднительно, поэтому организуется *представительная* выборка из n помидоров, среди которых первосортных окажется k штук. Отношение $\omega_в = \dfrac{k}{n}$ называется *выборочной долей*.

Выборочная доля является точечной оценкой генеральной доли и не внушает особого доверия, поскольку в разных выборках мы будем получать разные значения $\omega_в$, иногда далёкие от истины. В этой связи более предпочтительно оценить $\omega_Г$ интервалом.

Таким образом, наша задача состоит в том, чтобы найти доверительный интервал:

…– который с заранее заданной *надёжностью* γ накроет истинное значение $\omega_Г$ *генеральной доли*.

Далее для удобства я буду опускать подстрочный индекс у выборочной доли: ω.

Точность оценки Δ (или ***предельная ошибка доли***) рассчитывается по формуле …, где t_γ – *коэффициент доверия*, а μ – ***средняя ошибка доли***.

Для нахождения t_γ корректнее использовать *распределение Стьюдента* (таблицу или макет (пункт 2б)), но на практике в большинстве задач объём выборки $n > 30$ и в ходу *распределение нормальное* с лапласовским соотношением $2\Phi(t_\gamma) = \gamma$.

Средняя ошибка доли определяется так:

$\mu = \sqrt{\frac{\omega(1-\omega)}{n}\left(1-\frac{n}{N}\right)}$ – для бесповторной выборки;

…– для повторной выборки.

В том случае, если генеральная совокупность велика, а выборка малА, то для бесповторной выборки можно использовать и 2-ю формулу, ибо дробь $\frac{n}{N}$ будет близка к нулю. Как видите, формулы очень похожи, только вместо дисперсии у нас тут произведение $\omega(1-\omega)$, и чего томиться, сразу задача:

Пример 29

В целях изучения суточного пробега автомобилей автотранспортного предприятия проведено 10%-ное выборочное обследование 100 автомобилей методом случайного бесповторного отбора, в результате которого получены следующие данные:

Суточный пробег автомобиля, км	Число автомобилей
до 160	12
от 160 до 180	36
от 180 до 200	28
свыше 200	24
Итого	**100**

С вероятностью 0,954 требуется определить *долю* машин в генеральной совокупности с пробегом более 180 км.

Решение: вычислим количество автомобилей с пробегом более 180 км по выборке: $k = 28 + 24 = 52$. Таким образом:

$\omega = \frac{k}{n} = \frac{52}{100} = 0{,}52$ – *выборочная доля* автомобилей с пробегом более 180 километров.

Генеральную долю $\omega_Г$ таких автомобилей оценим с помощью *доверительного интервала*:

$\omega - \Delta < \omega_Г < \omega + \Delta$, где $\Delta = t_\gamma \cdot \mu$ – *предельная ошибка* доли.

Для уровня *доверительной вероятности* $\gamma = 0{,}954$ из соотношения $2\Phi(t_\gamma) = \gamma$ определяем знакомый *коэффициент доверия*:

…

.

…Студентам-экономистам почему-то любят предлагать нежные значения «гамма» (у них эта задача – чуть ли не обязательная по предмету). Причём, в методичках прямо так и пишут без пояснений, что вероятности $\gamma = 0{,}954$ соответствует коэффициент $t = 2$.
И никаких там подстрочных индексов, лапласов или пунктов 1* . Запомнил, и всё. Плохо.

Вычислим *среднюю ошибку* доли. Коль скоро выборка 10%-ная, то объём генеральной совокупности равен $N = \frac{100\%}{10\%} \cdot n = 10 \cdot 100 = 1000$ автомобилей, и для *бесповторной выборки*:

$$\mu = \sqrt{\frac{\omega(1-\omega)}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{0{,}52 \cdot 0{,}48}{100}\left(1 - \frac{100}{1000}\right)} = \sqrt{\frac{0{,}2496}{100} \cdot \frac{9}{10}} \approx 0{,}047$$

Таким образом, *точность оценки* составляет $\Delta = t_\gamma \cdot \mu \approx 2 \cdot 0{,}047 \approx 0{,}095$ и искомый доверительный интервал:

…

$0{,}425 < \omega_Г < 0{,}615$ – с вероятностью 95,4% данный интервал накрывает истинную *генеральную долю* $\omega_Г$ автомобилей с пробегом более 180 км.

Ответ: $0{,}425 < \omega_Г < 0{,}615$

Кстати, тут легко оценить и *абсолютное количество* таковых машин:

$0{,}425 \cdot N < K < 0{,}615 \cdot N$

$0{,}425 \cdot 1000 < K < 0{,}615 \cdot 1000$

$425 < K < 615$ – от 425 до 615 автомобилей в генеральной совокупности.

Но результат это, конечно, слабоватый. И помочь здесь может увеличение объёма выборки. Родственная формула уже выведена в предыдущем параграфе, и я просто заменю дисперсию произведением $\omega(1-\omega)$:

… – здесь по желаемой *предельной ошибке* Δ можно вычислить необходимый объём выборки.

И прямо сейчас у вас представится такая возможность.

На десерт:

Пример 30

Методом механического бесповторного отбора проведено однопроцентное обследование веса пирожных, изготовленных кондитерской фабрикой за сутки. Распределение веса пирожных по весу следующее:

Вес пирожных, г	96-98	98-100	100-102	102-104	Итого
Число пирожных	5	30	60	5	100

а) С вероятностью 0,9974 определить пределы, в которых будет находиться доля пирожных весом не менее 100 г, во всей суточной продукции

б) Сколько процентов пирожных нужно проверить, чтобы увеличить точность оценки в 7 раз? *(при той же доверительной вероятности)* Оценить целесообразность такого статистического исследования.

Краткое решение и ответ в конце книги.

Резюмируя по главе, читаем ВДУМЧИВО:

В результате выборочного исследования *генеральной совокупности* мы получаем различные *выборочные характеристики* (*выборочную среднюю*, *дисперсию*, *долю* и другие показатели).

Задача состоит в том, чтобы определить, ***насколько достоверно*** полученное *выборочное значение* $\theta_в$ характеризует соответствующее *генеральное значение* $\theta_Г$.

$\theta_в$ является *точечной оценкой* $\theta_Г$ и обладает тем недостатком, что в отдельно взятой выборке может оказаться далёким от истины. Поэтому для оценки генерального значения используют *интервальную оценку*:

– *доверительный интервал* $\theta_в - \Delta < \theta_Г < \theta_в + \Delta$, который с заранее заданной *доверительной вероятностью* γ (*надёжностью*) накрывает истинное значение $\theta_Г$

Точность оценки Δ зависит от γ, *объёма выборки* и способа отбора её элементов (*повторный* или *бесповторный*).

Планируя *статистическое исследование*, следует выбрать способ отбора, уровень надёжности γ, а также желаемую точность Δ, по которой легко определить необходимый объём выборки. Слишком высокая точность может быть не оправдана, так как для её достижения потребуется отобрать очень много элементов.

…Лаконично получилось, даже самому понравилось!

5. Статистические гипотезы

Есть ли жизнь после сессии? Далеко не каждая гипотеза является статистической.

5.1. Понятие статистической гипотезы

Пусть исследуется некоторый признак статистической совокупности. Успеваемость студентов, продолжительность жизни лампочек, точность измерений, да что угодно – хоть пятьдесят оттенков серого. Всё, что можно «оцифровать» и подсчитать.

Как проводится исследование? Обычно так: из генеральной совокупности извлекается *репрезентативная выборка* (всё понятно?) и на основании изучения этой выборки делается вывод обо всей совокупности. Напоминаю, что это основной метод математической статистики и называется он *выборочным методом*. В зависимости от исследования, могут проводиться неоднократные выборки, выборки из нескольких ген. совокупностей, да и вообще анализироваться произвольные статистические данные**.**

И в результате обработки этих данных появляются мысли, которые оформляются в ***статистические гипотезы*****.**

Статистической **называют гипотезу** о законе распределения статистической совокупности **либо** о числовых параметрах известных (!) распределений**.**

Например:

– рост танкистов распределен ***нормально*****;**

– дисперсии стрельбы двух танковых дивизий равны между собой, при этом известно, что точность стрельбы распределена нормально***.**

** Из многочисленных ранее проведённых исследований.*

В первом случае выдвигается гипотеза о законе распределения, во втором – о числовых характеристиках двух распределений, закон которых известен.

Откуда взялись эти гипотезы? В первом случае была проведена выборка танкистов (например, 100 человек) и в результате её исследования появилось обоснованное предположение, что рост ВСЕХ танкистов распределён нормально. Во втором случае исследовались *выборочные данные* по точности стрельбы двух дивизий, в результате чего возник интерес проверить – а одинакова ли генеральная результативность, или какая-то дивизия стреляет точнее?

В обеих гипотезах речь идёт о генеральных совокупностях, и выдвигаются эти гипотезы на основании анализа *выборочных* данных. Это распространенная схема, но она не единственна, бывают и другие статистические гипотезы.

А вот такая гипотеза статистической не является:

– жизнь ~~*в танке*~~ *на Марсе существует***.**

Поскольку в ней не идёт речи ни о распределении, ни о параметрах статистической совокупности.

5.2. Нулевая и альтернативная гипотезы

Выдвигаемую гипотезу называют ***нулевой*** и обозначают через H_0. Обычно это наиболее очевидная и правдоподобная гипотеза (хотя это вовсе не обязательно). И в противовес к ней рассматривают ***альтернативную*** или ***конкурирующую*** гипотезу H_1.

! Примечание: *название «нулевая» появилось исторически, оно случайно и к нулю не имеет никакого отношения.*

В примерах с танкистами (см. выше) альтернативные гипотезы очевидны (отрицают нулевую), но существуют и другие варианты, так, к гипотезе H_0: *генеральная средняя нормально распределённой совокупности равна* $a=10$, можно сформулировать разные конкурирующие гипотезы: $H_1: a\neq 10$, либо $H_1: a>10$, либо $H_1: a<10$, или конкретно $H_1: a=11$. Это зависит от условия и данных той или иной задачи.

Так как *нулевая гипотеза* выдвигается на основе анализа **выборочных** данных, то она может оказаться как правильной, так и неправильной. Более того, **мы не сможем на 100% гарантировать её истинность либо ложность даже после *статистической проверки***! Ибо любая, самая «надёжная» выборка все равно остаётся выборкой и может нас дезинформировать (пусть с очень малой вероятностью).

Проверка проводится с помощью ***статистических критериев*** – это специальные *случайные величины*, которые принимают различные действительные значения. В разных задачах критерии разные, и мы рассмотрим их в конкретных примерах.

В результате проверки нулевая гипотеза либо принимается, либо отвергается в пользу альтернативной. При этом есть риск допустить ошибки двух типов:

5.3. Ошибки первого и второго рода

Ошибка первого рода состоит в том, что гипотеза H_0 будет отвергнута, хотя на самом деле она правильная. Вероятность допустить такую ошибку называют ***уровнем значимости*** и обозначают буквой α («альфа»).

Ошибка второго рода состоит в том, что гипотеза H_0 будет принята, но на самом деле она неправильная. Вероятность совершить эту ошибку обозначают буквой β («бета»). Значение $1-\beta$ называют ***мощностью критерия*** – это вероятность отвержения неправильной гипотезы.

В практических задачах, как правило, задают *уровень значимости*, наиболее часто выбирают значения $\alpha=0{,}1$, $\alpha=0{,}05$, $\alpha=0{,}01$.

И тут возникает мысль, что чем меньше «альфа», тем вроде бы лучше. Но это только вроде: **при уменьшении вероятности** α - *отвергнуть правильную гипотезу* **растёт вероятность** β - *принять неверную гипотезу* (при прочих равных условиях). Поэтому перед исследователем стоит задача грамотно подобрать соотношение вероятностей α и β, при этом учитывается **тяжесть последствий**, которые повлекут за собой та и другая ошибки.

Понятие ошибок 1-го и 2-го рода используется не только в статистике, и для лучшего понимания я приведу пару нестатистических примеров.

Петя зарегистрировался в почтовике. По умолчанию, H_0 – он считается добропорядочным пользователем. Так считает антиспам фильтр. И вот Петя отправляет письмо. В большинстве случаев всё произойдёт, как должно произойти – нормальное письмо дойдёт до адресата (правильное принятие нулевой гипотезы), а спамное – попадёт в спам (правильное отвержение). Однако фильтр может совершить ошибку двух типов:

1) с вероятностью α ошибочно отклонить нулевую гипотезу *(счесть нормальное письмо за спам и Петю за спаммера)* или

2) с вероятностью β ошибочно принять нулевую гипотезу *(хотя Петя редиска)*.

Какая ошибка более «тяжелая»? Петино письмо может быть ОЧЕНЬ важным для адресата, и поэтому при настройке фильтра целесообразно уменьшить уровень значимости α, «пожертвовав» вероятностью β (увеличив её). В результате в основной ящик будут попадать все «подозрительные» письма, в том числе особо талантливых спаммеров. …Такое и почитать даже можно, ведь сделано с любовью :)

Существует примеры, где наоборот – более тяжкие последствия влечёт ошибка 2-го рода, и вероятность α следует увеличить (в пользу уменьшения вероятности β). Не хотел я приводить подобные примеры, и даже отшутился на сайте, но по какой-то мистике через пару месяцев сам столкнулся с непростой дилеммой. Видимо, таки, надо рассказать:

У человека появилась серьёзная болячка. В медицинской практике её принято лечить (основное «нулевое» решение). Лечение достаточно эффективно, однако не гарантирует результата и более того опасно (иногда приводит к серьёзному пожизненному увечью). С другой стороны, если не лечить, то возможны осложнения и долговременные функциональные нарушения.

Вопрос: что делать? И ответ не так-то прост – в разных ситуациях разные люди могут принять разные решения (упаси вас).

Если болезнь не особо «мешает жить», то более тяжёлые последствия повлечёт ошибка 2-го рода – когда человек соглашается на лечение, но получает фатальный результат (принимает, как оказалось, неверное «нулевое» решение). Если же…, нет, пожалуй, достаточно, возвращаемся к теме:

5.4. Процесс проверки статистической гипотезы

состоит из следующих этапов:

1) Обработка выборочных данных и выдвижение основной H_0 и конкурирующей H_1 гипотез.

2) Выбор статистического критерия K. Это **непрерывная случайная величина**, принимающая различные действительные значения. В разных задачах критерии разные.

3) Выбор уровня значимости α, о дилемме выбора этого значения я только что рассказал выше.

4) Нахождение ***критического значения*** $k_{кр.}$ – это значение случайной величины K, которое зависит от выбранного *уровня значимости* α и опционально от других параметров. Критическое значение определяет ***критическую область***. Она бывает левосторонней, правосторонней и двусторонней, и, например, может располагаться так *(красная штриховка)*:

…

Критическая область – это ***область отвержения нулевой гипотезы***. Незаштрихованную область называют ***областью принятия гипотезы***.

Следует отметить, что это только одна из графических моделей. Существуют статистические критерии, которые принимают далеко не все действительные значения.

5) Далее на основании выборочных данных рассчитывается ***наблюдаемое значение критерия***: $k_{набл.}$. **И выносится вердикт:**

– Если $k_{набл.}$ в критическую область НЕ попадает, то гипотеза H_0 *на уровне значимости* α **принимается**.

Однако это не означает, что нулевая гипотеза истинна

Ведь существует β-вероятность того, что мы совершили ошибку 2-го рода (приняли неверную гипотезу).

– Если $k_{набл.}$ в *критическую область* **попадает**, то гипотеза H_0 *на уровне значимости* α отвергается.

Однако это не означает, что нулевая гипотеза непременно ложна

Ведь существует α-вероятность того, что мы совершили ошибку 1-го рода (отвергли верную гипотезу). Так, если был выбран уровень значимости $\alpha = 0{,}05$, то *в среднем* в 5 случаях из 100 мы отвергнем правильную гипотезу.

…Ну а что делать? – такая вот статистика неточная наука :)

И по горячей информации сразу разберём одну из наиболее распространённых гипотез:

5.5. Гипотеза о генеральной средней нормального распределения

Постановка задачи такова: предполагается, что генеральная средняя a *нормального распределения* равна некоторому значению a_0. Это нулевая гипотеза:

$H_0 : a = a_0$

Для проверки гипотезы на уровне значимости α проводится выборка объема n и рассчитывается выборочная средняя $\bar{x}_в$. Исходя из полученного значения и специфики той или иной задачи, можно сформулировать следующие конкурирующие гипотезы:

1) …
2) …
3) …
4) $H_1 : a = a_1$, где a_1 – конкретное альтернативное значение *генеральной средней*.

При этом возможны две принципиально разные ситуации:

➢ **Если генеральная дисперсия σ^2 известна**

Тогда в качестве статистического критерия K рассматривают *случайную величину* …, где $\overline{X}$ – случайное значение *выборочной средней*. Почему случайное? Потому что в разных выборках мы будем получать разные значения $\bar{x}_в$, и заранее предугадать это значение невозможно.

Далее находим критическую область. Для конкурирующих гипотез $H_1 : a < a_0$ и $H_1 : a = a_1$ (случай $a_1 < a_0$) строится *левосторонняя область*, для гипотез $H_1 : a > a_0$ и $H_1 : a = a_1$ (случай $a_1 > a_0$) – *правосторонняя*, и для гипотезы $H_1 : a \neq a_0$ – *двусторонняя* – по той причине, что конкурирующее значение генеральной средней может оказаться как больше, так и меньше a_0-го.

Чтобы найти критическую область нужно отыскать критическое значение $u_{кр.}$. Оно определяется из соотношения … – для односторонней области *(лево- или право-)* и $\Phi(u_{кр.}) = \frac{1-\alpha}{2}$ – для двусторонней области, где α – выбранный уровень значимости, а $\Phi(u)$ – старая знакомая функция Лапласа.

Теперь на основании выборочных данных рассчитываем наблюдаемое значение критерия:

…

Это можно было сделать и раньше, но такой порядок более последователен и логичен.

Интерпретация результатов зависит от типа критической области:

1) Для левосторонней критической области. Если $u_{набл.} > -u_{кр.}$, то гипотеза H_0 на уровне значимости α принимается. Если $u_{набл.} < -u_{кр.}$, то отвергается. И картинки тут недавно были, просто заменю букву:

…

2) Правосторонняя критическая область. Если $u_{набл.} < u_{кр.}$, то гипотеза H_0 принимается, в случае $u_{набл.} > u_{кр.}$ *(красный цвет)* – отвергается:

…

3) Двусторонняя критическая область. Если $-u_{кр.} < u_{набл.} < u_{кр.}$ *(незаштрихованный интервал)*, то гипотеза H_0 принимается, в противном случае – отвергается:

…

Условие принятия гипотезы здесь часто записывают компактно – с помощью *модуля*: $|u_{набл.}| < u_{кр.}$

И немедленно приступаем к задачам, а то по студенческим меркам я тут уже на пол диссертации наговорил:)

Пример 31

Из нормальной генеральной совокупности с известной дисперсией $\sigma^2 = 3{,}2$ извлечена выборка объёма $n = 25$ и по ней найдена выборочная средняя $\bar{x}_в = 19{,}3$. Требуется на уровне значимости 0,01 проверить нулевую гипотезу $H_0 : a = 20$ против конкурирующей гипотезы $H_1 : a = 19$.

Прежде чем приступить к решению, пару слов о смысле такой задачи. Есть *генеральная совокупность* с известной дисперсией и есть веские основания полагать, что *генеральная средняя* равна 20 (нулевая гипотеза). В результате выборочной проверки получена *выборочная средняя* 19,3, и возникает **вопрос**: это результат случайный или же генеральная средняя и на самом деле меньше двадцати? – в частности, равна 19 (конкурирующая гипотеза).

Решение: **по условию, известна генеральная дисперсия** $\sigma^2 = 3{,}2$, **поэтому** для проверки гипотезы $H_0 : a = a_0 = 20$ используем случайную величину ….

Найдём *критическую область*. Для этого нужно найти *критическое значение*. Так как конкурирующее значение $H_1 : a = a_1 = 19$ **меньше** чем $a_0 = 20$, то критическая область будет левосторонней (см. теоретический материал выше).

Критическое значение определим из соотношения $\Phi(u_{кр.}) = \dfrac{1-2\alpha}{2}$. Для уровня значимости $\alpha = 0{,}01$:

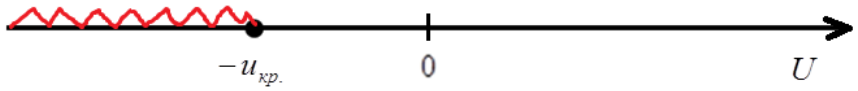
$$\Phi(u_{кр.}) = \frac{1-2\cdot 0{,}01}{2}$$

$$\Phi(u_{кр.}) = \frac{1-0{,}02}{2}$$

$$\Phi(u_{кр.}) = \frac{0{,}98}{2}$$

$$\Phi(u_{кр.}) = 0{,}49$$

По таблице значений функции Лапласа или с помощью Макета (пункт 1*) определяем, что этому значению функции соответствует аргумент $u_{кр.} \approx 2{,}33$. Таким образом, при $u < -u_{кр.}$ *(красная критическая область)* нулевая гипотеза отвергается, а при $u > -u_{кр.}$ – принимается:



В данном случае $-u_{кр.} \approx -2{,}33$.

