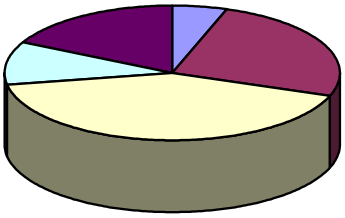
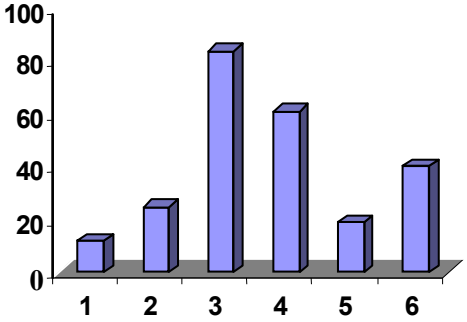
И.А. Палий

Прикладная статистика

Учебное пособие





Министерство образования РФ
Сибирская государственная автомобильно-дорожная академия
(СибАДИ)

И.А. ПАЛИЙ

ПРИКЛАДНАЯ СТАТИСТИКА

Учебное пособие

Допущено Министерством образования Российской Федерации
в качестве учебного пособия для студентов высших учебных заведений,
обучающихся по направлению 55000 – Технические науки и социально-
экономическим специальностям

Омск
Издательство СибАДИ
2003

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Жизнь – без начала и конца,
Нас всех подстерегает случай.
А. Блок. Народ и поэт

Статистика изучает случайные явления, которые, по своей сути, не поддаются однозначному описанию и прогнозированию. Например, нельзя абсолютно точно предсказать, сколько человек родится или умрет в стране за данный промежуток времени. Нельзя с точностью до копейки (цента, сантима) определить доход некоторой семьи за определенный промежуток времени (можно найти на дороге монетку в 10 копеек, выиграть в лотерею, получить неожиданное наследство, и, наоборот, можно потерять часть денег из-за болезни, или неверно принятого решения, или биржевого кризиса). Невозможно с точностью до минуты определить, какое время проработает купленный телевизор (компьютер, автомобиль) до первой поломки.

Жизнь человека, общества, цивилизации складывается из случайных явлений. Чтобы общество было устойчивым, а жизнь предсказуемой, важно не давать случаю слишком большой воли (любая попытка совсем исключить из жизни случай обречена на провал).

Современные задачи планирования, управления, прогнозирования невозможно решать, не располагая достоверными статистическими данными и не используя статистические методы обработки этих данных. Стремление объяснить настоящее и заглянуть в будущее всегда было свойственно человечеству, а для решения этих задач применялись различные методы. Статистика при описании случайных явлений использует язык науки – математику. Это значит, что реальные ситуации заменяются вероятностными схемами и анализируются методами теории вероятностей. Выразительная сила математики как языка очень велика.

Серьезные математические методы стали использоваться для анализа статистических наблюдений сравнительно недавно. Человечество осознало необходимость сбора статистических данных о различных сторонах жизни общества значительно раньше появления сопутствующего развитого математического аппарата. Но и сравнительно несложные методы сбора и анализа данных оказались важным инструментом, помогающим принимать разумные решения.

Любые статистические данные всегда неполны, и неточны, и другими быть не могут. Задача статистики заключается в том, чтобы дать обоснованные выводы о свойствах изучаемого явления, анализируя

неполные и неточные данные. Статистика доказала, что умеет справляться с подобными проблемами.

1.ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА ИЗ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

В одном мгновенье видеть вечность,
Огромный мир - в зерне песка,
В единой горсти - бесконечность
И небо - в чашечке цветка.
В. Блейк (перевод С. Маршака)

Понятия генеральной совокупности и выборки из нее являются первоначальными в статистике. Строгие определения пришли из теории вероятностей, хотя терминология математической статистики отличается от терминологии теории вероятностей. Вместо случайной величины X в теории вероятностей, в математической статистике говорят о генеральной совокупности X. Таким образом, понятие генеральной совокупности тождественно понятию случайной величины, т.е. включает в себя описание области определения (пространства элементарных исходов), множества значений, функциональной зависимости, закона распределения.

Вместо эксперимента, в результате которого случайная величина X приняла значение x (в теории вероятностей), в математической статистике говорят о случайном выборе из генеральной совокупности X значения x.

Вместо n независимых экспериментов, в результате которых случайная величина X приняла значения $x_1, x_2, \ldots, x_n$ (в теории вероятностей), в математической статистике говорят о случайной выборке объема n значений $x_1, x_2, \ldots, x_n$ из генеральной совокупности X.

При нестрогом подходе, под генеральной совокупностью понимают множество всех объектов некоторого наблюдения в совокупности с множеством всех значений этого наблюдения, соответствующих каждому объекту. А под выборкой объема n понимают множество из n объектов, реально подвергшихся наблюдению, в совокупности с n значениями наблюдения для каждого объекта. Например, социолог, изучающий мнение избирателей, под генеральной совокупностью понимает множество всех избирателей данной страны, а под выборкой объема n – множество из n человек, которых он опросил. Мы будем иметь в виду и такую точку зрения на генеральную совокупность.

Основная задача статистики – получить обоснованные выводы о свойствах генеральной совокупности, анализируя извлеченную из нее выборку $x_1, x_2, \ldots, x_n$. Более подробно: описать закон распределения генеральной совокупности; подобрать значения параметров этого закона,

оценить числовые характеристики генеральной совокупности; если генеральная совокупность – многомерная случайная величина, оценить всевозможные коэффициенты корреляции между ее составляющими; если имеется несколько выборок, извлеченных из разных генеральных совокупностей, определить, одинаково распределены эти генеральные совокупности или нет; одинаковы ли определенные числовые характеристики этих генеральных совокупностей или нет и т.д., и т.п.

Все перечисленные вопросы сформулированы на языке теории вероятностей. От статистики требуют ответы и на другие вопросы: можно ли утверждать, что новое лекарство эффективнее излечивает от некоторой болезни, чем старое? Какой будет численность населения страны в следующем году? Существует ли связь между значениями предела прочности и предела текучести различных марок стали? Чтобы ответы на подобные вопросы соответствовали действительности, нужно уметь строить подходящие вероятностные модели для реальных ситуаций. А для этого нужно уметь представить выборку в подходящем для изучения виде. Возникает задача описания и представления выборки.

Наконец, располагая сведениями о свойствах генеральной совокупности, можно предсказать свойства повторно извлеченных из нее выборок – заглянуть в будущее.

2. ВЫБОРКА, ЕЕ ПРЕДСТАВЛЕНИЕ И ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ

Все, что видим мы – видимость только одна.
Далеко от поверхности моря до дна.
Полагай несущественным явное в мире,
Ибо тайная сущность вещей - не видна.
О. Хайям (перевод Г. Плисецкого*)*

2.1. ПРЕДСТАВЛЕНИЕ ВЫБОРКИ

2.1.1. Таблица частот и интервальная таблица частот

Небольшие выборки удобно представлять в виде таблицы из двух строк. В первой строке записывают элементы выборки (они называются вариантами), расположенные в порядке возрастания. Во второй строке записываются частоты вариант. Частотой варианты называется число, равное количеству повторений варианты в выборке. Если n_i – частота варианты x_i, всего в выборке k разных вариант, то $n_1 + n_2 + ...+ n_k = n$, где n – объем выборки. Описанная таблица называется таблицей частот.

Рассмотрим пример. С производственной линии случайным образом 36 раз отбирали по 10 единиц некоторого изделия. Каждый раз отмечалось число дефектных изделий.

Получена выборка 1:

0	0	1	0	2	0	1	2	1	0	0	0	0	0	3	1	0	0
0	0	0	2	0	0	1	1	0	0	0	1	1	0	1	0	1	1

Здесь $n = 36$, в выборке представлены 4 варианты: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$.

Таблица частот выглядит следующим образом (табл. 2.1):

Таблица 2.1

x_i	0	1	2	3
n_i	21	11	3	1

Относительной частотой варианты x_i называется число ν_i, равное отношению n_i / n. Если сумма частот равна n, то сумма относительных частот равна $n/n = 1$.

Таблица относительных частот для этого примера такова (табл. 2.2):

Таблица 2.2

x_i	0	1	2	3
ν_i	21/36	11/36	3/36	1/36

Таблица относительных частот напоминает таблицу вероятностей дискретной случайной величины. Только вместо значений случайной величины пишут варианты выборки, а роль вероятностей исполняют относительные частоты.

Накопленной частотой $n_x^{нак}$ называется число вариант выборки, меньших данного числа x.

Относительной накопленной частотой $\nu_x^{нак}$ называется отношение $n_x^{нак}/n$. Найдем накопленные и относительные накопленные частоты вариант выборки для нашего примера (табл 2.3).

Таблица 2.3

x_i	0	1	2	3
$n_{xi}^{нак}$	0	21	32	35
$\nu_{xi}^{нак}$	0	21/36	32/36	35/36

Ясно, что $n_{x1}^{нак} = 0$, $\nu_{x1}^{нак} = 0$, т.к. нет ни одной варианты, меньшей x_1. Кроме того,

$$n_{xi}^{нак} = n_{xi-1}^{нак} + n_{i-1} = \sum_{j<i} n_j \; ; \; \nu_{xi}^{нак} = \nu_{xi-1}^{нак} + \nu_{i-1} = \sum_{j<i} \nu_j \; ,$$

отчего частоты и называются накопленными. Относительные накопленные частоты – это статистические аналоги значений функций распределения $F(x_i)$ дискретной случайной величины X. Действительно,

$$F(x_i) = P(x < x_i) = \sum_{j<i} P(x = x_j) = \sum_{j<i} P_j \; .$$

Если выборка извлечена из непрерывно распределенной генеральной совокупности, причем ее объем n достаточно велик, то в выборке представлено много значений, и такую выборку неразумно представлять в виде таблицы частот. Кроме того, при работе с непрерывно распределенными случайными величинами рассматривают не отдельные значения этих величин, а некоторые интервалы этих значений. Поэтому достаточно большую выборку, извлеченную из непрерывно распределенной генеральной совокупности, группируют по интервалам следующим образом. Весь диапазон значений вариант разбивают на разумное число интервалов одинаковой, как правило, ширины h. Чтобы не было недоразумений при подсчете числа вариант выборки, попавших в каждый интервал, левый конец каждого интервала считают закрытым, а правый – открытым, так что интервалы имеют вид $[x_{i-1}; x_i)$.

Частотой i-го интервала n_i называется число, равное количеству вариант выборки, попавших в этот интервал,

Относительной частотой i-го интервала ν_i называется отношение n_i /n. Кроме того, вычисляют накопленные и относительные накопленные частоты для **правых** границ интервалов.

Если всего интервалов k, очевидно :

$$\sum_{i=1}^{k} n_i = n; \sum_{i=1}^{k} \nu_i = 1; n_{x_k}^{нак} = n; \nu_{x_k}^{нак} = 1,$$

где x_k – правая граница последнего интервала, все варианты выборки меньше числа x_k.

Полученные числа заносят в таблицу, которая называется интервальной таблицей частот.

Рассмотрим пример. У 50 новорожденных измерили массу тела с точностью до 10г. Результаты (в кг) таковы (выборка 2):

3,7	3,85	3,7	3,78	3,6	4,45	4,2	3,87	3,33	3,76
3,75	4,03	3,75	4,18	3,8	4,75	3,25	4,1	3,55	3,35
3,38	3,3	4,15	3,95	3,5	3,88	3,71	3,15	4,15	3,8
4,22	3,75	3,58	3,55	4,08	4,03	3,24	4,05	3,56	3,05
3,58	3,98	3,88	3,78	4,05	3,4	3,8	3,06	4,38	4,2

Сгруппируем эту выборку. Наименьшая масса равна 3,05 кг, наибольшая масса равна 4,75 кг. “Упакуем” выборку в интервал [3 – 4,8], который разобьем на 6 интервалов шириной 0,3.

Интервальная таблица частот выглядит следующим образом (накопленные частоты считают для правых границ интервалов) (табл.2.4).

Таблица 2.4

$[x_{i-1}, x_i)$	[3-3,3)	[3,3-3,6)	[3,6-3,9)	[3,9-4,2)	[4,2-4,5)	[4,5-4,8)
n_i	5	11	17	11	5	1
ν_i	0,1	0,22	0,34	0,22	0,1	0,02
$n_{xi}^{нак}$	5	16	33	44	49	50
$\nu_{xi}^{нак}$	0,1	0,32	0,66	0,88	0,98	1,0

2.1.2. Графическое представление выборки. Полигон, гистограмма, кривая накопленных частот

Рисунки и графики – удобный и наглядный способ представления выборки. Выборку, извлеченную из дискретной генеральной совокупности, можно представить в виде полигона частот. На плоскости в прямоугольной системе координат строят точки с координатами (x_i, ν_i) и соединяют эти точки отрезками прямых. Полученная ломаная и называется полигоном частот. Полигон можно, конечно, построить и для сгруппированной выборки. Но такую выборку нагляднее всего представить в виде гистограммы. Гистограмма – это фигура, состоящая из прямоугольников. Основания прямоугольников – это интервалы, на которые разбита сгруппированная выборка. Высота i-го прямоугольника h_i определяется формулой

$$h_i = \nu_i / h, \;\; i = 1, 2, 3, \ldots, k.$$

Таким образом, высоты прямоугольников пропорциональны частотам интервалов, а сумма высот равна

$$\sum_{i=1}^{k} \nu_i / h = 1/h.$$

Поэтому площадь гистограммы равна $(1/h)*h = 1$.

Гистограмма – это аналог графика функции плотности вероятности *f(x)* непрерывной случайной величины, площадь под графиком *f(x)* равна 1. Кривая накопленных частот (кумулятивная кривая) – это статистический аналог графика функции распределения *F(x)* непрерывной случайной величины. Кривая накопленных частот строится так: точки с координатами $(x_i, \; \nu_{xi}^{нак})$ соединяют отрезками прямых. Кроме того, накопленные частоты для любого числа $x < x_1$ равны 0, накопленные частоты для любого числа $x > x_k$ равны 1. Чтобы найти накопленную частоту для некоторого $x_1 < x < x_k$, нужно воспользоваться линейной интерполяцией. На рис. 2.1, 2.2, 2.3 показаны полигон частот для выборки 1, гистограмма и кумулятивная кривая для выборки 2 соответственно.

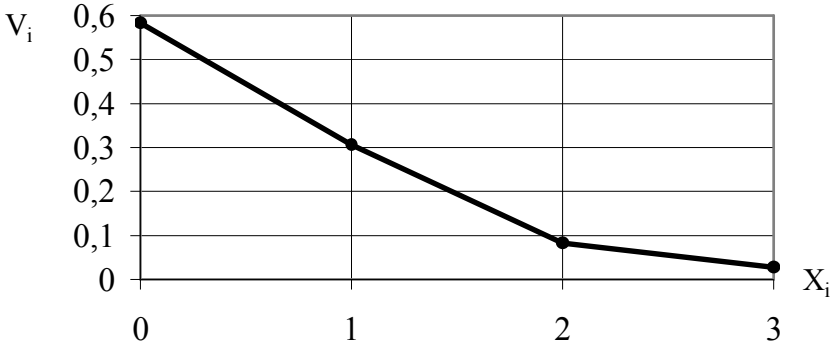


Рис. 2.1

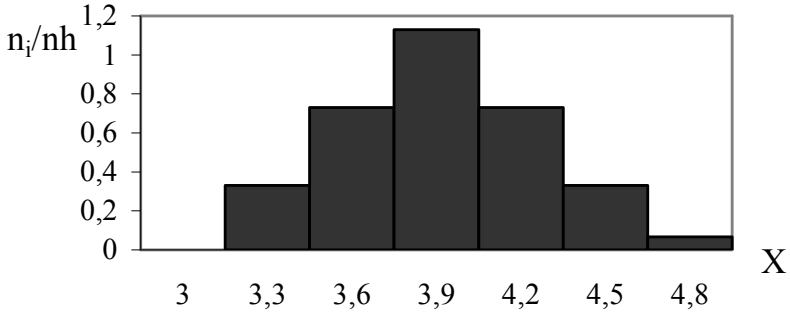


Рис. 2.2

$h_1 = 0,1/0,3 = 0,33$; $h_2 = 0,22/0,3 = 0,73$; $h_3 = 0,34/0,3 = 1,13$; $h_4 = h_2 =$

=0,73; $h_5 = h_1 = 0{,}33$; $h_6 = 0{,}02/0{,}3 = 0{,}067$.

Покажем, как, используя линейную интерполяцию, найти относительную накопленную частоту $v_x^{нак}$ для числа $x_1 < x < x_k$.

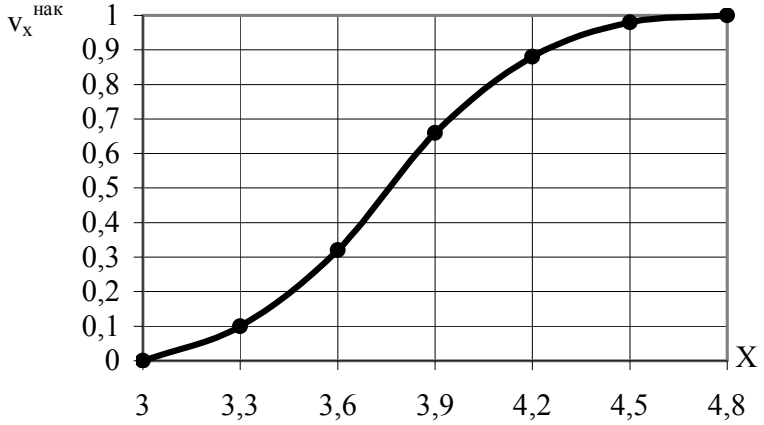


Рис. 2.3

Пусть x принадлежит интервалу $[x_{i-1}, x_i)$. Рассмотрим соответствующий участок кривой накопленных частот (рис.2.4).

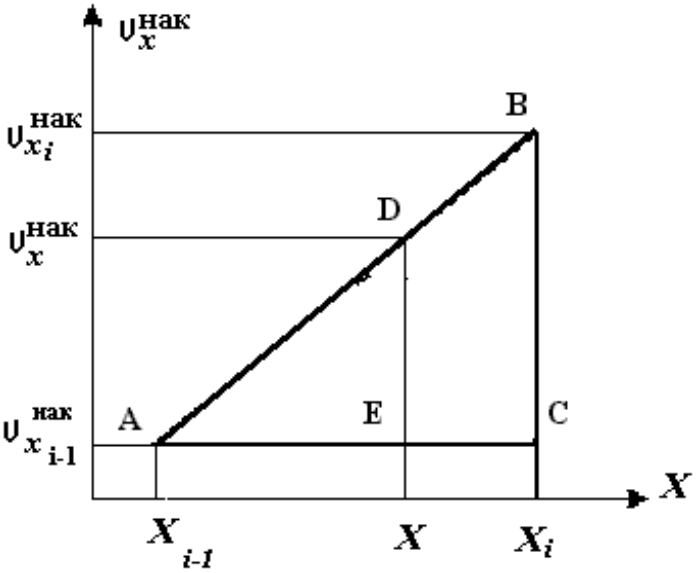


Рис. 2.4

Имеем: $AC = h$; $AE = x - x_{i-1}$; $BC = v_{x_i}^{нак} - v_{x_{i-1}}^{нак}$; $DE = v_x^{нак} - v_{x_{i-1}}^{нак}$;

$$\Delta ABC \sim \Delta ADE.$$

Из подобия треугольников следует, что

$$\frac{AC}{AE} = \frac{BC}{DE}, \text{ или } \frac{h}{x - x_{i-1}} = \frac{v_{x_i}^{нак} - v_{x_{i-1}}^{нак}}{v_x^{нак} - v_{x_{i-1}}^{нак}}.$$

Отсюда получаем

$$v_x^{нак} = \frac{(x - x_{i-1}) * (v_{x_i}^{нак} - v_{x_{i-1}}^{нак})}{h} + v_{x_{i-1}}^{нак}.$$

Например, в выборке 2 :

$$\nu_4^{нак} = 0{,}66 + [(4 - 3{,}9)*(0{,}88 - 0{,}66)]/0{,}3 = 0{,}73.$$

Точно так же решается и обратная задача: по известной частоте $\nu_x^{нак}$ найти число x. Имеем

$$x = \frac{h*(\nu_x^{нак} - \nu_{x_{i-1}}^{нак})}{\nu_{x_i}^{нак} - \nu_{x_{i-1}}^{нак}} + x_{i-1}.$$

Например, для выборки 2 относительную накопленную частоту 0,5 имеет число

$$x = \frac{0{,}3*(0{,}5 - 0{,}32)}{0{,}66 - 0{,}32} + 3{,}6 = 3{,}76.$$

Действительно, если $\nu_x^{нак} = 0{,}5$, то число x лежит внутри интервала [3,6; 3,9), так как $\nu_{3,6}^{нак} = 0{,}32 < 0{,}5$, а $\nu_{3,9}^{нак} = 0{,}66 > 0{,}5$.

2.2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

2.2.1. Выборочное среднее, мода, медиана

Выборочное среднее $\bar{x}$ – это среднее арифметическое вариант выборки. Если объем выборки равен n, то

$$\bar{x} = (1/n)\sum_{j=1}^{n} x_j = (1/n)\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \nu_i x_i,$$

где k - число различных вариант; n_i – частота варианты x_i, $i = 1, 2, 3, ..., k$.

Если выборка сгруппирована, то часто даже неизвестно, какие именно варианты попали в i-й интервал. Тогда частоту интервала n_i умножают на средину интервала. Конечно, при этом получается ошибка, так как варианты, попавшие в интервал, не обязаны все совпадать с числом $(x_i + x_{i-1})/2$. Но эта ошибка не может быть слишком большой, особенно при достаточно больших n. Ведь в среднем половина вариант, попавших в интервал $[x_{i-1}, x_i)$, будет меньше числа $(x_i + x_{i-1})/2$, а половина – больше, поэтому ошибки будут иметь разные знаки и, таким образом, компенсируют друг друга. Легко видеть, что формула для выборочного среднего $\bar{x}$ совпадает с формулой для вычисления математического ожидания дискретной случайной величины. Роль вероятностей играют относительные частоты ν_i.

Найдем выборочные средние для выборок, рассмотренных ранее.

1. Выборка 1.

$$\bar{x} = \sum_{i=1}^{4} \nu_i * x_i = 0*21/36 + 1*11/36 + 2*3/36 + 3*1/36 = 0{,}56.$$

Итак, в среднем из каждых 10 единиц товара 0,56 единицы дефектны.

2. Выборка 2.

Найдем сначала выборочное среднее непосредственно по выборке, а затем по сгруппированной выборке и сравним полученные числа.

В первом случае имеем:

$\bar{x}$ = 1/50*(3,7 + 3,85 + 3,7 + 3,78 + 3,6 + 4,45 + 4,2 + 3,87 + 3,33 + 3,76 + +3,75 + 4,03 + 3,75 +4.18+ 3,8 + 4,75 + 3,25 + 4,1 + 3,55 + 3,35 + 3,38 +3,3 + +4,15 + 3,95 + 3,5 + 3,88 + 3,71 + 3,15 + 4,15 + 3,8 + 4,22 + 3,75 + 3,58 + +3,55 + 4,08 + 4,03 + 3,24 + 4,05 + 3,56 + 3,05 + 3,58 + 3,98 + 3,88 + 3,78 + +4,05 + +3,4 + 3,8 + 3,06 + 4,38 + 4,2) = 3,78.

Средняя масса ребенка равна 3,78 кг.

Рассчитаем выборочное среднее по сгруппированной выборке.

$\bar{x}$ = 3,15*0,1 + 3,45*0,22 + 3,75*0,34 + 4,05*0,22 + 4,35*0,1 + 4,65*0,02 = =3,77.

Расхождение равно 10 граммам. Но ведь и массы детей определялись с точностью до 10 граммов, так что мы не превзошли ошибки округления. Сам же подсчет оказался намного проще.

В теории вероятностей модой $x_{мо}$ дискретной случайной величины называется такое её значение, которое имеет максимальную вероятность. Модой непрерывной случайной величины называется такое её значение, на котором достигается максимум функции плотности вероятности *f(x)*. Закон распределения называется унимодальным, если мода единственна. Соответственно вводится понятие моды и в статистике. Модой $\hat{x}$ (обозначают $\hat{x}$, читают "*x* с крышечкой") называется варианта x_i с наибольшей частотой (относительной частотой). В выборке 1 мода $\hat{x}$ = 0.

Если выборка сгруппирована, то сначала определяют модальный интервал, т.е. интервал с наибольшей частотой (относительной частотой). В качестве моды можно взять середину модального интервала. Эту оценку можно подправить с помощью простого дополнительного построения на гистограмме (рис. 2.5).

В выборке 2 модальный интервал – это интервал [3,6; 3,9). Тогда $\hat{x}$ = =3,75. Так как высоты прямоугольников слева и справа от интервала [3,6; 3,9) одинаковы, подправлять значение $\hat{x}$ не нужно.

В теории вероятностей медианой непрерывной случайной величины *X* называется такое число $x_{ме}$, когда $P(X < x_{ме}) = 0,5 = P(X > x_{ме})$. Соответственно в статистике медианой (обозначают $\tilde{x}$, читают "*x* с волной") называют такое число $\tilde{x}$, когда 50% вариант выборки меньше этого значения, а 50% больше его. Ясно, что для любой выборки можно подобрать бесконечно много медиан. Чтобы избежать неоднозначности,

будем называть медианой число $\tilde{x}$ такое, когда $v_{\tilde{x}}^{нак} = 0{,}5$, где 0,5 – ордината точки с абсциссой $\tilde{x}$ на кривой накопленных частот.

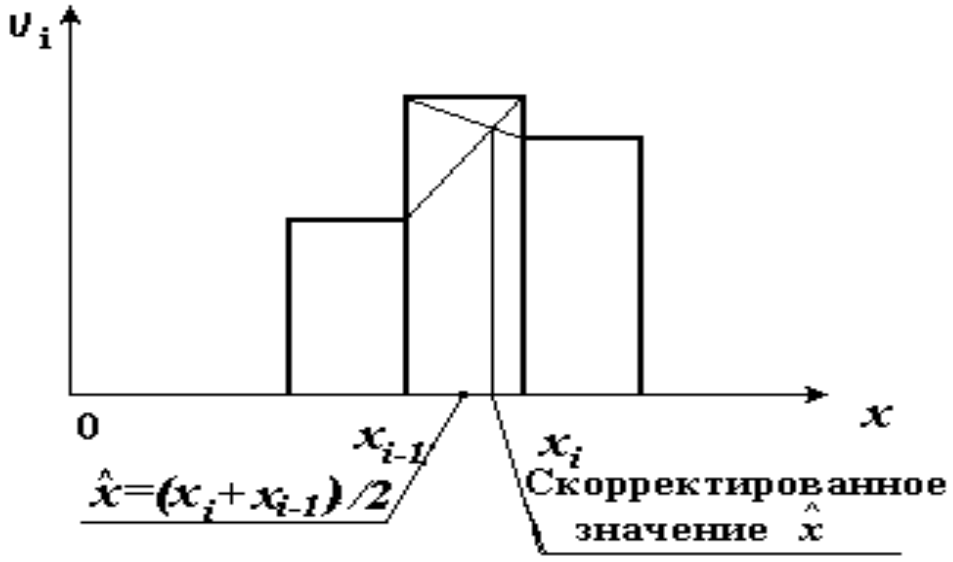


Рис 2.5

Чтобы найти медиану, нужно сначала найти медианный интервал $[x_{i-1}; x_i)$, где $v_x^{нак} < 0{,}5$; $v_{xi}^{нак} > 0{,}5$, тогда $\tilde{x} \in [x_{i-1}; x_i)$. Используя формулу, выведенную в пункте 2.1.2, получаем, что

$$x = x_{i-1} + \frac{h * (0{,}5 - v_{x_{i-1}}^{нак})}{v_{x_i}^{нак} - v_{x_{i-1}}^{нак}}.$$

В выборке 2 медианным интервалом является интервал [3,6; 3,9), так как $v_{3,6}^{нак} = 0{,}32$; $v_{3,9}^{нак} = 0{,}66$. Тогда

$$\tilde{x} = 3{,}6 + \frac{0{,}3 * (0{,}5 - 0{,}32)}{0{,}66 - 0{,}32} = 3{,}76.$$

2.2.2. Квартили, декатили, персентили

Медиана делит выборку на две части: половина вариант меньше медианы, половина – больше медианы. Можно найти три числа: Q_1, Q_2, Q_3, которые аналогичным образом делят выборку на 4 равные части. Эти числа называются квартилями. Число Q_2 совпадает с медианой $\tilde{x}$, число Q_1 называется нижней квартилью, число Q_3 называется верхней квартилью. В теории вероятностей квартилями непрерывной случайной величины X называются числа Q_1, Q_2, Q_3, определяемые из условия

$P(X < Q_1) = P(Q_1 < X < Q_2) = P(Q_2 < X < Q_3) = P(X > Q_3) = 0{,}25$.