Вычислим *наблюдаемое значение критерия*:

…

$u_{набл.} > -u_{кр.}$, поэтому **на уровне значимости** $\alpha = 0{,}01$ **нулевую гипотезу** $H_0 : a = 20$ **принимаем**.

Такой, вроде бы неожиданный результат, объясняется тем, что генеральное стандартное отклонение достаточно великО: $\sigma = \sqrt{3{,}2} \approx 1{,}79$, а посему нет оснований отвергать «главное» значение $a_0 = 20$ (несмотря на то, что выборочная средняя $\bar{x}_в = 19{,}3$ гораздо ближе к конкурирующему значению $a_1 = 19$). Иными словами, такое значение выборочной средней, вероятнее всего, объясняется естественным разбросом *вариант* x_i.

Ответ: на уровне значимости 0,01 нулевую гипотезу принимаем.

Что означает *«на уровне значимости 0,01»*? Это означает, что мы с 1%-ной вероятностью рисковали отвергнуть нулевую гипотезу, при условии, что она действительно справедлива. Не забываем, что **на самом деле она всё же может быть и неверной**, т.к. существует β-вероятность того, мы приняли неправильную гипотезу.

Примеры расчёта *мощности критерия* $1-\beta$ для заданного уровня значимости α и различных конкурирующих значений можно найти, например, в учебном пособии и задачнике В. Е. Гмурмана (поздние издания). Это более редкая задача, на которой я не останавливаюсь в своём курсе, ибо его цель – разобрать наиболее «ходовые» задачи **и, главное** – **заинтересовать вас математической статистикой!**

То была «обезличенная» задача, коих очень много, но мы будем менять мир к лучшему… физическими и химическими способами:) Заодно и понятнее будет, что здесь к чему:

Пример 32

По результатам $n=5$ измерений температуры в печи найдено $\bar{x}_в = 256°C$. Предполагается, что ошибка измерения есть нормальная случайная величина с $\sigma = 6°C$. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : a = 250°C$ против конкурирующей гипотезы $H_1 : a > 250°C$.

Сначала разберём, в чём жизненность этой ситуации. Есть печка. Для нормального технологического процесса нужна температура 250 градусов. Для проверки этой нормы 5 раз измерили температуру, получили 256 градусов. Из многократных предыдущих опытов известно, что среднеквадратическая погрешность измерений составляет 6 градусов (она обусловлена погрешностью самого термометра и другими случайными обстоятельствами)

И здесь не понятно, почему выборочный результат (256 градусов) получился больше нормы – то ли температура действительно выше и печь нуждается в регулировке, то ли это просто погрешность измерений, которую можно не принимать во внимание.

Решение: по условию, **известно генеральное среднее квадратическое отклонение** $\sigma = 6$, **поэтому** для проверки гипотезы $H_0 : a = a_0 = 250$ используем случайную величину ….

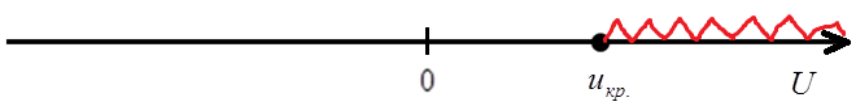
Найдём критическую область. Так как в конкурирующей гипотезе $H_1 : a > 250$ речь идёт о бОльших значениях температуры, то эта область будет правосторонней. Критическое значение определим из соотношения $\Phi(u_{кр.}) = \frac{1-2\alpha}{2}$. Для уровня значимости $\alpha = 0{,}05$:

$$\Phi(u_{кр.}) = \frac{1-2\cdot 0{,}05}{2}$$

$$\Phi(u_{кр.}) = \frac{1-0{,}1}{2}$$

$$\Phi(u_{кр.}) = \frac{0{,}9}{2} = 0{,}45$$

По таблице значений функции Лапласа или с помощью Макета (пункт 1*) определяем, что $u_{кр.} \approx 1{,}645$. Таким образом, при $u > u_{кр.}$ *(красный цвет)* нулевая гипотеза отвергается, а при $u < u_{кр.}$ – принимается:



Вычислим наблюдаемое значение критерия:

…

$u_{набл.} > u_{кр.}$, поэтому **на уровне значимости** $\alpha = 0{,}05$ **нулевую гипотезу** $H_0 : a = 250$ **отвергаем**.

Как бы сказали статистики, выборочный результат $\bar{x}_в = 256°C$ ***статистически значимо*** отличается от нормативного значения $250°C$, и печь нуждается в регулировке (для уменьшения температуры).

Ответ: на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : a = 250$ отвергаем.

Ещё раз осмыслим – что означает *«на уровне значимости 0,05»*? Это означает, что с вероятностью 5% мы отвергли правильную гипотезу (совершили ошибку 1-го рода). И тут остаётся взвесить риск – насколько критично чуть-чуть уменьшить температуру (если мы всё-таки ошиблись и температура на самом деле в норме). Если даже небольшое уменьшение температуры недопустимо, то имеет смысл провести повторное, более качественное исследование: увеличить количество замеров n, использовать более совершенный термометр, улучшить условия эксперимента и т.д.

Следующая задача для самостоятельного решения, и на всякий случай я ещё раз продублирую ссылку на таблицу значений функции Лапласа и Макет:

Пример 33

Средний вес таблетки сильнодействующего лекарства (номинал) должен быть равен 0,5 мг. Выборочная проверка $n = 100$ выпущенных таблеток показала, что средний вес таблетки равен $\bar{x}_в = 0{,}508$ мг. Многократными предварительными опытами на фармацевтическом заводе установлено, что вес таблеток распределен нормально со средним квадратическим отклонением $\sigma = 0{,}11$ мг. На уровне значимости $\alpha = 0{,}1$ проверить гипотезу о том, что средний вес таблеток действительно равен $H_0 : a = 0{,}5$.

РассмотрИте как конкурирующую гипотезу $H_1 : a > 0{,}5$, так и гипотезу $H_1 : a \neq 0{,}5$. И в самом деле – ведь полученное значение $\bar{x}_в = 0{,}508$ является случайным и в другой выборке оно может запросто оказаться и меньше чем 0,5.

Краткое решение, как обычно, в конце книги.

Кстати, это ещё один пример, где ошибка 2-го рода *(ошибочное принятие неверной нулевой гипотезы)*, может повлечь гораздо более тяжелые последствия (опасную передозировку). Поэтому в такой ситуации лучше включить паранойю и увеличить уровень значимости до $\alpha = 0{,}1$ – при этом мы будем чаще отвергать правильную нулевую гипотезу (совершать ошибку 1-го рода), но зато перестрахуемся и проведём более тщательное исследование.

Можно ли одновременно уменьшить вероятности ошибок 1-го и 2-го рода? (значения α и β)

Да можно. Если увеличить объём выборки. Что совершенно логично.

Теперь вторая ситуация. Та же задача, почти всё то же самое, но:

➢ Генеральная дисперсия НЕ известна

Если значение σ^2 не известно, то остаётся ориентироваться на исправленную выборочную дисперсию s^2 и критерий …, где $\overline{X}$ – случайное значение *выборочной средней*, а S – соответствующее исправленное стандартное отклонение. Данная случайная величина имеет *распределение Стьюдента* с $k = n-1$ степенями свободы. Алгоритм решения полностью сохраняется:

Пример 34

На основании $n = 7$ измерений найдено, что средняя высота сальниковой камеры равна $\bar{x}_в = 51$ мм и $s = 1{,}2$ мм. В предположении о нормальном распределении проверить на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : a = 50$ мм против конкурирующей гипотезы $H_1 : a \neq 50$ мм.

И начнём мы опять со смысла задачи. Согласно норме, высота сальниковой камеры должна равняться 50 мм. Но по выборке из 7 измерений получено *среднее значение* 51 мм и за неимением *генеральной дисперсии* вычислена *исправленная выборочная дисперсия*. Возникает вопрос: выборочный результат случаен или нет?

Решение: **так как генеральная дисперсия не известна**, то для проверки гипотезы $H_0 : a = a_0 = 50$ используем случайную величину ….

…

Конкурирующая гипотеза имеет вид $H_1 : a \neq a_0$, а значит, речь идёт о двусторонней критической области. *Критическое значение* можно найти по таблице распределения Стьюдента либо с помощью Макета (пункт 2в). Для уровня значимости $\alpha = 0{,}05$ и количества степеней свободы $k = n-1 = 7-1 = 6$:

$$t_{кр.} = t_{двуст.кр.}(\alpha; k) = t_{двуст.кр.}(0{,}05; 6) \approx 2{,}45$$

Таким образом, при $-t_{кр.} < t < t_{кр.}$ нулевая гипотеза принимается, и вне этого интервала (в критической области при $t > |t_{кр.}|$) – отвергается:

… …

Вычислим *наблюдаемое значение критерия*:

$$t_{набл.} = \frac{(\bar{x}_в - a_0)\sqrt{n}}{s} = \frac{(51-50)\sqrt{7}}{1{,}2} = \frac{(51-50)\sqrt{7}}{1{,}2} \approx 2{,}20$$ – полученное значение попало в *область принятия гипотезы* $(-2{,}45 < t < 2{,}45)$, поэтому **на уровне значимости 0,05 нулевую гипотезу принимаем**.

Ответ: на уровне значимости 0,05 гипотезу $H_0 : a = 50$ мм принимаем.

Иными словами, с точки зрения статистики, выборочный результат $\bar{x}_в = 51$ мм, скорее всего (! но это не точно), обусловлен погрешностью выборки, и на самом деле высота сальниковой камеры соответствует норме (50 мм).

Творческая задача для самостоятельного решения:

Пример 35

Нормативный расход автомобильного двигателя составляет 10 л на 100 км. После конструктивных изменений, направленных на уменьшение этого показателя, были получены следующие результаты 10 тестовых заездов:

Расход топлива, л /100 км.	9,64	10,12	9,7	9,45	10,33	9,93	9,34	10,03	9,63	9,33

На уровне значимости 0,05 выяснить, действительно ли расход топлива стал меньше.

Да, это не редкость – когда нужно не только проверить гипотезу, но и предварительно рассчитать выборочные значения. Следует отметить, что **даже при известной генеральной дисперсии, ориентироваться на неё тут нельзя**, ибо конструктивные изменения могут изменить не только генеральную среднюю, но и генеральную дисперсию.

И в лучших традициях книги, все числа уже забиты в Эксель – там же инструкция по расчётам выборочных показателей. Если кто-то что-то запамятовал, то **вот ролик о том, как провести эти вычисления быстро** *(Ютуб)*.

В данной задаче критическая область левосторонняя, и критическое значение $t_{кр.} = t_{одн.кр.}(\alpha; k)$ для односторонней области отыскивается по самой нижней строке таблицы или с помощью Макета (тот же пункт 2в). Постарайтесь грамотно оформить решение, образец в конце книги. Продолжаем.

Как отмечалось в начале главы, статистической является гипотеза либо о законе распределения статистической совокупности либо о числовых параметрах известных распределений, и начали мы со второй группы. Таких гипотез **воз и маленькая тележка**, и самые популярные из них я только что разобрал. Теперь перейдём к 1-му типу гипотез:

5.6. Гипотеза о законе распределения генеральной совокупности

Рассмотрим генеральную совокупность, распределение которой неизвестно. Однако есть основание полагать, что она распределена по некоторому закону Z (чаще всего, **нормально**). Это предположение (об этом поговорим позже) может появиться как до, так и в результате статистического исследования, когда мы извлекли и изучили выборку объёма n.

И нам требуется на уровне значимости α проверить нулевую гипотезу H_0 – о том, что **генеральная совокупность распределена по закону** Z против конкурирующей гипотезы H_1 о том, что **она по нему НЕ распределена**.

Как проверить эту гипотезу?

Постараюсь объяснить кратко. Как мы выяснили ранее, выборочные данные группируются в дискретный или интервальный вариационный ряд с *вариантами* x_i и соответствующими *частотами* n_i, – картинка на следующей странице:

x_i	n_i
x_1	n_1
x_2	n_2
x_3	n_3
…	…
x_m	n_m
$\sum =$	n

Так как эти данные взяты из практического опыта, то выборочный вариационный ряд называют ***эмпирическим рядом***, а частоты n_i – ***эмпирическими частотами***.

Далее строятся графики, рассчитываются выборочные характеристики (выборочная средняя $\bar{x}_в$, выборочная дисперсия $\sigma_в^2$ и другие). На основе некоторых выборочных характеристик по специальным формулам, **которые зависят от проверяемого закона** Z, строится ***теоретическое распределение***, где для тех же *вариант* $x_1, x_2, x_3, ..., x_m$ рассчитываются ***теоретические частоты*** $n'_1, n'_2, n'_3, ..., n'_m$. **Эти частоты моделируют закон** Z **и наилучшим образом приближают выборочные данные**. При этом $\sum n'_i$ чуть меньше либо равна $\sum n_i = n$.

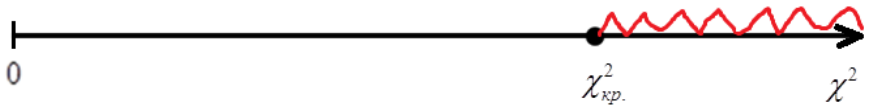
И возникает **вопрос**: *значимо* или *незначимо* различие между эмпирическими $n_1, n_2, n_3, ..., n_m$ и соответствующими теоретическими $n'_1, n'_2, n'_3, ..., n'_m$ частотами? Для ответа на это вопрос рассматривают различные статистические критерии, которые называют ***критериями согласия***, и наиболее популярный из них

5.7. Критерий согласия Пирсона

Спасибо, Карл: ……Всем понятно, почему величина χ^2 случайная? – По той причине, что в разных выборках мы будем получать разные, заранее непредсказуемые эмпирические частоты.

При достаточно большом n *(объёме выборки)* распределение этой *случайной величины* близкО к *распределению хи-квадрат* с количеством степеней свободы $k = m - r - 1$, где r – количество оцениваемых параметров закона Z.

Далее строится правосторонняя критическая область:



Критическое значение $\chi^2_{кр.} = \chi^2_{кр.}(\alpha; k)$ можно найти с помощью соответствующей таблицы или Макета (пункт 3б).

Наблюдаемое значение критерия рассчитывается по *эмпирическим* и найденным *теоретическим частотам*: …

Если $\chi^2_{набл.} < \chi^2_{кр.}$, то на уровне значимости α **нет оснований отвергать гипотезу** H_0 **о том, что генеральная совокупность распределена по закону** Z. То есть, различие между эмпирическими и теоретическими частотами *незначимо* и, скорее всего, обусловлено случайными факторами (случайностью самой выборки, способом отбора, группировки данных и т.д.)

Если $\chi^2_{набл.} > \chi^2_{кр.}$, то нулевую гипотезу отвергаем, иными словами эмпирические и теоретические частоты отличаются *значимо*, и это различие вряд ли случайно.

И, наконец, коровы, которые нас уже заждались. Реалистичность фактических данных оставлю на совести автора методички сельскохозяйственной академии:

Пример 36

По результатам выборочного исследования найдено распределение средних удоев молока в фермерском хозяйстве (литров) от одной коровы за день:

Литры	7,5-10,5	10,5-13,5	13,5-16,5	16,5-19,5	19,5-22,5	22,5-25,5	25,5-28,5	28,5-31,5	31,5 - 34,5
Коров	2	6	10	17	33	11	9	7	5

На уровне значимости 0,05 проверить гипотезу о том, что генеральная совокупность *(средний удой коров **всей** фермы)* распределена нормально. Построить эмпирическую гистограмму и теоретическую кривую.

…Если не любите молоко, то пусть это будет чай, сок, пиво или другой напиток, который вам нравится :) Чтобы было интереснее исследовать эту волшебную ферму.

Решение: на уровне значимости α проверим гипотезу H_0 о *нормальном распределении* генеральной совокупности против конкурирующей гипотезы H_1 о том, что она так НЕ распределена. Используем критерий согласия Пирсона $\chi^2 = \sum \frac{(n_i - n_i')^2}{n_i'}$.

Эмпирические частоты известны из предложенного интервального ряда, и осталось найти теоретические. Для этого нужно вычислить *выборочную среднюю* $\bar{x}_в$ и *выборочное стандартное отклонение* $\sigma_в$. Выберем в качестве *вариант* x_i середины частичных интервалов *(длина каждого интервала* $h = 3$*)* и **заполним расчётную таблицу**:

Интервалы	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
7,5-10,5	9	2	18	162
10,5-13,5	12	6	72	864
13,5-16,5	15	10	150	2250
16,5-19,5	18	17	306	5508
19,5-22,5	21	33	693	14553
22,5-25,5	24	11	264	6336
25,5-28,5	27	9	243	6561
28,5-31,5	30	7	210	6300
31,5-34,5	33	5	165	5445
Суммы:		100	2121	47979

Вычислим выборочную среднюю: $\bar{x}_в = \frac{\sum x_i n_i}{n} = \frac{2121}{100} = 21{,}21$ литра.

Выборочную дисперсию вычислим по формуле:
…

И выборочное стандартное отклонение: $\sigma_в = \sqrt{D_в} = \sqrt{29{,}9259} \approx 5{,}47$ литра, по причине большого объёма выборки его исправлением можно пренебречь.

Теоретические частоты рассчитываются по формуле:

…, где $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ – ***функция Гаусса***, а $z_i=\frac{x_i-\bar{x}_в}{\sigma_в}$.

Входные данные известны: $n=100$, $h=3$, $\bar{x}_в=21{,}21$, $\sigma_в\approx 5{,}47$ и мы заполняем ещё одну расчётную таблицу:

x_i	n_i	z_i	$f(z_i)$	n'_i
9	2	-2,2320	0,0330	1,81
12	6	-1,6836	0,0967	5,30
15	10	-1,1352	0,2095	11,49
18	17	-0,5868	0,3358	18,42
21	33	-0,0384	0,3986	21,86
24	11	0,5100	0,3503	19,21
27	9	1,0584	0,2279	12,50
30	7	1,6068	0,1097	6,02
33	5	2,1552	0,0391	2,14

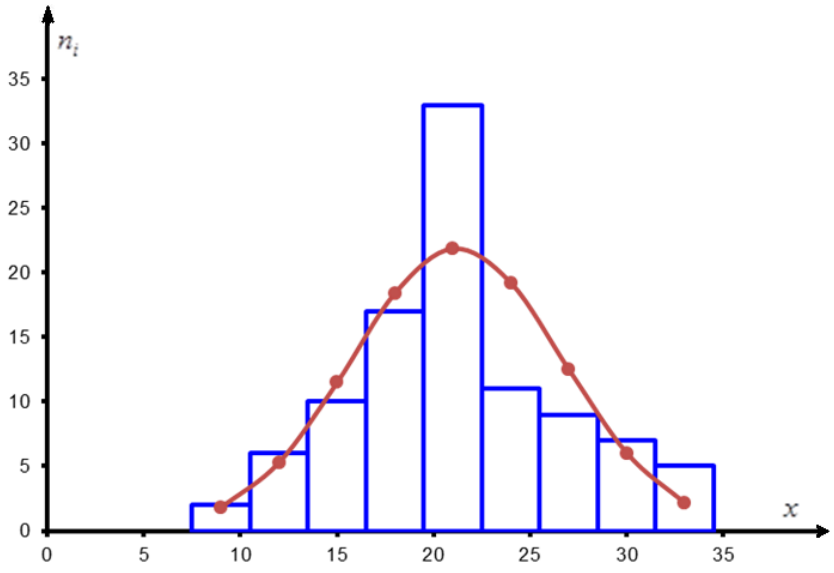
вычисления удобно проводить в Экселе и на всякий случай я распишу 1-ю строчку:

$z_1=\frac{x_1-\bar{x}_в}{\sigma_в}\approx\frac{9-21{,}21}{5{,}47}\approx -2{,}230$

$f(z_1)\approx f(-2{,}23)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(-2{,}23)^2}{2}}\approx 0{,}0330$ – здесь выгодно использовать встроенную экселевскую функцию *=НОРМРАСП(-2,23; 0; 1; 0)*, первый аргумент которой равен текущему значению z_i. За неимением Экселя и калькулятора пользуйтесь таблицей.

И, наконец, теоретическая частота: $n'_1=\frac{h\cdot n}{\sigma_в}\cdot f(z_1)\approx\frac{3\cdot 100}{5{,}47}\cdot 0{,}0330\approx 1{,}81$, довольно часто её округляют до целого значения, но без округления результат всё же точнее.