Точно так же можно найти 9 чисел: D_1, D_2, ..., D_9, которые разбивают выборку (площадь под графиком $f(x)$) на десять равных частей. Эти числа называются декатилями. Если разбить выборку (площадь под графиком

f(x)) на сто равных частей, точки деления называются персентилями. Их 99, они обозначаются $P_1, P_2, \ldots, P_{99}$. Ясно, что $P_{25} = Q_1$, $P_{50} = Q_2 = \tilde{x}$, $P_{75} = Q_3$. Числа $Q_1, Q_2, Q_3, P_1, P_2, \ldots, P_{99}$ находятся точно так же, как $\tilde{x}$. Например, $\nu_{Q1}^{нак} = 0{,}25$, тогда

$$Q_1 = x_{i-1} + \frac{h * (0{,}25 - \nu_{x_{i-1}}^{нак})}{\nu_{x_i}^{нак} - \nu_{x_{i-1}}^{нак}},$$

где $\nu_{x_{i-1}}^{нак} < 0{,}25$; $\nu_{x_i}^{нак} > 0{,}25$; $Q_1 \in [x_{i-1}, x_i,)$.

2.2.3. Измерение разброса: размах, выборочная дисперсия, выборочное среднее квадратическое отклонение (стандартное отклонение), коэффициент вариации

Размах R – простейшая мера разброса значений данной выборки. Если x_{max} – максимальная, x_{mix} – минимальная варианты, то $R = x_{max} - x_{mix}$. Этой величиной пользуются при работе с малыми выборками.

Более эффективные меры разброса должны учитывать все элементы выборки. Одна из самых распространенных мер называется выборочной дисперсией S^2. Она вычисляется точно так же, как дисперсия дискретной случайной величины. Следовательно, выборочная дисперсия оценивает средний разброс значений выборки относительно выборочного среднего.

$$S^2 = (1/n)\sum_{j=1}^{n}(x_j - \bar{x})^2 = (1/n)\sum_{j=1}^{n} x_j^2 - (\bar{x})^2 = (1/n)\sum_{i=1}^{k} n_i x_i^2 - (\bar{x})^2 =$$

$$= \sum_{i=1}^{k} \nu_i x_i^2 - \bar{x}^2,$$ где k - число разных вариант выборки.

Если выборка сгруппирована, частота i-го интервала n_i умножается на середину интервала – число $(x_i + x_{i-1})/2$. Соответственно корень квадратный из выборочной дисперсии называется выборочным средним квадратическим отклонением и обозначается S. Другое часто встречающееся название для S – стандартное отклонение; оно короче, поэтому мы будем чаще использовать его.

Найдем эти параметры для выборки 2.

$S^2 = 3{,}15^2 * 0{,}1 + 3{,}45^2 * 0{,}22 + 3{,}75^2 * 0{,}34 + 4{,}05^2 * 0{,}22 + 4{,}35^2 * 0{,}1 + +4{,}65^2 * 0{,}02 - (3{,}77)^2 = 0{,}127$; $S = 0{,}36$.

В среднем масса ребенка отличается от средней массы на 0,36 кг. В теории вероятностей для нормального закона распределения доказываются так называемые “правило двух сигм” и “правило трех сигм”: вычисляются вероятности того, что нормально распределенная случайная величина отклонится по модулю от своего математического

ожидания a не более чем на два или три средних квадратических отклонения σ.

$P(|X - a| < 2\sigma) = 0{,}9545$; $P(|X - a| < 3\sigma) = 0{,}9973$.

Эти правила приблизительно выполняются для большинства унимодальных законов распределения и соответственно выборок из таких генеральных совокупностей:

1. Более 95% значений выборки лежат в интервале $(\bar{x} - 2S, \bar{x} + 2S)$.
2. Более 99% значений выборки лежат в интервале $(\bar{x} - 3S, \bar{x} + 3S)$.

Для выборки 2 имеем :

$\bar{x} - 2S = 3{,}77 - 0{,}36 * 2 = 3{,}05$; $\bar{x} - 3S = 3{,}77 - 0{,}36 * 3 = 2{,}69$;

$\bar{x} + 2S = 3{,}77 + 0{,}36 * 2 = 4{,}49$; $\bar{x} + 3S = 3{,}77 + 0{,}36 * 3 = 4{,}85$.

В интервале (3,05; 4,49) лежат 48 значений (или 96%) выборки; в интервале (2,69; 4,85) лежат 100% значений выборки.

Коэффициент вариации V служит для сравнения стандартных отклонений нескольких выборок и вычисляется по формуле $V = S/\bar{x}$.

Если коэффициенты вариации оказались величинами одного порядка, то средние рассеяния данных относительно среднего в этих выборках можно считать примерно равными.

Рассмотрим простой пример. Пусть массы трех килограммовых пакетов с сахаром оказались такими: $x_1 = 0{,}995$ кг; $x_2 = 1$ кг; $x_3 = 1{,}005$ кг. Тогда $\bar{x}_1 = 1{,}00$ кг; $S_1 = 4{,}08 * 10^{-3}$ кг; $V_1 = 4{,}08 * 10^{-3}$.

Допустим так же, что масса некоторого железобетонного блока должна равняться 100 кг, а массы трех отобранных блоков оказались равными 99,5 кг, 100,00 кг и 100,5 кг. Отсюда $\bar{x}_2 = 100$ кг; $S_2 = 0{,}408$ кг; $V_2 = 4{,}08 * 10^{-3}$.

Пусть, наконец, некоторый студент, сдавая сессию, получил такие оценки: 4, 3, 5. Значит, $\bar{x}_3 = 4{,}0$; $S_3 = 0{,}82$; $V_3 = 0{,}21$.

Сравнивая три найденных коэффициента вариации, заключаем, что точности работы устройств, развешивающих сахар в пакеты и изготовляющих железобетонные блоки, одинаковы. Хотя в первом случае максимальное отклонение массы от номинала составило 5 г, а во втором случае в 100 раз больше – 500 г. Зато разброс оценок студента значительно больше: $V_3 \approx 50\ V_1$.

2.2.4. О симметричных и несимметричных распределениях

Закон распределения непрерывной случайной величины X называется симметричным, если график функции плотности вероятности $f(x)$ имеет ось симметрии, например, нормальный закон распределения симметричен. Для унимодального симметричного закона распределения очевидно равенство моды, медианы и математического ожидания. Если имеет место

небольшая асимметрия (рис 2.6.), то возможны только два случая: $x_{мо} < x_{ме} < M(X)$ или $M(X) < x_{ме} < x_{мо}$. То же справедливо и для выборок из подобных генеральных совокупностей. Значит, разность $(\bar{x} - \hat{x})$ можно использовать в качестве меры асимметрии: чем больше эта разность, тем больше асимметрия. Асимметрия называется положительной, когда $\bar{x} > \hat{x}$, и отрицательной, когда $\bar{x} < \hat{x}$.

Рис. 2.6

Для получения безразмерной меры разность $(\bar{x} - \hat{x})$ делят на S. Число $(\bar{x} - \hat{x})/S$ называется первым коэффициентом асимметрии Пирсона (К.Пирсон (1857-1936) – один из создателей современной математической статистики). Второй коэффициент асимметрии Пирсона приблизительно равен первому, только мода заменяется медианой. Второй коэффициент асимметрии равен числу $3(\bar{x} - \tilde{x})/S$. Коэффициент 3 появился из-за того, что обычно верна приближенная формула $(\bar{x} - \hat{x}) \approx 3(\bar{x} - \tilde{x})$. Для выборки 2 имеем:

1-й коэффициент асимметрии Пирсона равен (3,77 - 3,75)/0,36 = 0,056;

2-й коэффициент асимметрии Пирсона равен 3*(3,77 – 3,76)/0,36 = =0,083.

Наша выборка извлечена из генеральной совокупности с симметричным законом распределения.

В теории вероятностей коэффициент асимметрии определяется как отношение третьего центрального момента к кубу среднеквадратического отклонения.

2.2.5. Вычисление выборочного среднего и выборочной дисперсии для объединения двух выборок

Пусть из одной и той же генеральной совокупности X извлечены две

выборки объемов n_1 и n_2 и для каждой выборки отдельно вычислены выборочное среднее и выборочная дисперсия: x_1, x_2, S_1^2, S_2^2. Найдем параметры x и S^2 для объединения этих выборок .

1. $\bar{x} = (\sum_{j=1}^{n1+n2} x_j)/(n_1+n_2)$, тогда $(n_1+n_2)\bar{x} = \sum_{j=1}^{n1+n2} x_j = n_1\bar{x}_1 + n_2\bar{x}_2$.

Отсюда

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}.$$

Эта же формула применяется и тогда, когда выборки сгруппированы.

2. $(n_1+n_2)*S^2 = \sum_{j=1}^{n1+n2} x_j^2 - (n_1+n_2)\bar{x}^2 = \sum_{j=1}^{n1} x_j^2 + \sum_{j=n1+1}^{n2} x_j^2 - (n_1+n_2)\bar{x}^2 +$

$$+(-n_1\bar{x}_1^2 + n_1\bar{x}_1^2 - n_2\bar{x}_2^2 + n_2\bar{x}_2^2) = n_1S_1^2 + n_2S_2^2 + n_1\bar{x}_1^2 + n_2\bar{x}_2^2 - \frac{(n_1\bar{x}_1^2 + n_2\bar{x}_2^2)^2}{n_1+n_2}$$

.

Рассмотрим выражение

$$n_1\bar{x}_1^2 + n_2\bar{x}_2^2 - \frac{(n_1\bar{x}_1^2 + n_2\bar{x}_2^2)^2}{n_1+n_2}.$$

После приведения к общему знаменателю получаем, что оно равно

$$\frac{n_1n_2}{n_1+n_2} * (\bar{x}_1 - \bar{x}_2)^2.$$

Следовательно,

$$S^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1+n_2} + \frac{n_1n_2}{(n_1+n_2)^2} * (\bar{x}_1 - \bar{x}_2)^2.$$

Но если выборки извлечены из одной и той же генеральной совокупности, то числа $\bar{x}_1$ и $\bar{x}_2$ не должны сильно отличаться друг от друга. Кроме того, легко видеть, что

$$\frac{n_1n_2}{(n_1+n_2)^2} \le 1/4.$$

Поэтому членом $\frac{n_1n_2}{n_1+n_2} * (\bar{x}_1 - \bar{x}_2)^2$ можно пренебречь и положить

$$S^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1+n_2}.$$

Для примера разобьем выборку 2 на две части по 25 вариант в каждой. Как разбивать – все равно, главное, чтобы выбор был случайным. Пусть выборки будут такие:

1-я часть:

3,7 3,85 3,7 3,78 3,6 4,45 4,2 3,87 3,33 3,76

3,75 4,03 3,75 4,18 3,8 4,75 3,25 4,1 3,55 3,35
3,38 3,3 4,15 3,95 3,5

Для этой выборки $\bar{x}_1 = 3{,}8$; $S_1^2 = 0{,}132$.

2-я часть:

3,88 3,71 3,15 4,15 3,8 4,22 3,75 3,58 3,55 4,08
4,03 3,24 4,05 3,56 3,05 3,58 3,98 3,88 3,78 4,05
3,4 3,8 3,06 4,38 4,2

Для этой выборки $\bar{x}_2 = 3{,}76$; $S_2^2 = 0{,}131$. Тогда

$$\bar{x} = \frac{25 * 3{,}8 + 25 * 3{,}76}{50} = 3{,}78;$$

$$S^2 = \frac{25 * 0{,}132 + 25 * 0{,}131}{50} = 0{,}1315;\ S = 0{,}36.$$

Небольшие отличия $\bar{x}$ и S^2 от найденных ранее получились из-за того, что $\bar{x}_1$, $\bar{x}_2$, S_1^2, S_2^2 считались “в лоб”, для несгруппированных выборок.

2.2.6. Общая, межгрупповая и внутригрупповая дисперсии

Пусть из k выборок объемов $n_1, n_2, \ldots, n_k$ соответственно образована одна выборка объема $n = n_1 + n_2 + \ldots + n_k$. Обозначим через $\bar{x}$, $\bar{x}_1, \ldots, \bar{x}_k$, S^2, $S_1^2, \ldots, S_k^2$ выборочные средние и выборочные дисперсии объединенной выборки и исходных выборок соответственно. Обобщая формулы, рассмотренные выше, получим, что объединенная дисперсия равна

$$S^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{\sum_{i=1}^{k} S_i^2 n_i}{n} + \frac{\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i}{n}.$$

Величину S называют еще общей дисперсией. Величины $S_1^2, S_2^2, \ldots, S_k^2$ называют внутригрупповыми дисперсиями.

Величина $\frac{1}{n}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i$ называется межгрупповой дисперсией. Она показывает, насколько в среднем выборочные средние отдельных выборок отличаются от общего выборочного среднего. Тем самым оценивается, насколько внутригрупповые выборочные средние отличаются друг от друга. Мы разложили общую дисперсию на сумму межгрупповой дисперсии и среднего из внутригрупповых дисперсий.

2.2.7. Кривая Лоренца и показатели концентрации

С помощью кривой Лоренца представляют распределение некоторых ресурсов (капитала, земли, рабочей силы и т.п.) среди владельцев

ресурсов. Если значительная часть ресурсов сосредоточена у небольшой доли владельцев, говорят о высокой степени концентрации ресурсов.

Степень концентрации оценивают с помощью специальных коэффициентов. Неравномерность распределения ресурсов можно проследить и по кривой Лоренца, при построении этой кривой по горизонтальной оси откладывают накопленные доли владельцев ресурсов, а по вертикальной оси – относительные накопленные частоты объема ресурсов. Полученные точки соединяют отрезками.

Рассмотрим распределение в 1964 г. ферм в США, сгруппированных по величине занимаемых площадей (табл. 2.5).

Таблица 2.5

Площадь фермы, акр (1акр≈0,4га)	Число ферм 10^3	Общая площадь занимаемой земли, акр·10^3	Относительные частоты		Относительные накопленные частоты,%	
			Число ферм	Площадь земли	Число ферм	Площадь земли
[0-10)	183	778	0,057	0,0007	5,7	0,07
[10-50)	637	17325	0,202	0,0156	25,9	1,63
[50 - 100)	542	39589	0,172	0,0357	43,1	5,2
[100 - 180)	633	86592	0,201	0,0780	63,2	13,0
[180 - 260)	355	76857	0,112	0,0692	74,4	19,92
[260-500)	451	159598	0,143	0,1438	88,7	34,3
[500 - 1000)	210	144600	0,067	0,1302	95,4	47,32
≥1000	145	584848	0,046	0,5268	100,0	100,0
ВСЕГО	3156	1110187	1,00	1,00	–	–

Здесь ресурсы – это земля; владельцы ресурсов – фермы. Кривая Лоренца построена на рис. 2.7.

Если бы распределение земли было строго равномерным, то 5,7% ферм располагали бы 5,7% земли; 25,9% ферм располагали бы 25,7% земли и т.д., а кривая Лоренца стала бы биссектрисой координатного угла. Эта биссектриса называется линией равномерного распределения.

Чем сильнее кривая Лоренца отклоняется от линии равномерного распределения, тем выше концентрация ресурсов. В нашем случае 52,7% всей земли сконцентрировано у 4,6% крупных ферм. А на остальные 95,4% небольших ферм приходится менее половины угодий.

Степень концентрации можно оценить, вычисляя площадь фигуры А (см. рис.2.7), ограниченной линией равномерного распределения и кривой Лоренца. Если принять площадь квадрата за 1, то удвоенная площадь фигуры *A* равна разности 1 минус удвоенная площадь фигуры *B*.

Последняя легко считается как сумма площадей трапеций, составляющих фигуру В. Таким образом определяется коэффициент Джини:

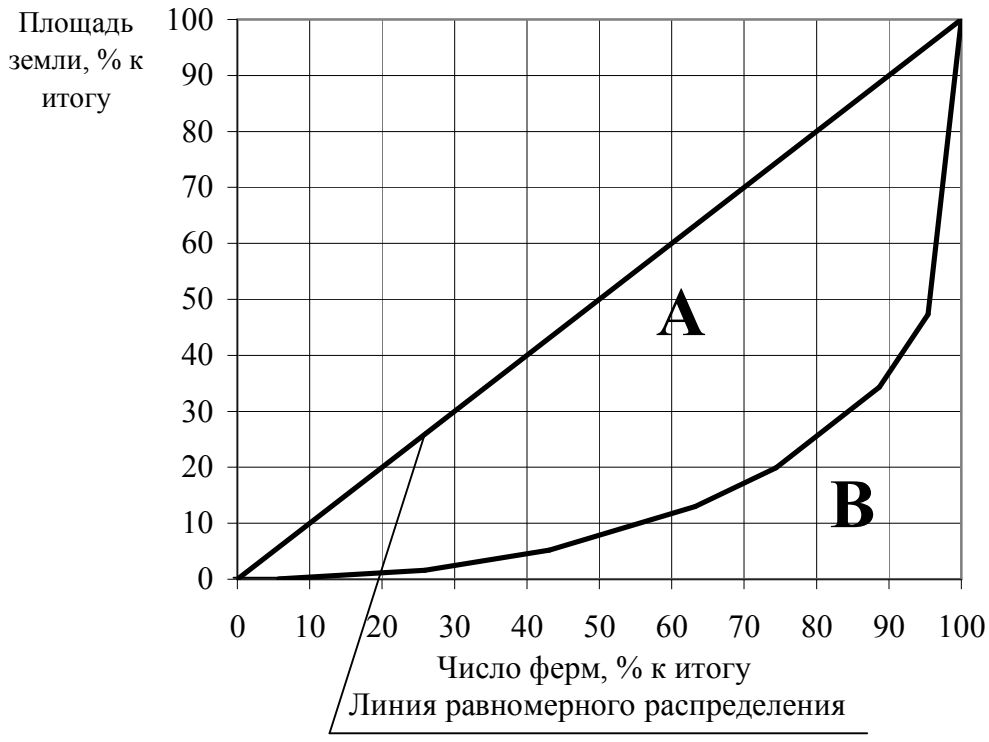


Рис. 2.7

$$G = 1 - 2\sum_{i=1}^{k} \nu_{xi}\nu_{yi-1}^{нак} - \sum_{i=1}^{k} \nu_{xi}\nu_{yi} = 1 - 2\sum_{i=1}^{k} \nu_{xi}\nu_{yi}^{нак} + \sum_{i=1}^{k} \nu_{xi}\nu_{yi},$$

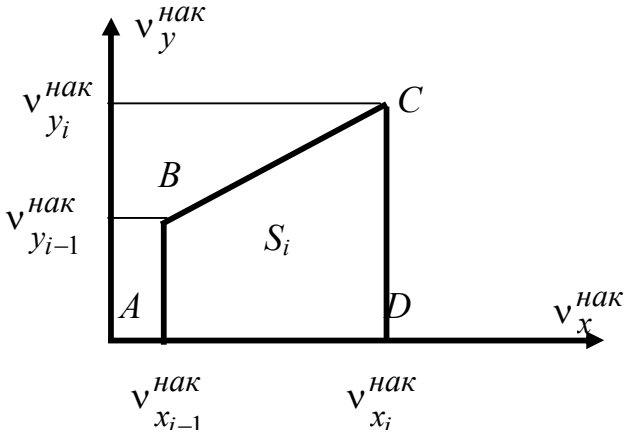
где k – число интервалов группировки;

ν_{xi} – относительная частота i-го интервала группировки владельцев ресурсов;

ν_{yi} – относительная частота i-го интервала группировки ресурсов;

$\nu_{yi}^{нак}$ – относительная накопленная частота i-го интервала группировки ресурсов.

На рис.2.8 показана i-я трапеция, составляющая фигуру B, и приведен расчет площади этой трапеции.



$$AB = \nu_{y_{i-1}}^{нак} = \nu_{y_i}^{нак} - \nu_{y_i};$$

$$CD = \nu_{y_i}^{нак};$$

$$AD = \nu_{x_i}^{нак} - \nu_{x_{i-1}}^{нак} = \nu_{x_i};$$

$$S_i = 0{,}5 \cdot (AB + BC) \cdot AD =$$

$$= 0{,}5 \cdot (2\nu_{y_i}^{нак} - \nu_{y_i}) \cdot \nu_{x_i} =$$

$$= 0{,}5 \cdot (2\nu_{y_{i-1}}^{нак} + \nu_{y_i}) \cdot \nu_{x_i}.$$

Рис. 2.8

Тогда

$$G = 1 - 2 \cdot S_B = 1 - 2 \cdot \sum_i S_i = 1 - \sum_i (2 \cdot \nu_{y_i}^{нак} - \nu_{y_{i-1}}) \cdot \nu_{x_i} =$$

$$1 - \sum_i (2 \cdot \nu_{y_{i-1}}^{нак} + \nu_{y_i}) \cdot \nu_{x_i} = 1 - 2\sum_i \nu_{x_i} \nu_{y_i}^{нак} + \sum_i \nu_{x_i} \nu_{y_i} =$$

$$1 - 2\sum_i \nu_{x_i} \nu_{y_{i-1}}^{нак} - \sum_i \nu_{x_i} \nu_{y_i}.$$

В нашем случае

G = 1 - 2(0,057*0,0007 + 0,202*0,0163 + 0,172*0,052 + 0,201*0,13 + +0,112*0,1992 + 0,143*0,343 + 0,067*0,4732 + 0,046*1) + (0,057*0,0007 + +0,202*0,0156 + 0,172*0,0357 + 0,201*0,078 + 0,112*0,0692 + 0,143* *0,1438 + 0,067*0,1302 + 0,046*0,5268) = 0,7113 (71,13%).

Другой коэффициент, оценивающий степень концентрации, называется коэффициентом Лоренца. Рассмотрим сумму

$$\sum_{i=1}^{k} \left| \nu_{xi} - \nu_{yi} \right|,$$

По известному свойству модуля

$$\sum_{i=1}^{k} \left| \nu_{xi} - \nu_{yi} \right| \le \sum_{i=1}^{k} \nu_{xi} + \sum_{i=1}^{k} \nu_{yi} = 1 + 1 = 2.$$

Число 2 получается в пределе, если практически 100% ресурсов сосредоточены у бесконечно малой доли владельцев. Поэтому, чем ближе к 2 эта сумма, тем выше концентрация ресурсов, тем неравномернее они распределены.

Коэффициент Лоренца определяется так:

$$L = \frac{\sum_{i=1}^{k} \left| \nu_{xi} - \nu_{yi} \right|}{2} * 100\%.$$

Для нашего случая получаем:

L = (1/2)*(|0,057 - 0,0007| + |0,202 - 0,0156| + |10,172 - 0,0357|+ +|0,201 - 0,0780| + |0,112 - 0,0692| + |0,143 - 0,1438| + |0,067 - 0,1302| + +|0,046 - 0,5268|)*100% = 54,5%.

Полученные значения коэффициентов Джини и Лоренца говорят о высокой степени концентрации земли на крупных фермах.

2.3. ЗАДАЧИ

1. Как изменятся выборочное среднее, мода, медиана и выборочная дисперсия, если каждый член выборки:

а) увеличить (уменьшить) на число *d*?

б) увеличить (уменьшить) в *k* раз?

В задачах 2 - 13 нужно представить выборку графически и найти её

числовые характеристики.

2. Диаметры 40 металлических шариков (мм):

8,53	8,59	8,51	8,59	8,41	8,46	8,57	8,62	8,45
8,51	8,46	8,55	8,61	8,68	8,52	8,43	8,40	8,41
8,54	8,47	8,53	8,55	8,43	8,47	8,59	8,63	8,56
8,42	8,58	8,60	8,52	8,56	8,56	8,60	8,54	8,61
8,42	8,54	8,57	8,68					

3. Продолжительность работы 30 электрических лампочек (часы /10):

51	56	69	31	56	49	51	53	74	51
63	48	53	51	64	50	59	84	55	82
55	72	70	54	51	77	98	62	73	55

4. Скорость автомобилей на некотором участке дороги (км/ч):

41	41	29	15	41	43	42	34	41	30
23	48	50	36	35	46	28	46	50	41
55	27	43	53	48	47	34	35	29	42
30	35	38	41	36	38	45	59	44	43

5. В «Северных прериях» Э. Сетон-Томпсон рассказывает, что из окна вагона поезда канадской Тихоокеанской железной дороги в районе Альберты он видел 26 стад антилоп. В книге указывается количество животных в каждом стаде:

8	14	7	18	3	9	4	1	6	12	2	8	1
3	4	6	18	4	25	4	34	6	5	6	16	4

6. Пятьюдесятью абитуриентами на вступительных экзаменах получены следующие баллы (из 20 возможных):

12	14	19	15	14	18	13	16	17	12	20	17	15
13	17	16	20	14	14	13	17	16	15	19	16	15
18	17	15	14	15	15	18	15	15	19	14	16	18
18	15	15	17	15	16	16	14	14	17	19		

7. Результаты исследования прочности 200 образцов бетона на сжатие:

Предел прочности (МПа)	[19,20)	[20,21)	[21,22)	[22,23)	[23,24)	[24,25)
Количество образцов	10	26	56	64	30	14

8. Продолжительности автомобильных рейсов, определенные по дорожным ведомостям:

Продолжительность рейса (суток)	[0,2)	[2,4)	[4,6)	[6,8)	[8,10)
Число рейсов	400	600	900	700	400

9. Распределение частот барометрического давления воздуха в городе

Ташкенте с мая по август 1897г.:

Давление (мм рт. ст.)	709	710	711	712	713	714	715	716	717
Количество дней	2	7	24	30	44	48	36	35	32
Давление (мм рт. ст.)	718	719	720	721	722	723	724	725	726
Количество дней	26	23	21	14	12	8	7	2	1

10. Следующее распределение частот было получено в результате эксперимента с разведением мышей:

Количество мышей в одном помете (шт.)	1	2	3	4	5	6	7	8	9
Частота	7	11	16	17	26	31	11	1	1

11. Длины початков кукурузы в дюймах (с точностью до половины дюйма):

Длина початка	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10
Частота	1	1	8	33	70	110	176	172	124	61	32	10	2

12. При подсчете количества простых чисел в восьмом миллионе весь интервал был разбит на 2000 групп по 500 последовательных чисел в каждой группе. Пусть X – количество простых чисел в группе, $N(x)$ – число групп, в которых по X простых чисел. В результате подсчетов получилась таблица

X	18	19	20	21	22	23	24	25	26	27	28	29	30	31
$N(x)$	1	4	5	6	11	18	48	63	70	102	141	149	165	188
X	32	33	34	35	36	37	38	39	40	41	42	43	44	–
$N(x)$	203	181	160	141	115	78	63	38	16	15	14	4	1	–

Показать, что, если бы простые числа были расположены случайно, дисперсия была бы значительно больше.

13. Приведенные ниже числа представляют собой затраты в долл. на питание 66 семей, каждая из которых состоит из 4 человек (данные конца 1960-х годов).

48	44	40	51	44	45	46	57	57	34	38	47
48	52	39	41	39	38	43	29	45	54	38	28
48	28	47	52	33	40	45	40	55	45	32	32
56	41	52	36	50	37	53	42	38	49	46	42
41	51	39	47	37	35	44	39	32	50	46	41
43	40	45	44	53	46						

14. Даны следующие 7 выборок объема 20, сгруппированных по одним

и тем же интервалам:

$[x_{i-1}, x_i)$	n_i^1	n_i^2	n_i^3	n_i^4	n_i^5	n_i^6	n_i^7
[12-15)	2	6	4	1	0	2	2
[15-18)	4	3	4	1	1	3	8
[18-21)	8	2	4	16	18	5	5
[21-24)	4	3	4	1	1	8	3
[24-27)	2	6	4	1	0	2	2

а) Не производя вычислений, на глаз, сравнить следующие пары стандартных отклонений: S_1 и S_2; S_2 и S_3; S_1 и S_4; S_4 и S_5; S_1 и S_6; S_2 и S_6; S_6 и S_7.

в) Вычислить стандартные отклонения.

15. Преподаватели *A* и *B* ведут разные курсы у одних и тех же студентов. Преподаватель *A*, оценивая знания студентов, предлагает им письменные работы и подсчитывает баллы, набранные студентами за ответы на вопросы в работах. Преподаватель *B* поступает так: всего нужно посетить 24 занятия, за каждое посещение начисляется 2 очка. Баллы, полученные пятью студентами у этих преподавателей, таковы:

Студент	1	2	3	4	5
Преподаватель *A*	69	70	77	62	58
Преподаватель *B*	48	42	44	46	46

Вычислить коэффициент вариации баллов у каждого преподавателя. Почему оценкам преподавателя *B* не следует доверять?

16. Следующие баллы получены пятью студентами у преподавателей *X, Y, Z,* ведущих три смежных дисциплины:

Студент	1	2	3	4	5
Преподаватель *X*	168	190	147	158	179
Преподаватель *Y*	36	44	37	38	40
Преподаватель *Z*	76	78	85	67	65

Вычислить коэффициенты вариации оценок. Можно ли утверждать, что системы оценок сходны по своим принципам?

17. Варианты выборки называют стандартизированными, если они преобразуются по следующему правилу:

$$x_i' = (x_i - \bar{x})/S,$$

где x_i – старое значение варианты;

x_i' – новое значение варианты;

$\bar{x}$, S – выборочное среднее и стандартное отклонение исходной выборки.

а) Показать, что выборочное среднее преобразованной выборки равно 0, а стандартное отклонение равно 1.

б) Стандартизировать баллы студентов из задачи 15 и сравнить

успеваемость каждого студента по каждой дисциплине.