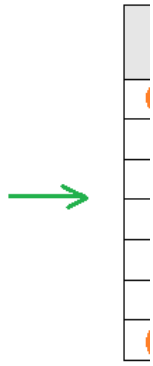
Построим эмпирическую гистограмму с высотой «ступенек» n_i и теоретическую кривую, которая проходит через точки (x_i, n'_i):



Нормальная кривая построена на основе выборочных данных *(выборочной средней и стандартного отклонения)*, и наилучшим образом приближает гистограмму. Дальнейшая задача состоит в том, чтобы оценить, насколько ЗНАЧИМО отличаются эмпирические частоты *(ступеньки гистограммы)* от соответствующих теоретических частот *(уровень коричневых точек)*.

Но перед тем как сравнивать теоретические и эмпирические частоты, следует объединить интервалы с малыми (меньше пяти) частотами. В данном случае объединяем два первых и два последних интервала, для этого суммируем частоты, обведённые красным цветом, и получаем оранжевые результаты:

n_i	n'_i
2	1,81
6	5,30
10	11,49
17	18,42
33	21,86
11	19,21
9	12,50
7	6,02
5	2,14



Это нужно для того, чтобы сгладить неоправданно большое расхождением между малыми частотами по краям выборки. Действие не обязательное, но крайне желательное, ибо студентов на моей памяти из-за этого заставляли переделывать задание.

Найдём *критическое значение* $\chi^2_{кр.} = \chi^2_{кр.}(\alpha; k)$ критерия согласия Пирсона. Количество степеней свободы определяется по формуле $k = m - r - 1$, где m – количество интервалов, а r – количество оцениваемых параметров рассматриваемого закона распределения.

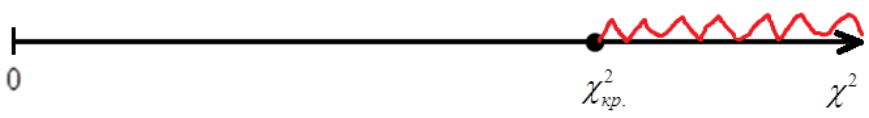
Так как мы объединяли интервалы, то теперь их не девять, а $m = 7$.
У нормального закона мы оцениваем $r = 2$ параметра.

Пояснение: *$\bar{x}_в$ – это оценка неизвестного генерального матоожидания, а $\sigma_в$ – это оценка неизвестного генерального стандартного отклонения, итого два оцениваемых параметра.*

Таким образом, $k = 7 - 2 - 1 = 4$ и для уровня значимости $\alpha = 0{,}05$:
$\chi^2_{кр.} = \chi^2_{кр.}(0{,}05;\ 4) \approx 9{,}4877$

Это значение можно найти по таблице критических значений распределения хи-квадрат или с помощью Макета (Пункт 3б).

При $\chi^2_{набл.} > \chi^2_{кр.}$ нулевая гипотеза отвергается, а при $\chi^2_{набл.} < \chi^2_{кр.}$ таких оснований нет *(заметьте, что формулировка не утверждает истинность гипотезы!)*:



Вычислим *наблюдаемое значение критерия … (суть – сумму расхождений между частотами)*, для этого заполним ещё одну расчётную табличку:

n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
8	7,12	0,1100
10	11,49	0,1923
17	18,42	0,1092
33	21,86	5,6746
11	19,21	3,5087
9	12,50	0,9778
12	8,16	1,8053
Сумма:		12,3779

На всякий пожарный пример расчёта:….

В нижней строке таблицы у нас получилось готовое значение $\chi^2_{набл.} \approx 12{,}3779 > \chi^2_{кр.}$, поэтому **на уровне значимости 0,05 гипотезу H_0 о нормальном распределении генеральной совокупности отвергаем**.

Иными словами **различие между эмпирическими и теоретическими частотами статистически значимо и вряд ли объяснимо случайными факторами**. При этом с вероятностью 5% мы совершили ошибку 1-го рода (то есть, ген. совокупность на самом деле распределена нормально, но мы отвергли верную нулевую гипотезу).

Ответ: на уровне значимости 0,05 гипотезу о нормальном распределении генеральной совокупности отвергаем

В чём может быть причина? Ведь по *теореме Ляпунова*, большинство коров не оказывают практически никакого влияния на удой других коров, и поэтому распределение ген. совокупности должно быть близкО к нормальному.

Причины могут быть разными. Например, неоднородный состав совокупности (коровы разной породы), или на ферме есть VIP-хлев, где коровы получают улучшенное питание :) А может быть, некоторые коровы больны и как раз оказывают существенное влияние на остальных, в связи с чем нарушается условие теоремы Ляпунова.

Интересно отметить, что при уменьшении уровня значимости до 0,01 критическое значение $\chi^2_{кр.} = \chi^2_{кр.}(0{,}01;\, 4) \approx 13{,}277$, и гипотеза о нормальном распределении уже принимается. Однако не нужно забывать, что здесь выросла β-вероятность того, что мы приняли неправильную гипотезу (совершили ошибку 2-го рода). С оценкой этой вероятности можно ознакомиться в специализированной литературе по статистике.

И, конечно, в случае сомнений имеет смысл увеличить объём выборки, чтобы провести повторное исследование.

Да, и **видео по вычислениям**! Хотя особой технической новизны тут нет.

Рассмотренная задача может встретиться в более простой или более сложной формулировке. В версии-«лайт» вам предложат готовые теоретические частоты, где остаётся только проверить гипотезу. Продвинутое же условие звучит примерно так:

На основании исследования выборки выдвинуть гипотезу о законе распределения генеральной совокупности

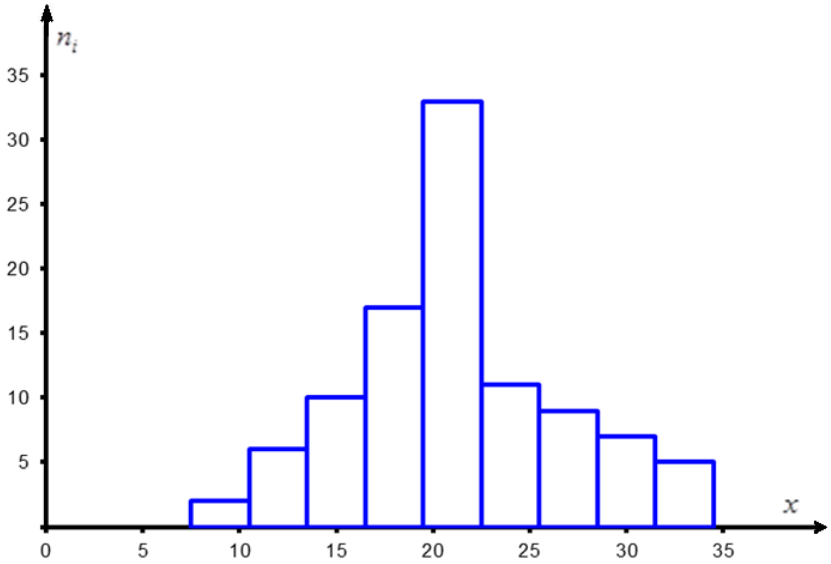
То есть, здесь **не говорится о том, что предполагаемый закон нормальный** (или какой-то другой) – этот вопрос вам предлагается проанализировать самостоятельно.

Каким образом это можно сделать?

Во-первых, гипотезу можно выдвинуть априорно, даже не исследуя выборку. В частности, на основании упомянутой выше *теоремы Ляпунова*: если каждый объект совокупности оказывается ничтожно малое влияние на всю совокупность, то её распределение близкО к нормальному.

Это утверждение носит статус теоремы! То есть, строго доказано в теории.

Но по условию, требуют опираться на *выборочные данные*, и здесь есть сразу несколько признаков, чтобы «вычислить» этот закон. Самый простой и наглядный способ – графический. Грубо говоря, чертим и смотрим. Интервальный вариационный ряд чаще всего изображают *гистограммой*, возвращаемся к нашим коровам:



Построенная гистограмма по форме напоминает колоколообразный **график плотности нормального распределения**, и это является веской причиной предположить, что генеральная совокупность распределена нормально. Да, здесь есть слишком высокий средний столбик, но, возможно, это просто случайность выборки.

Если столбики примерно одинаковы по высоте, то предполагаем, что генеральная совокупность распределена **равномерно**. Для **показательного распределения** тоже будет своя, характерная гистограмма.

В случае дискретных распределений тоже никаких проблем – строим полигон и смотрим, на что он похож.

Следующие признаки аналитические, приведу их для нормального распределения:

1) У нормального распределения *математическое ожидание* совпадает с *модой* и *медианой*. В нашем случае соответствующие выборочные показатели весьма близкИ друг к другу (матожидание оценивается выборочной средней):

$\bar{x}_в = 21{,}21, \quad M_0 \approx 20{,}76, \quad m_e \approx 20{,}86$ (литры)

Желающие могут рассчитать моду и медиану самостоятельно. Впрочем, желающими часто становятся поневоле, поскольку задача, которую мы рассматриваем, нередко идёт в комплексе со всеми этими заданиями.

2) Выполнение ***правила «трёх» сигм***. Практически все значения нормальной случайной величины находятся в интервале $(a-3\sigma; a+3\sigma)$. Найдём этот интервал для нашей выборки. Матожидание «а» оценивается *выборочной средней* $\bar{x}_в = 21{,}21$, а стандартное отклонение «сигма» – *выборочным стандартным отклонением* $\sigma_в \approx 5{,}47$.Таким образом, наш эмпирический интервал:

…
… – и в него действительно попадают все коровы!

3) Кроме того, есть ещё **коэффициенты асимметрии и эксцесса** нормального распределения, которые не вошли в этот курс

На практике в исследование желательно включить все пункты за исключением, возможно, третьего (т.к. *асимметрию* и *эксцесс* рассчитывают далеко не всегда).

Следует отметить, что перечисленные выше предпосылки ещё не означают, что соответствующая гипотеза будет принята, в чём мы недавно убедились. А если гипотеза и окажется принятой, то это всё равно на 100% не гарантирует нормальность распределения *(так как существует β-вероятность принять неверную гипотезу (ошибка 2-го рода))*.

И, конечно, задача для самостоятельного решения, передаю привет студентам Университета путей сообщения:

Пример 37

В результате проверки 500 контейнеров со стеклянными изделиями установлено, что число повреждённых изделий X имеет следующее эмпирическое распределение:

x_i	0	1	2	3	4	5	n
n_i	270	166	49	10	3	2	500

(x_i – количество повреждённых изделий в контейнере, n_i – количество контейнеров) …Здесь тоже представьте изделия по своему интересу :)

С помощью критерия согласия Пирсона на уровне значимости 0,05 проверить гипотезу о том, что случайная величина X – *число повреждённых изделий* распределена **по закону Пуассона**.

Перелистываем страницу и читаем инструкцию по решению.

Все числа забиты в Эксель, придерживайтесь следующего алгоритма:

1) Находим выборочную среднюю $\bar{x}_в$. Это значение будет точечной оценкой параметра «лямбда» теоретического распределения $p(i) \approx \frac{\lambda^i}{i!} \cdot e^{-\lambda}$.

2) Находим значения … для $i = 0,1,2,3,4,5$. Вычисления можно проводить на обычном калькуляторе, но удобнее использовать экселевскую функцию =ПУАССОН(*i; «икс выборочное»; 0*).

3) Находим теоретические частоты $n'_i = p(i) \cdot n$

4) Находим *критическое значение* $\chi^2_{кр.} = \chi^2_{кр.}(\alpha; k)$ *критерия согласия Пирсона*, где $k = m - r - 1$. В данной задаче мы объединяем две последние варианты ввиду их малых частот, следовательно, $m = 5$. Оценивается один параметр («лямбда»), поэтому $r = 1$.

5) Рассчитываем *наблюдаемое значение критерия* $\chi^2_{набл.} = \sum_{i=1}^{m} \frac{(n_i - n'_i)^2}{n'_i}$.

6) Делаем вывод.

Примерный образец чистового оформления задачи в конце книги.

Ну а в конце главы **ещё раз** быстренько **осознАем** то, что мы изучили ☺

Статистические гипотезы формулируются на основе исследования *выборки* и предназначены для *статистической проверки* предполагаемого закона *ген. совокупности* **либо** параметров распределений, законы которых известны.

Сначала выдвигается *нулевая* (обычно наиболее правдоподобная) гипотеза и *альтернативная* к ней (*конкурирующая*) гипотеза.

Нулевая гипотеза подлежит *статистической проверке*. Проверка осуществляется с помощью различных *статистических критериев* (специальных *случайных величин*) которые зависят от условия той или иной задачи.

В результате проверки *нулевая гипотеза* может быть принята либо отвергнута, **но это не означает её истинность или ложность!** Ибо всегда есть риск допустить ошибку:

– *Ошибка 1-го рода* состоит в том, что гипотеза будет отвергнута, но на самом деле она правильная (соответствует действительности). Вероятность совершить ошибку 1-го рода обозначают через α («альфа») и называют *уровнем значимости*; его задают заранее.

– *Ошибка 2-го рода* состоит в том, что гипотеза будет принята, но на самом деле она неверная. Вероятность совершить ошибку 2-го рода обозначают буквой β («бета»). Вероятность отвержения неверной гипотезы $1 - \beta$ называют *мощностью критерия*.

Уменьшая α, мы увеличиваем β (и наоборот), поэтому перед исследованием нужно подобрать оптимальное соотношение этих вероятностей – оно обычно зависит от **тяжести последствий**, которые влекут ошибки 1-го и 2-го рода. Чтобы **одновременно** уменьшить эти вероятности, нужно увеличить объём выборки.

6. Группировка данных

Этот элементарный материал я хотел включить в первую главу, но он там оказался «не в тему», поскольку сам открывает большую тему :)

Рассмотрим некоторую статистическую совокупность, например, множество студентов ВУЗа. Очевидно, данное множество можно исследовать как единое целое – подсчитать общее количество студентов, вычислить их средний возраст, среднюю успеваемость и другие показатели. Благо, *статистических данных* – море. Но всё это ***общие*** характеристики. А хотелось бы деталей. И в таких случаях совокупность целесообразно разделить на ***группы***, то есть выполнить ***группировку***.

***Группировка* – это разделение статистической совокупности** (не важно, генеральной или выборочной) **на группы по одному или бОльшему количеству признаков.** И разделить её можно по-разному –

6.1. Основные виды группировок

– во-первых, выделить **качественно однородные** группы. Например, разбить студентов ВУЗа на лиц М и Ж пола и есть ещё ныне модный пункт «не определился» (для кого это важно). Такую *группировку* называют ***типологической***. Или, как вы любите говорить, «типа логической» :) Кстати, студенты уже по факту разделены на факультеты – и это тоже пример типологической группировки, но уже по другому признаку. **Итак:**

***типологическая группировка* – это разделение неоднородной статистической совокупности на качественно однородные группы.**

Само собой, полученные *группы* исследуются по отдельности и сравниваются – как между собой, так и с *общими* показателями. При этом проводится ***структурная группировка* – это разделение качественно однородной совокупности по какому-либо** *вариационному* **признаку**. По росту, весу, уровню IQ, скорости движения, периоду полураспада и так далее. Признаков – тьма.

Да будет свет! – в качестве простейшего условного примера рассмотрим среднюю успеваемость студентов ВУЗа: $\bar{x} = 3{,}75$ (общая средняя). Однако это не слишком информативный показатель.

Гораздо интереснее провести *типологическую группировку*, например, разделить всех студентов на «физиков» и «лириков», и подсчитать ***групповые средние***: $\bar{x}_{Ф} = 3{,}6$, $\bar{x}_{Л} = 3{,}9$. Ну вот, теперь прекрасно видно, кому в универе жить хорошо :) Или рассчитать *групповые средние* по факультетам: $\bar{x}_1 = 3{,}66$; $\bar{x}_2 = 3{,}25$; $\bar{x}_3 = 4{,}05$; ..., $\bar{x}_k = 3{,}93$. И выяснить, почему это на 2-м факультете такая низкая успеваемость по сравнению со средней успеваемостью $\bar{x} = 3{,}75$ по ВУЗу.

Довольно часто грань между типологической и структурной группировкой стирается. Приведу избитый, но показательный пример с банками. Все банки можно разделить на мелкие, средние и крупные (*типологическая группировка*). Но с другой стороны, эти категории основаны на количественном показателе, мелкие – меньше одного литра, средние – от одного до трёх, и крупные – больше трёх литров. То есть, это одновременно и *структурная группировка*. Эксперты центробанки гарантируют ☺

Помимо перечисленных, существуют и другие виды группировок, в частности, аналитическая группировка и комбинационная группировка. Но о них позже, после практической разминки. Ранее мы уже неоднократно проводили группировку данных, давайте вспомним пару примеров:

Пример 4

По результатам выборочного исследования рабочих цеха были установлены их квалификационные разряды: 4, 5, 6, 4, 4, 2, 3, 5, 4, 4, 5, 2, 3, 3, 4, 5, 5, 2, 3, 6, 5, 4, 6, 4, 3.

...

В этой задаче дана однородная совокупность – рабочие цеха, и нами была проведена их *структурная группировка* по разряду, в результате чего нарисовался дискретный вариационный ряд:

x_i	n_i
2	3
3	5
4	8
5	6
6	3

где x_i – разряды, а n_i – количество рабочих того или иного разряда

Пример 6

По результатам исследования цены некоторого товара в различных торговых точках города, получены следующие данные (в некоторых денежных единицах):

7,5	7,6	8,7
6,1	10,6	9,8
7	6	8,3
6	8,2	8,5
7,4	7,1	9,5
6,8	9,6	6,3
6,3	8,5	5,8
7,5	9,2	7,2
7	8	7,5
7,5	8	6,5

...

В этом примере мы тоже провели *структурную группировку* (товаров по их цене) и получили интервальный вариационный ряд:

Диапазон цен	n_i
5,7-6,7	7
6,7-7,7	11
7,7-8,7	6
8,7-9,7	4
9,7-10,7	2

где n_i – количество товаров из того или иного ценового интервала.

И сейчас мы продолжим группировать данные. В предположении того, что студент сможет разделить собак и котов (типологическая группировка), ему обычно предлагают провести *структурную* и / или *аналитическую группировку*. Разберём их по порядку.

6.2. Структурная группировка

Напоминаю, это группировка качественно однородной совокупности по числовому признаку. Примеры только что были выше, и мы продолжаем. Суровая задача местного Политеха для студентов около- и машиностроительных специальностей:

Пример 38

В результате выборочного исследования 30 станков рассчитаны их относительные показатели металлоёмкости (т/кВт):

6	1,1818	1,6667
3,3333	3,75	0,4
0,3333	0,5556	2,6667
0,15	0,6923	1,6667
1,2609	2,5	1,2
0,875	2,1667	0,5
0,5789	1,4286	2
2	0,5	0,8571
2,1429	8	0,9333
0,8182	2,3333	6

Требуется:

а) вычислить *общую (выборочную) среднюю*;

б) выполнить *структурную равноинтервальную группировку*;

в) выполнить *структурную равнонаполненную группировку*;

г) выбрать наиболее удачную группировку и вычислить *выборочные средние*; результаты оформить в виде групповой таблицы;

д) по выбранной группировке построить *интервальный вариационный ряд*;

е) сделать краткие выводы.

Но прежде, немного о содержании. Согласно автору методички, относительная металлоемкость – это частное от деления веса станка на мощность его двигателя (тонн на киловатт). Разделили, например, 5 тонн на 2 кВт и получили 2,5 тонны на один кВт. Эти значения и представлены в таблице. Правильность и достоверность перечисленных фактов я снова оставлю на совести автора, да и, в конце концов, нам требуется обработать числа, а уж что это такое – не особо важно, хоть объём талии пчёлок. …И всё-таки математика немного шизофреническая наука :)

Решение:

ну, с пунктом **а)** справится даже неподготовленный человек. Очевидно, что для нахождения выборочной средней нужно просуммировать все значения и разделить полученный результат на *объём выборки*:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_{30}}{n} = \frac{6 + 3{,}3333 + 0{,}3333 + ... + 6}{30} = \frac{58{,}4913}{30} \approx 1{,}95 \text{ т/кВт}$$ *(не забываем указать размерность!)*

Эти и другие вычисления лучше проводить в Экселе, и чуть ниже будет ролик о том, как быстро выполнить все пункты задания. Ибо на калькуляторе щёлкать 30 слагаемых муторно (хотя, вариант вполне рабочий).

➢ Равноинтервальная группировка

б) Выполним *структурную равноинтервальную группировку*. Пугаться не нужно, это задание уже было – нам нужно построить обычный интервальный вариационный ряд с равными интервалами, и я кратко повторю алгоритм.

В условии ничего не сказано о количестве интервалов, и поэтому для определения их оптимального количества используем формулу Стерджеса:

… интервалов *(результат округляем влево)*.

Найдём минимальное $x_{min}=0,15$ и максимальное $x_{max}=8$ значения и вычислим размах вариации: $R=x_{max}-x_{min}=8-0,15=7,85$ т/кВт. Таким образом, длина *частичного интервала* составит: $h=\frac{R}{k}=\frac{7,85}{5}=1,57$ т/кВт. Теперь «нарезаем» интервалы и подсчитываем количество станков n_i в каждом из них:

Интервалы (т/кВт)	n_i
0,15 - 1,72	18
1,72 - 3,29	7
3,29 - 4,86	2
4,86 - 6,43	2
6,43 - 8	1

Контроль: $\sum n_i=18+7+2+2+1=30=n$, что и требовалось проверить.

И уже сейчас мы видим, что построенный вариационный ряд не слишком хорош – по той причине, что в трёх последних интервалах слишком мало станков, и считать по ним средние значения и другие показатели не вполне корректно. Во избежание этого недостатка используют разные методы, в частности, другой метод группировки:

➢ Равнонаполненная группировка

в) Это разбиение совокупности на группы с одинаковым (или примерно одинаковым) количеством объектов, станков в данном случае. Но интервалы здесь получатся разной длины.

Отсортируем числа по возрастанию и выделим 5 групп по $\frac{n}{k}=\frac{30}{5}=6$ станков в каждой:

0,15	0,5789	1,1818	2	2,6667
0,3333	0,6923	1,2	2	3,3333
0,4	0,8182	1,2609	2,1429	3,75
0,5	0,8571	1,4286	2,1667	6
0,5	0,875	1,6667	2,3333	6
0,5556	0,9333	1,6667	2,5	8

Формально всё выглядит тип-топ (и можно оставить так), но некоторые значения логичнее перенести в соседние группы. Так, значение 0,5789 *(верхняя строка)* явно ближе к 1-й группе, а значение 2,6667 – к предпоследней группе; туда их и перенесём:

0,15	0,6923	1,1818	2	3,3333
0,3333	0,8182	1,2	2	3,75
0,4	0,8571	1,2609	2,1429	6
0,5	0,875	1,4286	2,1667	6
0,5	0,9333	1,6667	2,3333	8
0,5556		1,6667	2,5	
0,5789			2,6667	

г) Очевидно, что равнонаполненная группировка более удачна, с ней и работаем. **По каждой группе** подсчитаем сумму, *объём* (количество станков) и *выборочную среднюю* – как результат деления суммы на соответствующий объём. Вычисления сведём в ***групповую таблицу***:

1
0,1
0,333
0,
0,
0,
0,555
0,578
3,017
Ко
7
Сре
0,431

…

И на всякий пожарный пример расчёта первой *групповой средней*:

$$\bar{x}_1 = \frac{0{,}15 + 3{,}3333 + 0{,}4 + 0{,}5 + 0{,}5 + 0{,}5556 + 0{,}5789}{7} = \frac{3{,}0178}{7} \approx 0{,}4311 \text{ т/кВт};$$

Да, кстати, **не забываем** предварительно проконтролировать объём выборки:

$\sum n_i = 7 + 5 + 6 + 7 + 5 = 30 = n$, ч.т.п.

д) Построим интервальный вариационный ряд по равнонаполненной группировке. Границы интервалов можно брать как *средние арифметические* «стыковых» значений, например: $\frac{0{,}5789 + 0{,}6923}{2} = 0{,}6356$ *(граница между 1-м и 2-м интервалом)*. Но вполне допустимо (и даже лучше) разметить интервалы «на глазок», выбирая удобные «круглые» значения:

Металлоемкость, т/кВт	Количество станков в группе, n_i
0,15-0,6	7
0,6-1,1	5
1,1-1,8	6
1,8-3	7
3-8	5

Полученный интервальный ряд имеет разную длину интервалов, но для него точно так же можно построить гистограмму, эмпирическую функцию распределения, а также рассчитать типовые характеристики. Правда, с модой проблема будет и для её нахождения таки лучше использовать равноинтервальную группировку *(пункт б)*.

Теперь **смотрим ролик** по быстрому и эффективному выполнению расчётов.

Выражаясь научно, мы выполнили ***статистическую сводку***.

Статистическая сводка – это комплекс действий по обработке статистических данных с целью получения обобщающих показателей и анализа стат. совокупности.

Причём, в пункте **а)** была ***простая*** *статическая сводка* (подсчёт общих показателей), которая переросла в *сводку* ***сложную***, включающую в себя *группировку данных*, расчёт групповых характеристик и сведение результатов в *групповую таблицу*. Но это ещё не всё:

е) Сделаем краткие выводы.

Я не случайно выделил данный пункт. Довольно часто в заданиях по статистике требуется сделать выводы – в них нужно **отразить основные результаты выполненных действий и особенности исследуемой совокупности**. Собственно, **это и есть цель статистического исследования – сделать выводы.**

И за нами дело не станет. Сказать здесь можно следующее. В результате исследования рассчитана средняя металлоёмкость $\bar{x} \approx 1{,}95$ т/кВт по выборке и средние значения по группам равнонаполненной (наиболее удачной) группировки. Большинство станков (18 шт. в первых трёх группах) имеют показатель металлоёмкости меньший, чем средняя металлоёмкость по выборке. Пять станков (группа 5) обладают значительно бОльшей металлоёмкостью, чем остальные, и причины этого требуют отдельного анализа (возможно, станки морально устарели).

Несколько строчек вполне достаточно, даже многовато получилось. Но это на пользу – грамотный аналитик или чиновник должен мастерски уметь «лить воду» ☺. …Вот видите, какой полезный курс….

Следующее задание для самостоятельного решения:

Пример 39

По результатам выборочного исследования 50 предприятий получены данные об их квартальной прибыли (числа в экселевском файле), млн. руб. Требуется:

1) вычислить среднюю прибыль;

2) провести равнонаполненную группировку и вычислить *групповые средние*;

3) построить соответствующий вариационный ряд;

4) сделать выводы.

Вообще, здесь удобно разбить выборку на 5 интервалов (и такой вариант вполне себе неплох), но от греха подальше лучше использовать формулу Стерджеса, что я и сделал в образце решения, который, как обычно, находится в конце книги. Ваш вариант решения может немного отличаться от моей версии. И выводы, разумеется, тоже.

Теперь вернёмся к пункту «бэ» Примера 38, где была выполнена не слишком удачная равноинтервальная группировка, скопирую табличку сверху:

Интервалы (т/кВт)	n_i
0,15 - 1,72	18
1,72 - 3,29	7
3,29 - 4,86	2
4,86 - 6,43	2
6,43 - 8	1

Как вы помните, от «куцых» последних интервалов мы избавились, выполнив равнонаполненную группировку. Но есть и другой метод «лечения», который называется

6.3. Перегруппировка

– это *вторичная группировка*, которая состоит в преобразовании уже построенного вариационного ряда. И одним из инструментов перегруппировки является ***укрупнение интервалов***. В примере выше можно просто объединить три последних интервала, и коль скоро нам известны *первичные* (исходные) данные, то заодно подкорректируем границы всех интервалов до удобных значений:

Интервалы (т/кВт)	n_i
0,15 - 1,72	18
1,72 - 3,29	7
3,29 - 4,86	2
4,86 - 6,43	2
6,43 - 8	1

…

Не так, конечно, получилось подробно, как в равнонаполненной группировке, но тоже вполне наглядно. При желании, к слову, первый интервал легко ***измельчить***, получив нечто близкое или даже совпадающее с этой группировкой. Благо, исходные числа в нашем распоряжении.

Но что делать, если *первичные данные* не известны?

Пример 40

Перегруппируйте данные о численности работающих на 55 предприятиях, образовав следующие группы: до 400, 400-1000, 1000-3000, 3000-6000, свыше 6000:

Численность рабочих, чел.	Число предприятий
до 100	4
100-500	8
500-1000	5
1000-2000	14
2000-5000	15
5000-10000	5
10000-20000	3
свыше 20000	1
Итого:	55

В этой задаче мы не знаем исходные *варианты* (конкретную численность рабочих по предприятиям), но **решение** есть! **Держите исходную таблицу перед глазами** *(распечатайте или перепишите на листок)* и ВНИМАТЕЛЬНО вникайте в суть:

1) Выделим новый промежуток «до 400» *(красный цвет на рисунке ниже)*. В него, понятно, войдёт интервал «до 100» (4 предприятия) и часть интервала «100-500», а именно часть, выделенная коричневым цветом:

…

Теперь длину коричневой части $(400-100=300)$ нужно сопоставить с длиной всего интервала «100-500», которая составляет $500-100=400$:

$\frac{300}{400}=\frac{3}{4}$ – таким образом, *три четверти* предприятий интервала «100-500» следует отнести в пользу промежутка «до 400»: $\frac{3}{4}\cdot 8=6$.

Итого в промежутке «до 400» оказывается $4+\frac{3}{4}\cdot 8=4+6=10$ предприятий.

…Вроде всё просто, а объяснить довольно сложно :) Соответственно, на кусок «400-500» останется $8-6=2$ предприятия. Выражаясь академично, этот принцип можно называть ***выделением пропорциональных долей***. Доли выделяются пропорционально длинам частей интервала

2) Выделим новый промежуток «400-1000». В него войдёт оставшийся старый «кусок» «400-500» с 2 предприятиями и старый интервал «500-1000» с 5 предприятиями:

…

Итого на промежутке «400-1000» оказалось $2+5=7$ предприятий.

3) Выделим новый промежуток «1000-3000». В него полностью войдёт старый интервал «1000-2000» с 14 предприятиями и *одна треть* старого интервала с «2000-5000» с $\frac{1}{3}\cdot 15=5$ предприятиями:

…

Нужную долю (одну треть) мы нашли как отношение длины коричневого интервала $(3000-2000=1000)$ к длине интервала «2000-5000» $(5000-2000=3000)$:

$\frac{1000}{3000}=\frac{1}{3}$

Таким образом, в промежуток «1000-3000» вошло:

$14+\frac{1}{3}\cdot 15=14+5=19$ предприятий.

4) В новый промежуток «3000-6000» входят *две трети* старого интервала «2000-5000» (см. рис. выше), что составляет $\frac{2}{3}\cdot 15=10$ предприятий (или $15-5=10$), и, кроме того, *одна пятая* старого интервала «5000-10000», к которой относится $\frac{1}{5}\cdot 5=1$ предприятие:

…

Одна пятая найдена как отношение длины коричневого интервала «5000-6000» к длине интервала «5000-10000»: $\frac{1000}{5000}=\frac{1}{5}$

Таким образом, в промежуток «3000-6000» вошло $\frac{2}{3}\cdot 15+\frac{1}{5}\cdot 5=10+1=11$ предприятий.

5) И, наконец, в последний новый промежуток «свыше 6000» входят *четыре пятых* старого интервала «5000-10000» (см. рис. выше) или $\frac{4}{5}\cdot 5=4$ предприятия, а также 3 предприятия старого интервала «10000-20000» и 1 предприятие интервала «свыше 20000».

Итого: $4+3+1=8$ предприятий

Перегруппировка завершена, новый вариационный ряд построен:

Численность рабочих, чел.	Число предприятий
до 400	10
400-1000	7
1000-3000	19
3000-6000	11
свыше 6000	8
Итого:	55

И обязательно проконтролируем *объем* выборки, мало ли что-то потерялось или мы где-то обсчитались:

$10+7+19+11+8=55$, в чём и требовалось убедиться.

Следует отметить, что *метод выделения долей*, строго говоря, не точен, и если в нашем распоряжении есть *первичные данные*, то, конечно, ориентируемся на них – в результате с высокой вероятностью получатся немного другие частоты по группам. Но для выборочной совокупности годится и долевая перегруппировка, поскольку от выборки к выборке мы всё равно будем получать разные значения и строить похожие, но всё же разные вариационные ряды.

Перегруппировка часто применятся для того чтобы сопоставить «родственные» совокупности с разными интервалами:

Пример 41

По результатам выборочного исследования двух ~~банок~~ банков получены данные о заработной плате их служащих:

Заработная плата, у.е.	Количество служащих, чел.
до 100	1
100-500	4
500-1000	10
1000-2000	15
2000-5000	32
5000 и более	3
Итого:	**65**

Заработная плата, у.е.	Количество служащих, чел.
до 1000	11
1000-1500	12
1500-2500	14
2500-4200	7
4200-6000	4
6000 и более	2
Всего	**50**

Сравнить уровень з/п в банках, выделив интервалы: до 500, 500-1000, 1000-2000, 2000-3000, 3000-4000, 4000-5000, свыше 5000, и рассчитав относительные частоты по каждому банку. Результаты представить в виде общей таблицы, сделать выводы.

И для удобства есть традиционный эксель-шаблон, не ленимся! Если трудно, то можно использовать рисунки с разметкой интервалов (по образцу предыдущего примера); в образце я ограничился аналитическим решением. Расширяем поле деятельности:

6.4. Аналитическая группировка

Данная группировка позволяет установить *наличие* и *характер* **зависимости** одного вариационного ряда от другого. Это может быть связь между признаками разных *статистических совокупностей* или (что чаще) между признаками одной совокупности:

Пример 42

Имеются выборочные данные о выпуске продукции (млн. руб.) и прибыли (млн. руб.) по 30 предприятиям за некоторый период:

Номер предприятия	Выпуск продукции	Прибыль	Номер предприятия	Выпуск продукции	Прибыль
1.	65	15,7	16.	52	14,6
2.	78	18	17.	62	14,8
3.	41	12,1	18.	69	16,1
4.	54	13,8	19.	85	16,7
5.	66	15,5	20.	70	15,8
6.	80	17,9	21.	71	16,4
7.	45	12,8	22.	64	15
8.	57	14,2	23.	72	16,5
9.	67	15,9	24.	88	18,5
10.	81	17,6	25.	73	16,4
11.	92	18,2	26.	74	16
12.	48	13	27.	96	19,1
13.	59	15,5	28.	75	16,3
14.	68	16,2	29.	101	19,6
15.	83	16,7	30.	76	17,2

Методом ***аналитической группировки*** установить наличие и характер зависимости между стоимостью произведенной продукции и *средней* прибылью предприятий. Результаты оформить в виде *групповой* и ***аналитической*** *таблицы*. Сделать выводы, куда ж без них.

Итак, по условию нам дано **два** вариационных ряда: X – выпуск продукции по предприятиям (в млн. руб.) и Y – прибыль по соответствующим предприятиям (тоже в млн. руб.). При этом очевидно, что **один показатель зависит от другого** – чем больше предприятие выпускает, тем, вероятно, больше у него прибыль. Но всегда ли это так? Нет не всегда. Ведь крупное предприятие может быть и убыточным, может не продать всю продукцию при увеличении её производства. Однако *общая тенденция* состоит в том, что при увеличении выпуска продукции, увеличивается и *средняя* прибыль по предприятиям. Ибо масштаб имеет значение, пекарни – это пекарни, а хлебзаводы – это заводы.

Такая *нежёсткая* зависимость называется ***корреляционной*****. Это зависимость, при которой изменение одного показателя влечёт изменение СРЕДНИХ значений другого показателя.** Этим корреляционная зависимость отличается от *функциональной*, где изменение аргумента оказывает чёткое и безусловное влияние на изменение функции.

Показатель X (выпуск продукции) называется ***факторным*** (причинным) или ***признаком-фактором***. Показатель Y (прибыль) называется ***результативным*** (зависимым, следственным) или ***признаком-результатом***.

Но не всё так просто. Дело в том, что **вышесказанное является лишь нашим предположением**. А вдруг в условии дано 30 каких-нибудь северокорейских заводов, где нет такой зависимости?

Именно поэтому по условию нужно установить **наличие** зависимости между выпуском продукции и прибылью и определить её *характер*. Под **характером** понимается *корреляционность* зависимости и её направление, при этом возможны следующие варианты:

– ***прямая связь*** («чем больше, тем больше» – наш случай);
– ***обратная связь*** («чем больше, тем меньше»);
– отсутствие связи («чем больше, тем так же хаотично»).

И установить всё это нужно методом *аналитической группировки*, которая позволяет выявить наличие (либо отсутствие) и направление *корреляционной связи* между *признаком-фактором* X и *признаком-результатом* Y.

И мы начинаем, наконец, оформлять **решение**:

Прежде всего, нужно определить *признак-фактор* и *признак-результат*. Самостоятельно, на основе логических рассуждений. Тут же высказываем предположение о наличии и направлении предполагаемой *корреляционной связи*. В нашей задаче можно записать примерно следующее:

Очевидно, что *средний размер* прибыли по предприятиям зависит от стоимости выпущенной продукции, при этом, чем больше выпущено продукции, тем выше может быть прибыль. Таким образом, выпуск продукции X является *признаком-фактором*, а прибыль предприятий Y – *признаком-результатом*; предполагаемая *корреляционная зависимость – прямая*.

Обращаю ваше внимание, что данная часть задания является если не обязательной, то строго желательной. Часто в условии прямо запрашивается этот пункт.

Теперь проверяем нашу гипотезу (предположение) методом *аналитической группировки*.

Как выполнить аналитическую группировку?

Сначала нужно упорядочить совокупность по признаку-фактору. Расположим предприятия по возрастанию выпуска продукции *(оранжевый цвет)*:

Номер предприятия	Вы про
3.	
7.	
12.	
16.	
4.	
8.	
13.	
17.	
22.	
1.	
5.	
9.	
14.	
18.	
20.	

…

В Экселе эта сортировка выполняется буквально в пару щелчков, и чуть ниже будет ролик о том, как быстро решить нашу задачу. Номера предприятий можно было опустить, но я оставил их для лучшего понимания выполненного действия. Заметьте, что зависимый показатель является *ведомым*, это означает, что числа в колонке «Прибыль» переместились вслед за числами в колонке «Выпуск продукции».

Теперь выполняем группировку совокупности – опять же по признаку-фактору (выпуску продукции). Поскольку в условии нет никаких указаний на этот счёт, то используем стандартную равноинтервальную группировку.

Размах вариации составляет:
$R = x_{\max} - x_{\min} = 101 - 41 = 60$ млн. руб.

Оптимальное количество интервалов определим по формуле Стерджеса, для объёма совокупности $n = 30$ оно составляет:

… интервалов *(округлили влево)*.

Таким образом, длина каждого интервала: $h = \frac{R}{k} = \frac{60}{5} = 12$ млн. руб., в результате чего у нас получаются следующие интервалы выпуска продукции:

41-53, 53-65, 65-77, 77-89 и 89-101 млн. руб.

Собственно, разносим предприятия по группам и начинаем заполнять групповую таблицу. Напоминаю, что значения, попадающие на «стык» интервалов следует относить в следующий интервал:

Группа 1, выпуск 41-53 млн. руб.	
Выпуск	**Прибыль**
41	12,1
45	12,8
48	13
52	14,6
4	
	52,5
	13,125

…

В нижней части таблицы подсчитываем количество n_i предприятий по каждой группе, сумму значений (**внимание!**) результативного признака (прибыли), и, наконец, *средние значения* $\bar{y}_i$ результативного признака по группам.

Основные результаты представим в виде *аналитической таблицы*:

№ группы	Выпуск продукции, млн. руб.	Кол-во предприятий в группе, n_i	Средняя прибыль на одно предприятие, млн. руб.
1	41-53	4	13,125
2	53-65	5	14,66
3	65-77	12	16,167
4	77-89	6	17,567
5	89-101	3	18,967

Теперь смотрим **короткое видео** о том, как это всё быстро подсчитать.

Сделаем **краткие выводы**: аналитическая группировка показала, что при увеличении стоимости выпущенной продукции наблюдается увеличение средней прибыли по предприятиям. Таким образом, между рассмотренными факторами действительно существует *прямая корреляционная зависимость*. Таким образом, наше предположение (см. начало задания) подтвердилось.

Обращаю внимание, что средние значения (правая колонка) не обязаны строго возрастать *(либо убывать)*, и если вы получите такой результат, то это будет нормально, ведь корреляционная зависимость не является жёсткой. ВажнА *общая тенденция* увеличения *(либо уменьшения)* средних значений. Но может статься, что по итогу вообще не проучится какой-то внятной закономерности. И это тоже результат исследования! С соответствующими выводами о слабой коррел. зависимости либо отсутствии таковой.

И здесь возникает закономерный вопрос: а насколько СИЛЬНО влияет факторный признак на результат, насколько теснА связь между факторами? На этот вопрос отвечают **специальные статические показатели**, о которых я рассказал на сайте. По причине редкости этой не самой простой задачи она не вошла в эту книгу.