18. В приведенной ниже таблице фермы США сгруппированы по величине занимаемых площадей

Площадь, занимаемая фермой, акр (1акр ≈ 0,4га)	Число ферм $\cdot 10^3$	
	1940	1964
<10	506	183
[10-50)	1780	637
[50 -100)	1291	542
[100-180)	1310	633
[180 - 260)	486	355
[260 - 500)	459	451
[500 -1000)	164	210
Площадь, занимаемая фермой, акр (1акр ≈ 0,4га)	Число ферм $\cdot 10^3$	
> 1000	101	145
Всего	6097	3156

а) Почему пришлось прибегнуть к интервалам разной ширины?

б) Какие изменения произошли в фермерском хозяйстве США?

19. Ниже приводятся распределения возрастных групп населения США и острова Самоа в 1960г.:

Остров Самоа		США	
Возраст, лет	Численность 10^3 чел.	Возраст, лет	Численность 10^3 чел.
<5	3709	<5	16243
[5-10)	3244	[5-15)	24429
[10-15)	2993	[15-25)	22220
[15-20)	2182	[25 – 35)	23878
[20 - 25)	1444	[35 – 45)	21535
[25-35)	2261	[45 – 55)	17398
[35-45)	1844	[55-65)	13327
[45 - 55)	1162	[65-75)	8432
[55 - 65)	672	≥ 75	3862
≥ 65	540	–	–
Всего	20051	–	151324

а) Найти Q_1, $\tilde{x}$, Q_3 в каждом случае и объяснить результаты.

б) Определить долю населения старше 55 лет в каждой стране.

20. Ниже приводятся два следующих распределения. Годовой денежный доход лиц, окончивших только среднюю школу, и лиц,

имеющих высшее образование (4-годичный колледж), данные налоговых деклараций за 1967 год.

Доход, долл.	% лиц с данным доходом	
	Среднее образование	Бакалавры
<2000	5,6	3,8
[2000 - 4000)	9,2	4,9
[4000-7000)	31,8	15,5
[7000-10000)	32,6	25,1
[10000 -15000)	16,2	29,4
≥15000	4,6	21,3
Всего	100	100

а) Найти Q_1, $\tilde{x}$, Q_3 для каждой выборки и объяснить результаты.

б) Подобрать разумные правые границы для последних интервалов, вычислить $\bar{x}$ и S для каждой выборки и объяснить результаты

21. Построить кривую Лоренца и найти коэффициент Джини для следующих данных:

Группы предприятий по численности занятых, чел.	[1 - 500)	[500-1000)	[1000-5000)	[5000-10000)	≥10000
Число предприятий	4941	1173	1408	202	94
Численность занятых, млн. чел.	0,99	0,84	2,92	1,36	1,81

22. Построить кривую Лоренца и найти коэффициент Джини для следующих данных:

Группы населения, ранжированные по уровню среднедушевого дохода (по 10% от общей численности населения)	1	2	3	4	5	6	7	8	9	10
Удельный вес в совокупном доходе, (%)	2,3	5,1	6,0	6,9	7,8	8,6	9,7	11,5	15,8	26,3

3. ОБРАБОТКА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

Музыку я разъял, как труп.
*Поверил я алгеброй гармони*ю.
А. Пушкин. Моцарт и Сальери

3.1. ДВУМЕРНЫЕ ВЫБОРКИ

До сих пор мы считали, что генеральная совокупность X – одномерная случайная величина. В результате эксперимента такая случайная величина принимает одно значение – x. Но генеральная совокупность может быть и многомерной случайной величиной. Здесь мы ограничимся случаем двумерных случайных величин (X, Y). Составляющие двумерного вектора – случайные величины X и Y - могут быть как зависимыми, так и независимыми. Значения двумерной случайной величины (X, Y) – это упорядоченные пары чисел (x, y). Выборка объема n из двумерной генеральной совокупности - это набор из n упорядоченных пар (x_i,y_i), i = =1,2,...,n. Такие выборки называются двумерными. Рассмотрим несколько примеров.

1. Генеральная совокупность (X, Y) – это множество предложений русского языка. Случайная величина X – число слов в предложении. Случайная величина Y - число букв в предложении. Ниже приводится текст из 10 предложений - отрывок из рассказа А.П. Чехова «Анна на шее». После каждого предложения в скобках указано количество слов (x_i) и количество букв (y_i) в данном предложении. Пробелы здесь не учитываются.

«Поехали на бал. (3,12) Вот и дворянское собрание, и подъезд со швейцаром. (8,41) Передняя с вешалками, шубы, снующие лакеи и декольтированные дамы, закрывающиеся веерами от сквозного ветра; пахнет светильным газом и солдатами. (19,122) Когда Аня, идя вверх по лестнице под руку с мужем, услышала музыку и увидела в громадном зеркале всю себя, освещенную множеством огней, то в душе ее проснулась радость и то самое предчувствие счастья, какое испытывала она в лунный вечер на полустанке. (41,203) Она шла гордая, самоуверенная, в первый раз чувствуя себя не девочкой, а дамой, и невольно походкою и манерами подражая своей покойной матери. (22,106) И в первый раз в жизни она чувствовала себя богатой и свободной. (12,52) Даже присутствие мужа не стесняло ее, так как, перейдя порог собрания, она уже угадала инстинктом, что близость старого мужа нисколько не унижает ее, а, наоборот, кладет на нее печать пикантной таинственности, которая так нравится мужчинам.

(35,197) В большом зале уже гремел оркестр, и начались танцы. (9,42) После казенной квартиры, охваченная впечатлениями света, пестроты, музыки, шума, Аня окинула взглядом залу и подумала: «Ах, как хорошо!» и сразу отличила в толпе всех своих знакомых, всех, кого она раньше отличала на вечерах и гуляньях, всех этих офицеров, учителей, адвокатов, чиновников, помещиков, его сиятельство, Артынова и дам высшего общества, разодетых, сильно декольтированных, красивых и безобразных, которые уже занимали свои позиции в избушках и павильонах благотворительного базара, чтобы начать торговлю в пользу бедных. (72,43) Громадный офицер в эполетах - она познакомилась с ним на Старо-Киевской улице, когда была гимназисткой, а теперь не помнила его фамилии – точно из-под земли вырос и пригласил ее на вальс, и она отлетела от мужа, и ей уже казалось, будто она плыла на парусной лодке, в сильную бурю, а муж остался далеко на берегу.» (53,247)

В табличном виде выборка выглядит так:

Предложение	1	2	3	4	5	6	7	8	9	10
Количество слов (x_i)	3	8	19	41	22	12	35	9	72	53
Количество букв (y_i)	12	41	122	203	106	52	197	42	439	247

2. Из большого мешка, содержащего монеты одинакового достоинства, случайным образом отобраны 10 монет. Каждая монета была взвешена, и для каждой определен ее возраст:

Монета	1	2	3	4	5	6	7	8	9	10
Время обращения, лет (x_i)	5	9	14	17	23	31	35	42	46	50
Вес, г (y_i)	2,82	2,85	2,80	2,80	2,79	2,78	2,77	2,79	2,75	2,72

3.Результаты подбрасывания двух кубиков:

№ подбрасывания	1	2	3	4	5	6	7	8	9	10
Число очков, выпавшее на 1-м кубике	4	6	5	1	1	5	1	5	6	6
Число очков, выпавшее на 2-м кубике	5	1	2	3	6	1	1	6	2	6

3.2. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ДВУМЕРНЫХ **ВЫБОРОК — ДИАГРАММЫ РАССЕЯНИЯ**

Графическое представление одномерной выборки – это гистограмма. Двумерные выборки удобно представлять с помощью так называемых диаграмм рассеяния. Каждый элемент двумерной выборки представляется точкой на плоскости с координатами (x_i, y_i), $i = 1,2,\ldots,n$. Диаграммы рассеяния, представляющие двумерные выборки из наших примеров, приведены на рис.3.1 – 3.3.

На рис. 3.1 хорошо видно, что точки на диаграмме рассеяния группируются относительно некоторой прямой, причем чем больше слов в предложении, тем больше в нем букв. В таком случае говорят, что между числом слов и числом букв в предложении существует положительная линейная корреляция (слово “корреляция” означает связь). Во втором случае (см. рис. 3.2) хорошо заметна отрицательная линейная корреляция между массой монеты и ее возрастом. Точки на третьей диаграмме рассеяния (см. рис. 3.3) расположены хаотически. Следует допустить отсутствие связи между числом очков, выпавшим на первом кубике, и числом очков, выпавшим на втором. Другими словами разумно предположить, что случайные величины X и Y- числа очков, выпавшие на первом и втором кубике соответственно, независимы.

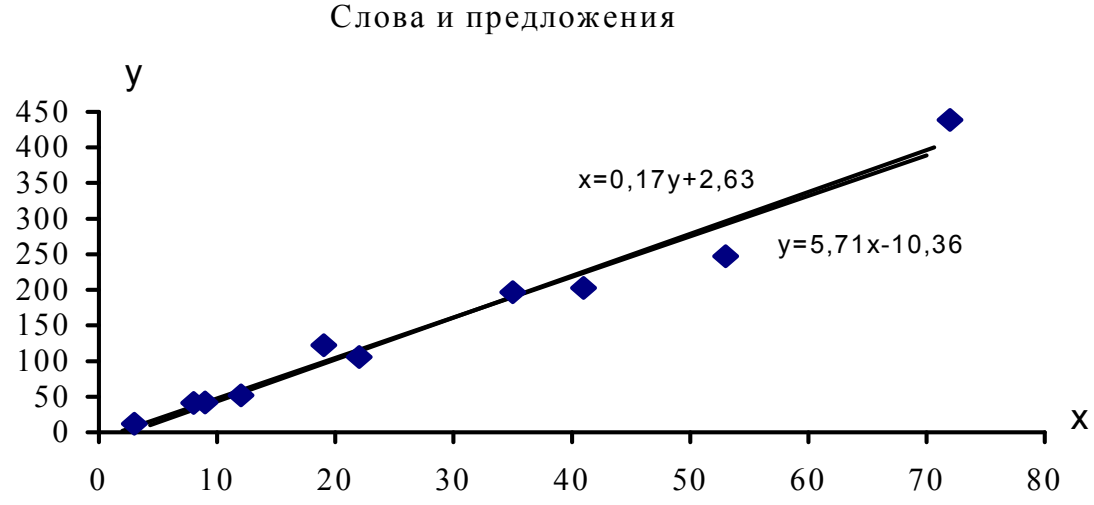


Рис. 3.1

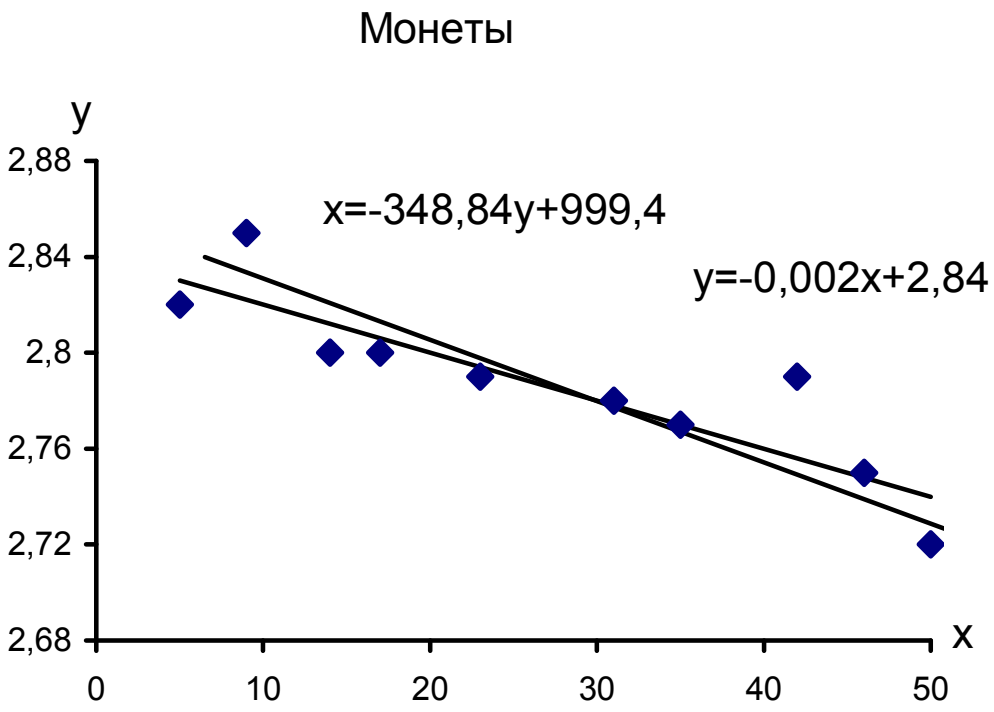

Монеты
y
2,88
2,84
2,8
2,76
2,72
2,68
x=-348,84y+999,4
y=-0,002x+2,84
x
0
10
20
30
40
50

Рис. 3.2

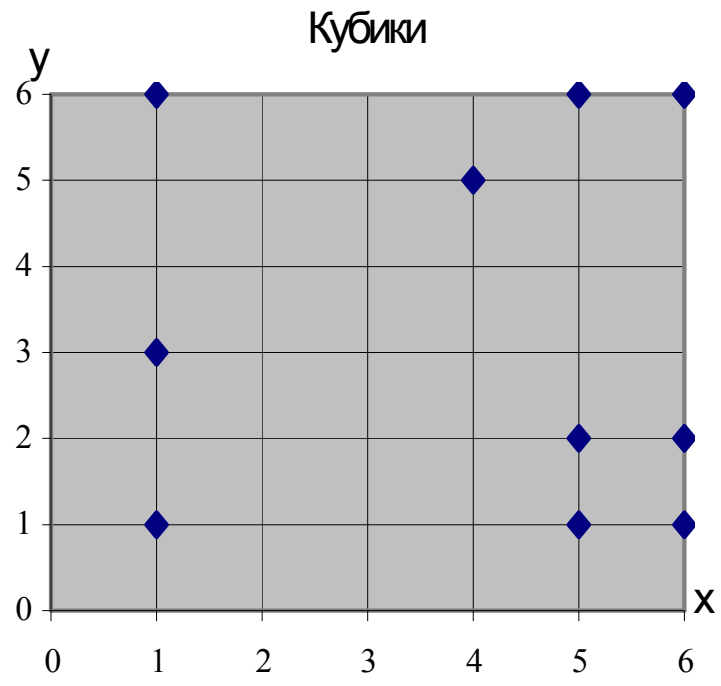

Кубики
y
6
5
4
3
2
1
0
x
0
1
2
3
4
5
6

Рис. 3.3

3.3. ВЫБОРОЧНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ — ЧИСЛОВАЯ ХАРАКТЕРИСТИКА ДВУМЕРНОЙ ВЫБОРКИ

В теории вероятностей числовой мерой линейной связи между случайными величинами X и Y служит коэффициент корреляции $\rho(X,Y)$, определяемый по формуле

$$\rho(X,Y)=\frac{M(XY)-MX)\cdot M(Y)}{\sigma(X)\cdot\sigma(Y)}.$$

Коэффициент корреляции обладает следующими свойствами:

1. Если X и Y независимы, то $\rho(X,Y) = 0$.
2. $|\rho(X,Y)| \le 1$.
3. $|\rho(X,Y)| = 1$ тогда и только тогда, когда случайные величины X и Y связаны линейной зависимостью $Y = aX + b$.

В математической статистике аналогом является выборочный коэффициент корреляции r, определяемый по формуле

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\,\bar{y}}{S_x S_y}.$$

Нетрудно убедиться в следующих свойствах выборочного коэффициента корреляции:

1. $|r| \le 1$.
2. $|r| = 1$ тогда и только тогда, когда точки (x_i, y_i) лежат на одной прямой.
3. Если точки (x_i, y_i) расположены на диаграмме рассеяния хаотически, то значение r весьма близко к нулю.

Вычислим значение выборочного коэффициента корреляции для наших трех случаев. Для удобства будем использовать таблицы.

Пример с текстом (табл. 3.1).

Таблица 3.1

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	3	12	36	9	144
2	8	41	328	64	1681
3	19	122	2318	261	14884
4	41	203	8323	1681	41209
5	22	106	2332	484	11236
6	12	52	624	144	2704
7	35	197	6895	1225	38809
8	9	42	378	81	1764
9	72	439	31608	5184	192721
10	53	247	13091	2809	61009
Сумма	274	1461	65933	12042	366161

Отсюда:

$\overline{x} = 27{,}4$; $S_x^2 = 1204{,}2 - 27{,}4^2 = 453{,}44$; $S_x = 21{,}3$;

$\overline{y} = 146{,}1$; $S_y^2 = 36616{,}1 - 146^2 = 15270{,}9$; $S_y = 123{,}58$;

$$\frac{1}{10}\sum_{i=1}^{10} x_i y_i = 6593{,}3;\ r = \frac{6593{,}3 - 27{,}4 \cdot 146{,}1}{21{,}3 \cdot 123{,}58} = 0{,}984\ .$$

Это значение весьма близко к единице. Число букв и число слов в предложении почти линейно зависят друг от друга.

Пример с монетами (табл. 3.2)

Таблица 3.2

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	5	2,82	14,1	25	7,95
2	9	2,85	25,65	81	8,12
3	14	2,80	39,2	196	7,84
4	17	2,80	47,6	289	7,84
5	23	2,79	64,17	529	7,78
6	31	2,78	86,18	961	7,73
7	35	2,77	96,95	1225	7,67
8	42	2,79	117,18	1764	7,78
9	46	2,75	126,5	2116	7,56
10	50	2,72	136	2500	7,40
Сумма	272	27,87	753,53	9686	77.67

$$r = \frac{75{,}353 - 27{,}2 \cdot 2{,}787}{15{,}13 \cdot 0{,}036} = -0{,}83\ .$$

Такое значение r указывает на достаточно сильную отрицательную линейную зависимость между возрастом монеты и ее массой.

Пример с кубиками (табл. 3.3).

Таблица 3.3

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	4	5	20	16	25
2	6	1	6	36	1
3	5	2	10	25	4
4	1	3	3	1	9
5	1	6	6	1	36
6	5	1	5	25	1
7	1	1	1	1	1

Окончание табл. 3.3

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
8	5	6	30	25	36
9	6	2	12	36	4
10	6	6	36	36	36
Сумма	40	33	129	202	153

$$r = \frac{12{,}9 - 4 \cdot 3{,}3}{2{,}05 \cdot 2{,}1} = -0{,}07$$

Такое маленькое значение r указывает на отсутствие связи между результатами бросаний кубиков, что соответствует интуитивному представлению о независимости бросаний.

В дальнейшем выражение $\frac{1}{n}\sum_{1}^{n} x_i y_i - \bar{x}\bar{y}$ будем обозначать через S_{xy} и назовем его выборочной ковариацией.

3.4. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Обратимся к примеру с текстом. На рис. 3.1 хорошо видно, что точки (x_i, y_i) группируются около прямой. Естественным образом возникает задача подбора уравнения этой прямой. Например, для того, чтобы предсказать, сколько примерно букв будет содержать предложение с заданным количеством слов, можно подобрать два уравнения:

$y = ax + b$ (независимая переменная - число слов, функция – число букв);

$x = cy + d$ (независимая переменная - число букв, функция – число слов).

Каждое из таких уравнений называется уравнением регрессии. (Слово “прогресс” означает развитие, движение вперед, слово “регресс” означает упрощение, движение назад). В случае уравнения $y = ax + b$ говорят о регрессии y на x; в случае уравнения $x = cy + d$ говорят о регрессии x на y.

В нашем примере каждая из переменных, как x, так и y, может быть объявлена независимой. Возможны ситуации, когда независимая переменная определяется однозначно. Например, можно исследовать растворимость некоторого вещества (переменная y) в зависимости от температуры растворителя (переменная x). Здесь x – независимая переменная, ее значение можно установить заранее, а y – статистически зависимая переменная. Исследуется только зависимость y от x.

Допустим, мы хотим подобрать коэффициенты уравнения $y = ax + b$ так, чтобы это уравнение наилучшим образом соответствовало экспериментальным данным (x_i, y_i); $i = 1,2,\ldots,n$. Но ведь понятие «наилучшим образом» не является строгим. Между точками на рис. 3.1

можно провести бесконечно много «хороших» прямых. Какая же из них «лучшая»?

Общепринятым способом определения неизвестных коэффициентов уравнения регрессии является метод наименьших квадратов, разработанный А. Лежандром (1806) и К. Гауссом (1821). Идея метода наименьших квадратов такова. Пусть нужно подобрать неизвестные коэффициенты $a_1, a_2,\ldots,a_k$ уравнения регрессии $y = f(a_1,a_2,...,a_k, x)$. Рассмотрим экспериментальную точку (x_i,y_i) и вычислим отклонение ординаты y_i точки от теоретического значения $f(a_1,a_2,...,a_k, x_i)$ (рис.3.4).

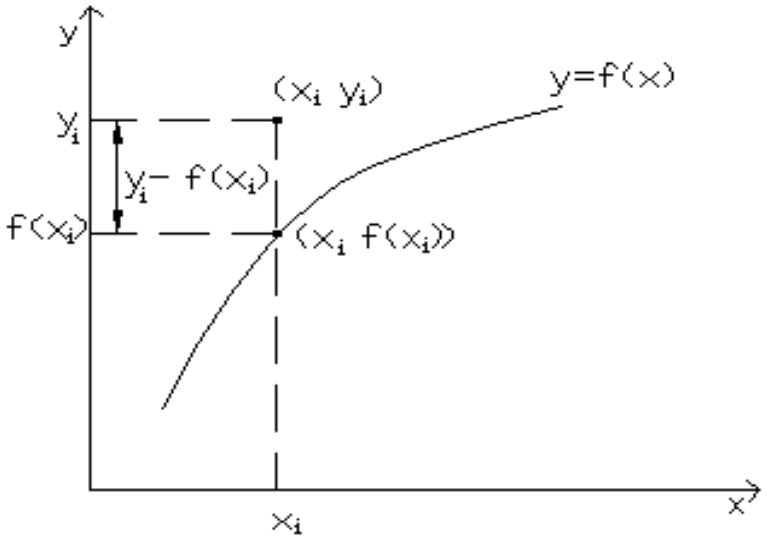


Рис. 3.4

$$d_i = y_i - f(a_1,a_2,...,a_k, x_i), \quad i = 1,2,...,n.$$

Неизвестные значения $a_1,a_2,...,a_k$ подберем из условия минимизации суммы квадратов отклонений d_i:

$$S = \sum_{i=1}^{n} d^2{}_i = \sum_{i=1}^{n} [y_i - f(a_1,a_2,..a_k,x_i)]^2 \longrightarrow \min.$$

Если теперь приравнять нулю частные производные $\frac{\partial S}{\partial a_1},\ldots,\frac{\partial S}{\partial a_k}$, получится система из k уравнений для определения k неизвестных чисел $a_1,a_2,...,a_k$.

Составим эту систему и решим ее в случае линейного уравнения регрессии. Нужно определить два неизвестных коэффициента a и b уравнения прямой $y = ax + b$. Имеем

$$S = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \longrightarrow \min;$$

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} x_i \left[y_i - (ax_i + b) \right] = 0; \quad \frac{\partial S}{\partial b} = -2\sum_{i=1}^{n} \left[y_i - (ax_i + b) \right] = 0.$$

Раскрывая скобки, получаем:

$$\begin{cases} a\sum_{i=1}^{n} x^2{}_i + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \end{cases}$$

Разделим второе уравнение системы на n. Уравнение примет вид $\overline{y} = a\overline{x} + b$, откуда $b = \overline{y} - a\overline{x}$.

Разделим на n первое уравнение системы и подставим в него полученное выражение b через a. После несложных преобразований имеем:

$$aS_x^{\,2} = S_{xy} \Rightarrow a = \frac{S_{xy}}{S_x^{\,2}}.$$

Итак,

$$a = S_{xy} / S_x^2; \qquad b = \overline{y} - a\overline{x}.$$

Уравнение $y = ax + b$ можно переписать в виде

$$(y - \overline{y}) = \frac{S_{xy}}{S_y^{\,2}}(x - \overline{x}),$$

следовательно, наша прямая проходит через точку $(\overline{x}; \overline{y})$.

Аналогично определяют коэффициенты c и d линейного уравнения регрессии x на y, $x = cy + d$.

$$c = S_{xy} / S_y^2 \quad d = \overline{x} - c\overline{y}.$$

Само уравнение можно записать так:

$$(x - \overline{x}) = \frac{S_{xy}}{S_y^{\,2}}(y - \overline{y}).$$

В этом случае минимизируется сумма квадратов отклонений по координате x:

$$S = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \left[x_i - \left(cy_i - d \right)^2 \right] \longrightarrow \min.$$

Найдем коэффициенты линейных уравнений регрессии y на x и x на y для примеров с текстом и монетами. Все необходимые расчеты уже были сделаны при вычислении коэффициентов корреляции (см. пункт 4.3).

Пример с текстом:

$\overline{x}$ =27,4; $\overline{y}$ =146,1; $\frac{1}{10}\sum_{i=1}^{10} x_i y_i = 6593,3$;

$$S_x^2 = 453,44; \quad S_y^2 = 15270,9.$$

Тогда

$$S_{xy} = 6593,3 - 27,4 * 146,1 = 2590,3;$$

$$a = \frac{S_{xy}}{S_x^2} = \frac{2590,16}{453,44} = 5,71;$$

$$b = \overline{y} - a\overline{x} = 146,1 - 5,71 * 27,4 = -10,42.$$

Уравнение регрессии y на x таково: $y = 5,71x - 10,42$.

Вычислим несколько значений y для разных x.

x	10	20	30	40	50	60	70
y	46,7	103,8	160,9	218,0	275,1	332,2	389,3

Найдем коэффициенты c и d уравнения регрессии x на y.

$$c = \frac{S_{xy}}{S_y^2} = \frac{2590,16}{15270,9} = 0,17; \; d = \overline{x} - c\overline{y} = 2,56.$$

Тогда $x = 0,17y + 2,56$.

Вычислим несколько значений x для разных y.

y	10	50	100	200	300	400
x	4,3	11,1	19,6	36,6	53,6	70,6

Эти прямые приведены на рис. 4.1. Прямые почти совпадают – еще одно доказательство сильной линейной зависимости между числом слов и числом букв в предложении.

Пример с монетами.

$\overline{x}$ =27,2; $\overline{y}$ =2,787; $\frac{1}{10}\sum_{i=1}^{10} x_i y_i = 75,353$; $S_x^2 = 228,76$; $S_y^2 = 0,00129$;

$$S_{xy} = 75,353 - 27,2 * 2,787 = -0,45; \; a = \frac{S_{xy}}{S_x^2} = \frac{-0,45}{228,76} = -0,002;$$

$b = \overline{y} - a\overline{x} = 2,787 + 0,002 * 27,2 = 2,84$. Тогда y=-0,002x + 2,84.

Коэффициент a отрицателен и очень мал. Несколько значений y:

x	5	20	35	50
y	2,83	2,80	2,77	2,74

$$c = \frac{S_{xy}}{S_y^2} = \frac{-0,45}{0,00129} = -348,84; \; d = \overline{x} - c\overline{y} = 999,4.$$

Уравнение регрессии x на y: $x = -348{,}84y + 999{,}4$.

Несколько значений x:

y	2,85	2,80	2,79	2,78	2,77	2,75
x	5,2	22,6	26,1	29,6	33,1	40,1

Эти прямые показаны на рис. 3.2. Прямые не так близки, как в случае с текстом, масса монеты не столь жестко связана с ее возрастом, как число слов и букв в предложении.

3.5. ДРУГИЕ УРАВНЕНИЯ РЕГРЕССИИ

3.5.1. Парабола второго порядка

Уравнение имеет вид $y = ax^2 + bx + c$.

Метод наименьших квадратов дает такую систему линейных уравнений относительно неизвестных коэффициентов a, b, c:

$$\begin{cases} a\sum_i x_i^4 + b\sum_i x_i^3 + c\sum_i x_i^2 = \sum_i x_i^2 y_i; \\ a\sum_i x_i^3 + b\sum_i x_i^2 + c\sum_i x_i = \sum_i x_i y_i; \\ a\sum_i x_i^2 + b\sum_i x_i + cn = \sum_i y_i. \end{cases}$$

3.5.2. Показательная функция

Уравнение имеет вид $y = bx^a$.

Прологарифмируем левую и правую части, для определенности вычислим натуральные логарифмы

$$ln(y) = a \cdot ln(x) + ln(b).$$

Обозначим $ln(y)$ через y_1, $ln(x)$ через x_1, $ln(b)$ через b_1. Получаем уравнение относительно неизвестных коэффициентов a и b_1:

$$y_1 = ax_1 + b_1$$

Определив по методу наименьших квадратов числа a и b_1, найдем $b = e^{b1}$.

3.5.3. Степенная функция

Уравнение имеет вид $y = ba^x$.

Прологарифмировав левую и правую части, получим линейное

уравнение относительно неизвестных параметров

$$y_1 = a_1x + b_1,$$

где $y_1 = ln(y)$, $a_1 = ln(a)$, $b_1 = ln(b)$.

После определения параметров a_1 и b_1 находим числа a и b:

$$a = e^{a_1}, \qquad b = e^{b_1}.$$

3.5.4. Гиперболическая функция

Уравнение имеет вид $y = \dfrac{1}{ax+b}$.