Следующее задание для самостоятельного решения:

Пример 43

По результатам выборочного исследования 20 банков известны процентные ставки и соответствующие суммы выданных кредитов:

Номер банка	Процентная ставка, %	Сумма кредита млн. руб.	Номер банка	Процентная ставка, %	Сумма кредита млн. руб.
1	17,5	13,5	11	22,4	5,2
2	20,8	7,6	12	16,1	17,9
3	13,6	25,52	13	17,9	12,3
4	24	2,5	14	21,7	5,4
5	17,5	13,24	15	18	12,18
6	15	20,15	16	16,4	17,1
7	21,1	6,1	17	26	11
8	17,6	13,36	18	18,4	12,12
9	15,8	19,62	19	16,7	16,45
10	18,8	11,9	20	12,2	26,5

Требуется:

1) Определить факторный и результативный признак и выдвинуть предположение о наличии и направлении корреляционной связи между показателями.

2) Методом аналитической группировки проверить наличие корреляционной связи, выборку разбить на 4 группы с равным количеством банков в каждой. Результаты представить в виде групповой и аналитической таблицы. Сделать выводы.

Обратите внимание, что во 2-м пункте **вам прямо указано**, как следует выполнять группировку – **в таких случаях не нужно проявлять самодеятельность** – **строго** следуем указаниям условия. А если решение получится не слишком удачное, то это уже проблемы автора задачи.

Все числа забиты в Эксель и вам осталось быстренько выполнить действия. Решение для сверки в конце книги.

Что ещё можно сказать по теме?

В некоторых задачах *результативных* признаков может быть несколько, как правило, два, например: X – выпуск продукции, Y_1 – прибыль и Y_2 – себестоимость производства. Никаких проблем – сортируем совокупность по признаку-фактору X (выпуску продукции), при этом в Экселе нужно выделить не два, а уже три столбца, о чём я недавно рассказывал в видеоролике. Далее выполняем группировку и рассчитываем *средние значения* прибыли и себестоимости по каждой группе. Делаем выводы. Заметим, кстати, что корреляционная связь $X \to Y_2$, вероятно, *обратная*, поскольку при увеличении выпуска продукции, издержки могут падать (ввиду автоматизации процесса при массовом производстве).

И в заключение параграфа хочу сказать, что показатели вам могут быть предложены самые разные, поэтому при решении подобных задач следует «включать голову» и элементарную логику.

6.5. Комбинационная группировка

Комбинационная группировка **– это группировка статистической совокупности совместно по двум или бОльшему количеству признаков**. Она позволяет выявить устройство совокупности и установить взаимосвязи между её признаками.

Рассмотрим выборку, состоящую из $n=100$ котов, среди которых оказалось 20 грациозных *(менее 4 кг)*, 50 обычных *(4-6 кг)* и 30 толстых *(более 6 кг)*. По существу, перед нами структурная группировка животных по их массе, и это первый признак статической совокупности. Теперь возьмём какой-нибудь второй признак, например, разделим всех котов на злых и добрых :) Признак, кстати, качественный, но при желании его можно «оцифровать», рассмотрев некую экспертную шкалу доброты.

В результате исследования выяснилось, что среди тощих котов 14 злых и 6 добрых, среди обычных – 24 злых и 26 добрых и среди толстых – 7 злых и 23 добрых.

Очевидно, что между этими признаками есть связь. Чем больше масса кота, тем *более вероятно*, что он окажется добрым. Ибо с лишним весом, полным желудком и ~~отрезаны …~~ злиться весьма проблематично. Однако и среди толстых котов тоже есть особи с проблемным характером. Такая *нежёсткая* зависимость называется…, вспоминаем… – правильно! **Корреляционной.**

Полученные данные обычно сводят в ***комбинационную таблицу***:

Масса кота, кг, X	X_1	X_2	Итого, m_j
менее 4			
4 - 6			
более 6			
Итого, m_j			

…

Внимательно изучаем таблицу и обозначения! Это **очень**, ОЧЕНЬ **важно для практики**:

1) *Признак-фактор* X (причину) и его категории располагают в левом столбце *(зелёный цвет)*, а *признак-результат* Y (следствие) и его категории – в «шапке» таблицы *(жёлтый цвет)*. Встречается и расположение наоборот (что с моей точки зрения удобнее), но в практических задачах почему-то в ходу первый вариант. Но мы не будем комплексовать, попробуем и так, и так.

2) В основной части таблицы *(серый цвет)* располагаются собственно результаты группировки – ***совместные групповые частоты*** n_{ij}. Итак, у нас в наличии есть:

$n_{11}=14$ тощих и злых и $n_{12}=6$ тощих и добрых котов;

$n_{21}=24$ обычных и злых и $n_{22}=26$ обычных и добрых котов;

$n_{31}=7$ толстых и злых и $n_{32}=23$ толстых и добрых котов;

Итого: 6 групп.

! Справка: *первый подстрочный индекс означает номер строки (рассматриваем серую область), а второй – номер столбца. Так, значение,* $n_{12}=6$ *расположено в 1-й строке, 2-м столбце, а значение* $n_{31}=7$ *– в 3-й строке, 1-м столбце.*

Сумма всех групповых частот равна *объёму* статистической совокупности:

$$\sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij} = n_{11}+n_{12}+n_{21}+n_{22}+n_{31}+n_{32} = 14+6+24+26+7+23 = 100 = n$$

! Справка: *значок двойного суммирования работает следующим образом: сначала переменная «и» принимает значение* $i=1$ *и переменная «жи» пробегает все свои значения (от 1 до 2), в результате чего получается сумма* $n_{11}+n_{12}$. *Затем первая переменная принимает значение* $i=2$ *и «жи» снова пробегает все свои значения:* $n_{21}+n_{22}$. *И, наконец, для* $i=3$ *получаем сумму* $n_{31}+n_{32}$.

Часто для краткости **пишут** $\sum\sum n_{ij} = n$ или даже используют одинарный значок суммы: $\sum n_{ij} = n$

Заканчиваем разбор таблицы:

3) В правом столбце *(зелёный цвет)* располагаются суммы по строкам (по группам признака-фактора). В нашей совокупности имеется $n_1 = 20$ грациозных, $n_2 = 50$ обычных и $n_3 = 30$ толстых котов. Итого: $\sum n_i = n_1+n_2+n_3 = 20+50+30 = 100$ особей.

В нижней строке *(жёлтый цвет)* подсчитываем суммы по столбцам (по категориям признака-результата): $m_1 = 14+24+7 = 45$ злых и $m_2 = 6+26+23 = 55$ добрых котов. Итого: $\sum m_j = m_1+m_2 = 45+55 = 100$, в чём и требовалось убедиться.

Общая котосумма (объём совокупности) находится в правом нижнем углу: $n = 100$.

Если вы что-то не очень поняли,
то ещё раз ВДУМЧИВО перечитайте объяснения!

Может ли в *комбинационной группировке* быть бОльшее количество факторов? Легко. Так, в нашем примере можно добавить фактор Z – жилищные условия кота (бездомный или домашний). В результате получится *трёхмерная* комбинационная группировка с группами:

тощие, злые и бездомные коты;
тощие, злые и домашние коты;
тощие, добрые и бездомные коты;
тощие, добрые и домашние коты;
обычные, злые и бездомные коты;
…
и так далее, всего 12 групп. Самостоятельно перечислите и представьте все остальные семейства – целый мир получится :)

И, завершая занимательное котоведение, призываю вас не кастрировать своих (и чужих) котов и не топить котят. И мир станет гармоничнее! …Простите за отступление, Майкл Джексон любил детей, а я люблю котов. Да и студентов тоже не тяну за хвосты :)

Поэтому переходим к стандартной студенческой задаче, в которой предлагается простейшая *двумерная комбинационная группировка*:

Пример 44

Имеются выборочные данные о выпуске продукции (млн. руб.) и сумме прибыли (млн. руб.) по 30 предприятиям:

Номер предприятия	Выпуск продукции	Прибыль	Номер предприятия	Выпуск продукции	Прибыль
1.	65	15,7	16.	52	14,6
2.	78	18	17.	62	14,8
3.	41	12,1	18.	69	16,1
4.	54	13,8	19.	85	16,7
5.	66	15,5	20.	70	15,8
6.	80	17,9	21.	71	16,4
7.	45	12,8	22.	64	15
8.	57	14,2	23.	72	16,5
9.	67	15,9	24.	88	18,5
10.	81	17,6	25.	73	16,4
11.	92	18,2	26.	74	16
12.	48	13	27.	96	19,1
13.	59	15,5	28.	75	16,3
14.	68	16,2	29.	101	19,6
15.	83	16,7	30.	76	17,2

Определить *признак-фактор* и *признак-результат* и высказать предположение о наличии и направлении *корреляционной зависимости* между признаками. Выполнить *комбинационную группировку*, разбив значения признака-фактора на 5 равных интервалов, а значения признака-результата – на 3 интервала. Сделать выводы.

Числовые данные я взял из Примера 42, где мы выяснили, что *признаком-фактором* (причиной) является X – выпуск продукции, а *признаком-результатом* (следствием) Y – прибыль предприятий. При увеличении выпуска продукции, очевидно, растёт *средняя* прибыль предприятий, таким образом, предполагаемая корреляционная зависимость – *прямая* («чем больше, тем больше»). И снова подчёркиваю нежёсткость этой зависимости: отдельно взятое предприятие может выпускать много, но «сидеть» в убытках, и наоборот – есть предприятия с небольшим объёмом выпуска, но высокой маржой (прибылью). Однако это всё отклонения от общей тенденции.

Начало **решения** совпадает с началом Примера 42. Упорядочим предприятия **по возрастанию признака-фактора**:

Номер предприятия	Выпуск продукции	Прибыль	Номер предприятия	Выпуск продукции	Прибыль
3.	41	12,1	21.	71	16,4
7.	45	12,8	23.	72	16,5
12.	48	13	25.	73	16,4
16.	52	14,6	26.	74	16
4.	54	13,8	28.	75	16,3
8.	57	14,2	30.	76	17,2
13.	59	15,5	2.	78	18
17.	62	14,8	6.	80	17,9
22.	64	15	10.	81	17,6
1.	65	15,7	15.	83	16,7
5.	66	15,5	19.	85	16,7
9.	67	15,9	24.	88	18,5
14.	68	16,2	11.	92	18,2
18.	69	16,1	27.	96	19,1
20.	70	15,8	29.	101	19,6

Далее мы нашли размах вариации $R = x_{max} - x_{min} = 101 - 41 = 60$ млн. руб. и разбили значения *признака-фактора* на 5 равных интервалов. Длина каждого интервала составила $h = \frac{R}{k} = \frac{60}{5} = 12$ млн. руб., после чего у нас получились следующие группы:

Группа 41-53
Выпуск
41
45
48
52

…

Теперь в каждой группе нужно выделить подгруппы, условно говоря, предприятия с небольшой, средней и высокой прибылью (3 интервала по условию). Для этого берём исходные значения *признака-результата* (прибыли) и сортируем их по возрастанию. Для компактности расположу упорядоченные значения в три колонки:

Прибыль		
12,1	15,7	16,7
12,8	15,8	16,7
13	15,9	17,2
13,8	16	17,6
14,2	16,1	17,9
14,6	16,2	18
14,8	16,3	18,2
15	16,4	18,5
15,5	16,4	19,1
15,5	16,5	19,6

С простейшей экселевской сортировкой, полагаю, проблем уже ни у кого нет. А если таки-есть, то снова гляньте **тут**~~, царь тут~~. Я зануда, в полезном смысле этого слова :)

Вычислим размах вариации: $R_y = y_{max} - y_{min} = 19{,}6 - 12{,}1 = 7{,}5$ млн. руб. и длину каждого интервала: $h_y = \frac{R_y}{k_y} = \frac{7{,}5}{3} = 2{,}5$,… ай, как удачно разделилось! В результате получилось три интервала прибыли: 12,1-14,6; 14,6-17,1 и 17,1-19,6 млн. руб.

Теперь в *групповой таблице* красными галочками помечаем предприятия 1-го интервала, зелёными – предприятия 2-го интервала и синими – 3-го интервала:

Группа 1, 41-53 мл
Выпуск
41
45
48
52
4

…

По каждой из 5 групп подсчитываем количество предприятий с небольшой *(красной)*, средней *(зелёной)* и высокой *(синей)* прибылью. Результаты сведём в *комбинационную таблицу*, при этом значения признака-фактора удобно расположить по горизонтали – в «шапке» таблицы, а значения признака-результата слева по вертикали:

Прибыль Y, млн. руб.	Выпуск продук		
	41-53	53-65	65
12,1-14,6	3	2	
14,6-17,1	1	3	1
17,1-19,6			

…

Следует заметить, что построение комбинационной группировки можно автоматизировать (например, в MS Excel), но в простейших учебных примерах проще выполнить ручной подсчёт частот.

Да, всем ли понятны значения (частоты) в серой области? Частота $n_{12}=2$ означает, что у нас есть два предприятия с выпуском продукции 53-65 млн. руб. и невысокой прибылью (12,1-14,6 млн. руб.). Частота $n_{34}=4$ означает, что в выборке 4 предприятия с выпуском продукции 77-89 млн. руб. и высокой прибылью (17,1-19,6 млн. руб.).

Для самоконтроля подсчитываем суммы по серым столбцам:

$n_1=3+1=4,\quad n_2=2+3=5,\quad n_3=11+1=12,\quad n_4=2+4=6,\quad n_5=3$, всего: $\sum n_i=4+5+6+12+3=30=n$ предприятий. Результаты заносим в нижнюю строку *(см. таблицу ниже)*.

И самое интересное – суммы по серым строкам:

$m_1=3+2=5$ предприятий с небольшой прибылью;

$m_2=1+3+11+2=17$ предприятий со средней прибылью;

$m_3=1+4+3=8$ предприятий с высокой прибылью.

Итого: $\sum m_j=5+17+8=30=n$, что и требовалось проверить. Результаты заносим в правый столбец. Таким образом, итоговая комбинационная таблица выглядит следующим образом:

Прибыль Y, млн. руб.	Выпуск п	
	41-53	53-65
12,1-14,6	3	2
14,6-17,1	1	3
17,1-19,6		
Итого, n_i	4	5

…

Сделаем выводы. **На основании чего? Смотрим, как располагаются частоты** (числа в серой области).

Если *частоты* имеют тенденцию располагаться по диагонали **от левого верхнего до правого нижнего угла**, то между признаками существует *прямая корреляционная зависимость* («чем больше, тем больше»). Это наш случай – по таблице хорошо видно, что с увеличением выпуска продукции растут и средние прибыли предприятий. Готово.

Если частоты имеют тенденцию располагаться по диагонали **от левого нижнего до правого верхнего угла**, то между признаками существует *обратная корреляционная зависимость* («чем больше, тем меньше»).

И, наконец, если частоты расположены хаотично, без явной закономерности, то корреляционная зависимость отсутствует либо является слабой.

И здесь опять возникает **вопрос**: насколько СИЛЬНО влияет признак-фактор на признак-результат**.** Ответ на этот вопрос дают **эмпирические показатели**, и один из них (линейный коэффициент корреляции) мы разберём в этой книге.

На практике в большинстве случаев вам предложат готовую *комбинационную таблицу*, и поэтому задания на *комбинационную группировку* не будет, полагаю, что в случае чего она не вызовет у вас затруднений.

Систематизируем основную информацию по теме, я, кстати, пересказываю её «своими словами» – то, что понял сам ☺:

Группировка статистической совокупности – это её разделение по определённому признаку (или признакам). Данная процедура проводится для более качественного, удобного и детального изучения совокупности. Существуют разные виды группировок:

Типологическая группировка – это разделение неоднородной совокупности на качественно однородные группы.

Структурная группировка – это разделение однородной совокупности на группы по какому-либо варьируемому (числовому) признаку. С технической точки зрения она может быть *равноинтервальной* либо *равнонаполненной* (стандартные варианты).

Аналитическая группировка предназначена для выявления взаимосвязи между вариационными рядами и позволяет установить наличие (отсутствие) *корреляционной зависимости** одного ряда от другого, а также её направление (**подробнее в следующей главе*). Направление может быть *прямым* (чем больше, тем больше) либо *обратным* (чем больше, тем меньше).

Комбинационная группировка – это группировка совокупности **совместно** по двум или бОльшему количеству признаков. Она наглядно показывает структуру совокупности, а также корреляционную зависимость между признаками либо её отсутствие.

7. Элементы корреляционно-регрессионного анализа

«Элементы» – это значит, «чуть-чуть»» ☺ И второе слово заголовка мы только что разобрали. Зависимость является ***корреляционной*, если изменение значений одного показателя влечёт изменение средних значений другого показателя**. По своему направлению она может быть *прямой* («чем больше, тем больше») либо *обратной* («чем больше, тем меньше»). Давайте пару свеженьких примеров:

1) Очевидно, что **чем больше** рост человека, **тем больше** может быть его масса. Однако это не является жёстким правилом, поскольку среди низкорослых людей есть кряжистые, а среди высоких – грациозные. Но общая тенденция состоит в том, что с увеличением роста людей растёт и их *средняя* масса. Это пример *прямой* зависимости.

2) И пример зависимости *обратной*. **Чем больше** процентная ставка в банке, **тем меньше** *средний* размер выданных кредитов – так как заёмщикам становится «не по карману» обслуживать большие займы. Да, конечно, среди них есть более обеспеченные люди, которым ссудят крупную сумму, но это исключение из правила (общей тенденции).

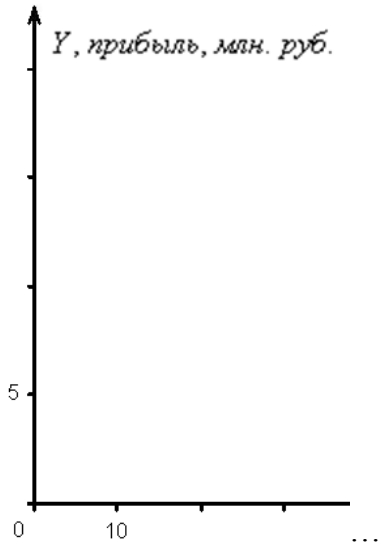
Наличие и направление *корреляционной зависимости* (если она есть) легко установить графически:

7.1. Графическое изображение эмпирических данных

Вернёмся к «избитому» ~~ещё не до конца~~ Примеру 42, где дано 30 предприятий с известными значениями x_i выпуска продукции (признак-фактор X) и соответствующими значениями прибыли y_i (признак-результат Y). По *первичным данным* строится

➢ Диаграмма рассеяния

– это множество точек $(x_i; y_i)$ в декартовой системе координат, абсциссы x_i которых соответствуют значениям признака-фактора X, а ординаты y_i – *соответствующим* значениям признака-результата Y. Вот наши 30 предприятий:



И тут не нужно быть экспертом, чтобы понять, что при увеличении выпуска продукции растут и прибыли предприятий.

Если зависимость *обратная* («чем больше, тем меньше»), то точки имеют тенденцию располагаться наоборот – от левого верхнего угла к правому нижнему. И такой пример будет позже. Если точки распределены по диаграмме примерно равномерно (нет явной закономерности), то корреляционная зависимость слабА либо отсутствует.

Минимальное количество точек должно равняться пяти-шести, в противном случае ***корреляционный анализ*** становится некорректным. А если точек много (30-50 и больше), то этот анализ усложняется и диаграмма «замусоривается». В таких случаях *первичные данные* подвергают группировке, как правило, комбинационной:

Прибыль Y, млн. руб.	Выпуск продукции X, млн. руб.					Итого, m_j
	41-53	53-65	65-77	77-89	89-101	
12,1-14,6	3	2				5
14,6-17,1	1	3	11	2		17
17,1-19,6			1	4	3	8
Итого, n_i	4	5	12	6	3	**30**

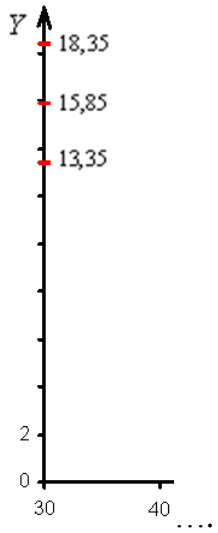
После чего *комбинационную таблицу* упрощают. А именно переходят от интервальных вариационных рядов («шапка» таблицы и левый столбец) к дискретным, выбрав в качестве *вариант* x_i и y_j середины соответствующих интервалов:

Прибыль y_j, млн. руб.	Выпуск продукции x_i, млн. руб.					Итого, m_j
	47	59	71	83	95	
13,35	3	2				5
15,85	1	3	11	2		17
18,35			1	4	3	8
Итого, n_i	4	5	12	6	3	**30**

И, наконец, для сгруппированных данных строят

➢ **Корреляционное поле**

– это множество точек с абсциссами x_i и ординатами y_j, которые соответствуют ненулевым значениям частот n_{ij} комбинационной таблицы:



Мысленно сопоставьте таблицу с картинкой! При этом сами частоты n_{ij} (числа в серых ячейках) на чертеже никак не отмечаются. И по внешнему виду *корреляционного поля* легко понять, что зависимость здесь прямая («чем больше, тем больше»).

Далее. Построенные чертежи наводят нас на светлую мысль, что эмпирические точки **было бы удобно** приблизить некоторой *функцией*, которая удачно характеризует зависимость. И здесь мы подошли к **третьему слову** заголовка главы – «регрессионного».

В статистическом смысле ***регрессия*** **– это зависимость** *средних значений* $\bar{y}_i$ *признака-результата* **от соответствующих значений** x_i *признака-фактора*. Термин «регрессия» появился исторически, и желающие могут найти эту историю в Сети. Если быть лаконичным, то полученные средние значения «игрек» регрессивно возвращают нас к первопричине – соответствующим исходным значениям «икс».

И дело за тем, чтобы найти функцию, которая для различных значений «икс» определяла бы нам средние значения «игрек». В случае *несгруппированных* данных это не самая простая задача, а вот для *комбинационной группировки* есть очевидное решение:

7.2. Эмпирические линии регрессии

По каждой группе *признака-фактора* (5 групп) рассчитаем *средние* значения $\bar{y}_i$ *признака-результата*, результаты удобно свести в дополнительную строку таблицы:

Прибыль y_j, млн. руб.	Выпуск продукции x_i, млн. руб.					Итого, m_j
	47	59	71	83	95	
13,35	3	2				5
15,85	1	3	11	2		17
18,35			1	4	3	8
Итого, n_i	4	5	12	6	3	30
Средние, $\bar{y}_i$	13,975	14,85	16,058	17,5167	18,35	

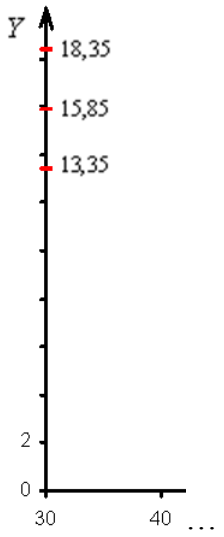
Так, при выпуске $x = x_1 = 47$ млн. средняя прибыль составляет:

$$\bar{y}_1 = \frac{y_1 n_{11} + y_2 n_{21} + y_3 n_{31}}{n_1} = \frac{13{,}35 \cdot 3 + 15{,}85 \cdot 1 + 18{,}35 \cdot 0}{4} = 13{,}975 \text{ млн. руб.}$$

и давайте ещё в качестве закрепляющего примера приведу расчёт для $x = x_4 = 83$:

$$\bar{y}_4 = \frac{y_1 n_{14} + y_2 n_{24} + y_3 n_{34}}{n_4} = \frac{13{,}35 \cdot 0 + 15{,}85 \cdot 2 + 18{,}35 \cdot 4}{6} \approx 17{,}5167 \text{ млн. руб.}$$

Эмпирическая линия регрессии *Y к X (именно так!)* **– это** *ломаная*, **соединяющая точки** $(x_i; \bar{y}_i)$:



Построенная *ломаная* проходит максимально близко к точкам корреляционного поля, при этом учитываются **весомость** частот n_{ij}, на основе которых были вычислены значения $\bar{y}_i$ *(см. вычисления перед чертежом).*

Эмпирическая линия регрессии используется не только для наглядного изображения корреляционной зависимости, но и для ***интерполяции*** промежуточных значений…, сейчас объясню…. Рассматривая различные промежуточные значения выпуска продукции («иксы», отличные от x_i) **с помощью** *ломаной* мы можем достаточно точно оценить соответствующие *средние* значения прибыли («игреки средние»).

Но это ещё не всё. Встречаются ситуации, где признаки X, Y взаимно влияют друг на друга. Приведу философский пример, адаптированный к современным реалиям:)

X – количество произведённых куриц на птицефабрике;
Y – количество произведённых яиц.

Совершенно понятно, что здесь, как признак X влияет на Y, так и наоборот, и поэтому можно рассмотреть вторую корреляционную зависимость – «икса» от игрека. А также построить вторую эмпирическую линию регрессии.

Эмпирическая линия регрессии X ***к*** Y *(именно так!)* **– это** *ломаная*, **соединяющая точки** $(\bar{x}_j; y_j)$, где $\bar{x}_j$ *– средние* значения признака X для различных значений y_j (комбинационной таблицы) признака Y.

И в качестве тренировки

Задание

По данным вышеприведённого примера (30 предприятий) построить эмпирическую линию регрессии X к Y.

Не ленимся! Формулы будут «зеркальными» и вычисления легко провести на обычном калькуляторе. Решение и чертёж в конце книги. Чуть позже я научу вас строить *корреляционное поле* в Экселе.

Наверное, вы заметили, что звенья ломаных расположились почти по прямой, и ещё более ярко эта тенденция прослеживается на диаграмме рассеяния, где точки «выстроились» примерно вдоль прямой линии.

В этой связи возникает **заманчивая идея**: а нельзя ли приблизить эмпирические точки *линейной функцией*?

Можно! (с)

Более того, во многих случаях это будет удачным решением! А, главное, технически простЫм.

Следующий параграф планировалась более 10 лет назад и вот, наконец, я здесь…. И вы здесь! И это замечательно! Даже не то слово. Это корреляционно:

7.3. Модель пАрной линейной регрессии

…И в этот момент я благоговейно улыбаюсь – как здОрово, что все мы здесь сегодня собрались:

Пример 45

Имеются выборочные данные по $n=8$ студентам: X – количество прогулов за некоторый период времени и Y – суммарная успеваемость за этот период:

X	12	9	8	14	15	11	10	15
Y	42	107	100	60	78	79	90	54

Сразу обращаю внимание, что в условии приведены *несгруппированные* данные. Помимо этого варианта, есть задачи, где изначально дана комбинационная таблица, и их мы тоже разберём. Сначала одно, затем другое. **Требуется**:

– высказать предположение о наличии и характере зависимости *признака-результата* Y от *признака-фактора* X;

– построить *диаграмму рассеяния* и сделать вывод о *форме* зависимости;

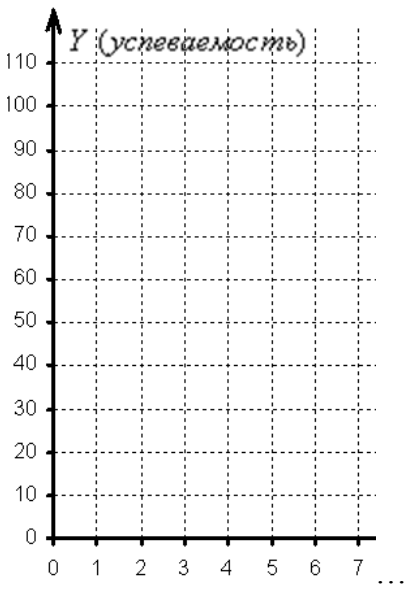
– найти уравнение ***линейной регрессии*** Y на X, выполнить чертёж;

– вычислить ***линейный коэффициент корреляции***, сделать вывод;

– вычислить ***коэффициент детерминации***, сделать вывод,

Решение: Очевидно, что чем больше студент прогуливает, тем *более вероятно*, что у него плохая успеваемость. Но всегда ли это так? Нет, не всегда. Успеваемость зависит от многих факторов. Один студент может посещать все пары, но все равно учиться посредственно, а другой – учиться неплохо даже при большом количестве прогулов (не рекомендация! ☺ – за всю жизнь я встретил двух-трёх человек с такими способностями). Однако *общая тенденция* состоит в том, что с увеличением количества прогулов **средняя** успеваемость студентов будет падать.

Таким образом, предполагаем наличие *обратной корреляционной зависимости* успеваемости Y от количества прогулов X. Гипотезу проще всего проверить графически, построим диаграмму рассеяния:



Обратите, кстати, внимание как раз на тот момент, что при одном и том же количестве прогулов (15) двое студентов имеют существенно разные результаты.

По *диаграмме рассеяния* хорошо видно, что с увеличением числа прогулов успеваемость *преимущественно* падает, что подтверждает наличие *обратной* корреляционной зависимости успеваемости от количества прогулов. Более того, почти все точки «выстроились» примерно вдоль прямой, что даёт основание предположить, что данная зависимость близкА к линейной.

И здесь я анонсирую дальнейшие действия: нам предстоит найти уравнение **прямой**, ТАКОЙ, которая проходит максимально близко к эмпирическим точкам, а также оценить ***тесноту*** (силу) корреляционной линейной зависимости – насколько близко расположены точки к построенной прямой.

Технически существует два пути решения:

– сначала найти уравнение прямой и затем оценить тесноту зависимости;

– сначала найти тесноту и затем составить уравнение.

В практически задачах чаще встречается второй вариант, но я начну с первого, он более последователен. Построим:

➢ **Уравнение линейной регрессии Y на X**

Это и есть та самая оптимальная прямая $y = ax + b$, которая проходит максимально близко к точкам. Обычно её находят **методом наименьших квадратов**, и мы пойдём знакомым путём. Заполним расчётную таблицу:

Прогулы, x_i	Плюшки, y_i	$x_i y_i$	x_i^2	y_i^2
12	42	504	144	1764
9	107	963	81	11449
8	100	800	64	10000
14	60	840	196	3600
15	78	1170	225	6084
11	79	869	121	6241
10	90	900	100	8100
15	54	810	225	2916
$\sum = 94$	**610**	**6856**	**1156**	**50154**

Коэффициенты «а» и «бэ» функции $y = ax + b$ найдём из решения системы:

…, в нашей задаче:

$$\begin{cases} 1156a + 94b = 6856 \\ 94a + 8b = 610 \end{cases}$$

Сократим оба уравнения на 2, всё попроще будет:

$$\begin{cases} 578a + 47b = 3428 \\ 47a + 4b = 305 \end{cases}$$

Систему выгоднее решить **по формулам Крамера**:

$\Delta = \begin{vmatrix} 578 & 47 \\ 47 & 4 \end{vmatrix} = 578 \cdot 4 - 47 \cdot 47 = 2312 - 2209 = 103 \neq 0$, значит, система имеет единственное решение.

$$\Delta_a = \begin{vmatrix} 3428 & 47 \\ 305 & 4 \end{vmatrix} = 3428 \cdot 4 - 305 \cdot 47 = 13712 - 14335 = -623$$

$$a = \frac{\Delta_a}{\Delta} = \frac{-623}{103} \approx -6{,}0485$$

$$\Delta_b = \begin{vmatrix} 578 & 3428 \\ 47 & 305 \end{vmatrix} = 578 \cdot 305 - 47 \cdot 3428 = 176290 - 161116 = 15174$$

$$b = \frac{\Delta_b}{\Delta} = \frac{15174}{103} \approx 147{,}32$$

И проверка forever, подставим полученные значения $a \approx -6{,}0485$, $b \approx 147{,}32$ в левую часть каждого уравнения **исходной** системы:

$1156 \cdot (-6{,}0485) + 94 \cdot 147{,}32 = -6992{,}066 + 13848{,}08 \approx 6856$

$94 \cdot (-6{,}0485) + 8 \cdot 147{,}32 = -568{,}559 + 1178{,}56 \approx 610$

– в результате получены соответствующие правые части, значит, система решена верно.

Таким образом, **искомое уравнение *регрессии***:

$y = -6{,}0485x + 147{,}32$

и на самом деле «игрек» **правильнее записать** с чертой:

$\bar{y} = -6{,}0485x + 147{,}32$ – по той причине, что для различных «икс» мы будем получать *средние* (среднеожидаемые) значения «игрек». Но дабы избежать «накладок» с обозначениями, да и просто для чистоты я буду часто записывать голый «игрек».

Полученное уравнение показывает, что с увеличением количества прогулов («икс») на 1 единицу суммарная успеваемость падает *в среднем* на 6,0485 – примерно на 6 баллов. Об этом нам рассказал коэффициент «а». И **ещё раз обращаю внимание на тот факт**, что найденная функция возвращает нам **средние** или **среднеожидаемые** значения «игрек» для различных значений «икс».

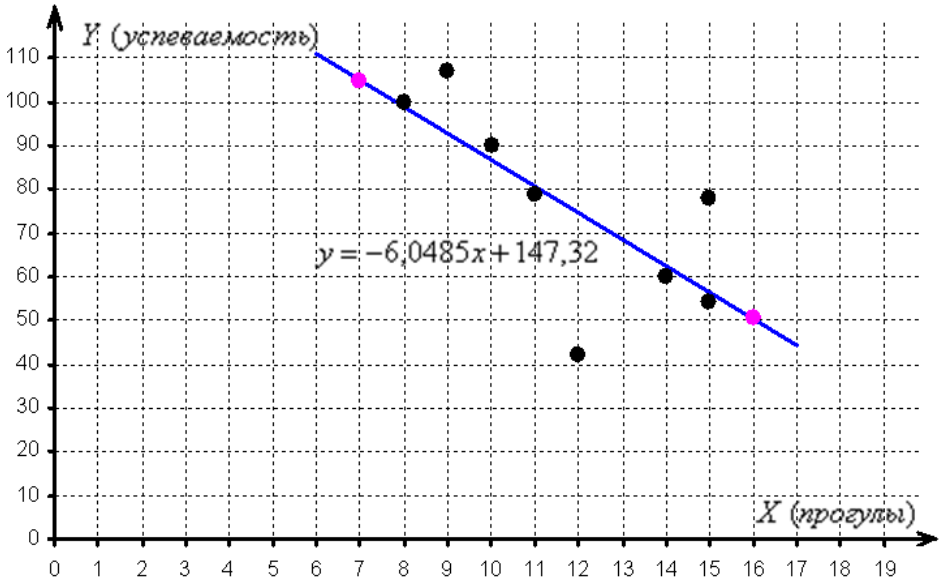
А почему это регрессия **именно «Y на X»** и о происхождении самого термина «регрессия» я рассказал чуть ранее, в параграфе эмпирические линии регрессии. Если кратко, то полученные с помощью уравнения *средние значения* успеваемости («игреки») регрессивно возвращают нас к первопричине – количеству прогулов. Вообще, регрессия – не слишком позитивное слово, но какое уж есть.

Линию регрессии изобразим на том же чертеже, вместе с *диаграммой рассеяния*. Для того чтобы построить прямую, достаточно знать две точки, выберем пару удобных значений «икс» и вычислим соответствующие «игреки»:

$x = 7 \quad \Rightarrow \quad y = -6{,}0485 \cdot 7 + 147{,}32 \approx 104{,}981;$

$x = 16 \quad \Rightarrow \quad y = -6{,}0485 \cdot 16 + 147{,}32 \approx 50{,}5437.$

Отметим найденные точки на чертеже (малиновый цвет) и проведём линию регрессии:



Говорят, что уравнение регрессии ***аппроксимирует*** (приближает) эмпирические данные (точки), и с помощью него можно *интерполировать* (оценивать) неизвестные промежуточные значения, так при количестве прогулов $x=13$ *среднеожидаемая* успеваемость ориентировочно составит $y=-6{,}0485\cdot 13+147{,}32\approx 68{,}7$ балла.

И, конечно, осуществимо ***прогнозирование***, так при $x=5$ *среднеожидаемая* успеваемость составит $y=-6{,}0485\cdot 5+147{,}32\approx 117$ баллов. Единственное, нежелательно брать «иксы», которые расположены слишком далеко от эмпирических точек, поскольку прогноз, скорее всего, не будет соответствовать действительности. Например, при $x=0$ соответствующее значение $y=147{,}32$ может вообще оказаться невозможным, ибо у успеваемости есть свой фиксированный «потолок». И, разумеется, «икс» или «игрек» в нашей задаче не могут быть отрицательными.

Второй вопрос касается *тесноты* зависимости. Очевидно, что чем ближе расположены эмпирические точки к прямой, тем теснее линейная корреляционная зависимость – тем уравнение регрессии достовернее отражает ситуацию, и тем качественнее полученная модель. И наоборот, если многие точки разбросаны вдали от прямой, то признак Y зависит от X вовсе не линейно (если вообще зависит) и линейная функция плохо отражает реальную картину. **Прояснить данный вопрос** нам поможет:

➢ Линейный коэффициент корреляции

Этот коэффициент как раз и оценивает тесноту линейной корреляционной зависимости и более того, указывает её *направление* (прямая или обратная). **Его полное название:** ***выборочный линейный коэффициент пАрной корреляции Пирсона*** :)

– *выборочный* – потому что мы рассматриваем выборочную совокупность;
– *линейный* – потому что он оценивает тесноту линейной корреляционной зависимости;
– *пАрной* – потому что у нас два признака (бывает хуже);
– и «*Пирсона*» – в честь английского статистика Карла Пирсона, это он автор понятия «корреляция».

А в зависимости от фантазии автора задачи вам может встретиться любая комбинация прокомментированных слов. Теперь нас не застанешь врасплох, Карл.

Линейный коэффициент корреляции вычислим **по формуле**:

…, где: $\overline{xy}$ – *среднее значение* произведения признаков, $\bar{x}$, $\bar{y}$ – средние значения признаков и σ_x, σ_y – стандартные отклонения признаков. Числитель формулы имеет особый смысл, о котором я расскажу позже, когда мы будем разбирать второй способ решения.

Осталось разгрести всё это добро :) Впрочем, все нужные суммы уже рассчитаны в таблице выше. Вычислим средние значения:

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{6856}{8} = 857, \quad \bar{x} = \frac{\sum x_i}{n} = \frac{94}{8} = 11{,}75, \quad \bar{y} = \frac{\sum y_i}{n} = \frac{610}{8} = 76{,}25.$$

Стандартные отклонения найдём как корни из соответствующих дисперсий, вычисленных по формуле:

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{1156}{8} - (11{,}75)^2} \approx 2{,}5372;$$

$$\sigma_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2} = \sqrt{\frac{50154}{8} - (76{,}25)^2} \approx 21{,}3351.$$

Таким образом, коэффициент корреляции:

…

И **расшифровка**: коэффициент корреляции может изменяться в пределах $-1 \le r \le 1$ и чем он ближе **по модулю** к единице, тем теснее линейная корреляционная зависимость – тем ближе расположены точки к прямой, тем качественнее и достовернее линейная модель. Если $r = -1$ или $r = 1$, то речь идёт о строгой линейной зависимости, при которой все эмпирические точки окажутся на построенной прямой. Наоборот, чем ближе r к нулю, тем точки рассеяны дальше, тем линейная зависимость выражена меньше. Однако в последнем случае зависимость всё равно может быть! – например, *нелинейной* или какой-нибудь более загадочной. Но до этого мы ещё дойдём. А у кого не хватит сил, донесём :)

Для оценки *тесноты* связи используют так называемую ***шкалу Чеддока***, в разных источниках она может иметь немного разные градации, например, такую

Диапазон значений $\lvert r \rvert$	Линейная корреляционная зависимость Y от X
0-0,1	практически отсутствует
0,1-0,3	слабая
0,3-0,5	умеренная
0,5-0,7	заметная
0,7-0,9	сильная
0,9-0,99	очень сильная
0,99-1	практически функциональная

при этом если $r < 0$, то корреляционная связь *обратная*, а если $r > 0$, то *прямая*.

У нас $r \approx -0{,}72$, таким образом**, существует сильная обратная корреляционная зависимость Y – суммарной успеваемости от X – количества прогулов**. Немудрено.

➢ Коэффициент детерминации

Если объяснять просто, то значения успеваемости *варьируются* (колеблются) под воздействием множества факторов, как неслучайных, так и случайных. И возникает вопрос, а насколько ВЕСОМО влияние именно прогулов?

$R = r^2$ – *коэффициент детерминации* **показывает** *долю вариации* **признака-результата** Y **, которая обусловлена воздействием признака-фактора** X .

Очевидно, что …. В нашей задаче: $R = r^2 \approx (-0{,}7193)^2 \approx 0{,}5174$ – таким образом, **в рамках построенной модели** успеваемость на 51,74% *(умножили* R *на 100)* зависит от количества прогулов. Оставшаяся часть вариации успеваемости (48,26%) обусловлена другими причинами. И красный цвет тут был не случаен:

Результаты подобных задач не являются какой-то «абсолютной истиной», это всего лишь одна из математических моделей!

Так, если бы эмпирические точки располагались примерно вдоль параболы, то, разумеется, мы бы получили плохую модель с низкими показателями r, R . Но это вовсе не означает, что корреляционная связь слабая, она просто *нелинейная*. **Ну а теперь:**

➢ Второй способ решения

Где мы сначала находим *коэффициент корреляции*, а затем уравнение *регрессии*.

Линейный коэффициент корреляции вычислим по формуле:

… , где σ_x, σ_y – стандартные отклонения признаков X, Y .