Положив $y_1 = \dfrac{1}{y}$, получим линейное уравнение относительно a и b:

$$y_1 = ax + b.$$

О более сложных уравнениях регрессии можно прочитать в специальной литературе по корреляционному и регрессионному анализу.

3.5.5. О квазилинейном уравнении регрессии

Уравнение регрессии будем называть **квазилинейным**, если оно имеет вид

$$y(a_1, a_2, ..., a_k, x) = a_1 f_1(x) + a_2 f_2(x) + ... + a_{k-1} f_{k-1}(x) + a_k.$$

Здесь $a_1, a_2, ..., a_k$ – неизвестные параметры уравнения регрессии, $f_1(x)$, $f_2(x), ..., f_{k-1}(x)$ – заданные функции аргумента x.

Это уравнение линейно относительно неизвестных параметров, метод наименьших квадратов дает такую линейную систему уравнений для определения значений $a_1, a_2, ..., a_k$

$$\begin{cases} a_1\sum\limits_i f_1^2(x_i) + a_2\sum\limits_i f_1(x_i)f_2(x_i) + \cdots + a_{k-1}\sum\limits_i f_1(x_i\}f_{k-1}(x_i) + a_k\sum\limits_i f_1(x_i) = \sum\limits_i f_1(x_i)y_i; \\ a_1\sum\limits_i f_1(x_i)f_2(x_i) + a_2\sum\limits_i f_2^2(x_i) + \cdots + a_{k-1}\sum\limits_i f_2(x_i)f_{k-1}(x_i) + a_k\sum\limits_i f_2(x_i) = \sum\limits_i f_2(x_i)y_i; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_1\sum\limits_i f_1(x_i)f_{k-1}(x_i) + a_2\sum\limits_i f_2(x_i)f_{k-1}(x_i) + \cdots + a_{k-1}\sum\limits_i f_{k-1}^2(x_i) + a_k\sum\limits_i f_{k-1}(x_i) = \sum\limits_i f_{k-1}(x_i)y_i; \\ a_1\sum\limits_i f_1(x_i) + a_2\sum\limits_i f_2(x_i) + \cdots + a_{k-1}\sum\limits_i f_{k-1}(x_i) + a_k n = \sum\limits_i y_i. \end{cases}$$

Обозначим теоретические значения $y(a_1, a_2, \cdots, a_k, x_i)$ через $\hat{y}_i(x_i)$ или просто $\hat{y}_i$.

Левая часть последнего уравнения системы – сумма теоретических значений величины *y*, правая часть этого уравнения – сумма выборочных (экспериментальных) значений этой величины. Таким образом, в случае квазилинейного уравнения регрессии, суммы теоретических и экспериментальных значений величины *y* равны, $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$.

Умножим теперь первое уравнение системы на a_1, второе – на a_2,..., последнее, *k*-е уравнение, умножим на a_k. и сложим все уравнения. В результате получим равенство

$$\sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} y_i \hat{y}_i \quad \text{или} \quad \sum_{i=1}^{n} \hat{y}_i (\hat{y}_i - y_i) = 0.$$

Рассмотрим разность $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$. Обозначим через u_i разность $y_i - \hat{y}_i$. Из доказанных свойств величин $\hat{y}_i$ вытекает, что

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = 0; \quad \sum_{i=1}^{n} u_i (\hat{y}_i - \bar{y}) = 0; \quad \bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}.$$

Отсюда следует равенство

$$\frac{1}{n}\sum_{i} (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i} u_i^2 + \frac{1}{n}\sum_{i} (\hat{y}_i - \bar{y})^2.$$

Другими словами $s_y^2 = s_u^2 + s_{\hat{y}}^2$,

где s_y^2 – дисперсия экспериментальных значений y_i; $s_{\hat{y}}^2$ – дисперсия теоретических значений $\hat{y}_i$. Она называется **объясненной** дисперсией, ведь значения $\hat{y}_i$ однозначно определяются уравнением регрессии и обладают дисперсией только в том смысле, что разным значениям аргумента *x* соответствуют разные значения функции $\hat{y}_i(x)$. Число s_u^2 называется **остаточной (необъясненной)** дисперсией. Это – дисперсия разностей (остатков, отклонений) $y_i - \hat{y}_i$. Эти разности не имеют никакого отношения к уравнению регрессии и поэтому не могут быть объяснены с точки зрения уравнения регрессии. Чем сильнее экспериментальные значения отклоняются от теоретических, тем больше число s_u^2, тем хуже уравнение регрессии соответствует экспериментальным данным (объясняет экспериментальные данные).

Из сказанного вытекает, что всегда $s_y^2 \geq s_{\hat{y}}^2$, и равенство достигается, если $y_i = \hat{y}_i$, $i = 1,2,\cdots,n$.

3.5.6. Пример построения нелинейного уравнения регрессии

В качестве примера рассмотрим данные из табл. 3.4, где указаны объемы производства (x_i, 1000т) и фермерская цена (y_i долл. за 1т), скорректированная на индекс потребительских цен вишни в США в 1954 - 1969 гг.

Таблица 3.4

Год	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
x_i	204	260	168	239	192	218	185	266	276	150	344	248	200	198	228	278
y_i	267	174	228	208	225	243	227	217	163	345	154	165	299	325	294	188

Как правило, зависимость между ценой и объемом производства товара нелинейна. Диаграмма рассеяния для данного примера показана на рис. 3.5. Какой-либо отчетливой зависимости между значениями величин x и y на диаграмме рассеяния не видно. Но о приблизительно линейной или параболической зависимости сказать все же можно. Подкрепим эти рассуждения расчетами.

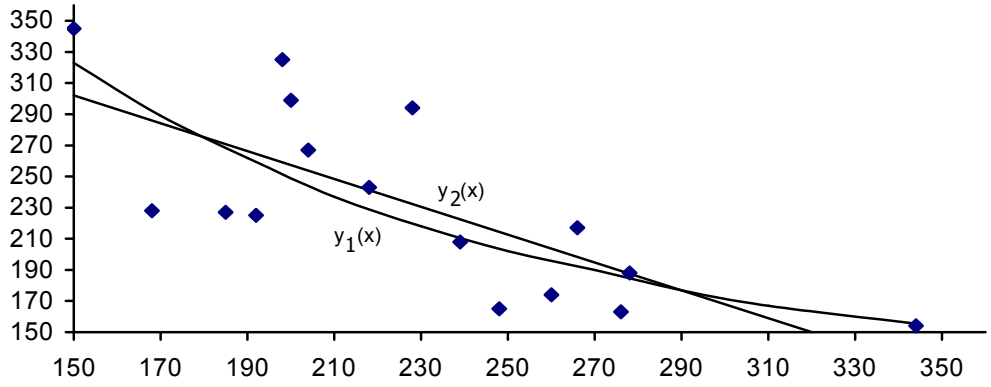


Рис. 3.5

Если вычислить по этим данным выборочный коэффициент корреляции, то получим, что r = -0,738, а это достаточно близко к 1. Ниже мы постараемся обосновать, почему парабола все-таки несколько лучше описывает эти данные, чем прямая. Коэффициенты системы линейных уравнений таковы:

$$n = 16; \quad \sum x_i = 3654; \quad \sum x_i^2 = 870918; \quad \sum x_i^3 = 216509904;$$

$$\sum x_i^4 = 560635921000; \quad \sum y_i = 3722; \quad \sum x_i y_i = 817695;$$

$$\sum x_i^2 y_i = 187221051.$$

Система для определения коэффициентов a, b, c параболического уравнения регрессии $y = ax^2 + bx + c$ получилась такой:

$$\begin{cases} 56063921000a + 216509904b + 870918c = 187221051; \\ 216509904a + 870918b + 3654c = 817659; \\ 870918a + 3654b + 16c = 3722. \end{cases}$$

Решение этой системы:

a = 0,00173; b = -1,723; c = 532,00.

Следовательно, $y = 0{,}00173x^2 - 1{,}723x + 532$.

Коэффициент a близок к нулю, это означает, что полученная парабола не слишком отличается от прямой линии.

Линейное уравнение регрессии, полученное по методу наименьших квадратов, таково: y = -0,887x +435,18.

Графики функций $y_1(x) = -0{,}00173x^2 - 1{,}723x + 532$ и $y_2(x) = -0{,}887x + 435{,}18$ показаны на рис. 3.5.

Если теперь рассчитать суммы квадратов отклонений:

$$S_1 = \sum_1^{16} [y_1(x_i) - y_i]^2, \qquad S_2 = \sum_1^{16} [y_2(x_i) - y_i]^2,$$

которые минимизируются при использовании метода наименьших квадратов, то, после округления, S_1 = 23953; S_2 = 23481. Разница, конечно, невелика, но рассеяние экспериментальных точек вокруг параболы все - таки меньше, чем вокруг прямой.

3.6. РАСЧЕТ КОЭФФИЦИЕНТОВ ЛИНЕЙНОГО УРАВНЕНИЯ РЕГРЕССИИ ПО СГРУППИРОВАННЫМ ДАННЫМ

При большом объеме n двумерной выборки ее группируют, получая т.н. корреляционную таблицу (табл. 3.5). Каждый из диапазонов значений составляющих двумерной выборки разбивают на несколько интервалов, как правило, одинаковой ширины. Затем подсчитывают частоты n_{ij} каждого из получившихся прямоугольников группировки – число пар двумерной выборки, попавших в данный прямоугольник.

Обозначения:

k – число интервалов группировки по составляющей x двумерной выборки;

x_i – середина i-го интервала группировки по составляющей x;

n_i – частота i-го интервала группировки по составляющей x, i = 1,2,..,k;

m - число интервалов группировки по составляющей y;

y_j – середина j-го интервала группировки по составляющей y;
l_j – частота j-го интервала группировки по составляющей y, $j = 1,2,...,m$;
n_{ij} – частоты прямоугольников группировки;
n – объем двумерной выборки.

Таблица 3.5

Середины интервалов x_i	Середины интервалов y_i $y_1\ y_2 \dots y_j \dots y_m$	Сумма частот
x_1	$n_{11}\ n_{12} \dots n_{1j} \dots n_{1m}$	n_1
x_2	$n_{21}\ n_{22} \dots n_{2j} \dots n_{2m}$	n_2
……..	……..	……..
x_i	$n_{i1}\ n_{i2} \dots n_{ij} \dots n_{im}$	n_i
……..	……..	……..
x_k	$n_{k1}\ n_{k2} \dots n_{kj} \dots n_{km}$	n_k
Сумма частот	$l_1\ \ l_2 \dots l_j \dots l_m$	n

Следующие соотношения очевидны:

$$\sum_i n_i = \sum_j l_j = \sum_i \sum_j n_{ij} = n; \sum_j n_{ij} = n_i; \qquad \sum_i n_{ij} = l_j.$$

Расчеты, выполненные по сгруппированной выборке, отличаются, конечно, от расчетов, выполненных непосредственно по исходным данным. Разница получается вследствие перехода к серединам интервалов. Но она, как правило, невелика, а вычисления по сгруппированной выборке получаются намного проще.

3.7. ИНДЕКС КОРРЕЛЯЦИИ

Выборочный коэффициент корреляции r является мерой линейной связи между составляющими двумерной выборки. Если такая связь существует, но не является линейной, значение r не может служить ее мерой. Чтобы оценить, насколько хорошо соответствует экспериментальным данным некоторое квазилинейное уравнение регрессии $y = f(x)$, используют индекс корреляции R_{yx}, определяемый формулой

$$R_{yx} = \sqrt{\frac{\sum_i (\hat{y}(x_i) - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}} = \sqrt{\frac{s_{\hat{y}}^2}{s_y^2}} = \sqrt{\frac{n\sum_i \hat{y}(x_i)^2 - \left(\sum_i y_i\right)^2}{n\sum_i y_i^2 - \left(\sum_i y_i\right)^2}}.$$

Если экспериментальные числа y_i совпадают с теоретическими значениями $y(x_i)$ (точки (x_i, y_i) на диаграмме рассеяния лежат на кривой $y = f(x)$), то $R_{yx} = 1$.

Так как всегда $s_y^2 \geq s_{\hat{y}}^2$, то $0 \leq R \leq 1$.

Чем ближе к 1 число R_{yx} , тем точнее уравнение регрессии соответствует экспериментальным данным, тем сильнее связь между значениями составляющих двумерной выборки.

Пример. Найдем индекс корреляции между объемом производства вишни и ценой вишни (пункт 3.5.5) при описании зависимости многочленом второго порядка. Расчетные данные:

$$n = 16; \quad \sum_j y_i = 3722; \quad \bar{y} = 232{,}625; \quad \left(\sum_j y_i\right)^2 = 13853284;$$

$$\sum_j y_i^2 = 918446; \quad n\sum_j \left(y(x_i) - \bar{y}\right)^2 = 471442{,}88;$$

R_{yx}=0,748, что несколько больше, чем модуль выборочного коэффициента корреляции r (r = -0,738). Мы получили подтверждение, что параболическое уравнение лучше соответствует опытным данным, чем линейное.

Индекс корреляции не позволяет определить, положительной или отрицательной является корреляция между величинами y и x (растут или убывают значения y с ростом x). Это можно сделать по виду диаграммы рассеяния и графика соответствующего уравнения регрессии.

В заключение отметим, что, построив уравнение регрессии x на y ($x = g(y)$), можно рассчитать индекс корреляции $R_{xy} \neq R_{yx}$, т.е. оценить, как x зависит от y.

3.8. ИНДЕКС ФЕХНЕРА И КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ

Здесь будут описаны два способа оценки степени связи между составляющими двумерной выборки без использования уравнения регрессии. Прежде всего, постараемся уточнить, что подразумевается под

термином «связь». Ведь если нет уравнения $y = f(x)$, связывающего аргумент x и зависимую переменную y, понятие «связь» становится расплывчатым. Будем говорить, что между составляющими двумерной выборки существует положительная корреляция (связь), если с ростом значений x значения y проявляют тенденцию к возрастанию. Соответственно говорят об отрицательной корреляции между x и y, если с ростом значений x значения y проявляют тенденцию к убыванию. Конечно, и формулировку «проявлять тенденцию к» нельзя назвать строгой. Но на интуитивном уровне она представляется понятной.

Г.Фехнер (1801 - 1887), немецкий психолог, предложил очень простой способ оценки степени такого рода связи. Для определения индекса Фехнера вычисляют средние $\overline{x}$, $\overline{y}$, а затем для каждой пары (x_i, y_i) определяют знаки отклонений $x_i - \overline{x}$, $y_i - \overline{y}$. Для каждой пары (x_i, y_i) возможны четыре сочетания знаков: + +; + —; — +; — —. Обозначим через V количество совпадений, через W – количество несовпадений знаков. Половину случаев $x_i = \overline{x}$ или $y_i = \overline{y}$ относят к V, половину – к W. Индекс Фехнера i определяется формулой $i = (V-W)/(V+W)$.

Ясно, что $-1 \le i \le 1$ и что при $i > 0$ имеем положительную корреляцию, при $i < 0$ – отрицательную, при $i = 0$ связь в указанном нами смысле отсутствует. Найдем индексы Фехнера для примеров из §3.1.

Пример с текстом. Пары знаков получаются такими:

(- -), (- -), (- -), (+ +), (- -), (- -) (+ +) (- -) (++), (++).

Отсюда $V = 10$, $W = 0$, $i_1 = 1$.

Пример с монетами. Пары знаков следующие:

(- +), (- +), (- +), (- +), (- +), (+ -), (+ -) (+ +), (+ -), (+ -).

Значит $V = 1$, $W = 9$, $i_2 = -0{,}8$.

Пример с кубиками. Последовательность пар знаков:

(0 +), (+ -), (+ -), (- -), (- +), (+ -), (- -). (+ +), (+ -), (+ +).

Если просто не учитывать первую пару ($x_1 = \overline{x} = 4$), то $V = 4$, $W = 5$, $i_3 = -0{,}11$. Если поделить единицу пополам, то $V = 4{,}5$; $W = 5{,}5$, $i_3 = -0{,}1$.

Корреляционное отношение как мера тесноты связи между составляющими двумерной выборки было предложено К. Пирсоном. Оно вычисляется по корреляционной таблице, а расчетная формула аналогична формуле для индекса корреляции. В дополнение к обозначениям §3.6 введем еще одно. Через $\overline{y}_i$ обозначим т.н. частное среднее значений y для i-го значения x:

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{m} y_j n_{ij} \quad , i = 1,2, \dots , k.$$

По аналогии с индексом корреляции, корреляционное отношение η_{yx} вводится так:

$$\eta_{yx} = \sqrt{\frac{\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 n_i}{\sum_{j=1}^{m}(y_j - \bar{y})^2 l_j}} = \sqrt{\frac{n\sum_{i=1}^{n}\bar{y}_i{}^2 n_i - \left(\sum_{j=1}^{m} y_j l_j\right)^2}{n\sum_{j=1}^{m} y_j{}^2 l_j - \left(\sum_{j=1}^{m} y_j l_j\right)^2}}.$$

Напомним, что

k – число интервалов группировки по составляющей x двумерной выборки;

x_i – середина i-го интервала группировки по составляющей x;

n_i – частота i-го интервала группировки по составляющей x, $i = 1,2,..,k$;

y_j – середина j-го интервала группировки по составляющей y;

m - число интервалов группировки по составляющей y;

l_j – частота j-го интервала группировки по составляющей y, $j = 1,2,...,m$;

n_{ij} – частоты прямоугольников группировки;

n – объем двумерной выборки.

Если все точки на диаграмме рассеяния сгруппированной выборки лежат на горизонтальной прямой, то все частные средние $\bar{y}_i$ равны $\bar{y}$.

$$\bar{y}_i = \bar{y}, \quad i = 1,2, \ldots, k \quad \Rightarrow \quad \eta_{yx} = 0.$$

Тогда говорят об отсутствии связи между значениями x и y. Если все точки на диаграмме рассеяния сгруппированной выборки лежат на некоторой прямой (кроме горизонтальной), то $\eta_{yx} = 1$. В остальных случаях $0 < \eta_{yx} < 1$.

Величина η_{yx} зависит от группировки. Как правило, с ростом числа интервалов группировки по переменной x корреляционное отношение растет. По аналогии с числом η_{yx} можно рассчитать число $\eta_{xy} \neq \eta_{yx}$, если считать x зависимой переменной, а y – независимой переменной.

Пример. На металлообрабатывающем заводе у 60 марок стали провели замеры предела текучести $F(x$, кг/мм$^2)$ и предела прочности $\sigma_в$ (y, кг/мм^2). В итоге получили 60 пар значений, представленных в табл. 3.6. Предполагается, что большие значения F обуславливают большие значения $\sigma_в$; марки стали с низким пределом текучести имеют и низкий предел прочности. Для обоснования гипотезы о высокой положительной корреляции между пределом прочности и пределом текучести сгруппируем выборку (табл. 3.7) и рассчитаем числовые характеристики.

Таблица 3.6

F	$\sigma_в$	F	$\sigma_в$	F	$\sigma_в$	F	$\sigma_в$
x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
154	178	51	95	98	140	44	69
133	164	101	114	97	115	92	116
58	75	169	209	105	101	141	157
145	161	87	101	71	93	155	193
94	107	88	139	39	69	136	155
113	141	83	98	122	147	82	81
86	97	106	III	33	52	136	163
121	127	92	104	78	117	72	79
119	138	85	103	114	138	66	81
112	125	112	118	125	149	42	61
85	97	98	102	73	76	113	123
41	72	103	108	77	85	42	85
96	113	99	119	47	61	133	147
45	88	104	128	68	85	153	179
99	109	107	118	137	142	85	91

Внешний вид табл. 3.7 несколько отличается от вида табл. 3.5, иллюстрирующей двумерную группировку. Табл. 3.7 построена так, чтобы можно было легко вообразить диаграмму рассеяния, не строя ее саму.

Имеем:

$n = 60$; $k = 7$; $m = 8$; $h_x = h_y = 20$ (длины интервалов группировки).

$$\overline{y} = \frac{1}{60}(200*2 + 180*2 + 160*5 + 140*9 + 120*13 + 100*14 + 80*10 + 60*5) = 114{,}7;$$

Таблица 3.7

Предел прочности, кг/мм^2	y_j	Предел текучести $[x_{i-1}, x_i]$, кг/мм^2							m_j
		[30 – 50)	[50 – 70)	[70 – 90)	[90 – 110)	[110 – 130)	[120 – 150)	[150 – 170)	
		x_i							
		40	60	80	100	120	140	160	
[190 – 210)	200							2	2
[170 – 190)	180							2	2
[150 – 170)	160						5		5
[130 – 150)	140			1	1	5	2		9
[110 – 130)	120			1	8	4			13
[90 – 110)	100		1	7	6				14
[70 – 90)	80	3	3	4					10
[50 – 70)	60	5							5
n_i		8	4	13	15	9	7	4	60

$\overline{y}_1 = \frac{1}{8}(60*5 + 80*3) = 67{,}5;\quad \overline{y}_2 = \frac{1}{4}(80*4 + 100) = 105;$

$\overline{y}_3 = \frac{1}{13}(160 + 120 + 100*7 + 80*4) = 100;\quad \overline{y}_4 = 113{,}3;$

$\overline{y}_5 = 131{,}1;\qquad \overline{y}_6 = 154{,}3;\qquad \overline{y}_7 = 190;$

$$\sum_{i=1}^{7}\left(\overline{y}_i - \overline{y}\right)^2 n_i = 8*(67{,}5 - 114{,}7)^2 + 4*(105 - 114{,}7)^2 + 13*(100 - 114{,}7)^2 +$$
$$+15*(113{,}3-114{,}7)^2 + 9*(131{,}1-114{,}7)^2 +7*(154{,}3-114{,}7)^2 +4*(190 - 114{,}7)^2 =$$
$$= 57115{,}8;$$

$$\sum_{i=1}^{8}\left(y_j - \overline{y}\right)^2 l_j = 2*(200 - 114{,}7)^2 + 2*(180 - 114{,}7)^2 + 5*(160 - 114{,}7)^2 +$$
$$+9*(140 - 114{,}7)^2 + 13*(120 - 114{,}7)^2 + 14*(100 - 114{,}7)^2 +10*(80 - 114{,}7)^2 +$$
$$+5*(60 - 114{,}7)^2 = 69493{,}4;$$

$\eta_{yx} = 0{,}82.$

Для справки: коэффициент корреляции $r = 0{,}92$, предел прочности и предел текучести связаны сильной линейной зависимостью.

3.9.ЗАДАЧИ

1**.** Как выражаются коэффициенты линейного уравнения регрессии через выборочный коэффициент корреляции r ?

2**.** Показать, что выборочный коэффициент корреляции r не изменится, если значения x_i, y_i подвергнуть преобразованию: $x_i = x_i + a$; $y_i = y_i + b$; $i = 1,2, \ldots, n$. Как изменится выборочный коэффициент корреляции r, если все числа x_i умножить на одно и то же число d, все числа y умножить на одно и то же число b, $i = 1,2, \ldots, n$?

3**.** В соответствии с методом наименьших квадратов составить систему уравнений для определения коэффициентов следующих уравнений регрессии:

$y = a + be^x$, $\qquad y = a + b\cdot\sin\omega x + c\cdot\cos\omega x$

(ω – заданное число), $y = a + \frac{b}{x}$.

В задачах 4 - 19 нужно найти числовые характеристики выборки и определить (если $r \geq 0{,}7$) коэффициенты линейного уравнения регрессии x на y, если y можно принять за независимую переменную.

4**.** Результаты тестирования (баллы) 10 студентов. Первый тест проверяет память (x), второй - способность к логическому мышлению (y):

x_i	5	8	7	10	4	7	9	6	8	6
y_i	7	9	6	9	6	7	10	7	6	8

5. Оценка за тест по способностям (x) шести продавцов–практикантов и результаты их работы за первый год (y) в сотнях фунтов проданного товара:

x_i	25	42	33	54	29	36
y_i	42	73	50	90	45	48

6. Снашивание (x) и твердость (y) резины в условных единицах:

x_i	21	15	12	22	5
y_i	5	6	7	4	8

7. Масса поросят (y) в килограммах в зависимости от возраста (x) в неделях:

x_i	1	2	3	4	5	6	7	8
y_i	2,5	3,9	5,2	6,3	7,5	9,0	10,8	13,1

8. В книге «Основы химии» Д.И.Менделеева приводятся данные о растворимости азотнокислого натрия $NaNO_3$ в зависимости от температуры воды. Указывается, сколько условных частей $NaNO_3$ (y) растворяется в 100 частях воды при соответствующих температурах в ^oC (x):

x_i	0	4	10	15	21 .	29	36	51	68
y_i	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

9. Средняя температура января в г. Саратове (x) и в г. Алатыре (Чувашия) (y) измерялась в течение 13 лет:

Год	1891	1892	1893	1894	1895	1896	1897
x_i	-19,2	-14,8	-19,6	-11,1	-9,4	-16,9	-13,7
y_i	-21,8	-15,4	-20,8	-11,3	-11,6	-19,2	-13,0
Год	1899	1911	1912	1913	1914	1915	—
x_i	-4,9	-13,9	-9,4	-8,3	-7,9	-5,3	—
y_i	-7,4	-15,1	-14,4	-4,1	-10,5	-7,2	—

10. Средняя температура июня в г. Москве (x) и в г. Ярославле (y) измерялась в течение 40 лет:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
12,0	10,8	13,9	10,1	15,0	13,8	17,2	13,9	18,1	16,0
12,0	11,3	11,2	10,0	15,0	16,0	16,9	14,8	18,4	17,8
12,0	12,0	14,0	10,0	15,5	13,9	16,9	15,0	19,2	15,0
12,0	13,0	14,0	12,0	15,9	14,7	17,0	16,0	19,3	16,1
12,8	10,9	13,0	12,4	16,0	13,0	16,8	17,0	20,0	17,0
13,8	10,0	15,0	11,0	15,9	15,0	17,5	16,0	20,1	17,7
13,1	11,5	14,9	13,0	16,0	16,0	18,0	14,0	14,0	14,8
13,0	13,0	15,9	14,2	16,9	12,9	18,0	14,0	14,0	15,2

11. Объем продажи (x) в миллиардах долларов и чистый доход (y) в миллионах долларов 20 фирм в США:

x_i	8,9	8,4	7,4	7,2	7,0	6,1	5,9	5,8	5,5	4,8
y_i	441	278	456	934	89	611	770	53	243	217
x_i	4,4	4,2	4,2	4,1	3,8	3,8	3,6	3,5	3,3	3,2
y_i	454	291	321	51	111	2	356	150	237	151

Определяется ли доход объемом продажи?

12. Среднегодовые уровни воды в озере Виктория - Ньянза (x) относительно некоторого фиксированного значения и числа солнечных пятен (y) за 1902 - 1921 гг.:

Год	x_i	y_i	Год	x_i	y_i
1902	-10	5	1912	-11	4
1903	13	24	1913	-3	1
1904	18	42	1914	-2	10
1905	15	63	1915	4	47
1906	29	*54*	1916	15	57
1907	21	62	1917	35	104
1908	10	49	1918	27	81
1909	8	44	1919	8	64
1910	1	19	1920	3	38
1911	-7	6	1921	-5	25

13. Число айсбергов, наблюдавшихся помесячно к югу от Ньюфаундленда (x) и к югу от Большой отмели (y) за 1920 г.:

x_i	3	10	36	83	130	68	25	13	9	4	3	2
y_i	0	1	4	9	18	13	3	2	1	0	0	0

14. Число разводов на 1000 жителей в 20 штатах США (y), средний доход на семью (x_1) в тыс. долл.; процент городского населения (x_2):

y_i	x_{1i}	x_{2i}	y_i	x_{1i}	x_{2i}
1,2	4,9	38,5	3,6	4,9	75
1,1	6,3	83,6	3,9	5,2	47,5
0,4	6,4	85,4	4,0	5,9	56,8
2,4	6,2	73,4	2,7	5,8	73,7
2,7	5,8	62,4	3,0	5,4	65,7
2,1	6,2	73,4	2,4	5,9	74,9
1,2	4,2	39,3	1,2	4,9	51,3
1,5	4,9	54,3	3,3	6,2	68,1
1,9	5,0	55,8	3,2	5,9	62,2
1,6	4,6	62,9	3,1	6,7	86,4

15**.** На сталелитейном заводе обследовали 15 плавок определенного сорта стали. Учитывался угар кремния (x), измеряемый в процентах, и выход стали (y), также измеряемый в процентах.

x_i	7,9	0,9	3,7	8,1	6,9	0,8	6,0	7,2	8,8	10,2	11,2	0,5
y_i	70,3	85,0	100,0	78,1	77,9	98,4	59,2	86,8	70,1	42,2	81,9	97,1
x_i	4,6	9,7	1,0									
y_i	68,2	92,1	91,2									

16**.** Продолжительность послеоперационного лечения в клинике (y) в днях и возраст больных (x) в годах, оперировавшихся по поводу грыжи:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
78	9	68	7	79	3	75	7
60	4	79	11	51	5	02	0
68	7	80	4	57	8	65	16
62	35	48	9	51	8	42	3
76	9	35	2	48	3	54	2
76	7	58	4	48	5	43	3
64	5	40	3	66	8	04	3
64	19	19	4	71	2	52	8

17**.** При исследовании некоторой химической реакции через каждые 5 минут определялось количество вещества (y) в %, оставшееся в системе. Подобрать коэффициенты уравнения $y = a + bx + cx^2$, где x – время после начала реакции в минутах.

x_i	0	7	12	17	22	27	32	37
y_i	100	87,3	72,9	63,2	54,7	47,5	41,4	36,3

18**.** Барометрическое давление связано с высотой следующим соотношением: $p / p_0 = e^{\frac{-k}{T}z}$,

где p - барометрическое давление на высоте z;

T - температура;

p_0 и k - параметры.