Член в числителе называют ***корреляционным моментом*** или ***коэффициентом ковариации*** (совместной вариации) признаков, он рассчитывается следующим образом: $\operatorname{cov}(X;Y) = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$, где n – *объём* статистической совокупности, а $\bar{x}, \bar{y}$ – средние значения признаков. Данный коэффициент показывает, насколько *согласованно* отклоняются пАрные значения $(x_i; y_i)$ от своих *средних* в ту или иную сторону. Формулу можно упростить, в результате чего получится ранее использованная версия, без подробных выкладок: $\operatorname{cov}(X;Y) = \dfrac{\sum x_i y_i}{n} - \dfrac{\sum x_i}{n} \cdot \dfrac{\sum y_i}{n} = \overline{xy} - \bar{x} \cdot \bar{y}$. Но сейчас мы пойдём другим путём. Заполним расчётную таблицу:

Прогулы, x_i	Плюшки, y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
12	42	0,25	-34,25	-8,5625	0,0625	1173,063
9	107	-2,75	30,75	-84,5625	7,5625	945,5625
8	100	-3,75	23,75	-89,0625	14,0625	564,0625
14	60	2,25	-16,25	-36,5625	5,0625	264,0625
15	78	3,25	1,75	5,6875	10,5625	3,0625
11	79	-0,75	2,75	-2,0625	0,5625	7,5625
10	90	-1,75	13,75	-24,0625	3,0625	189,0625
15	54	3,25	-22,25	-72,3125	10,5625	495,0625
94	**610**	**:суммы:**		**-311,5**	**51,5**	**3641,5**

В ходе заполнения таблицы сначала рассчитываем левые нижние суммы и средние значения признаков: $\bar{x}=\frac{\sum x_i}{n}=\frac{94}{8}=11{,}75,\quad \bar{y}=\frac{\sum y_i}{n}=\frac{610}{8}=76{,}25$ и только потом заполняем оставшиеся столбцы таблицы. **Кино будет!**

Вычислим *коэффициент ковариации*:

…

.

Стандартные отклонения вычислим как квадратные корни из дисперсий:

$$\sigma_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}=\sqrt{\frac{51{,}5}{8}}\approx 2{,}5372;$$

$$\sigma_y=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n}}=\sqrt{\frac{3641{,}5}{8}}\approx 21{,}3351.$$

Таким образом, *коэффициент корреляции*:

…

И если нам известны значения r, σ_x, σ_y, то коэффициенты уравнения $y=ax+b$ регрессии легко рассчитать по следующим **формулам**:

$$a=\frac{r\cdot\sigma_y}{\sigma_x}\approx\frac{-0{,}7193\cdot 21{,}3351}{2{,}53722}\approx -6{,}0485;$$

$$b=\bar{y}-a\bar{x}\approx 76{,}25-(-6{,}0485)\cdot 11{,}75\approx 147{,}32$$

Таким образом, искомое уравнение:

$y=-6{,}0485x+147{,}32$

Полученные результаты, естественно, совпали с результатами и выводами, сделанными ранее.

Теперь смотрим **ролик о том**, как это всё быстро подсчитать и построить.

Если под рукой нет Экселя, ничего страшного, разобранную задачу не так трудно решить в обычной клетчатой тетради.

Какой способ выбрать? Ориентируйтесь на свой учебный план и методичку. По умолчанию лучше использовать 2-й способ решения, он несколько короче, и, вероятно, потому и встречается чаще. Кстати, если вам нужно построить ТОЛЬКО уравнение регрессии, то уместен 1-й способ, ибо там мы находим это уравнение в первую очередь.

Следующая задача много-много лет назад была предложена курсантам местной школы милиции (тогда ещё милиции), и это чуть ли не первая задача по теме, которая встретилась в моей профессиональной карьере. Поэтому я безмерно рад предложить её вам сейчас:

Пример 46

В результате $n = 7$ независимых опытов получены 7 пар чисел:

X	0	-1	-3	-5	1	3	4
Y	2	0	-2	-4	9	5	7

…да, числа могут быть и отрицательными.

По данным наблюдений вычислить *линейный коэффициент корреляции* и *детерминации*, сделать выводы. Найти параметры *линейной регрессии* Y на X, пояснить их смысл. Изобразить *диаграмму рассеяния* и график регрессии. Вычислить $y(2)$, $y(-6)$ и пояснить, что означают полученные результаты.

Все данные уже забиты в Эксель, и вам осталось аккуратно выполнить расчёты. В образце я решил задачу вторым, более распространённым способом. И, конечно же, выполните проверку первым путём.

Следует отметить, что в целях экономии места и времени я специально подобрал задачи с малым объёмом выборки. На практике обычно предлагают 10 или 20 пар чисел, реже 30, и максимальная выборка, которая мне встречалась в студенческих работах – 100. …Соврал немного, 80. Ну а минимальное значение, напоминаю, 5-6 пар. **Теперь узнАем,**

➢ Как решить задачу в случае комбинационной группировки

Это когда в условии дана комбинационная таблица:

Пример 47

Имеются выборочные данные по 40 предприятиям региона:

Стоимость промышленно-производственных основных фондов, млрд. руб.	Суточная переработка сырья, тыс. ц			
	4-6	6-8	8-10	10-12
2,5-3,5	2			
3,5-4,5	6	3		
4,5-5,5	2	5	7	
5,5-6,5		2	2	3
6,5-7,5			1	7

Требуется:

1) Определить *признак-фактор* X и *признак-результат* Y и высказать предположение о наличии и *направлении зависимости* Y от X. Построить *корреляционное поле* и выдвинуть гипотезу о возможной форме зависимости.

2) Найти *коэффициенты корреляции* и *детерминации*, сделать выводы.

3) Найти уравнение *регрессии* Y на X и изобразить соответствующую линию на чертеже. Спрогнозировать среднюю суточную переработку сырья, когда стоимость основных фондов предприятий достигнет 9 млрд. руб.

Решение: 1) Определим *признак-фактор* и *признак-результат*. Очевидно, что чем больше стоимость основных фондов, тем крупнее предприятие и тем больше сырья оно способно переработать. Однако это не является непреложным правилом, ибо любое, самое крупное предприятие может неэффективно работать или даже простаивать. Тем не менее, *общая тенденция* состоит в том, что при увеличении стоимости фондов предприятий их *средняя* суточная переработка растёт. Такая зависимость называется… Правильно!

Таким образом, предполагаем наличие *прямой корреляционной зависимости* суточной переработки сырья (признак-результат Y) от стоимости основных фондов (признак-фактор X).

Частоты *комбинационной таблицы* располагаются преимущественно по диагонали – от левого верхнего до правого нижнего угла, что подтверждает *прямое* направление зависимости («чем больше, тем больше»).

Теперь определим *форму* зависимости *(линейная, квадратичная, экспоненциальная или какая-то другая)*. Простейший способ – графический, построили корреляционное поле и посмотрели. Для этого нужно немного модифицировать исходную таблицу, а именно перейти от интервальных вариационных рядов *(левый столбец и шапка таблицы)* к дискретным, выбрав в качестве вариант x_i и y_j середины соответствующих интервалов:

X – Стоимость осн. фондов, млрд. руб., x_i	Y -
3	
4	
5	
6	
7	
Итого, m_j	

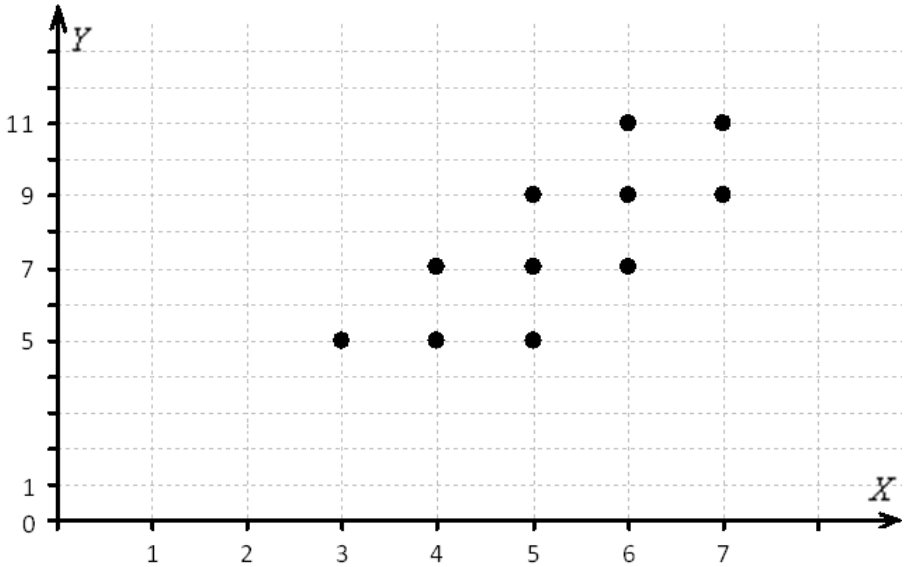
…

Заодно подсчитаем суммы частот по серым строкам (n_i) и суммы частот по серым столбцам (m_j), не забыв убедиться в том, что итоговые суммы равны *объёму* выборки:

$$\sum n_i = 2+9+14+7+8 = 40 = n, \qquad \sum m_j = 10+10+10+10 = 40 = n.$$

Довольно часто значения n_i и m_j уже подсчитаны и приведены в условии, но так бывает не во всех задачах, и поэтому я насыщаю решение всеми возможными действиями.

Обратите внимание, что значения x_i *признака-фактора* расположены **по вертикали** в левом столбце, а значения y_j *признака-результата* – **по горизонтали** в «шапке» таблицы. Именно такое расположение (а не наоборот) чаще всего встречается на практике, хотя оно не сильно удобно, в частности для построения корреляционного поля:



Ранее мы строили эмпирические линии регрессии – это простейший способ изобразить *форму* корреляционной зависимости. Однако гораздо удобнее привлечь на помощь функции. Анализируя чертёж, приходим к выводу, что эмпирические точки $(x_i; y_j)$ «выстроились» примерно по прямой, что позволяет предположить наличие *линейной* корреляционной зависимости Y – суточной переработки сырья от X – стоимости основных фондов.

Дальнейшие действия состоят в том, чтобы отыскать уравнение линейной регрессии $y = ax + b$, график которой проходит максимально близко к эмпирическим точкам *(с учётом их «весов» – частот n_{ij} в серых полях комбинационной таблицы)*, а также оценить *тесноту* линейной корреляционной зависимости – насколько близко расположены точки к построенной прямой. Эта теснота оценивается с помощью линейного коэффициента корреляции, с него и начнём:

2) Коэффициент корреляции вычислим по знакомой формуле $r_{xy} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$.

Лично я привык в первую очередь находить средние $\bar{x}, \bar{y}$ и стандартные отклонения σ_x, σ_y. Эти расчёты мы проводили неоднократно.

Сначала разберёмся с *признаком-фактором* X. Для этого из комбинационной таблицы *(см. выше)* выпишем значения x_i, n_i и заполним расчётную таблицу:

…

Вычислим *среднее значение* $\bar{x} = \dfrac{\sum x_i n_i}{n} = \dfrac{210}{40} = 5{,}25$ млрд. руб. и среднее квадратическое отклонение, как корень из дисперсии, вычисленной по формуле:

$$\sigma_x = \sqrt{\frac{\sum x_i^2 n_i}{n} - \bar{x}^2} = \sqrt{\frac{1156}{40} - (5{,}25)^2} = \sqrt{28{,}9 - 27{,}5625} = \sqrt{1{,}3375} \approx 1{,}1565$$

Аналогично, берём игрековые значения из комбинационной таблицы и заполняем расчетную таблицу для *признака-результата* Y:

…

после чего рассчитываем нужные показатели:

$$\bar{y} = \frac{\sum y_j m_j}{n} = \frac{320}{40} = 8 \text{ тыс. ц;}$$

$$\sigma_y = \sqrt{\frac{\sum y_j^2 m_j}{n} - \bar{y}^2} = \sqrt{\frac{2760}{40} - 8^2} = \sqrt{69 - 64} = \sqrt{5} \approx 2{,}2361.$$

Теперь найдём среднее значение $\overline{xy}$ произведения признаков. Для этого вычислим все возможные произведения x_i и y_j на соответствующие ненулевые частоты n_{ij}, наглядно распишу парочку штук:

x_i	y_j			
	5	7	9	11
3	2			
4	6	3		
5	2	5	7	
6		2	2	3
7			1	7

$4 \cdot 5 \cdot 6 =$

30			
120	84		
50	175	315	
	84	108	198
		63	539

$7 \cdot 9 \cdot 1''$

Вычислим сумму этих произведений:

$$\sum\sum x_i y_j n_{ij} = 30 + 120 + 84 + 50 + 175 + 315 + 84 + 108 + 198 + 63 + 539 = 1766$$

и искомую среднюю: $\overline{xy} = \frac{\sum\sum x_i y_j n_{ij}}{n} = \frac{1766}{40} = 44{,}15$.

И мы счастливы:

… – в результате получено положительное число и, согласно шкале Чеддока, существует ***сильная*** *прямая* линейная корреляционная зависимость Y суточной переработки сырья от X стоимости фондов.

Вычислим коэффициент детерминации:

… , таким образом, **в рамках построенной модели** 69,12% вариации суточной переработки сырья обусловлено стоимостью основных фондов. Остальные $100 - 69{,}12 = 30{,}88\%$ вариации обусловлено другими факторами.

3) Найдём уравнение $y = ax + b$ *линейной регрессии* Y на X. Здесь можно использовать уже известные формулы $a = \frac{r \cdot \sigma_y}{\sigma_x}$, $b = \bar{y} - a\bar{x}$, но есть более академичный вариант. Искомое уравнение имеет вид:

$y - \bar{y} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x})$, в данной задаче *(вычисления приближённые)*:

$$y - 8 = 0{,}8314 \cdot \frac{2{,}2361}{1{,}1565}(x - 5{,}25)$$

$$y - 8 = 1{,}6075(x - 5{,}25)$$

$y - 8 = 1{,}6075x - 8{,}4394$, примерно:

$$y = 1{,}61x - 0{,}44$$

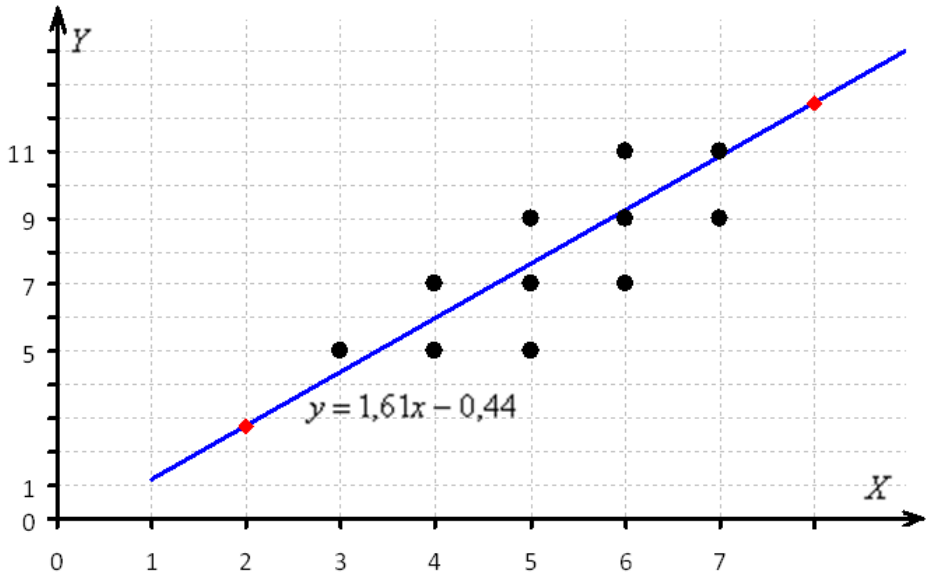
Полученное уравнение показывает, что при увеличении стоимости основных фондов на 1 млрд. руб. суточная переработка сырья увеличивается *в среднем* на 1,61 тысяч центнеров (смысл коэффициента «а»). **Напоминаю, что функция регрессии возвращает нам среднеожидаемые значения «игрек»**.

Найдём пару удобных точек для построения графика:

$x = 2 \quad \Rightarrow \quad y \approx 1{,}61 \cdot 2 - 0{,}44 = 2{,}78;$

$x = 8 \quad \Rightarrow \quad y \approx 1{,}61 \cdot 8 - 0{,}44 = 12{,}420,$

отметим их на чертеже *(красный цвет)* и аккуратно проведём линию регрессии на том же чертеже:



С помощью уравнения спрогнозируем *среднюю* суточную переработку сырья при стоимости основных фондов в 9 млрд. руб.:

$x = 9 \quad \Rightarrow \quad y \approx 1{,}61 \cdot 9 - 0{,}44 = 14{,}05$ тыс. ц.

Теперь **видео** о том, как быстро расправиться с этой задачей:

Помимо рассмотренного, существует второе уравнение –

➢ **Уравнение линейной регрессии X на Y**

– его можно составить *(в том числе и для несгруппированных данных)* по формуле:

…, после чего свести к виду:

$x(y) = cy + d$ – полученное уравнение позволяет нам узнать *средние* значения «икс», соответствующие различным значениям «игрек»

Чисто формально эта регрессия всегда существует, так в рассмотренной задаче признак X явно не зависит от Y, но вот линейная корреляционная зависимость есть! Причём, такой же тесноты. Из этого следуют важные факты, о которых мы поговорим в следующем параграфе. Кроме того, существуют ситуации, где признаки взаимно влияют друг на друга, уже известный вам пример:

X – количество произведённых куриц на птицефабрике;

Y – количество произведённых яиц.

Здесь в уравнении регрессии X на Y – самый что ни на есть здравый смысл.

График регрессии $x(y) = cy + d$ тоже можно изобразить на чертеже, и примечателен тот факт, что он будет пересекать график «традиционной» регрессии $y(x) = ax + b$ в точности в точке $(\bar{x}; \bar{y})$.

Следующая задачка для самостоятельного решения:

Пример 48

Известны следующие данные:

Y	X			
	45	55	65	75
10			5	3
20		7	15	8
30	3	9		

Найти линейный коэффициент корреляции и уравнение регрессии Y на X, а также X на Y. Построить корреляционное поле, линии регрессии и определить их точку пересечения. Вычислить $y(50)$ и $x(7)$. По каждому пункту сделать выводы.

Все числа уже в Экселе и вам остаётся выполнить вычисления, ничего страшного, если получится не очень красиво, важнА тренировка.

Обратите внимание, что в этом примере ничего не сказано о признаках X, Y, но нам ничего и не нужно о них знать, ведь задачу можно решить чисто формально – вне зависимости от того, где здесь признак-фактор, а где результат, и **есть ли вообще причинно-следственная связь между признаками**.

7.4. Корреляционная зависимость и причинно-следственная связь

Это разные вещи. …Да, вот так буднично, даже жирным шрифтом не выделил.

Если между признаками X, Y существует сильная корреляционная зависимость, то это ещё не значит, что между ними есть взаимосвязь.

Так, если мы возьмём два произвольных вариационных ряда, которые примерно одинаково растут (или убывают), то **в любом случае** получатся высокие *по модулю* значения r, R. При этом **между признаками может вообще не быть никакой причинно-следственной связи**, а-ля X – сезонное размножение сусликов в Монголии и Y – скорость свободного падения кирпича с Пизанской башни.

Поэтому причинно-следственная зависимость признака Y от X должна быть **предварительно обоснована если не экспертным путём, то хотя бы здравым смыслом**. Именно поэтому во всех содержательных задачах мы **обосновали причинно-следственную связь между признаками**. И это нужно обязательно делать, если вы проводите самостоятельное исследование. Пользуясь случаем, рекомендую эту тему для ваших научных и практических работ. Корреляционно-регрессионный анализ особо популярен в гуманитарных науках: социологии, психологии, etc и даже в истории.

Кроме того, величина Y может зависеть от X косвенно, опосредованно, и удачный тому пример есть в Википедии: очевидно, что между уличным травматизмом и количеством ДТП существует выраженная корреляционная зависимость, однако, эти показатели прямо не зависят друг от друга, у них есть общая причина – погодные условия (гололед, туман и т.д.). Поэтому логика и ещё раз логика.

С другой стороны, если корреляционная зависимость слабА или отсутствует, то это ещё не значит, что между признаками нет причинно-следственной связи.

Во-первых, эмпирические точки могут располагаться вдоль параболы, экспоненты или другой кривой, и, разумеется, в этих случаях мы получим малые значения линейных коэффициентов r, R. Но они будут высокими в рамках *нелинейных* моделей! На практике оптимальную модель подбирают аналитическим путём – строят различные кривые и находят *коэффициенты детерминации*. Где коэффициент R выше – та модель и удачнее. Быстрый способ узнать коэффициент «эр» для основных функций – Эксель (см. **ролик**). **Напоминаю, что при этом нужно обосновать причинно-следственную связь между признаками**. Но это ещё не всё. Есть куча зависимостей, где корреляцией даже не пахнет.