По методу наименьших квадратов оценить значения параметров k/T и p_0 по результатам наблюдений, проведенных при постоянной температуре:

Z_i,м	1000	1100	1200	1400	1500	1600
p_i, мм рт. ст.	640	595	504	363	310	267

19. Для исследования зависимости давления p насыщенного пара (Н/см^2) от удельного объема V (м^3/кг) составлена таблица опытных данных:

V_i	3,334	1,630	0,866	0,423	0,265	0,170	0,115
p_i	0,482	1,034	2,027	4,247	7,164	11,480	17,600

Подобрать коэффициенты функциональной зависимости $p = aV^b$.

20. Функциональная зависимость удельного сопротивления кристаллического кварца ρ (Ом·см) от абсолютной температуры T (K) имеет вид $\rho=10^{a\frac{1}{T}+b}$.

Используя опытные данные, оценить значения параметров a и b.

ρ_i	$5*10^{16}$	$4*10^{15}$	$3*10^{14}$	$2*10^{13}$	$2*10^{12}$	$1{,}5*10^{11}$	10^{10}
T_i	335	365	400	445	500	570	670

21. Получена выборка наблюдений переменных x и y:

x_i	1	2	3	5	6	7	8
y_i	62,1	87,2	109,3	127,3	134,3	136,2	136,9

Для представления этих данных предлагается выбрать лучшую из предложенных моделей:

$$1)\ y=\frac{x}{a+bx}.\ \ 2)\ y=ba^x.\ \ 3)\ y=bx^a.\ \ 4)\ y=a\ln(x)+b.$$

Оценить значения параметров a и b.

22. На заводе производят некоторый материал, твердость которого хотят повысить. Для этого увеличивают содержание некоторого химического вещества. Ниже приведены данные для 20 случайно отобранных образцов. Значения y – твердость образца (условные единицы), значения x – процентное содержание химического вещества относительно некоторого уровня.

x_i	18	18	18	6	20	9	11	22	17	17
y_i	72,2	80,1	69,8	58,2	79,7	45,6	58,6	85,4	80,1	66,7
x_i	19	14	22	8	22	11	24	14	24	5
y_i	79,1	56,4	82,4	55,2	107,8	34,4	115,4	73,5	99,5	56,8

Подобать коэффициенты линейного и параболического уравнений регрессии. Какое из уравнений больше соответствует экспериментальным данным?

4. ВРЕМЕННЫЕ РЯДЫ

Порвалась дней связующая нить,
Как мне обрывки их соединить!
В. Шекспир. Гамлет
(пер. Б.Пастернака)

4.1. ЧТО ТАКОЕ ВРЕМЕННОЙ РЯД

Временной ряд - это последовательность данных, отнесённых к определённым промежуткам (или моментам) времени: годам, месяцам, дням и т.п. Промежутки времени обычно имеют одинаковую ширину, их нумеруют натуральными числами, начиная с нуля. Соответственно элементы временного ряда обозначают $y_0, y_1, y_2, ...,y_t,...$ Временные ряды встречаются в самых разных областях человеческой деятельности.

В табл. 4.1 - 4.6 и на рис. 4.1 - 4.6 приведены различные временные ряды и их графические представления.

Таблица 4.1

Население США (млн. чел.) в 1965 - 1984 гг.

Год	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
y_t	194	197	199	201	203	205	208	210	212	214
Год	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
y_t	216	218	220	223	225	228	230	232	235	237

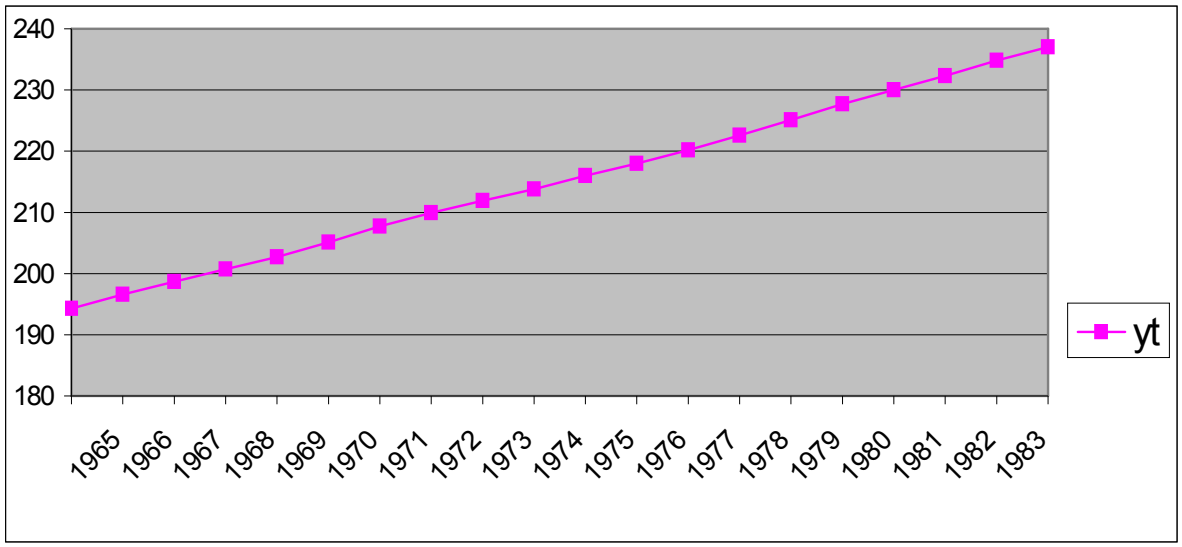
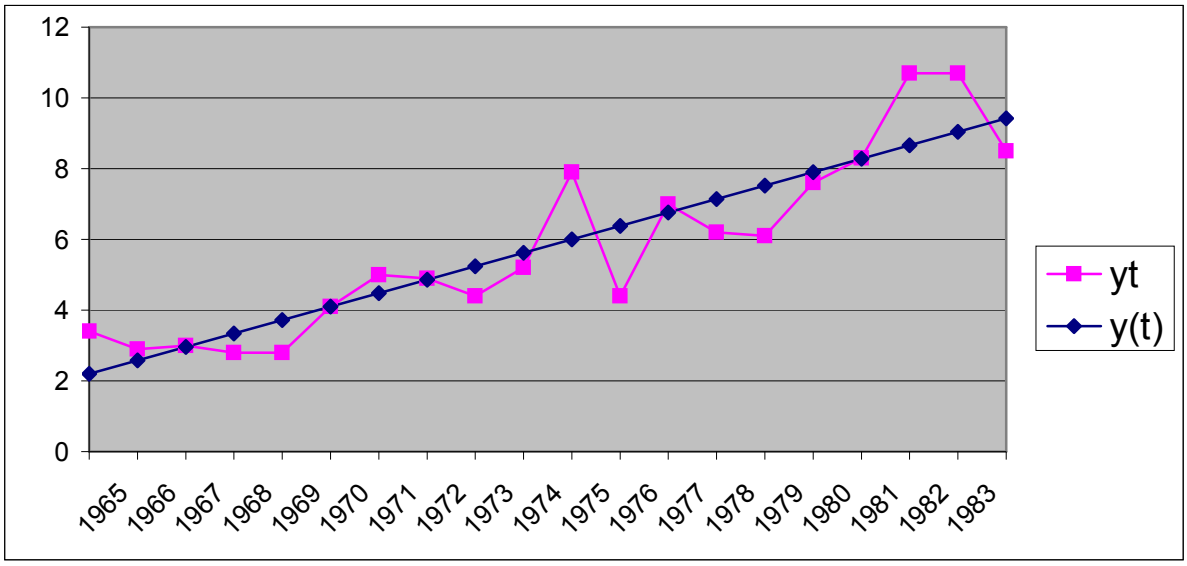


Рис.4.1. Население США в 1965 - 1984 гг.

Таблица 4.2

Безработица в США (млн. чел.) в 1965 - 1984 гг.

Год	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
y_t	3,4	2,9	3	2,8	2,8	4,1	5	4,9	4,4	5,2
Год	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
y_t	7,9	4,4	7	6,2	6,1	7,6	8,3	10,7	10,7	8,5



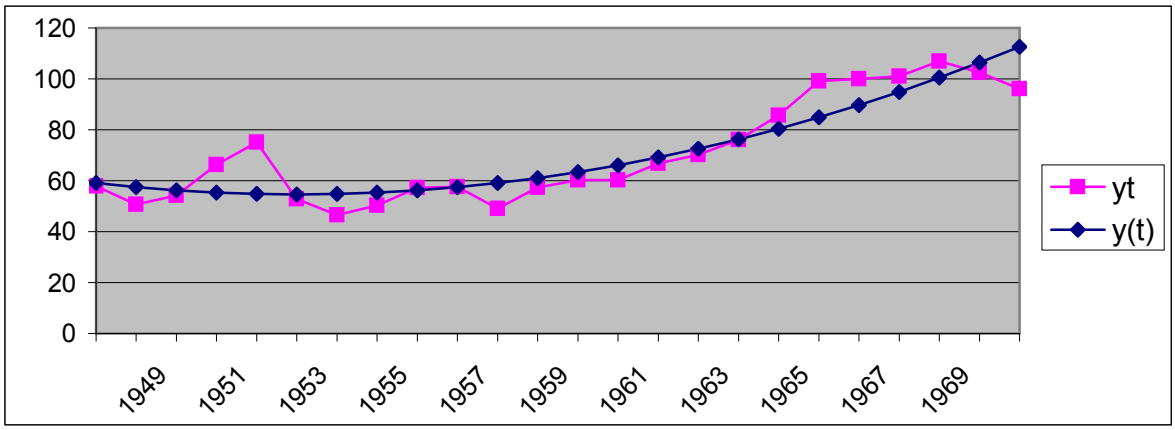
$$y(t) = 0,38t + 2,2.$$

Рис. 4.2. Безработица в США в 1965 - 1984 гг.

Таблица 4.3

Индекс выпуска оборудования для частнопредпринимательского сектора — компонента индекса промышленного производства в США в 1948 - 1971 гг. (1967 г. = 100)

Год	1948	1949	1950	1951	1952	1953
y_t	57,9	50,7	54,2	66,3	75,1	52,8
Год	1954	1955	1956	1957	1958	1959
y_t	46,5	50,3	57,2	57,6	49,1	57,4
Год	1960	1961	1962	1963	1964	1965
y_t	60,3	60,2	66,8	70,2	76,1	85,7
Год	1966	1967	1968	1969	1970	1971
y_t	99,1	100	101	107	103	96,1



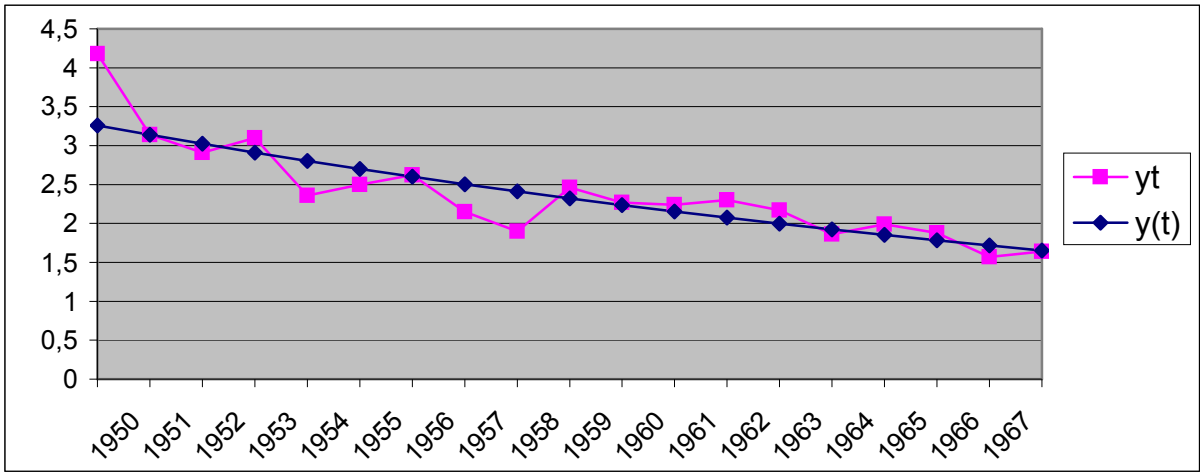
$$y(t) = 0{,}179t^2 - 1{,}79t + 59{,}1.$$

Рис. 4.3. Индекс выпуска оборудования для частнопредпринимательского сектора в США в 1948 - 1971 гг.

Таблица 4.4

Душевое потребление шерсти в США в 1950 -1968 гг. (потребление делится на численность населения континентальной территории США), фунт

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
y_t	4,18	3,14	2,91	3,1	2,36	2,5	2,62	2,15	1,9	2,46
Год	1960	1961	1962	1963	1964	1965	1966	1967	1968	
y_t	2,27	2,24	2,3	2,17	1,86	1,99	1,88	1,57	1,64	



$$y(t) = 3{,}26*0{,}963^t.$$

Рис. 4.4. Душевое потребление шерсти в США в 1950-1968 гг.

Таблица 4.5

Затраты на новые здания, сооружения, оборудование в США (млрд. долл.) в 1966 - 1971 гг.

Год	1966				1967				1968			
Квартал	1	2	3	4	1	2	3	4	1	2	3	4
y_t	13,3	16,1	15,9	18,2	14,5	16,7	16,2	18,1	15,1	16,9	16,8	19
Год	1969				1970				1971			
Квартал	1	2	3	4	1	2	3	4	1	2	3	4
y_t	16	18,8	19,3	21,5	17,5	20,3	20,3	21,7	17,7	20,6	20,1	23

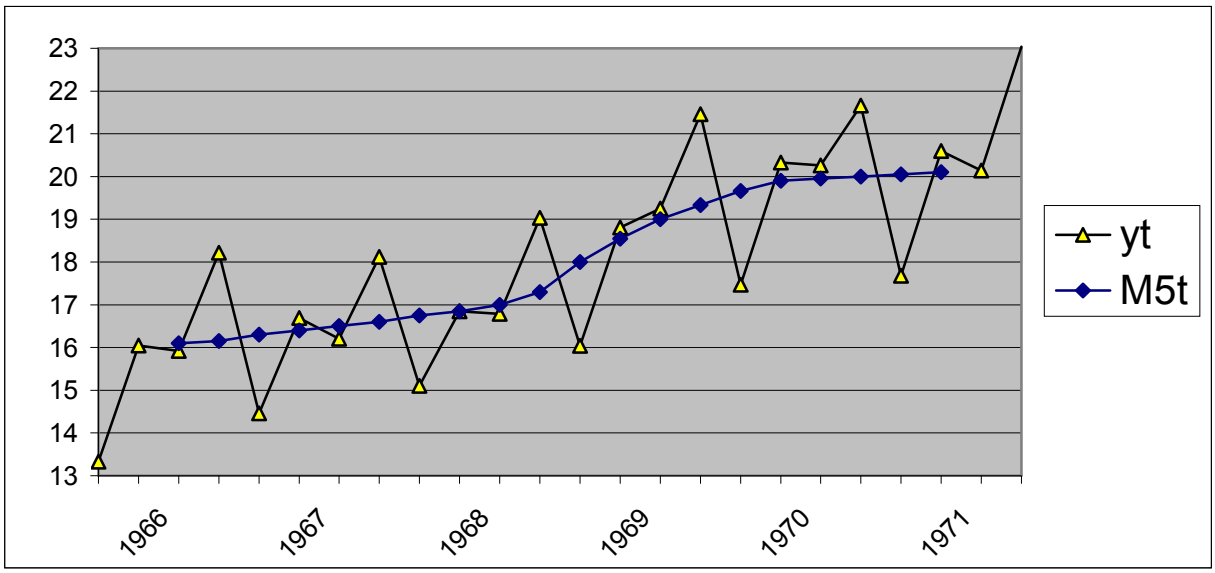


Рис. 4.5. Затраты на новые здания, сооружения, оборудование в США в 1966 - 1971 гг.

Таблица 4.6

Урожайность ячменя в Англии и Уэльсе в 1900 - 1924 гг. (в центнерах на акр, центнер в Англии = 50,8 кг)

Год	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912
y_t	14,9	14,5	16,6	15,1	14,6	16	16,8	16,8	15,5	17,3	15,5	15,5	14,2
Год	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	
y_t	15,8	15,7	14,1	14,8	14,4	15,6	13,9	14,7	14,3	14	14,5	15,4	

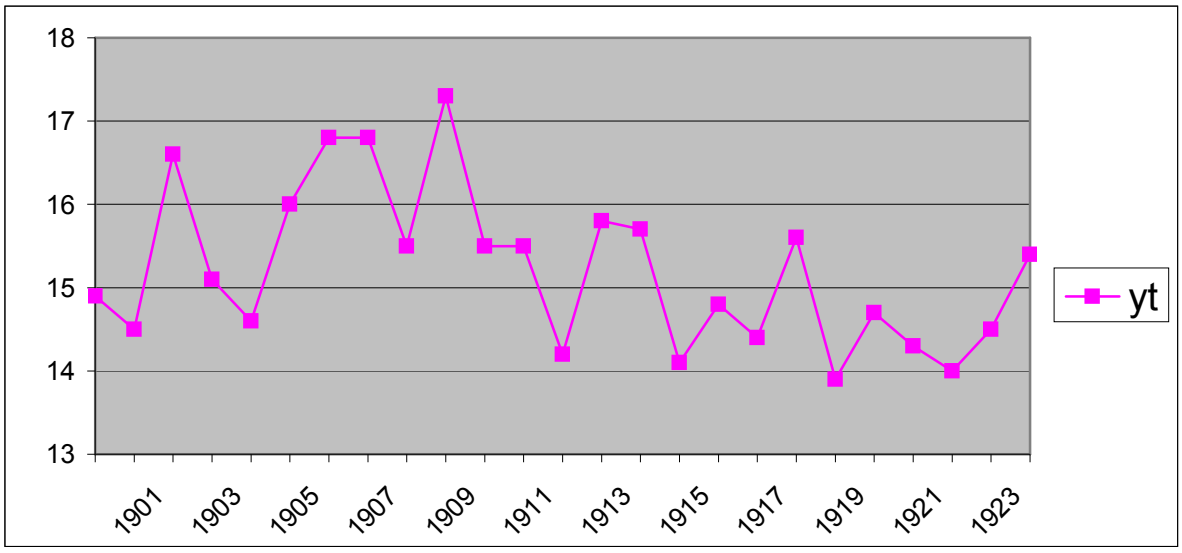


Рис. 4.6. Урожайность ячменя в Англии в 1900 - 1924 гг.

Даже немногих приведённых примеров достаточно, чтобы убедиться, какими разными могут быть временные ряды. График населения почти неотличим от прямой. В графике безработицы просматривается движение вдоль некоторой прямой. Но одновременно с этим движением присутствуют и сильные отклонения, которым желательно дать объяснение. График расходов на новые здания и сооружения содержит хорошо просматривающиеся годовые циклы. График урожайности ячменя состоит, кажется, из одних только колебаний, притом достаточно случайных.

Все пять приведённых временных рядов содержат значения, получившиеся накапливанием данных за определённый промежуток времени (год). Достаточно распространены данные и другого типа - значения некоторого процесса, определённые в фиксированные моменты времени.

4.2. ПОНЯТИЕ ОБ АНАЛИЗЕ ВРЕМЕННЫХ РЯДОВ

Анализ временных рядов проводится с целью понять и описать механизм, порождающий значение ряда, а затем попытаться предсказать поведение ряда хотя бы в ближайшем будущем.

Методы анализа временных рядов разрабатывались (и продолжают разрабатываться) высокопрофессиональными математиками. Чтобы познакомиться с ними подробно, нужно иметь уровень математической подготовки, намного превосходящий тот, на который рассчитано данное учебное пособие. Мы отметим только самые простые методы изучения поведения временного ряда.

4.2.1. О значениях временного ряда

В общем случае значения временного ряда представляют в виде суммы четырёх компонент:

а) тренда или долгосрочного движения (английское слово *"trend"* переводится как "тенденция");

б) сезонной компоненты;

в) колебаний относительно тренда, которые можно хотя бы приближённо считать регулярными;

г) остатка или случайной компоненты.

Первые три слагаемых относятся к так называемым детерминированным (определённым, закономерным) составляющим временного ряда. Их поведение предсказуемо, их значения могут быть вычислены при каждом t как некоторая функция момента времени t.

Случайная компонента не является определённой функцией времени. Её поведение во времени описывается при помощи подходящих вероятностных моделей.

4.2.2. Тренды временных рядов

Под трендом временного ряда будем понимать гладкую функцию, описывающую долгосрочное поведение временного ряда.

Рассмотрим три простейших уравнения тренда.

1. Линейная функция $y(t) = at + b$.
2. Полином второго порядка (парабола) $y(t) = at^2 + bt + c$.
3. Показательная функция $y(t) = ba^t$.

Подходящее уравнение тренда часто можно подобрать уже по графику временного ряда. Так, график роста населения США почти не отличается от прямой. Наличие линейного тренда просматривается в ряде безработицы в США, хотя отклонения от тренда значений этого ряда гораздо сильнее, чем на рис. 4.1. Глядя на рис. 4.3, можно предположить, что в ряде индекса выпуска оборудования для частнопредпринимательского сектора США присутствует параболический тренд.

А рис. 4.4 наводит на мысль о наличии показательного тренда в ряде душевого потребления хлопка в США.

Тренд используют при прогнозировании поведения ряда в будущем. Чтобы прогноз оказался достаточно точным, нужно уметь описывать и отклонения значений ряда от тренда. Вследствие того, что рассмотрение моделей случайной компоненты временных рядов требует использования достаточно серьёзной математики, мы вынуждены отказаться от изложения этих моделей.

Для получения уравнения тренда чаще всего используют метод наименьших квадратов, который был изложен в главе 3. Время t рассматривается как независимая переменная, а значения временного ряда полагают функцией времени. При этом вычисления несколько упрощаются из-за того, что значения аргумента t – натуральные числа. Если в ряде n значений, то t = 0,1,2,…, n-1. В дальнейшем будут использованы хорошо известные формулы сумм целых степеней последовательных натуральных чисел:

$$\sum_1^{n-1} t = \frac{n(n-1)}{2}; \qquad \sum_1^{n-1} t^2 = \frac{n(n-1)(2n-1)}{6};$$

$$\sum_1^{n-1} t^3 = \frac{n^2(n-1)^2}{4}; \qquad \sum_1^{n-1} t^4 = \frac{n-1}{30} n(2n-1)(3n^2-3n-1).$$

4.2.2.1 Линейный тренд

Пусть $y(t)=at+b$, тогда

$$a = S_{ty} / S_t^2; \quad b = \bar{y} - a\bar{t},$$

при этом

$$\bar{t} = \frac{n(n-1)}{2n} = \frac{n-1}{2}; \quad S_t^2 = \frac{1}{n}\sum_1^{n-1} t^2 - (\frac{n-1}{2})^2 = \frac{n^2-1}{12}.$$

Пример. Рассчитаем коэффициенты a и b линейного уравнения тренда для ряда роста населения США (табл. 4.1).

Здесь n=20; $\bar{t} = 9{,}5$; $S_t^2 = 33{,}25$; $\sum_1^{n-1} t y_t = 42386{,}9$; $S_{ty} = 74{,}42$; $\bar{y} = 215{,}255$;

a=2,24; $b = 193{,}9 \approx 194$.

Уравнение тренда получилось таким: y = 2,24t + 194.

Эта прямая настолько хорошо сливается с графиком временного ряда, что на рис. 4.1 она не показана. Если рассчитать трендовые значения в 1964 (t= -1) и в 1985 (t= 20) годах, получим: y(-1)=191,8; y(20)=238,8. Реальные значения таковы: y_{1964}=191,9; y_{1985}=239,3. Совпадение просто поразительное! Когда ряд так хорошо себя ведёт, коэффициенты a и b очень точно можно оценить просто “на глаз”, выбрав пару подходящих точек на графике. Возьмём t =0 и t =10 (y_0=194,3, y_{10}=216). Получаем следующую систему для определения чисел a и b:

$$\begin{cases} b = 194{,}3; \\ 10a + b = 216. \end{cases}$$

Отсюда a=2,17.

Пример. Построим уравнение линейного тренда для безработицы в США. Этот ряд сильно колеблется, но линейный тренд всё-таки достаточно хорошо виден. Сначала оценим коэффициенты *a* и *b* приближённо. Возьмём t=2 (y_2=3,0) и t=15 (y_2=7,6). Имеем систему

$$\begin{cases} 2a+b=3; \\ 15a+b=7{,}6. \end{cases}$$

Отсюда a=0,35; b=2,3.

Проведём расчёты по методу наименьших квадратов. В данном случае n=20; $\bar{t}=9{,}5$; $S_t^2=33{,}25$; $\bar{y}=5{,}795$; S_{ty}=12,72; a=0,38; b=2,2.

Уравнение тренда: y=0,38 t+2,2.

В 1964 году безработица в США равнялась 3,8 млн. чел., а в 1985г. – 8,3 млн. чел. Эти значения сильно отличаются от трендовых, особенно первое из них, т.к. $y(-1)$=1,78; $y(20)$=9,8.

Прямая y=0,38t+2,2 построена на рис.4.2.

4.2.2.2. Параболический тренд

Рассмотрим уравнение $y(t)=at^2+bt+c$.

Составим систему уравнений относительно неизвестных параметров *a, b, c,* получающуюся по методу наименьших квадратов:

$$\begin{cases} a\sum\limits_1^{n-1} t^4 + b\sum\limits_1^{n-1} t^3 + c\sum\limits_1^{n-1} t^2 = \sum\limits_1^{n-1} t^2 y_t^2; \\ a\sum\limits_1^{n-1} t^3 + b\sum\limits_1^{n-1} t^2 + c\sum\limits_1^{n-1} t = \sum\limits_1^{n-1} t y_t; \\ a\sum\limits_1^{n-1} t^2 + b\sum\limits_1^{n-1} t + nc = \sum\limits_1^{n-1} y_t. \end{cases}$$

Эту систему проще всего решать методом исключения неизвестных.

Пример. Рассчитаем коэффициенты *a, b, c* для параболического тренда ряда из табл.4.3 (индекс выпуска оборудования для частнопредпринимательского сектора США в 1948 – 1971 гг.).

Здесь $n=24$; $\sum\limits_0^{23} t=276$; $\sum\limits_0^{23} t^2=4324$; $\sum\limits_0^{23} t^3=76176$; $\sum\limits_0^{23} t^4=1431244$;

$\sum\limits_0^{23} y_t=1700$; $\sum\limits_0^{23} t y_t=22234{,}8$; $\sum\limits_0^{23} t^2 y_t=375898{,}2$.

Система линейных уравнений

$$\begin{cases} 1431244a + 76176b + 4324c = 375898{,}2; \\ 76176a + 4324b + 276c = 22234{,}8; \\ 4324a + 276b + 24c = 1700. \end{cases}$$

имеет следующее решение: a=0,179; b=-1,79; c=59,1.

График функции $y(t)$=0,179t^2-1,79t+59,1 показан на рис. 4.3.

Попробуем оценить значения параметров *a, b, c* приближённо. Положим $c = y_0$ = 57,9. Визуально координата t вершины параболы равна 6.

Тогда $b/(2a)$=-6, b=-12a. Осталось выбрать подходящую точку на графике временного ряда, которая не слишком сильно, по нашему мнению, отклоняется от тренда. Возьмём t=5, y_5=52,8.

52,8=25a-60a+57,9 => a=0,146; b=-1,75.

Приближённые оценки неплохо совпали с рассчитанными. Значения параметра a вычислялись с тремя знаками после запятой, значения параметра b - с двумя, а значения параметра c - только с одним из-за того, что a умножается на t^2, b умножается на t, а c – просто свободный член.

4.2.2.3. Показательная функция

Отношение $((y_{t+1} - y_t)/y_t)$*100% для некоторого момента времени t называют ***темпом прироста временного ряда***. Если значения y_t временного ряда есть значения показательной функции $y_1 = ba^t$, то их темпы прироста постоянны и равны:

$$\frac{ba^{t+1} - ba^t}{ba^t} \times 100\% = (a-1) \times 100\%.$$

Поэтому, когда значения временного ряда имеют приблизительно одинаковые темпы прироста, для ряда подбирают показательный тренд. В качестве примера рассмотрим ряд из табл. 4.4. (душевое потребление шерсти в США в 1950 – 1968 гг.).

Рассчитаем несколько темпов прироста. Пусть t = 1, 9, 12.

$$\frac{y_2 - y_1}{y_1} \cdot 100\% = \frac{2{,}91 - 3{,}14}{3{,}14} \cdot 100\% = -7{,}3\%;$$

$$\frac{y_{10} - y_9}{y_9} \cdot 100\% = \frac{2{,}27 - 2{,}46}{2{,}46} \cdot 100\% = -7{,}7\%;$$

$$\frac{y_{13} - y_{12}}{y_{12}} \cdot 100\% = \frac{2{,}17 - 2{,}30}{2{,}30} \cdot 100\% = -5{,}7\%.$$

Эти примеры не назовёшь идеально совпадающими, но по рис. 4.4 видно, что поведение ряда не соответствует линейному тренду, поэтому будем искать параметры показательной функции.