Представьте, что вы с разной силой дёргаете ручку игрового автомата, на котором крутятся бананчики, вишенки, семёрки и другие картинки. Есть ли причинно-следственная связь между вашими действиями и тем, что выпало на автомате? Безусловно. Но вот корреляционной зависимости (выпавших картинок от ваших усилий) нет никакой. Частоты в комбинационной таблице будут расположены хаотично, а при большом количестве испытаний примерно равномерно, и коэффициент R в любой вменяемой модели устремится к нулю.

Таким образом, к некоторым (и даже многим) зависимостям вообще нельзя применять метод корреляционного анализа. Или же можно, но работать он будет плохо.

Основная предпосылка использования корреляционно-регрессионного анализа состоит в том, что при изменении одного признака – другой должен гипотетически *(по нашему предположению и обоснованию)* **возрастать или убывать.**

Ещё раз перечитайте и хорошо ОСМЫСЛИТЕ вышесказанное!

…Молодцы! Теперь проконтролируйте, всё ли вам понятно в этих фразах:

Основная предпосылка использования *корреляционно-регрессионного анализа* состоит в том, что при изменении одного признака – другой должен по крайне мере гипотетически возрастать либо убывать. При этом **необходимо обосновать причинно-следственную связь между признаками.**

Корреляционный анализ оценивает *тесноту* зависимости *признака-результата* от *признака-фактора* (или факторов), а *регрессионный анализ – форму* зависимости, путём нахождения оптимальной *аппроксимирующей* функции, график которой проходит максимально близко к эмпирическим точкам. Подбор вида функции проще всего осуществить графически, визуально анализируя *диаграмму рассеяния* или *корреляционное поле*; также анализируются *коэффициенты детерминации*.

Наиболее распространена модель *пАрной линейной регрессии*, где теснота зависимости оценивается с помощью *линейного коэффициента корреляции*, а форма – с помощью *уравнения(ий) линейной регрессии*, которое задаёт прямую. С помощью этого уравнения *интерполируют* и *прогнозируют* *среднеожидаемые* значения признака-результата при различных значениях признака-фактора. Есть и другие модели регрессии, в том числе множественные (с несколькими признаками-факторами)

С расширенным курсом матстата можно ознакомиться на **mathprofi.ru**. Всех благ!

8. Решения и ответы

Пример 3. ***Решение: а)*** *Используем простую среднюю:*
ц/га – в среднем по трём областям.

б) *При анализе исходных данных бросается в глаза их неоднородность, так урожайность в 3-й области велика, но её посевная площадь мала. Поэтому урожайность уместно «взвесить» по площадям. Используем средневзвешенную (по площади) среднюю:*

ц/га в среднем по трём областям.

в) *Здесь урожайность тоже следует переоценить через посевную площадь, используя формулу Посевная площадь = Валовой сбор / Урожайность:*

ц/га в среднем по трём областям. Такой вид средней называют ***средней гармонической*****.**

Пример 5. ***Решение:*** *заполним расчётную таблицу:*

и составим эмпирическую функцию распределения:

Выполним чертёж:

Пример 7. ***Решение****: заполним расчётную таблицу:*

Построим гистограмму относительных частот:

полигон относительных частот:

и эмпирическую функцию распределения:

Пример 9. ***Решение****: заполним расчётную таблицу:*

Вычислим среднюю:

– две с половиной пуговицы, Карл!

По правому столбцу определяем варианту, которая делит совокупность на 2 равные части:

(именно здесь накопленная частота «перевалили» за 0,5).

Примечание: *кроме того, медиану легко усмотреть и устно – поскольку половина совокупности равна , а сумма первых двух частот*
, то совершенно понятно, что 250-й и 251-й пиджак упорядоченного ряда соответствует варианте .

Пример 11. ***Решение****: поскольку длина внутренних интервалов равна мин., то длины крайних интервалов полагаем такими же. Заполним расчётную таблицу:*

Вычислим выборочную среднюю: мин.

Моду вычислим по формуле, в данном случае:
– нижняя граница модального интервала;
– длина модального интервала;
– частота модального интервала;
– частота предшествующего интервала;
– частота следующего интервала.

Таким образом:

мин.

Анализируя накопленные частоты, приходим к выводу, что медианным является интервал (именно он содержит 50-ю и 51-ю варианты, делящие ряд пополам).

Медиану вычислим по формуле, в данном случае:
– нижняя граница медианного интервала;
– длина этого интервала;
– объём статистической совокупности;
– частота медианного интервала;
– накопленная частота предыдущего интервала.

Таким образом:

мин.

***Ответ**: среднее время изготовления детали характеризуется следующими центральными характеристиками:*

Задание. ***Генеральная дисперсия*** *– это среднее арифметическое квадратов отклонений всех вариант генеральной совокупности от её средней:*

, где – объём генеральной совокупности.

Для сформированного вариационного ряда формула принимает вид:

, где – либо варианты дискретного ряда, либо середины частичных интервалов интервального ряда, а – соответствующие частоты.

Пример 14. ***Решение**: найдём размах вариации: мин.*

Вычислим объём совокупности , произведения , их сумму и выборочную среднюю

мин.
Рассчитаем, произведения и их суммы. Вычисления сведём в таблицу:

Среднее линейное отклонение:
мин.

Выборочная дисперсия:
мин. в квадрате.

Несмещённой оценкой генеральной дисперсии является исправленная выборочная дисперсия: мин. в квадрате.

Несмещённость означает, что если в схожих условиях проводить аналогичные выборки, то полученные значения будут безо всякой закономерности варьироваться вокруг генерального значения.

***Ответ**:*

Пример 17. ***Решение**:*

***а)** Используем формулу. По условию, „. Таким образом, получаем и решаем уравнение:*

***б)** Используем формулу. По условию, „. Таким образом, получаем уравнение:*
, из которого находим

***Ответ**: а), б)*

Пример 18. ***Решение**: вычислим сумму вариант и сумму их квадратов:*

Найдём среднюю:

тонны – среднемесячный объем производства за полугодие.
Дисперсию вычислим по формуле:

Среднее квадратическое отклонение:
тонн.

Коэффициент вариации:

***Ответ**: тонны, тонн,*

***Краткие выводы:** за первое полугодие среднемесячный объём производства труб составил тонны. Низкие показатели вариации говорят о стабильной ситуации на производстве.*

Пример 20. ***Решение****:*

1) *По условию, точность оценки равна. Поскольку дисперсия, то среднее квадратическое отклонение составляет.*

Из формулы найдём коэффициент доверия:

Вычислим соответствующую доверительную вероятность:

– таким образом, утверждать, что генеральная средняя отличается от не более чем на (т.е. находится в доверительном интервале от 90 до 96) можно с вероятностью 86,64%.

2) *Для доверительной вероятности:*

– этому значению функции Лапласа соответствует аргумент:.

Вычислим точность оценки:

Определим доверительный интервал:

– данный интервал с вероятностью 99% накрывает истинное значение.

Ответ*: а), б)*

Пример 22. ***Решение****: доверительный интервал для оценки истинного значения измеряемой величины имеет вид:*

Для заданного уровня надёжности и количества степеней свободы по таблице распределения Стьюдента находим:.

Вычислим точность оценки: сек.

Таким образом, искомый доверительный интервал:

– данный интервал с вероятностью 99,9% накрывает истинное значение среднего времени изготовления одного диода.

Ответ:

Пример 24. ***Решение****: вычислим исправленное среднеквадратическое отклонение:*

1) *Определим доверительный интервал, где.*

Для уровня доверительной вероятности и объёма выборки по соответствующей таблице найдём.

Вычислим точность оценки:

Таким образом:

– с вероятностью данный интервал накроет генеральное среднее значение.

2) *Найдём доверительный интервал для генерального отклонения.*

а) *С помощью распределения :*

Вычислим и с помощью соответствующей функции Экселя (пункт 3б) найдём:

Таким образом:

– искомый интервал, накрывающий генеральное значение с вероятностью.

***б)** Дадим интервальную оценку приближенно, с помощью формулы:*

Коэффициент доверия найдём из соотношения. В данном случае:
и по таблице значений функции Лапласа либо
с помощью Экселя (пункт 1), выясняем, что.*
Таким образом:

– искомый интервал.

***Ответ**: 1),*
2) – с помощью распределения и – приближённо.

Пример 26. ***Решение**: поскольку опыты не зависят друг от друга, то результат каждого из них не оказывает влияние на другие результаты, а значит, по теореме Ляпунова, генеральная совокупность всех возможных результатов измерений распределена нормально. Искомый доверительный интервал имеет вид:*

, где, а – это выборочная средняя, которая будет получена в результате предстоящей серии опытов.

По условию, длина интервала должна равняться 0,5, таким образом, точность оценки: ед. Доверительная вероятность составляет, из соотношения находим:
.
Составим и решим уравнение:, откуда выразим:
.
Проверка:, что и требовалось проверить.

Таким образом, чтобы доверительный интервал с надёжностью накрыл истинное значение генеральной средней, нужно провести не менее 208 измерений.

***Ответ**: не менее 208.*

Пример 28. ***Решение**: вычислим исправленную выборочную дисперсию:*

***а)** Вычислим предельную ошибку выборки. Так как, то коэффициент доверия можно найти из соотношения. По условию, , следовательно:*

.

По таблице значений функции Лапласа определяем, что этому значению функции соответствует аргумент

Поскольку выборка 10%-ная бесповторная, то объём генеральной совокупности равен:

Вычислим среднюю ошибку выборки:

Таким образом, предельная ошибка:

и искомый доверительный интервал:

– данный интервал с вероятностью 0,954 накрывает среднее значение генеральной совокупности.

б) *Если выборка повторная, то средняя ошибка выборки рассчитывается по формуле:*

, точность оценки:
и соответствующий доверительный интервал:

– в результате получен более широкий интервал, таким образом, повторный отбор даёт чуть менее точную оценку.

! Следует, однако, заметить*, что в некоторых исследованиях предпочтительна именно повторная выборка (невзирая на более высокое значение).*

Ответ*: а), б), с математической точки зрения лучшую точность обеспечивает бесповторная выборка.*

Пример 30. ***Решение****: вычислим количество пирожных весом не менее 100 грамм:*
. Таким образом:

– выборочная доля таковых пирожных.

Соответствующую генеральную долю оценим с помощью доверительного интервала:
, где – предельная ошибка доли.

Коэффициент доверия найдём из соотношения:

По таблице значений функции Лапласа определяем, что этому значению функции соответствует аргумент

Вычислим среднюю ошибку доли. Поскольку выборка 1%-ная и бесповторная, то:

Таким образом, точность оценки и искомый доверительный интервал:

– данный интервал практически достоверно (99,74%) накрывает долю пирожных весом не менее 100 грамм во всей суточной партии.

б) *Улучшим точность оценки в 7 раз:*
и вычислим объём выборки, которую следует организовать, чтобы обеспечить эту точность. Учитывая, что объём генеральной совокупности составляет
:

Таким образом, для того, чтобы с вероятностью 99,73% можно было утверждать, что выборочная доля пирожных весом не менее 100 грамм будет отличаться от истинного значения не более чем на 0,02, следует выбрать не менее 3386 пирожных, что составляет примерно треть генеральной совокупности.

С точки зрения трудозатрат, такое исследование вряд ли оправдано.

Ответ*: а), б), скорее нет, чем да.*

Пример 33. ***Решение: поскольку известно ген. стандартное отклонение****, то для проверки гипотезы используем случайную величину.*

а) *Рассмотрим конкурирующую гипотезу. Так как альтернативные значения генеральной средней больше чем 0,5, то находим правостороннюю критическую область. Критическое значение определим из соотношения. Для уровня значимости* $\alpha = 0{,}1$*:*

При гипотеза принимается, а при
(в критической области) – отвергается:

Вычислим наблюдаемое значение критерия:

, таким образом, ***на уровне значимости 0,1 гипотезу принимаем****.*

б) *Рассмотрим конкурирующую гипотезу. В данном случае критическая область двусторонняя. Критическое значение найдём из соотношения. Для:*

При гипотеза принимается, а при (красная критическая область) – отвергается:

Наблюдаемое значение критерия вычислено в предыдущем пункте, и ***оно попадает в область принятия гипотезы****.*

Ответ: *в обоих случаях нулевую гипотезу на уровне значимости 0,1 принимаем.*

Пример 35. ***Решение****: вычислим сумму вариант, выборочную среднюю, квадраты отклонений и их сумму.*

Вычисления удобно свести в таблицу:

Выборочная дисперсия:.
Исправленная выборочная дисперсия: .
Исправленное стандартное отклонение:.

На уровне значимости 0,05 проверим нулевую гипотезу против конкурирующей гипотезы. ***Так как генеральная дисперсия не известна****, то для проверки используем случайную величину.*

Найдём критическую область. Поскольку в конкурирующей гипотезе речь идёт о меньших значениях, то она будет левосторонней. Для уровня значимости и количества степеней свободы по таблице критических значений распределения Стьюдента, найдём критическое значение для односторонней области:

При нулевая гипотеза отвергается, а при – принимается:

Вычислим наблюдаемое значение критерия:
поэтому ***на уровне значимости 0,05 нулевую гипотезу отвергаем****, иными словами, выборочное значение статистически значимо отличается от 10.*

Ответ*: на уровне значимости 0,05 можно утверждать, что после конструктивных изменений расход топлива стал меньше.*

Пример 37. ***Решение****: проверим гипотезу о том, что генеральная совокупность распределена по закону Пуассона. Используем критерий согласия Пирсона. Вычислим произведения* $x_i n_i$*, выборочную среднюю и теоретические частоты по формуле,где.*

Вычисления сведём в таблицу:

Объединяем две последние варианты ввиду их малых частот и находим критическое значение для уровня значимости и количества степеней свободы:

Вычислим наблюдаемое значение критерия:

Таким образом, , поэтому ***на уровне значимости нет оснований отвергать гипотезу о том, что генеральная совокупность распределена по закону Пуассона****.*

Примечание: *обратите внимание, что такая формулировка подчёркивает тот факт, что принятая нулевая гипотеза может оказаться и неверной. Ибо существует - вероятность того, что в действительности ген. совокупность вовсе не распределена по закону Пуассона.*

Ответ: *на уровне значимости 0,05 нулевую гипотезу принимаем.*

Пример 39. ***Решение:***

1) *Вычислим среднюю квартальную прибыль предприятий:*

млн. руб.

2) *Проведём равнонаполненную группировку. Оптимальное количество интервалов определим по формуле Стерджеса: и, округляя влево, получаем 6. Таким образом, в каждом интервале будет содержаться – от 7 до 9 предпр.*

Упорядочим совокупность по возрастанию и выделим в ней следующие группы; здесь же – в групповой таблице вычислим суммы и групповые средние:

Промежуточный контроль: , ч.т.п.

3) *Построим интервальный вариационный ряд:*

4) *Средняя прибыль предприятий за квартал составила млн. руб. Прибыль варьируется в пределах от 82 до 124 млн. руб. и равнонаполненная группировка показала, что распределение предприятий по данному показателю близкО к равномерному. То есть, практически нет предприятий со слишком большой или слишком малой прибылью.*

З.Ы. Возможно, вы заметили что-то ещё! ;-)

Пример 41. ***Решение:*** ***1)*** *выполним перегруппировку по 1-му банку:*

– В новый промежуток «до 500» войдут интервалы «до 100» и «100-500»:
чел.

– Новые промежутки «500-1000, 1000-2000» совпадают со старыми интервалами.

– Новые промежутки «2000-3000, 3000-4000, 4000-5000» полностью входят в старый интервал «2000-5000». Делим частоту этого интервала на 3:

– в каждый новый промежуток.

В промежутки «2000-3000, 3000-4000» относим по 11 человек, а в «4000-5000» – 10 человек (предполагая то, что людей с бОльшей заработной платой – меньше)

– Новый промежуток «5000 и более» совпадает со старым интервалом.

2) *Выполним перегруппировку второго вариационного ряда:*

– Старый интервал «до 1000» разобьём на два новых равных промежутка, при этом в промежуток «до 500» отнесём 5 человек, а в промежуток «500-1000» – 6 человек (предполагая, что людей с более низкой з/п – чуть меньше)

– В новый промежуток «1000-2000» входит интервал «1000-1500» и половина интервала «1500-2500», в людях это составит:

чел.

– В новый промежуток «2000-3000» входит половина интервала «1500-2500» и интервала «2500-4200», в людях это составляет:

чел.

– В новый промежуток «3000-4000» входит интервала «2500-4200», в людях это составляет: чел.

– В новый промежуток «4000-5000» входит интервала «2500-4200» и интервала «4200-6000», в людях это составит:

чел.

– И в новый промежуток «свыше 5000» входит интервала «4200-6000» и интервал «свыше 6000», в людях это составит: чел.

Результаты сведём в единую таблицу, при этом рассчитаем относительные частоты по каждому банку:

Для обоих банков характерна зарплата от 1000 до 2000 у.е., однако в 1-м банке чуть более высокий уровень заработной платы – значительное количество сотрудников получает более 2000 у.е. Но, скорее всего, основная их масса имеет з/п в диапазоне 2000-3000, здесь требуется дополнительное исследование первичных данных, поскольку формальное разбиение интервала «2000-5000» на три равных интервала не очень удачно.

З.Ы. Возможно, вы заметили что-то ещё! ;-)

Пример 43. ***Решение:***

1) *Очевидно, что чем больше процентная ставка, тем в среднем будет меньше сумма кредита, поскольку при высоких ставках заёмщику труднее расплачиваться по обязательствам. Таким образом, процентная ставка – признак-фактор, а сумма кредита – признак-результат. Предполагаемая корреляционная зависимость – обратная («чем больше, тем меньше»).*

2) *Проверим выдвинутое предположение методом аналитической группировки. Упорядочим выборочную совокупность по возрастанию процентной ставки и разобьём её на группы по банков в каждой группе:*

По каждой группе вычислим сумму кредитов (строка «Итого») и средние значения кредита млн. руб. (разделив суммы на объёмы групп, то есть на 5).

Результаты сведём в аналитическую таблицу:

Таким образом, при увеличении процентных ставок средние значения выданных кредитов уменьшаются, что подтверждает обратную корреляционную зависимость суммы кредита от процентной ставки.

Задание. ***Решение****: для наглядности скопирую комбинационную таблицу:*

Вычислим «икс среднее» по каждой из трёх групп. При
:

;
при
:

;
и при

.

Изобразим на чертеже то же самое корреляционное поле (оно у нас одно) и эмпирическую линию регрессии к – ломаную, соединяющую точки:

Пример 46. ***Решение:*** *вычислим суммы и средние значения признаков, и заполним расчётную таблицу:*

Вычислим коэффициент ковариации:

.

Вычислим средние квадратические отклонения:

Вычислим коэффициент корреляции:

, таким образом, существует сильная прямая корреляционная зависимость от.

Вычислим коэффициент детерминации:
– таким образом, 77,19% вариации признака-результата обусловлено влиянием фактора. Остальная вариация (22,81%) обусловлена другими факторами.

Вычислим коэффициенты линейной регрессии:

Таким образом, искомое уравнение регрессии:

Данное уравнение показывает, что с увеличением значения «икс» на одну единицу «игрек» увеличивается в среднем примерно на 1,32 единицы (смысл коэффициента «а»).

При среднеожидаемое значение «игрек» составит примерно 2,62 ед. (смысл коэффициента «бэ»).
Найдём пару точек для построения прямой:

и выполним чертёж:

Вычислим:
– среднеожидаемое значение «игрек» при (интерполированный результат);

– среднеожидаемое значение «игрек» при (спрогнозированный результат).

Пример 48. ***Решение****: вычислим частоты по каждому признаку:*

Линейный коэффициент корреляции найдём по формуле.
Заполним расчётную таблицу для признака:

Вычислим среднее значение:

и среднее квадратическое отклонение:

Заполним расчётную таблицу для признака:

Вычислим и
.

Вычислим произведения:

их сумму и среднюю.

Вычислим линейный коэффициент корреляции:

, таким образом, существует заметная (см. шкалу Чеддока) обратная линейная корреляционная зависимость между признаками (в обе стороны).

Составим уравнение линейной регрессии на (здесь и далее вычисления приближённые):

Полученное уравнение показывает, что при увеличении значения «икс» на 1 ед. среднеожидаемое значение «игрек» в среднем уменьшается на 0,47 единицы.

Составим уравнение линейной регрессии на:

Полученное уравнение показывает, что при увеличении значения «игрек» на 1 ед. среднеожидаемое значение «икс» в среднем уменьшается примерно на 0,87 единицы.

Найдём точки для построения графиков:

построим корреляционное поле и изобразим линии регрессии:

Линии регрессии пересекаются в точке

Вычислим:
– среднеожидаемое значение «игрек» при ;
– среднеожидаемое значение «икс» при .

Примечание: *вычисления местами не очень точные из-за округлений.*