Уравнение $y(t)=ba^t$ приводится к линейному путём почленного логарифмирования.

Параметры линейного уравнения $y_1(t)=a_1t+b_1$,
где $y_1(t)= ln(y(t)); a_1= ln(a); b_1= ln(b)$, найдём по методу наименьших квадратов. Здесь n=19; $\bar{t}=9$; $S_t^2=30$; $\bar{y}_1=\frac{1}{n}\sum_{t=0}^{18}\ln y_t=0{,}839$; $\sum_{t=0}^{18} t\ln y_t=121{,}883$; $S_{ty1}=-1{,}135$; a_1=-0,0378; b_1=1,1795. Отсюда $a=e^{a_1}=0{,}963$; $b=e^{b_1}=3{,}26$.

График функции $y(t)=3{,}26\times 0{,}963^t$ показан на рис. 4.4. Он почти не отличается от прямой линии, что свидетельствует, возможно, о том, что тренд подобран не слишком удачно. Зато трендовое значение при t=19 (прогноз для 1969 г.) y(19)=1,59 почти не отличается от реального y_{19}=1,54.

4.2.2.4. Исключение трендовой составляющей

Для изучения случайной составляющей временного ряда из него нужно удалить трендовую, сезонную и циклическую составляющие. После определения уравнения тренда его можно удалить из ряда, вычислив разности (остатки) y_t-$y(t)$ или отношения $(y_t/y(t))$100% (процентные отклонения от тренда).

На рис 4.7 и 4.8 показаны графики рядов остатков и процентных отклонений от тренда для ряда из табл. 4.2 (безработица в США).

Расчётные значения приведены в табл. 4.7.

Таблица 4.7

Год	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
t	0	1	2	3	4	5	6	7	8	9
y_t	3,4	2,9	3,0	2,8	2,8	4,1	5,0	4,9	4,4	5,2
y(t)	2,2	2,6	3,0	3,3	3,7	4,1	4,5	4,9	5,2	5,6
y_t-y(t)	1,2	0,3	0	-0,5	-0,9	0	0,5	0	-0,8	-0,4
$(y_t/y(t))$100%	154,5	112,4	101,4	83,8	75,2	100,0	111,6	100,8	84,0	92,5
Год	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
t	10	11	12	13	14	15	16	17	18	19
y_t	7,9	4,4	7	6,2	6,1	7,6	8,3	10,7	10,7	8,5
y(t)	6,0	6,4	6,8	7,1	7,5	7,9	8,3	8,7	9,0	9,4
y_t-y(t)	1,9	-2,0	0,2	-0,9	-1,4	-0,3	0	2,0	1,7	-0,9
$(y_t/y(t))$100%	131,7	69,0	103,6	86,8	81,1	96,2	100,2	123,6	118,4	90,2

Процентные отношения позволяют сравнить между собой амплитуды отклонений от тренда. Например, абсолютные значения остатков при t=0 и t=14 почти равны (1,2 и -1,4 соответственно). Но в первом случае остаток 1,3 означает отклонение от тренда на 54,5%, а во втором - только на 19%.

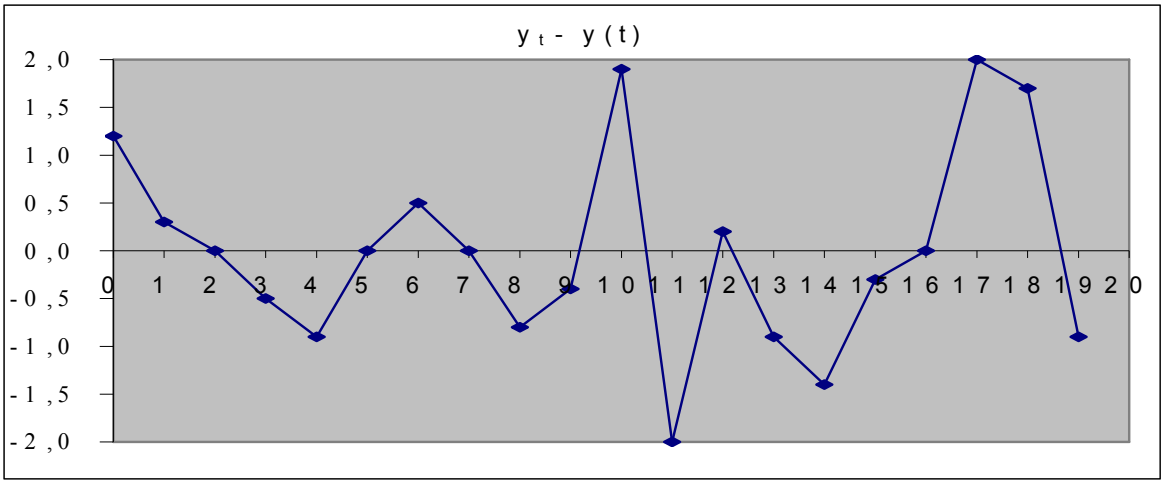


Рис.4.7. Остатки ряда безработицы в США

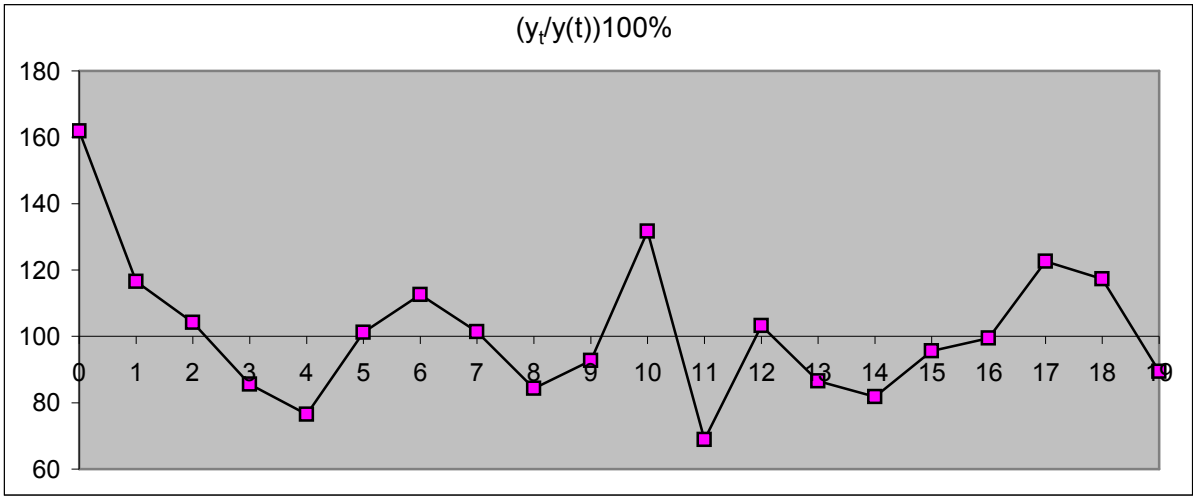


Рис. 4.8. Процентные отклонения от тренда ряда безработицы в США

4.2.2.5. Скользящие средние

Скользящие средние вычисляют, когда неясно, какую подходящую функцию нужно подобрать для тренда. Значение скользящей средней в момент времени t - это среднее арифметическое подряд идущих значений временного ряда на интервале времени, в центре которого - точка t. Обозначим через M_t^p значение скользящей средней в момент времени t, определённое по p точкам.

Например:

$$M_t^3 = \frac{y_{t-1} + y_t + y_{t+1}}{3}; \quad M_t^5 = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5} \text{ и т. п.}$$

Из сказанного ясно, что число p точек должно быть нечётным, иначе среднее арифметическое нужно отнести к дробному моменту времени. На практике, однако, нередки случаи, когда p удобно выбрать чётным. Тогда вычисляют не простое среднее арифметическое, а с некоторыми весами. Например, если положить p=4, можно рассчитать M_t^4 так:

$$M_t^4 = \frac{M_{t-0,5}^4 + M_{t+0,5}^4}{2} = \frac{1}{2}\left(\frac{y_{t-2} + y_{t-1} + y_t + y_{t+1}}{4} + \frac{y_{t-1} + y_t + y_{t+1} + y_{t+2}}{4}\right) =$$

$$= \frac{y_{t-2} + 2y_{t-1} + 2y_t + 2y_{t+1} + y_{t+2}}{8}.$$

Расчёт скользящей средней по четырём точкам свёлся к расчёту по пяти точкам с набором весов (1, 2, 2, 2, 1).

Ряд скользящих средних гораздо более гладкий, чем исходный, ведь отклонения исходного ряда усредняются. Поэтому скользящую среднюю можно считать одной из разновидностей тренда: сглаживая исходный ряд, она даёт представление о тенденции поведения ряда.

4.2.3. Сезонные колебания и индексы сезонности

Говорят, что во временном ряде присутствует сезонная компонента, если в нём наблюдается последовательность почти повторяющихся циклов одинакового периода. В качестве примеров приведём рост объёмов продаж в преддверии Рождества и Нового года, сезонное падение цен на сельхозпродукцию, всплески объёмов перевозок пассажиров городским транспортом утром и в конце рабочего дня и т.п. Сезонная компонента часто присутствует в экономических рядах. Такую компоненту нужно уметь выделять и устранять из ряда, чтобы данные стали равными между собой. Но сначала из ряда необходимо удалить трендовую составляющую. Как правило, для рядов с сезонными колебаниями стирают взвешенную скользящую среднюю. Дело в том, что длины сезонных циклов, как правило, чётны (4 квартала в году, 24 часа в сутках и т.п.), именно поэтому скользящая средняя получается взвешенной, т. к. величину интервала сглаживания целесообразно выбирать равной периоду сезонности.

В качестве примера рассмотрим ряд из табл. 4.5 (затраты на новые здания, сооружения, оборудование в США в 1966 – 1971 гг.). На графике этого ряда (см. рис. 4.5) отчётливо видны сезонные колебания (регулярные падения затрат в первом квартале и рост в четвёртом). Исключим из ряда трендовую составляющую. Для него неплохо подойдёт линейный тренд, но мы построим взвешенную скользящую среднюю по 5 точкам с набором весов (1, 2, 2, 2, 1), т.к. год состоит из четырёх кварталов.

Результаты расчётов приведены в табл. 4.8.

Таблица 4.8

Год	Квартал	t	y_t	M_t^5 (1, 2, 2, 2, 1)	$\frac{y_t}{M_t^5}\times 100\%$	Ряд, очищенный от сезонных колебаний
	1	0	13,33	-	-	14,9
1966	2	1	16,05	-	-	15,8
	3	2	15,92	16,02	99,38	16,0
	4	3	18,22	16,24	112,19	16,6
	1	4	14,46	16,36	88,39	16,3
1967	2	5	16,69	16,38	101,89	16,5
	3	6	16,20	16,45	98,48	16,2
	4	7	18,12	16,55	109,49	16,5
	1	8	15,10	16,64	90,75	17,0
1968	2	9	16,85	16,83	100,12	16,6
	3	10	16,79	17,06	98,42	16,8
	4	11	19,03	17,42	109,24	17,3
	1	12	16,04	17,98	89,21	18,0
1969	2	13	18,81	18,59	101,18	18,6
	3	14	19,25	19,07	100,94	19,3
	4	15	21,46	19,44	110,39	19,5
	1	16	17,47	19,75	88,41	19,6
1970	2	17	20,33	19,91	102,10	20,0
	3	18	20,26	19,96	101,50	20,3
	4	19	21,66	20,02	108,19	19,7
	1	20	17,68	20,04	88,22	19,9
1971	2	21	20,60	20,19	102,03	20,3
	3	22	20,14	-	-	20,2
	4	23	23,04	-	-	20,9

Скользящая средняя построена на рис. 4.5.

Очистим ряд от тренда, рассчитав процентные отклонения $(y_t / M_t^5)\times 100\%$. Эти отклонения распределим по годам и кварталам (табл. 4.9).

Таблица 4.9

Квартал / Год	1	2	3	4
1966	-	-	99,38	112,19
1967	88,39	101,89	98,48	109,49
1968	90,75	100,12	98,42	109,24
1969	89,21	101,18	100,94	110,39
1970	88,41	102,11	101,50	108,19
1971	88,22	102,03	-	-
$\sum$	444,98	507,33	498,72	549,5

Найдём средние значения отклонений от тренда для каждого квартала и всего временного ряда.

Обозначим эти числа $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$, $\bar{y}_4$ и $\bar{y}$ соответственно.

$$\bar{y}_1 = \frac{444{,}98}{5} = 88{,}996;\ \bar{y}_2 = \frac{507{,}33}{5} = 101{,}466;\ \bar{y}_3 = \frac{498{,}72}{5} = 99{,}744;$$

$$\bar{y}_4 = \frac{549{,}5}{5} = 109{,}9;$$

$$\bar{y} = \frac{444{,}98 + 507{,}33 + 498{,}72 + 549{,}5}{20} = \frac{2000{,}53}{20} = 100{,}027.$$

Примем значение $\bar{y}$ за базовое. В среднем значения временного ряда почти не отклоняются от тренда. Зато средние величины отклонений в 1 и 4 кварталах весьма велики. Они объясняются величинами сезонных эффектов.

Рассчитаем, сколько процентов от $\bar{y}$ составляют $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$, $\bar{y}_4$. Эти величины называют индексами сезонности.

$$SI_1 = \frac{\bar{y}_1}{\bar{y}} \cdot 100\% = \frac{88{,}996}{100{,}03} = 88{,}97\%;\quad SI_2 = \frac{\bar{y}_2}{\bar{y}} \cdot 100\% = \frac{101{,}466}{100{,}03} = 101{,}44\%;$$

$$SI_3 = \frac{\bar{y}_3}{\bar{y}} \cdot 100\% = \frac{99{,}744}{100{,}03} = 99{,}72\%;\quad SI_4 = \frac{\bar{y}_4}{\bar{y}} \cdot 100\% = \frac{109{,}9}{100{,}03} = 109{,}87\%.$$

Теперь можно удалить сезонную компоненту из временного ряда. Для этого каждое значение ряда делится на соответствующий индекс сезонности, выраженный в долях. В нашем случае все значения первого квартала делятся на число 0,8997; все значения второго квартала – на число 1,0144; данные третьего квартала – на число 0,9972, а данные четвёртого квартала нужно поделить на 1,0987. Ряд, очищенный от сезонных колебаний, приведён в табл. 4.8. Эти числа почти не отличаются от трендовых значений. Можно сделать вывод, что собственно случайная компонента этого временного ряда невелика, а влияние сезонных эффектов на величину отклонений от тренда не меняется от года к году.

Если для этого временного ряда подобрать линейный тренд, его уравнение будет таким: $y(t) = 0{,}278t + 14{,}86$. Значения индексов сезонности, рассчитанные с применением этого уравнения незначительно отличаются от только что найденных:

$$SI_1 = 89{,}04\%,\ SI_2 = 101{,}7\%,\ SI_3 = 99{,}44\%,\ SI_4 = 109{,}82\%.$$

Располагая уравнением тренда, можно оценить значения временного ряда в следующем году, скорректировав трендовые значения на индексы сезонности. Прогноз получается таким.

$y_{24} = y(24) \cdot SI_1 = 21{,}53 \cdot 0{,}8904 = 19{,}17;\quad y_{25} = y(25) \cdot SI_2 = 21{,}81 \cdot 1{,}017 =$

$= 22{,}18;\quad y_{26} = y(26) \cdot SI_3 = 22{,}09 \cdot 0{,}9944 = 21{,}97;\quad y_{27} = y(27) \cdot SI_4 =$

$= 22{,}37 \cdot 1{,}0982 = 24{,}56.$

В заключение отметим, что по самому определению индексов сезонности их сумма равна 100*k,* где *k* – число сезонов, на которое делится выбранная единица времени (в нашем случае *k*=4, год поделён на 4 квартала).

4.3. Задачи

Для нижеприведённых временных рядов нужно выполнить следующие действия:

1. Представить ряд графически.
2. Подобрать подходящее уравнение тренда по методу наименьших квадратов или подходящую скользящую среднюю, если характер тренда неясен.
3. Удалить трендовую составляющую из временного ряда и построить график остатков.
4. Сравнить поведение ряда остатков индекса промышленного производства США (подъёмы, спады, поворотные точки, где подъёмы и спады сменяют друг друга и, т. п.) с поведением остальных рядов остатков. Можно ли утверждать о существовании закономерностей в поведении рядов?

1. Индекс промышленного производства в США в 1950 – 1985 гг. (1977 г. = 100)

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	33,1	35,9	37,2	40,4	38,2	43,0	44,9	45,5	42,6
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	47,7	48,8	49,1	53,2	56,3	60,1	66,1	72,0	73,5
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	77,6	81,2	78,5	79,6	87,3	94,4	93,0	84,8	92,6
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	100,0	106,5	110,1	108,6	111,0	103,1	109,2	121,4	123,7

2. Индекс потребительских цен США в 1950 – 1985 гг. (1982 – 1984 гг. = 100)

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	24,1	26,0	26,5	26,7	26,9	26,8	27,2	28,1	28,9
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	29,1	29,6	29,9	30,2	30,6	31,0	31,5	32,4	33,4
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	34,8	36,7	38,8	40,5	41,8	44,4	49,3	53,8	56,9
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	60,6	65,2	72,6	82,4	90,9	96,5	99,6	103,9	107,6

3. Население США (млн. чел.) в 1950 – 1985 гг.

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	152,3	154,9	157,6	160,2	163,0	165,9	168,9	172,0	174,9
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	177,8	180,7	183,7	186,5	189,2	191,9	194,3	196,6	198,7
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	200,7	202,7	205,1	207,7	209,9	211,9	213,8	216,0	218,0
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	220,2	222,6	225,1	227,7	230,0	232,3	234,8	237,0	239,3

4. Рабочая сила в США (млн. чел.) в 1950 – 1985 гг.

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	62,2	62,0	62,1	63,0	63,6	65,0	66,6	66,9	67,6
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	68,4	69,6	70,5	70,6	71,8	73,1	74,5	75,8	77,3
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	78,7	80,7	82,8	84,4	87,0	89,4	91,9	93,8	96,2
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	99,0	102,3	105,0	106,9	108,7	110,2	111,6	113,5	115,5

5. Безработица в США (млн. чел.) в 1950 – 1985 гг.

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	5,3	3,3	3,0	2,9	5,5	4,4	4,1	4,3	6,8
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	5,5	5,5	6,7	5,5	5,7	5,2	3,4	2,9	3,0
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	2,8	2,8	4,1	5,0	4,9	4,4	5,2	7,9	4,4
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	7,0	6,2	6,1	7,6	8,3	10,7	10,7	8,5	8,3

6. Индекс производительности труда в США в 1950 – 1985 гг. (1977 г. = 100)

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	51,7	53,8	55,4	57,5	58,4	60,1	60,9	62,5	64,4
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	66,5	67,6	70,0	72,5	74,5	78,7	81,0	83,2	85,5
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	87,8	87,8	88,4	91,3	94,1	95,9	93,9	95,7	98,3
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	100,0	100,8	99,6	99,3	100,7	100,3	103,0	105,5	107,7

7. Реальный валовой национальный продукт (ВНП) США (в ценах 1982 г., млрд. долл.)

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	1203,7	1328,2	1328,2	1435,3	1416,2	1494,9	1525,6	1551,1	1539,2
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	1629,1	1665,3	1708,7	1799,4	1873,3	1973,3	2087,6	2208,3	2271,4
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	2365,6	2423,3	2416,2	2484,8	2605,8	2744,1	2729,3	2695,0	2826,7
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	2958,6	3115,2	3192,4	3187,1	3248,8	3166,0	3279,1	3501,4	3618,7

8. Реальный доход после уплаты налогов на душу населения в США (в ценах 1982 г.) в 1950 – 1985 гг.

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958
y_t	791,8	819,0	844,3	880,0	894,0	944,5	989,4	1012,1	1028,8
Год	1959	1960	1961	1962	1963	1964	1965	1966	1967
y_t	1067,2	1091,1	1123,2	1170,2	1207,3	1291,0	1365,7	1431,3	1493,2
Год	1968	1969	1970	1971	1972	1973	1974	1975	1976
y_t	1551,3	1599,8	1668,1	1728,4	1797,4	1916,3	1896,6	1931,7	2001,0
Год	1977	1978	1979	1980	1981	1982	1983	1984	1985
y_t	2066,6	2167,4	2212,6	2214,3	2248,6	2261,5	2428,1	2668,8	2838,7

9. Средняя цена за фунт стриженной шерсти, выплачиваемая фермерам США в 1950 – 1969 гг.

Год	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
Цена, цент	62,1	97,1	54,1	54,9	53,2	42,8	44,3	53,7	36,4	43,3
Год	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Цена, цент	42,0	42,9	47,7	48,5	53,2	47,1	52,1	39,8	40,5	41,8

Сравнить поведение этого ряда с поведением ряда потребления шерсти (см. табл. 4.4.)

10. Душевое производственное потребление хлопка, ацетата и других искусственных волокон (потребление делится на численность населения континентальной территории США, фунт) в 1950 – 1969 гг.

Год	1950	1951	1952	1953	1954	1955	1956
Хлопок	30,87	31,55	28,48	27,92	25,41	26,51	25,94
Ацетатный шёлк	8,9	8,26	7,74	7,66	7,1	8,58	7,13
Прочие искусств. волокна	0,93	1,27	1,59	1,75	2,02	2,61	2,88
Год	1957	1958	1959	1960	1961	1962	1963
Хлопок	23,7	22,21	24,47	23,19	22,21	22,43	21,33
Ацетатный шёлк	6,87	6,47	7,07	5,84	6,14	6,77	7,6
Прочие искусств. волокна	3,31	3,3	4,19	4,22	4,69	5,76	6,67

Год	1964	1965	1966	1967	1968	1969	
Хлопок	22,09	23,01	23,52	22,12	20,61	19,32	
Ацетатный шёлк	7,89	7,97	8,08	7,53	8,39	7,95	
Прочие искусств. волокна	8,11	10,05	11,62	13,06	17,09	18,51	

В задачах 11 – 14 выполнить следующие действия:

1. Представить ряд графически и убедиться в присутствии сезонных эффектов.
2. Подобрать подходящий тренд и удалить из ряда трендовую составляющую.
3. Рассчитать индексы сезонности и очистить ряд от сезонных колебаний.
4. Предсказать, используя уравнение тренда и найденные значения индексов сезонности, значения временных рядов (по сезонам) в следущем году.

11. Индекс производства автомобилей США (одна из компонент индекса промышленного производства, 1957 – 1959 гг. = 100)

Месяц / Год	Январь	Февраль	Март	Апрель	Май	Июнь
1968	179,5	173,8	193,4	183,5	202,4	208,3
1969	187,7	181,5	184,8	164,6	165,3	181,0
1970	146,2	140,4	152,2	162,4	173,2	185,0
Месяц / Год	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
1968	134,1	45,6	165,0	207,4	212,2	192,0
1969	94,7	91,9	175,0	186,0	172,3	155,3
1970	98,0	68,9	108,5	88,0	87,0	137,6

12. Расстояния, пройденные авиалайнерами Соединённого Королевства за месяц, тыс. миль

Месяц / Год	Январь	Февраль	Март	Апрель	Май	Июнь
1963	6827	6178	7084	8162	8462	9644
1964	7269	6775	7819	8371	9069	10248
1965	8350	7829	8829	9948	10638	11253
1966	8186	7444	8484	9864	10252	12282
1967	8334	7899	9994	10078	10801	12950
1968	8639	8772	10894	10455	11179	10588
1969	9491	8919	11607	8852	12537	14759
1970	10840	10436	13589	13402	13103	14933
Месяц / Год	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
1963	10466	10748	9963	8194	6848	7027

1964	11030	10882	10330	9109	7685	7602
1965	11424	11391	10665	9396	7775	7933
1966	11637	11577	12417	9637	8094	9280
1967	12222	12246	13281	10366	8730	9614
1968	10794	12270	13812	10857	9290	10925
1969	13667	13731	15110	12185	10645	12161
1970	14147	14057	16234	12389	11595	12772

13. Значения квартальных выручек (тыс. долл.) авиакомпании, специализирующейся на зарубежных рейсах

Год Квартал	1966	1967	1968	1969	1970	1971
1	-	468	508	555	622	693
2	-	866	753	875	1035	1138
3	1037	1327	1239	1382	1629	1690
4	495	546	530	595	687	773

14. Производство молока в России с января 1992 г. по октябрь 1996 г. (тыс. тонн в месяц)

Месяц Год	Январь	Февраль	Март	Апрель	Май	Июнь
1992	2015	2123	2624	2891	3335	4071
1993	1759	1773	2361	2649	3203	3936
1994	1510	1484	1988	2211	2559	3209
1995	1172	1226	1651	1859	2392	2864
1996	1038	1104	1439	1521	1827	2446
Месяц Год	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
1992	4040	3392	2467	2092	1494	1562
1993	3861	3321	2438	1760	1299	1345
1994	3204	2687	2031	1506	1050	1054
1995	2714	2420	1925	1338	984	1020
1996	2369	2081	1577	1081	-	-

5. ПОНЯТИЕ ОБ ИНДЕКСАХ

А для низкой жизни были числа,
Как домашний подъярёмный скот,
Потому что все оттенки смысла
Умное число передаёт.

Н. Гумилёв

Индексы **–** это простое и удобное средство сравнения данных между собой. Дело в том, что вычисление разности между двумя числами даёт мало информации. Например, если в течение недели курс доллара вырос с четырёх до пяти рублей за доллар в первом случае и с 4000 до 4001 рубля за доллар во втором случае, то оба раза разница между курсами составляла бы 1 рубль. Но в первом случае произошла катастрофа финансовой системы, а во втором говорить просто не о чем. Если же выписать процентные разницы, то они окажутся равными 25 и 0,25% и говорят сами за себя.

Индексы – это проценты, дадим их строгие определения.

5.1. ИНДИВИДУАЛЬНЫЕ (ЧАСТНЫЕ) ИНДЕКСЫ

Пусть имеется некоторый упорядоченный набор данных, например, временной ряд. Выберем одно из его значений за ***базовое.*** Это значение обозначается буквой C и называется константой ***базового периода.*** А теперь выразим все остальные данные в процентах к числу C. Эта операция и называется индексированием. Таким образом, формула для вычисления индекса I_t имеет вид

$$I_t = \frac{y_t}{C} \cdot 100\%,$$

где y_t — исходные значения из временного ряда, I_t – соответствующий ему индекс.

Для примера вычислим индексы оптовых цен и объёмов производства яблок в США в 1960 – 1969 гг. Данные за 1960 г. примем за базовые, этот факт обозначается так: 1960 = 100. Такую запись применяют в литературе по экономической статистике. Тем самым показывают, что значения базового периода приняты за 100%. Исходные данные и значения индексов приведены в табл. 5.1. Цены обозначены буквой p в центах за фунт; объёмы производства обозначены буквой q и изменяются в миллиардах фунтов.

Таблица 5.1

Год	*q*	Iq	*p*	Ip
1960	4,91	100,0	4,84	100,0
1961	5,63	114,7	4,15	85,7
1962	5,68	115,7	4,32	89,3
1963	5,72	116,5	4,21	87,0
1964	6,24	127,1	4,00	82,6
1965	5,99	122,0	4,35	89,9
1966	5,65	115,1	4,46	92,1
1967	5,39	109,8	5,56	114,9
1968	5,44	110,8	6,11	126,2
1969	6,72	136,9	4,09	84,5

Пояснение к вычислению индексов:

$I_{q\ 1961}$ = (5,63/4,91)*100% = 114,7; $I_{q\ 1962}$ = (5,68/4,91) = 115,7;

$I_{p\ 1961}$ = (4,15/4,84)*100% = 85,7; $I_{p\ 1962}$ = (4,32/4,84) = 89,3.

Графики индексных рядов цен и объёмов производства показаны на рис.5.1.

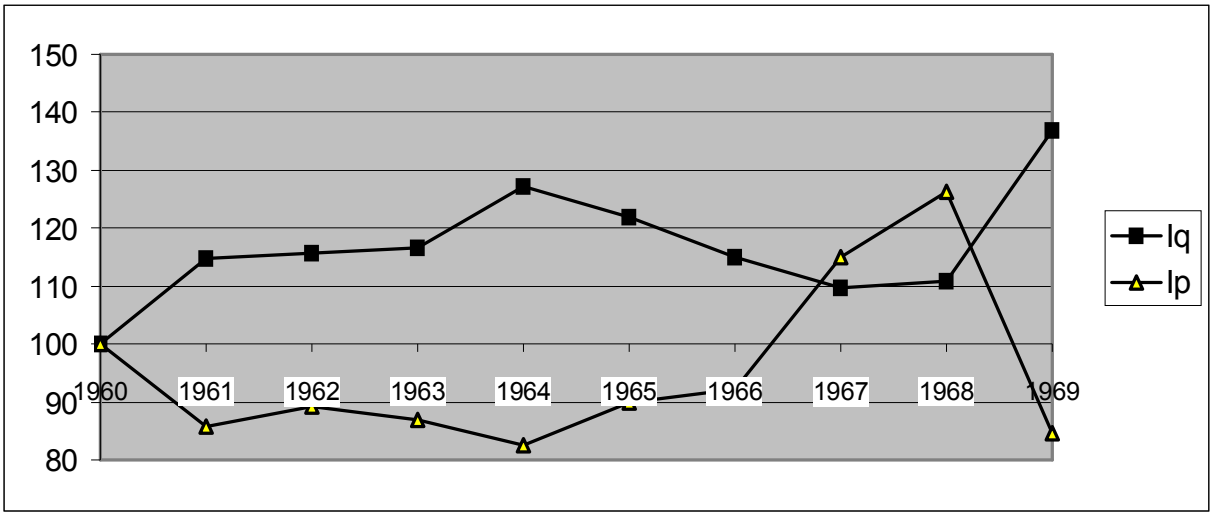


Рис. 5.1

На рис.5.1 видно, что рост объёмов производства яблок влечёт падение оптовых цен на яблоки, наоборот, падение объёмов производства влечёт рост цен.

Между прочим построить в одной системе координат графики двух рядов разной размерности (цены и объёмы производства) невозможно, а графики индексных (процентных) рядов строятся естественным образом. После чего появляется возможность сравнивать поведение этих рядов во времени.

Часто в качестве базового периода выбирают несколько подряд идущих периодов временного ряда, тогда константа базового периода - это среднее арифметическое соответствующих значений временного ряда.

Данные из индексного ряда легко сравниваются только со 100% базового периода. Но между собой их нужно сравнивать аккуратно. Например, если индексы цен 1965 и 1967 гг. из табл. 5.1 равны соответственно 89,9 и 114,9%, то неправильно было бы говорить, что цены в 1967 г. выросли по сравнению с ценами 1965 г. на 25 %. Такая разница называется разницей в пунктах (цены 1967 г. выше цен 1965 г. на 25 пунктов). Чтобы узнать правильную процентную разницу, нужно произвести сдвиг базового периода. Если цены 1965 г. принять за 100%, то цены 1965 г. составляют (114,9/89,9)*100% =127,8%. Таким образом, цены 1967 г. выросли на 27,8% в сравнении с ценами 1965 года. Если цены 1967г. принять за 100%, то цены 1965г. составляют (89,9/114,9)*100% = =78,2% цен 1967 г. Значит, цены 1965 г. ниже цен 1967 г. на 21,8%.

Проиллюстрируем разницу между пунктами и процентами на таком примере. Пусть в течение двух недель курс доллара вырос с 5000 рублей за один доллар до 5020 рублей за доллар. Разница в пунктах равна 20 пунктам, процентная же разница составляет ((5020-5000)/5000)·100% = =0,4%. Это немного (по крайней мере, в российских условиях). Если же за те же две недели курс доллара вырос с 5 до 25 рублей за один доллар, то разница в 20 пунктов соответствует процентной разнице в 400%. Это означает крах финансовой системы для любой страны.

5.2. ОБЩИЕ ИНДЕКСЫ

5.2.1. Агрегатные индексы

Агрегатные индексы вычисляются сразу по нескольким временным рядам. Каждыз.й из рядов умножается на некоторый вес, а затем для каждого момента времени эти произведения складываются. Ряд сумм индексируется. Схема расчёта агрегатных индексов приведена в табл. 5.2.

Таблица 5.2

t	w_1y_{t1}	w_1y_{t2}	…	w_ny_{tn}	$y_t=\sum w_iy_{ti}$	I_t
0	w_1y_{01}	w_1y_{02}		w_ny_{0n}	$y_0=\sum w_iy_{0i}$	I_0
1	w_1y_{11}	w_1y_{12}		w_ny_{1n}	$y_1=\sum w_iy_{1i}$	I_1
2	w_1y_{21}	w_1y_{22}		w_ny_{2n}	$y_2=\sum w_iy_{2i}$	I_2
3	w_1y_{31}	w_1y_{32}		w_ny_{3n}	$y_3=\sum w_iy_{3i}$	I_3
…	…	…		…	…	…

Здесь t – периоды или моменты времени; y_{tj} – члены временных рядов; j – номер временного ряда; j = 1,2, …, n; w_j – вес, на который умножаются

члены *j*-го временного ряда. Результирующий ряд y_t индексируется, $I_t=(y_t/C)*100\%$, где $C=y_k$ или $C=\frac{1}{m}\sum_{t=k}^{k+m} y_t$.

Веса определяют или вычисляют многими различными способами. Чаще всего в качестве весов для индексов цен пользуются объёмами производства или продаж, а для индексов объёма производства – ценами. При этом, как правило, веса совпадают со значениями базового периода.

5.2.2. Средние индексы

Средние индексы вычисляют по нескольким индивидуальным индексным рядам. Ряды индексов умножаются на некоторые веса, а затем сумма этих произведений для каждого значения *t* делится на сумму всех весов w_j. Чтобы средний индекс совпадал с соответствующим агрегатным, весами индивидуальных индексов должны быть слагаемые знаменателя агрегатного индекса.

5.2.3. Индексы цен

При вычислении индексов цен в качестве весов выбирают объёмы производства и продажи товаров. Пусть временные ряды содержат только два периода: ***базовый***, нумеруемый цифрой 0, и ***отчётный*** (текущий), нумеруемый цифрой 1. Для расчёта индексов цен используют две основные формулы – Пааше и Ласпейреса.

Весами в ***индексе цен Пааше*** выступает количество продукции отчётного периода, поэтому такой индекс цен вычисляется по формуле

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\%.$$

В числителе дроби стоит стоимость продукции отчётного периода, в знаменателе – стоимость тех же товаров в ценах базового периода. Такой индекс цен показывает, насколько товары в отчётном периоде стали дороже (дешевле), чем в базовом.

Весами в ***индексе цен Ласпейреса*** является количество продукции базового периода, такой индекс цен вычисляется по формуле

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100\%.$$

Индекс цен Ласпейреса показывает, во сколько бы раз товары базового периода подорожали (подешевели) из-за изменения цен на них в текущем периоде.

На практике индекс цен, рассчитанный по формуле Пааше, обычно чуть меньше индекса, рассчитанного по формуле Ласпейреса.

Рассчитаем индексы цен по формулам Пааше и Ласпейреса для такого условного примера (табл. 5.3).

Таблица 5.3

Товар	Цена за 1 единицу		Продано единиц	
	p_0	p_1	q_0	q_1
1	1,8	1,9	1000	1300
2	3,5	3,7	1500	2000
3	5,8	5,5	500	550

Индекс цен Пааше равен

$$I_p^n = \frac{1{,}9 \cdot 1300 + 3{,}7 \cdot 2000 + 5{,}5 \cdot 550}{1{,}8 \cdot 1300 + 3{,}5 \cdot 2000 + 5{,}8 \cdot 550} \cdot 100\% = 102{,}9\%.$$

Индекс цен Ласпейреса равен

$$I_p^л = \frac{1{,}9 \cdot 1000 + 3{,}7 \cdot 1500 + 5{,}5 \cdot 500}{1{,}8 \cdot 1000 + 3{,}5 \cdot 1500 + 5{,}8 \cdot 500} \cdot 100\% = 102{,}5\%/$$

5.2.4. Дефлятирование стоимостных величин

Дефлятированием стоимостных величин называется деление стоимостной величины на подходящий агрегатный индекс цен, вследствие чего устраняется влияние инфляции, стоимостные величины становятся выраженными в постоянных денежных единицах и сравнимыми между собой. Дефлятируют, например, валовой национальный продукт (ВНП) при помощи специального индекса-дефлятора, доходы населения (при помощи индекса потребительских цен), доходы крупных фирм и корпораций (при помощи индекса оптовых цен).

Валовой национальный продукт – это стоимость товаров и услуг, произведённых экономикой за некоторый промежуток времени (квартал, год). В табл. 5.4. показано дефлятирование ВНП США. В результате получается значение ВНП в долларах 1958 г. (этот год принят за базовый). Значения ВНП, выраженные в постоянных долларах, можно, в отличие от значений ВНП, выраженных в текущих долларах, сравнить между собой. Из табл. 5.3. видно, что, из-за падения цен в 1929 – 1933 гг., ВНП реально уменьшился в 1,44 раза, а не в 1,85 раза, если проводить сравнение в текущих долларах. Вследствие роста цен в 1941 – 1971 гг., ВНП реально вырос только в 2,8 раза. Если сравнивать в текущих долларах, получится ложный вывод о росте ВНП в 1971 г. в сравнении с 1941 г. в 8,41 раза.

Таблица 5.4

Год	ВНП, текущие доллары (млрд .долл.)	Дефлятор скрытых цен (1958г.=100)	ВНП в постоянных долларах (1958 г.)
1929	103,1	50,6	203,8
1933	55,6	39,3	141,5
1941	124,5	47,2	263,8
1950	284,8	80,2	355,1
1958	447,3	100	447,3
1965	681,2	110,9	614,2
1971	1046,8	141,6	739,3

С помощью индекса потребительских цен дефлятируют заработную плату и пенсии, чтобы определить реальные доходы населения. Такие оценки всегда приближённые, ведь люди тратят свои деньги по-разному, структура их расходов не может в точности совпадать с той, которая принята при расчёте индекса потребительских цен.

В качестве примера дефлятирования рассчитаем, на сколько уменьшатся реальные доходы, если за год потребительские цены выросли в среднем на 50%, а средний рост доходов составил 20%. Рост цен на 50% означает их увеличение в 1,5 раза. Доходы же выросли только в 1,2 раза. Поэтому реальные средние доходы, выраженные в денежных единицах начала года, составляют (1,2/1,5)·100% = 80% от прежних доходов, т.е. уменьшились на 20%. Неправильно было бы сказать, что доходы уменьшились на 50% – 20% = 30%. Если бы цены выросли на 20 процентов, а доходы на 50 процентов (в жизни так может случиться только с отдельными семьями, а не со всей совокупностью населения!), то рост средних доходов составил бы [(1,5/1,2) – 1]·100% = 25% (снова не 30%).

5.3. ЗАДАЧИ

1. Один и тот же временной ряд индексировался по разным базовым периодам. Заполнить пробелы в представленных индексных рядах.

Год	1962	1963	1964	1965	1966
I_t, 1961=100	92,3	124,9	175,8	-	-
I_t, 1967=100	-	-	334	222,7	161,3

2. Ниже приводятся оптовые цены (обозначены буквой p, центы за фунт) и объёмы производства (обозначены буквой q, миллиарды фунтов) яблок и персиков в США в 1960 – 1969 годах. Вычислить индивидуальные и агрегатные индексы цен и объёмов производства.

Год	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Яблоки										
q	4,91	5,63	5,68	5,72	6,24	5,99	5,65	5,39	5,44	6,72
p	4,84	4,15	4,32	4,21	4	4,35	4,46	5,56	6,11	4,09
Персики										
q	3,56	3,7	3,55	3,51	3,43	3,35	3,38	2,68	3,59	3,67
p	3,84	3,95	3,87	4,35	4,59	4,45	5,27	6,36	5,44	5,35

3. Построить парные графики цен и объёмов производства индексных рядов из задачи 2. Подтверждают ли графики утверждение, что, при прочих равных условиях, большие объёмы производства соответствуют меньшим ценам, и наоборот?

4. В таблице указаны среднегодовые оклады служащих некоторой фирмы и индексы потребительских цен.

Год	1966	1967	1968	1969	1970	1971
Средний оклад, долл.	10670	11060	11910	12790	14400	14720
Индекс потребительских цен, 1967=100	97,2	100	104,2	109,8	116,3	121,3

Вычислить реальные оклады. Допустим, что номинальные оклады выросли на 3%, а индекс потребительских цен вырос на 2 %. Можно ли утверждать, что реальные оклады выросли ровно на 1%? Каков точный ответ?

5. Найти агрегатные индексы цен по формулам Пааше и Ласпейреса.

Товар	Единица измерения	Цена, руб.		Количество проданных товаров	
		Апрель	Май	Апрель	Май
Чай	Пачка	16,38	17,04	1000	5000
Кофе	Банка	69,25	73,4	2000	2500
Сыр	кг	50,4	52,4	400	500

6. Написать формулы Пааше и Ласпейреса для вычисления индексов объема проданных (произведенных) товаров, когда в качестве весов для объема проданных товаров берутся цены. Найти индексы объема продаж для предыдущего примера.

6. ПРОВЕРКА ГИПОТЕЗЫ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО КРИТЕРИЮ ПИРСОНА (КРИТЕРИЮ χ^2)

Те, что веруют слепо, - пути не найдут.
Тех, кто мыслит, - сомнения вечно гнетут.
Опасаюсь, что голос раздастся однажды:
«О, невежды! Дорога не там и не тут?»
О. Хайям (перевод Г. Плисецкого)

6.1. ПРИМЕР

Рассмотрим такую ситуацию. 200 электронных ламп, выбранных наудачу из большой партии, испытывались на продолжительность работы. Результаты (в часах) таковы (табл. 6.1):

Таблица 6.1

$[x_{i-1};x_i)$	[0;300)	[300;600)	[600;900)	[900;1200)	[1200;1500)	[1500;1800)
n_i	53	41	30	22	16	12
$[x_{i-1};x_i)$	[1800;2100)	[2100;2400)	[2400;2700)	[2700;3000)	[3000;3300)	—
n_i	9	7	5	3	2	—

Хотелось бы дать разумный ответ на такие вопросы: какую продолжительность работы следует ожидать, если взять наудачу лампу из этой же партии? Какова вероятность, что лампа проработает не менее 1000 часов? Какова вероятность того, что лампа проработает менее 200 часов? Ответить на эти вопросы легко, если известен закон распределения случайной величины X – времени работы лампы. Но его-то мы не знаем. Мы располагаем только выборкой (правда, достаточно большой, n = 200) из генеральной совокупности X. Попробуем, пользуясь этой выборкой, подобрать подходящий закон распределения.

Построим прежде всего гистограмму (рис. 6.1).

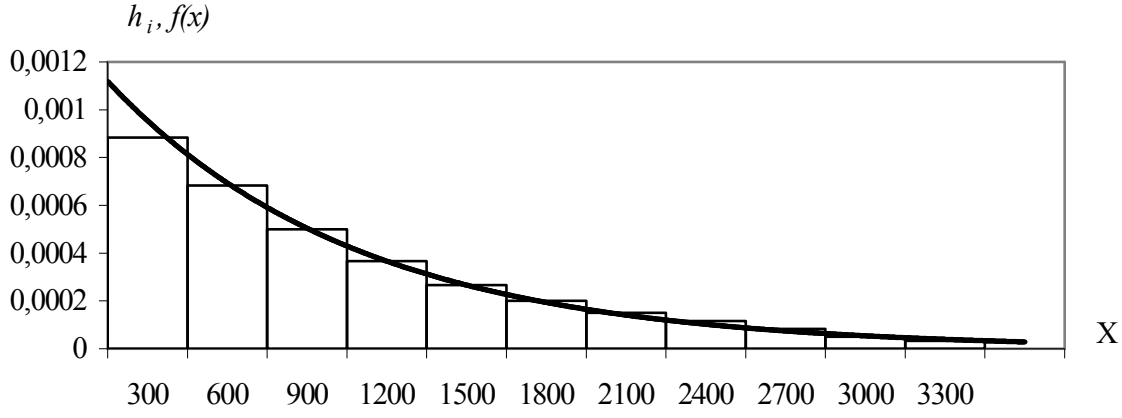


Рис. 6.1

Высоты прямоугольников таковы:

$$h_1 = \frac{53}{n*h} = \frac{53}{200*300} = 0{,}00088;\ h_2 = \frac{41}{n*h} = \frac{41}{60000} = 0{,}00068;$$

$$h_3 = \frac{30}{n*h} = \frac{30}{60000} = 0{,}0005;\ h_4 = 0{,}00037;\ h_5 = 0{,}00027;\ h_6 = 0{,}0002;$$

$h_7 = 0{,}00015$; $h_8 = 0{,}00012$; $h_9 = 0{,}00008$; $h_{10} = 0{,}00005$; $h_{11} = 0{,}00003$.

Гистограмма – аналог графика функции плотности вероятности. В нашем случае гистограмма очень похожа на график функции плотности показательного закона. Мы вправе предположить, что большая выборка хорошо представляет генеральную совокупность и что если гистограмма похожа на график экспоненты, то это означает, что выборка извлечена из генеральной совокупности, распределенной по показательному закону с функцией плотности вероятности

$$f(x) = \lambda e^{-\lambda x}.$$

Однако показательный закон зависит от одного параметра – числа λ. Чтобы полностью описать закон, нужно знать, чему равно λ. Подберем значение λ по выборке, причем поступим самым бесхитростным способом. Как известно, математическое ожидание случайной величины, имеющей показательное распределение, $M(X) = 1/\lambda$. Если наша выборка хорошо представляет генеральную совокупность, мы вправе полагать, что значение выборочного среднего $\bar{x}$ не слишком отличается от $M(X)$. Поэтому найдем $\bar{x}$ и положим $\lambda = 1/\bar{x}$.

$$\bar{x} = \frac{1}{200}(150 \cdot 53 + 450 \cdot 41 + 750 \cdot 30 + 1050 \cdot 22 + 1350 \cdot 16 + 1650 \cdot 12 +$$

$$+1950 \cdot 9 + 2250 \cdot 7 + 2550 \cdot 5 + 2850 \cdot 3 + 3150 \cdot 2) = 871{,}5(\text{ч}).$$

Тогда $\lambda = 1/\bar{x} \approx 0{,}00115$, $f(x) = 0{,}00115e^{-0{,}00115x}, x \geq 0$.

Вычислим значения $f(x)$ на границах интервалов (табл. 1.2) и построим график функции плотности вероятности прямо на гистограмме (см. рис. 6.1).

Таблица 6.2

x_i	0	300	600	900	1200	1500
$f(x_i)$	0,00115	0,00081	0,00058	0,00041	0,00029	0,0002
x_i	1800	2100	2400	2700	3000	3300
$f(x_i)$	0,000115	0,0001	0,00007	0,00005	0,000037	0,000026

Не следует увлекаться слишком большим количеством значащих цифр, ведь все наши данные достаточно приближенные.

Кривая функции плотности вероятности $f(x)$ очень «ладно» легла на гистограмму. Такое хорошее совпадение гистограммы и графика $f(x)$ прибавляет уверенности в том, что закон распределения генеральной совокупности X выбран достаточно точно.

Попробуем теперь оценить числом расхождение между экспериментальными данными и тем, что должно быть «по теории».

Мы можем вычислить теоретическую вероятность p_i попадания случайной величины X, распределенной по показательному закону с функцией плотности $f(x) = 0{,}00115e^{-0{,}00115x}$, $x \geq 0$ в интервал $[x_{i-1}, x_i)$.

$$p(x_{i-1} < X < x_i) = e^{-\lambda x_{i-1}} - e^{-\lambda x_i} = e^{-0{,}00115x_{i-1}} - e^{-0{,}00115x_i}.$$

Зная вероятность p_i, можно вычислить математическое ожидание числа попаданий случайной величины X в интервал $[x_{i-1}, x_i)$ в результате n независимых испытаний, оно равно np_i. Теперь можно найти разность $n_i - np_i$ между числом вариант выборки, попавших в интервал $[x_{i-1}, x_i)$, и ожидаемым числом попаданий. Чтобы оценить суммарное расхождение между теоретическими и опытными данными, нужно сложить все полученные разности. Чтобы положительные и отрицательные разности не уничтожили друг друга, возведем их в квадрат. Кроме того, важно не абсолютное значение $n_i - np_i$, а относительное $(n_i - np_i)/np_i$. Действительно, если $n_i = 0$, $np_i = 1$, это совсем не одно и то же, что в случае, когда $n_i = 10$, $np_i = 11$. Относительное отклонение в первом случае равно 1, а во втором – только 1/11.

Итак, вычислим прежде всего вероятности p_i.

$p_1 = P(0 < X < 300) = e^{-\lambda*0} - e^{-\lambda*300} = e^{0} - e^{-0{,}345} = 1 - 0{,}708 = 0{,}2918;$

$p_2 = P(300 < X < 600) = e^{-\lambda*300} - e^{-\lambda*600} = 0{,}7082 - 0{,}5016 = 0{,}2066;$

$p_3 = P(600 < X < 900) = e^{-\lambda*600} - e^{-\lambda*900} = 0{,}1464;$

$p_4 = 0{,}1036$; $p_5 = 0{,}0734$; $p_6 = 0{,}052$; $p_7 = 0{,}0368$; $p_8 = 0{,}0261$;

$p_9 = 0{,}0185$; $p_{10} = 0{,}0131$; $p_{11} = 0{,}0092$.

Дальнейшие вычисления приведены в табл. 6.3.

Таблица 6.3

$[x_{i-1}; x_i)$	p_i	np_i	n_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
[0;300)	0,2918	58,36	53	-5,36	0,490
[300;600)	0,2066	41,32	41	-0,32	0,002
[600;900)	0,1464	29,28	30	0,72	0,018
[900;1200)	0,1036	20,72	22	1,28	0,079
[1200;1500)	0,0734	14,68	16	1,32	0,119
[1500;1800)	0,0520	10,40	12	1,60	0,246
[1800;2100)	0,0368	7,36	9	1,64	0,365
[2100;2400)	0,0261	5,22	7	1,78	0,607
[2400;2700)	0,0185	3,70	5	1,30	0,457
[2700;3000)	0,0131	2,62	3	0,38	0,056
[3000;3300)	0,0092	1,84	2	0,16	0,014
–	$\sum p_i = 0{,}9775$	$\sum np_i = 195{,}5$	$\sum n_i = 200$	–	$\chi^2 = 2{,}45$

Сумма вероятностей p_i равна 0,9775. Это значит, что интервал [0; 3300) охватывает практически все возможные значения выбранного нами теоретического закона. Сумма чисел последнего столбца традиционно

обозначается буквой χ^2 (читается «хи - квадрат»). В нашем случае

$$\chi^2 = \sum_{i=1}^{11} \frac{(n_i - np_i)^2}{np_i} = 2{,}45 .$$

Много это или мало?

6.2. НЕМНОГО ТЕОРИИ

Только что мы находили число χ^2.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

где k – число интервалов; n_i – частота i-го интервала;

p_i – теоретическая вероятность попадания случайной величины X (генеральной совокупности) в i-й интервал;

n – число независимых испытаний (объем выборки);

np_i – математическое ожидание числа попаданий случайной величины X в i-й интервал

Но на приведенную формулу можно посмотреть и по-другому. Вместо числа n_i рассмотрим случайную величину n_i (в математической статистике случайные величины и их значения часто обозначаются одними и теми же маленькими буквами). Случайная величина n_i - это число появлений «успеха» в n независимых испытаниях, где под «успехом» понимается попадание случайной величины X в i-й интервал. Таким образом, вероятность «успеха» равна p_i, а случайная величина n_i имеет биномиальное распределение с параметрами n и p_i. В частности, $M(n_i) = np_i$. Рассмотрим теперь случайную величину χ^2, функцию от случайных величин $n_1, n_2, \ldots, n_k$, определяемую формулой

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} .$$

Еще раз подчеркнем, что в этой формуле n и p_i – это числа, а n_i – это случайные величины. Имея выборку, мы можем найти значения случайных величин n_i, которые они приняли в результате n независимых испытаний, и вычислить затем значение $\chi_{эксп}$ – экспериментальное значение случайной величины χ^2. Можно доказать, что если закон распределения генеральной совокупности X подобран правильно, то с ростом n случайную величину χ^2 можно считать распределенной по так называемому закону распределения χ^2. Это непрерывное распределение, формулу функции плотности вероятности которого мы не будем здесь приводить. Распределение зависит от одного параметра r, который называется числом степеней свободы. В нашем случае

$$r = k\text{-}1\text{-}S,$$

где k – число интервалов;

S – число параметров закона распределения, вычисленных по выборке.

Возникает естественный вопрос: каким должно быть число n, чтобы его можно было считать «достаточно большим» и пользоваться распределением χ^2? Желательно, чтобы n было таким большим, чтобы все произведения np_i были не меньше 5 (рекомендация всех учебников по статистике). На самом деле, как показывает практика, вполне достаточно выполнения неравенств $np_i \geq 1$, $n \geq 50$.

Примерный график функции плотности вероятности случайной величины χ^2 показан на рис. 6.2.

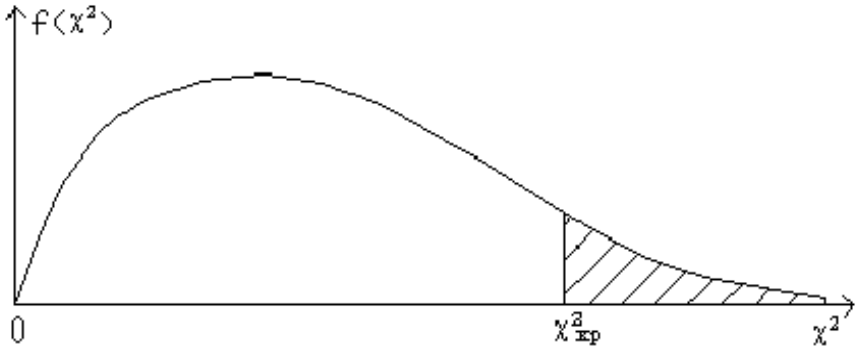


Рис.6.2

Если закон распределения генеральной совокупности X подобран правильно, экспериментальное значение $\chi_{эксп}$, вычисленное на основании выборки, не может быть слишком большим. Зададимся достаточно большой вероятностью β (β = 0,9; 0,95; 0,99), так что события с вероятностью $\alpha = 1 - \beta$ будем считать практически невозможными. Вероятность α называют уровнем значимости.

С точки зрения подтверждения выдвинутой нами гипотезы о законе распределения генеральной совокупности X мы должны считать практически невозможными большие значения случайной величины χ^2. Мы считаем практически невозможными значения случайной величины χ^2 из интервала $(\chi^2_{кр}, \infty)$, где число $\chi^2_{кр}$ определяется из условия (см. рис.6.2)

$$p(\chi^2 > \chi^2_{кр}) = \alpha .$$

Для распределения χ^2 составлены специальные таблицы (приложение 3). По ним можно найти число $\chi^2_{кр}$, зная α и число степеней свободы r. Число $\chi^2_{кр}$ сравнивают с числом $\chi^2_{эксп}$. Если оказывается, что $\chi^2_{эксп} < \chi^2_{кр}$, то говорят, что с точки зрения принятия выдвинутой гипотезы о законе распределения генеральной совокупности X произошло достоверное событие. Гипотеза считается не противоречащей опытным данным и принимается. Если же оказывается, что $\chi^2_{эксп} > \chi^2_{кр}$, то

выдвинутая гипотеза отвергается, считается, что она противоречит опытным данным.

В нашем случае $k = 11$, $r = 11 - 1 - 1 = 9$ (по выборке был определен один параметр - λ). Если положить $\beta = 0{,}95$ ($\alpha = 0{,}05$ - наиболее употребительное значение уровня значимости), то по таблице распределения χ^2 находим, что $\chi^2_{кр} = 16{,}92$. Между тем $\chi^2_{эксп} = 2{,}45 < \chi^2_{кр}$. Так что мы можем считать, что случайная величина X имеет показательное распределение с параметром $\lambda = 0{,}00115$. Если бы мы объединили три последних интервала в один, то имели бы: $r = 9 - 1 - 1 = 7$; $\chi^2_{кр} = 14{,}07$; $\chi^2_{эксп} = 2{,}33 < \chi^2_{кр}$.

Случайная величина χ^2 называется критерием χ^2. Критерий χ^2 был предложен Карлом Пирсоном в 1900 г. До этого времени совпадение экспериментальных результатов с теоретическими оценивалось по тому, как они выглядят на графике.

Нам осталось ответить на вопросы, поставленные в пункте 6.1. Мы считаем справедливым показательный закон с параметром $\lambda = 0{,}00115$. Следовательно, $M(X) = 1/\lambda \approx 870$ (ч).

$$p(X > 1000) = e^{-\lambda * 1000} - e^{-\infty} = e^{-1{,}15} \approx 0{,}32;$$
$$p(X < 200) = e^{0} - e^{-\lambda * 200} = 1 - e^{-0{,}23} \approx 0{,}21.$$

1.3. ДРУГИЕ ПРИМЕРЫ

6.3.1. Проверка гипотезы о нормальном законе распределения

Заказчику необходимы валы с допустимым отклонением диаметра от номинального размера ±0,1 мкм. Прежде чем покупать партию из 1000 валов, он приобрел партию из 200 валов, чтобы оценить ожидаемую долю неподходящих ему изделий. Результаты измерений представлены в табл. 6.4.

Таблица 6.4

200 отклонений диаметра вала от номинального размера (мкм)

Середина интервала	-0,14	-0,12	-0,10	-0,08	-0,06	-0,04	-0,02
Частота	3	8	11	20	27	36	29
Середина интервала	0,00	0,02	0,04	0,06	0,08	0,10	0,12
Частота	18	17	17	8	4	1	1

Здесь $h = 0{,}02$ мкм; $n = 200$; $nh = 4$.

Гистограмма показана на рис.6.3. Высоты гистограммы таковы:

$h_1 = 0{,}75$; $h_2 = 2$; $h_3 = 2{,}75$; $h_4 = 5$; $h_5 = 6{,}75$; $h_6 = 9$; $h_7 = 7{,}25$; $h_8 = 4{,}5$; $h_9 = h_{10} = 4{,}25$; $h_{11} = 2$; $h_{12} = 1$; $h_{13} = h_{14} = 0{,}25$.

Числовые характеристики: $\bar{x} = -0{,}028$ (мкм); $S = 0{,}05$ (мкм).

Судя по гистограмме, можно заключить, что случайная величина X – отклонение диаметра вала от номинального – имеет нормальное распределение. Функция плотности нормального закона зависит от двух параметров – a и σ: $f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{\frac{-(x-a)^2}{2 \cdot \sigma^2}}$.

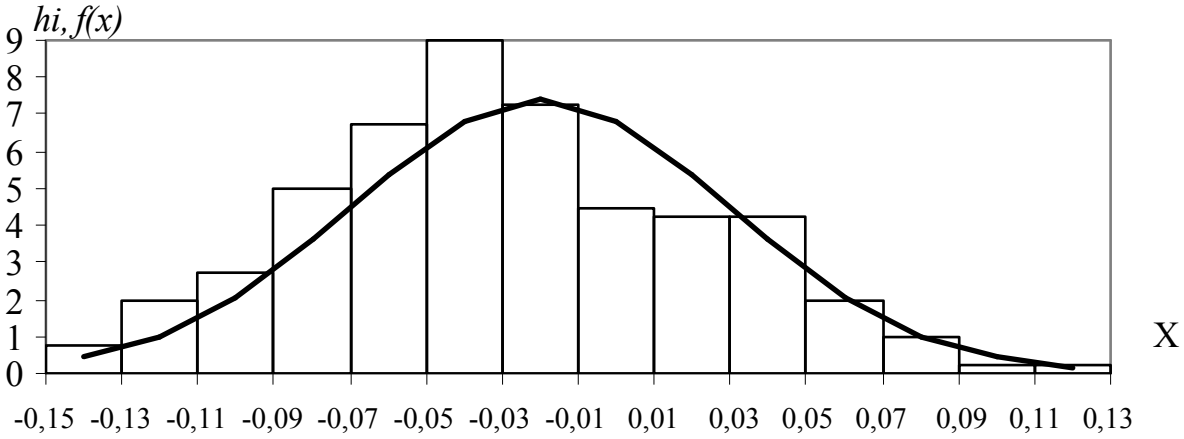


Рис. 6.3

Как известно, $M(X) = a$, $\sigma(X) = \sigma$. Для определения a и σ положим, что $a = \bar{x}$, $\sigma = S$. Отсюда $a = -0{,}03$; $\sigma = 0{,}05$ (значение $\bar{x}$ округлено, исходя из соображений здравого смысла). Тогда

$$f(x) = \frac{1}{0{,}05 \cdot \sqrt{2\pi}} e^{\frac{-(x+0{,}03)^2}{2 \cdot 0{,}0025}} = 8 \cdot e^{-200(x+0{,}03)^2}.$$

Значения функции плотности вероятности на границах интервалов таковы (табл. 1.5):

Таблица 6.5

x_i	-0,15	-0,13	-0,11	-0,09	-0,07	-0,05	-0,03	-0,01
$f(x_i)$	0,45	1,08	2,22	3,89	5,81	7,38	8,00	7,38
x_i	0,01	0,03	0,05	0,07	0,09	0,11	0,13	–
$f(x_i)$	5,81	3,89	2,22	1,08	0,45	0,16	0,05	–

График функции плотности вероятности показан на рис. 6.3.

Вычислим теоретические вероятности попадания в интервалы. Формула вычисления вероятности попадания в интервал $[x_{i-1}; x_i)$ нормально распределенной случайной величины X такова:

$$p(x_{i-1} < X < x_i) = \Phi\left(\frac{x_i - a}{\sigma}\right) - \Phi\left(\frac{x_{i-1} - a}{\sigma}\right),$$

где $\Phi(x)$ – функция Лапласа.

Значения функции Лапласа приведены в приложении 1. Отсюда:

$$p_1(-0{,}15<X<-0{,}13)=\Phi\left(\frac{-0{,}13+0{,}03}{0{,}05}\right)-\Phi\left(\frac{-0{,}15+0{,}03}{0{,}05}\right)=-0{,}477-(-0{,}492)=0{,}015;$$

$$p_2(-0{,}13<X<-0{,}11)=\Phi\left(\frac{-0{,}11+0{,}03}{0{,}05}\right)-\Phi\left(\frac{-0{,}13+0{,}03}{0{,}05}\right)=-0{,}445-(-0{,}477)=0{,}032.$$

Дальнейшие вычисления приведены в табл.6.6.

Из-за того, что значение параметра a случайно совпало с одной из границ, значения вероятностей P_i оказались симметричны относительно интервала (-0,05; -0,01). Два последних интервала [0,09; 0,11) и [0,11; 0,13) объединены ввиду их малочисленности.

Положим $\alpha = 0{,}05$. Число степеней свободы $r = 13 - 2 - 1 = 10$, $\chi^2_{кр} = 18{,}3 > \chi^2_{эксп} = 8{,}06$. Нет оснований отвергнуть выдвинутую нами гипотезу о нормальном законе распределения отклонений диаметра вала от номинального значения.

Таблица 6.6

$[x_{i-1};\ x_i)$	$\left(\frac{x_i-a}{\sigma}\right)$	$\Phi\left(\frac{x_i-a}{\sigma}\right)$	p_i	np_i	n_i	n_i-np_i	$\frac{(n_i-np_i)^2}{np_i}$
–	-2,4	-0,492	–	–	–	–	–
[-0,15;-0,13)	-2	-0,477	0,015	3	3	0	0
[-0,13;-0,11)	-1,6	-0,445	0,032	6,4	8	1,6	0,4
[-0,11; -,09)	-1,2	-0,387	0,058	11,6	11	-0,6	0,03
[0,09;-0,07)	-0,8	-0,288	0,099	19,8	20	0,2	0,002
[-0,07,-0,05)	-0,4	-0,155	0,133	26,6	27	0,4	0,006
[-0,05;-0,03)	0	0,000	0,155	31	36	5	0,81
[-0,03;-0,01)	0,4	0,155	0,155	31	29	-2	0,13
[-0,01;0,01)	0,8	0,288	0,133	26,6	18	-8,6	2,78
[0,01;0,03)	1,2	0,387	0,099	19,8	17	-2,8	0,4
[0,03;0,05)	1,6	0,445	0,058	11,6	17	5,4	2,51
[0,05;0,07)	2	0,477	0,032	6,4	8	1,6	0,4
[0,07;0,09)	2,4	0,492	0,015	3	4	1	0,33
[0,09;0,10	2,8	0,497	0,005	1,0	1		
[0,11;0,13)	3,2	0,499	0,002	0,4	1	0,6	0,26
∑	–	–	0,991	198,2	200	–	8,06

Оценим долю валов, подходящих заказчику. Вероятность того, что диаметр вала соответствует требованиям заказчика равна $p(-0{,}1<X<0{,}1) =$

$$=\Phi\left(\frac{0{,}1+0{,}03}{0{,}05}\right)-\Phi\left(\frac{-0{,}1+0{,}03}{0{,}05}\right)=\Phi(2{,}6)+\Phi(1{,}4)=0{,}495+0{,}419=0{,}914.$$

В среднем около 9 % валов окажутся непригодными для заказчика.

6.3.2. Проверка гипотезы о равномерном законе распределения

В течение 10 часов регистрировали время прибытия машин к бензоколонке (табл. 6.7).

Таблица 6.7

Время прибытия (часы)	[8-9)	[9-10)	[10-11)	[11-12)	[12-13)	[13-14)	[14-15)	[15-16)	[16-17)	[17-18)
n_i	22	30	22	16	28	13	17	20	17	15

При уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что время прибытия машин – случайная величина, имеющая равномерное распределение.

Построим гистограмму. Так как $n = 200$, $h = 1$, то высоты гистограммы таковы:

$h_1 = \frac{22}{200} = 0{,}11$; $h_2 = \frac{30}{200} = 0{,}15$; $h_3 = 0{,}11$; $h_4 = 0{,}08$; $h_5 = 0{,}14$; $h_6 = 0{,}065$; $h_7 = 0{,}085$; $h_8 = 0{,}1$; $h_9 = 0{,}085$; $h_{10} = 0{,}075$.

Гистограмма приведена на рис. 6.4.

Если мы считаем, что время прибытия машин имеет равномерное распределение, мы должны определить два параметра (a и b) равномерного закона. Как известно, функция плотности вероятности $f(x)$ равномерного закона такова:

$$f(x) = \begin{cases} \frac{1}{(b-a)}, a < x < b \\ 0, x \notin (a,b) \end{cases}.$$

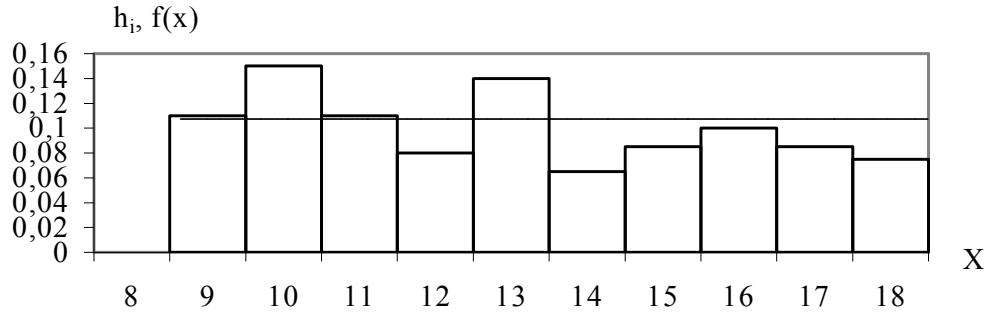


Рис.6.4

При этом $M(x) = \frac{a+b}{2}$; $D(x) = \frac{(b-a)^2}{12}$; $\sigma(x) = \frac{b-a}{2\sqrt{3}}$.

Так что для определения a и b можно записать два уравнения:

$$\begin{cases} \dfrac{a+b}{2} = \overline{x}; \\ \dfrac{b-a}{2\sqrt{3}} = S, \end{cases}$$

откуда $a = \overline{x} - S\sqrt{3}$; $b = \overline{x} + S\sqrt{3}$.

Но мы поступим проще и разумнее. Наша выборка расположена на интервале (8,18), поэтому положим: $a = 8$, $b = 18$, $f(x) = 0{,}1$ ($x \in (8,18)$).

График функции плотности вероятности $f(x)$ также показан на рис.6.4.Все теоретические вероятности p_i одинаковы и равны $\dfrac{h}{(b-a)} = 0{,}1$. Дальнейшие расчеты представлены в табл.6.8.

Таблица 6.8

$[x_{i-1}; x_i)$	p_i	np_i	n_i	$n_i - np_i$	$\dfrac{(n_i - np_i)^2}{np_i}$
[8;9)	0,1	20	22	2	0,2
[9;10)	0,1	20	30	10	5
[10;11)	0,1	20	22	2	0,2
[11;12)	0,1	20	16	-4	0,8
[12,13)	0,1	20	28	8	3,2
[13;14)	0,1	20	13	-7	2,45
[14;15)	0,1	20	17	-3	0,45
[15,16)	0,1	20	20	0	0
[16;17)	0,1	20	17	-3	0,45
[17;18)	0,1	20	15	-5	1,25
–	$\sum p_i = 1$	$\sum np_i = 200$	$\sum n_i = 200$	–	$\chi^2_{эксп} = 14$

Итак, $\chi^2_{эксп} = 14$. Найдем $\chi^2_{кр}$. Мы не определяли по выборке параметров закона - время работы бензоколонки задано заранее. Поэтому число степеней свободы $r = 10 - 1 = 9$. Тогда $\chi^2_{кр} = 16{,}9 > \chi^2_{эксп}$. Выдвинутую гипотезу можно принять.

6.3.3. Проверка гипотезы о биномиальном законе распределения

Семь монет подбрасывались 1536 раз. Каждый раз отмечалось число X выпавших гербов (табл. 6.9).

Таблица 6.9

x_i	0	1	2	3	4	5	6	7
n_i	12	78	270	456	386	252	69	13

При уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что монеты правильные.

Если все монеты правильные, то вероятность выпадения герба для каждой из них равна $p = 0{,}5$. Тогда случайная величина X – число выпавших гербов при бросании семи монет – имеет биномиальное распределение с параметрами $n = 7$ и $p = 0{,}5$. Биномиальное распределение дискретно, поэтому нужно вычислить теоретические вероятности p_i каждого из 8 возможных значений случайной величины X. Эти вероятности считают по формуле Бернулли:

$p(X = 0) = C_7^0 p^0 q^7 = 0{,}5^7 = 0{,}0078$; $p(X = 1) = C_7^1 p^1 q^6 = 7*0{,}5^7 = 0{,}055$;
$p(X = 2) = C_7^2 p^2 q^5 = 21*0{,}5^7 = 0{,}164$; $p(X = 3) = C_7^3 p^3 q^4 = 35*0{,}5^7 = 0{,}273$;
$p(X = 4) = C_7^4 p^4 q^3 = 35*0{,}5^7 = 0{,}273$; $p(X = 5) = C_7^5 p^5 q^2 = 21*0{,}5^7 = 0{,}164$;
$p(X = 6) = C_7^6 p^6 q^1 = 7*0{,}5^7 = 0{,}055$; $p(X = 7) = C_7^7 p^7 q^0 = 0{,}5^7 = 0{,}0078$.

Теперь можно вычислить математические ожидания чисел появлений каждого из значений случайной величины X при 1536 бросаниях семи монет, сравнить их с экспериментальными данными и вычислить $\chi^2_{эксп}$. Результаты сведены в табл. 6.10.

Таблица 6.10

x_i	p_i	np_i	n_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
0	0,0078	12	12	0	0
1	0,055	84	78	-6	0,43
2	0,164	252	270	18	1,29
3	0,273	420	456	36	3,09
4	0,273	420	386	-34	2,75
5	0,164	252	252	0	0
6	0,055	84	69	-15	2,68
7	0,0078	12	13	1	0,08
–	$\sum p_i = 1$	$\sum np_i = 1536$	$\sum n_i = 1536$	–	$\chi^2_{эксп} = 10{,}32$

Найдем $\chi^2_{кр}$. В случае дискретной случайной величины при подсчете r вместо числа интервалов берут число различных значений x_i. В нашем случае $r = 8 - 1 = 7$, так как ни одного параметра по выборке мы не находим. Тогда $\chi^2_{кр} = 14{,}1 > \chi^2_{эксп} = 10{,}32$. Нет оснований опровергнуть гипотезу о правильности монет.

6.3.4. Проверка гипотезы о законе распределения Пуассона

В таблице приведены числа n_i участков равной площади (0,25 км^2) южной части Лондона, на каждый из которых приходилось по x_i попаданий самолетов-снарядов во время второй мировой войны (табл. 6.11).

Таблица 6.11

x_i	0	1	2	3	4	5 и больше
n_i	229	211	93	35	7	1

Всего n = 576 участков. При уровне значимости α = 0,05 проверить гипотезу о том, что случайная величина X – число самолетов-снарядов, попавших на участок, имеет распределение Пуассона.

Вероятность того, что случайная величина X, имеющая распределение Пуассона, примет значение i, равна

$$p(X=i)=\frac{\lambda^i}{i!}e^{-\lambda},$$

где $\lambda > 0$ - параметр закона, $i = 0,1,2, \ldots$.

Оценим значение параметра λ по выборке. Так как $M(X) = \lambda$, то положим $\lambda = \bar{x}$, $\bar{x}=\frac{1}{576}(0\cdot 229+1\cdot 211+2\cdot 93+3\cdot 35+4\cdot 7+5\cdot 1)=0{,}93$.

Положим λ = 0,93. Теперь можно найти вероятности $p_i = p(X = i)$, $i = 0,1,2,3,4,5$.

$$p_0 = p(X=0)=\frac{\lambda^0}{0!}e^{-\lambda}=0{,}395; \qquad p_1 = p(X=1)=\frac{\lambda^1}{1!}e^{-\lambda}=0{,}367;$$

$$p_2 = p(X=2)=\frac{\lambda^2}{2!}e^{-\lambda}=0{,}170; \qquad p_3 = p(X=3)=\frac{\lambda^3}{3!}e^{-\lambda}=0{,}053;$$

$$p_4 = p(X=4)=0{,}012;\ p_5 = p(X\ge 5)=1-p_o-p_1-p_2-p_3-p_4=0{,}003.$$

Остальные вычисления сведены в табл. 6.12.

Таблица 6.12

i	p_i	np_i	n_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
0	0,395	227,5	229	1,5	0,01
1	0,367	211,4	211	-0,4	0,001
2	0,170	97,9	93	-4,9	0,25
3	0,053	30,5	35	4,5	0,66
4	0,012	6,9	7	-0,6	0,04
≥5	0,003	1,7	1		
–	$\sum p_i = 1$	$\sum np_i = 576$	$\sum n_i = 576$	–	$\chi^2_{эксп} = 0{,}96$

Два последних значения n_4 и n_5, np_4 и np_5 объединены, чтобы обеспечить выполнение условия $np_i \ge 5$. Таким образом, осталось 5 разных значений случайной величины: 0, 1, 2, 3 и все, что больше или равно 4. Число степеней свободы равно r = 5 - 1 - 1 = 3, так как по выборке было определено значение параметра λ. Тогда $\chi^2_{кр} = 7{,}8 > \chi^2_{эксп} = 0{,}96$. И в этом случае можно считать справедливой выдвинутую гипотезу.

6.3.5. Последний пример

Согласно закону Геллина, предложенному им в 1855 г., вероятности

рождения двоен, троен и четверней есть соответственно p, p^2, p^3, где p – число, постоянное для данной группы населения. На основании приведенных ниже данных проверить, выполняется ли закон Геллина для многоплодных рождений среди японцев и белого населения США. В табл.6.13 через ν_2, ν_3, ν_4 обозначены относительные частоты рождений двоен, троен и четверней соответственно за указанные периоды.

Таблица 6.13

Годы	Население	Число рождений	ν_2	ν_3	ν_4
1922-1936	Белые США	27939615	0,01129	0,0001088	0,00000177
1926-1931	Японцы	1226106	0,00697	0,0000473	–

Прежде всего оценим по нашим выборкам неизвестные значения p. Положим, что сумма частот $\nu_2 + \nu_3 + \nu_4$ равна сумме $p + p^2 + p^3 = \dfrac{p(1-p^3)}{1-p} \approx \dfrac{p}{1-p}$, так как ясно, что p – очень маленькое число. Для белого населения США имеем:

$$\frac{p}{1-p} = 0{,}01129 + 0{,}0001088 + 0{,}00000177 = 0{,}01140057 \approx 0{,}0114;$$

$$p = \frac{0{,}0114}{1+0{,}0114} \approx 0{,}0113;\ p^2 \approx 0{,}000128;\ p^3 \approx 0{,}000001.$$

Теперь можно воспользоваться критерием χ^2. Нужно определить, извлечена ли выборка из генеральной совокупности X, имеющей такой закон распределения (табл. 6.14).

Таблица 6.14

x_i	1	2	3	4
p_i	$1- p - p^2 - p^3$	p	p^2	p^3

Здесь $p = 0{,}0113$.

Все вычисления сведем в табл. 6.15. Частоты n_1, n_2, n_3, n_4 равны соответственно:

$n_1 = n\nu_1 = 27939615 * (1 - \nu_2 - \nu_3 - \nu_4) = 27621087{,}5$;

$n_2 = n\nu_2 = 27939615 * 0{,}01129 = 315438{,}25$; $n_3 = n\nu_3 = 3039{,}8$; $n_4 = n\nu_4 = 49{,}45$.

Таблица 6.15

x_i	p_i	np_i	n_i	$n_i - np_i$	$\dfrac{(n_i - np_i)^2}{np_i}$
1	0,988571	27620293	27621088	795	0,02
2	0,0113	315717	315438	-279	0,25
3	0,000128	3576	3040	-536	80,34
4	0,000001	28	49	–	15,75
–	$\sum p_i = 1$	$\sum np_i = 27939615$	$\sum n_i = 27939615$	–	$\chi^2_{эксп} = 96{,}4$

Число степеней свободы r равно $r = 4 - 1 - 1 = 2$, $\chi^2_{кр} = 6{,}0 << \chi^2_{эксп}$. Расхождение велико, предложенный закон должен быть отвергнут.

Проделаем те же вычисления в случае с японцами.

$v_2 + v_3 + v_4 \approx 0{,}00702$.

Тогда $p = \dfrac{0{,}0070}{1 + 0{,}0070} \approx 0{,}00697$; $p^2 = 0{,}0000486$; $p^3 = 0{,}00000034$;

$n_1 = nv_1 = 1226106*(1 - v_2 - v_3 - v_4) = 1217502$; $n_2 = nv_2 = 8545{,}96$;

$n_3 = nv_3 = 57{,}99$; $n_4 = nv_4 = 0$.

Найдем $\chi^2_{эксп}$ (табл. 6.16).

Таблица 6.16

x_i	p_i	np_i	n_i	n_i-np_i	$\dfrac{(n_i - np_i)^2}{np_i}$
1	0,993	1217502	1217502	0	0
2	0,007	8544	9545,96	1,96	0
3	0,0000486	$\begin{cases}59{,}54\\0{,}41\end{cases}$	$\begin{cases}57{,}99\\0\end{cases}$	-1,96	0,06
4	0,00000034				
–	$\sum p_i = 1$	$\sum np_i = 27939615$	$\sum n_i = 27939615$	–	$\chi^2_{эксп} = 0{,}06$

$\chi^2_{кр} = 3{,}8 > \chi^2_{эксп} = 0{,}06$. В этом случае гипотеза не отвергается.

6.4. ЗАДАЧИ

Во всех задачах на проверку гипотезы о законе распределения генеральной совокупности принять уровень значимости $\alpha = 0{,}05$, если не оговорено противное.

1. 100 раз подбрасывались 4 монеты. Каждый раз отмечалось число x_i выпавших цифр:

x_i	0	1	2	3	4
n_i	8	20	42	22	8

Можно ли считать, что случайная величина X – число выпавших цифр при бросании 4-х монет – имеет биномиальное распределение?

2. В библиотеке случайно отобрано 200 выборок по 5 книг в каждой. Регистрировалось число поврежденных книг (подчеркивания, помарки, вырванные страницы и т.п.):

x_i	0	1	2	3	4	5
n_i	1	2	72	77	34	14

Проверить гипотезу о том, что случайная величина X – число поврежденных книг в выборке из 5 книг – имеет биномиальное

распределение.

3. На некотором заводе были обследованы рабочие, получившие на производстве незначительные увечья. За 52 недели результаты оказались такими:

Число рабочих, получивших увечья за неделю (x_i)	0	1	2	3
Число недель, в течение которых увечья получили x_i рабочих	31	17	3	1

Можно ли эти данные аппроксимировать законом распределения Пуассона?

4. Было проверено 500 одинаковых контейнеров со стеклянными изделиями. В каждом контейнере нашли число поврежденных изделий:

x_i	0	1	2	3	4	5	6	7
n_i	199	169	87	31	9	3	1	1

Можно ли утверждать, что случайная величина X – число поврежденных изделий в контейнере – имеет распределение Пуассона?

5. Ниже приводятся ставшие классическими данные Борткевича о числе лиц, убитых ударом копыта в 10 прусских армейских корпусах за 20 лет (1875-1894):

Число смертей в одном корпусе за год (i)	0	1	2	3	4
Число случаев, когда произошло i смертей	109	65	22	3	1

Проверить гипотезу о том, что число смертей в одном корпусе за год подчиняется закону Пуассона.

6. По данным шведской статистики, в Швеции в 1935 г. родилось 88273 ребенка, причем распределение рождений по месяцам таково:

Месяц	Январь	Февраль	Март	Апрель	Май	Июнь
Число рождений в этом месяце	7280	6957	7883	7884	7892	7609
Месяц	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
Число рождений в этом месяце	7585	7393	7203	6903	6552	7132

Совместимы ли эти данные с гипотезой о том, что день рождения наудачу выбранного человека с равной вероятностью приходится на любой из 365 дней года?

7. Ниже приводятся результаты опыта с подбрасыванием костей. Количество граней с 6 очками при 4096 подбрасываниях 12 костей:

Число выпадений 6 очков	0	1	2	3	4	5	6	7 и более
n_i	447	1145	1181	796	380	115	24	8

Проверить гипотезу о правильности костей.

В задачах 8 - 16 проверить по критерию Пирсона одну из трех гипотез о законе распределения генеральной совокупности: равномерном, нормальном или показательном законе.

8. Регистрировалось время прихода 800 посетителей выставки (начало отсчета – момент открытия выставки). Результаты указаны в таблице; в первой строке – интервалы времени, во второй – количество посетителей, пришедших в течение данного интервала времени:

$[x_{i-1}; x_i)$	[0-1)	[1-2)	[2-3)	[3-4)	[4-5)	[5-6)	[6-7)	[7-8)
n_i	368	212	109	51	23	18	13	6

9. Результаты обследования роста 1000 человек:

Рост, см	n_i	Рост, см	n_i	Рост, см	n_i
(143 -146)	1	[158-161)	120	[173-176)	64
[146-149)	2	[161 -164)	181	[176 -179)	28
[149- 152)	8	[164 -167)	201	[179 -182)	10
[152-155)	26	[167-170)	170	[182-185)	3
[155-158)	65	[170-173)	120	[185-188)	1

10. Результаты испытаний прочности партии стальной проволоки диаметром 1,4 мм:

Предел прочности, кг/мм^2	Число мотков проволоки	Предел прочности, кг/мм^2	Число мотков проволоки
[45 -150)	10	[165 -170)	12
[150 155)	24	[170-175)	7
[155 –160)	28	[175 -180)	5
[160-165)	22		

11. Результаты взвешивания 800 стальных шариков:

Масса, граммы	Частота	Масса, граммы	Частота
[20,0-20,5)	91	[22,5-23,0)	83
[20,5-21,0)	76	[23,0-23,5)	79
[21,0-21,5)	75	[23,5-24,0)	73
[21,5-22,0)	74	[24,0-24,5)	80
[22,0-22,5)	92	[24,5-25,0)	77

1.4.12. При изготовлении стального листа для автомобильных корпусов некоторые места, подверженные ржавчине и коррозии, следует гальванизировать, т.е. обычный стальной лист целиком покрыть тонким ровным слоем цинка. Заказчику необходимо найти металлургический завод, который имеет возможность провести гальванизацию таким образом, чтобы плотность слоя покрытия была не меньше 91,5 г/м^2. На

одном заводе собраны следующие данные о цинковом покрытии стальных листов:

Плотность покрытия, г/м^2	Число стальных листов	Плотность покрытия, г/м^2	Число стальных листов
[84-99)	4	[144 -159)	10
[99-114)	10	[159-174)	4
[114-129)	18	[174-189)	1
[129-144)	18	[189-204)	1

Оценить долю листов, которая не будет удовлетворять требованиям заказчика.

13. Результаты наблюдения за среднесуточной температурой воздуха в течение 320 суток:

Температура воздуха, ° С	Частота	Температура воздуха, ° С	Частота
[- 40…-30)	5	[0…20)	81
[-30…-20)	11	[20…30)	36
[-20…-10)	25	[30…40)	20
[-10…0)	42	[40…50)	8
[0…10)	88	[50…60)	4

14. Результаты испытаний 1000 элементов на время безотказной работы (часы):

Время работы	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)
Частота	365	245	150	100	70	45	25

Положить $\alpha = 0{,}01$.

15. Цифры 0,1,2,…,9 среди 800 первых десятичных знаков числа π появились 74, 92, 83, 79, 80, 73, 77, 75, 76, 91 раз соответственно. Согласуются ли эти данные с утверждением, что цифры в десятичном представлении числа π распределены равномерно?

16. Для проверки точности хода специальных маятниковых часов в выбранные наудачу моменты времени фиксировались углы отклонения оси маятника от вертикали. Амплитуда колебаний поддерживалась равной A = 15°. Результаты 1000 таких измерений, разбитые на интервалы в 3°, приведены в таблице.

Середина интервала	-13,5	-10,5	-7,5	-4,5	-1,5	1,5	4,5	7,5	10,5	13,5
Частота	188	88	64	86	62	74	76	81	100	181

Проверить гипотезу о согласии наблюдений с законом распределения арксинуса. Функция плотности этого закона имеет вид

$$f(x) = \frac{1}{\pi\sqrt{a^2 - x^2}}; \quad -a < x < a.$$

7.ПОНЯТИЕ О ТОЧЕЧНЫХ И ИНТЕРВАЛЬНЫХ ОЦЕНКАХ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Много лет размышлял я над жизнью земной,
Непонятного нет для меня под луной,
Мне известно, что мне ничего не известно! -
Вот последняя правда, открытая мной.
О. Хайям (перевод Г. Плисецкого*)*

7.1. ВЫБОРОЧНЫЕ СТАТИСТИКИ

Выборочной статистикой называется произвольная числовая функция $f(x_1, x_2, \ldots, x_n)$, вычисляемая для значений $x_1, x_2, \ldots, x_n$, образующих выборку. Если вместо чисел $x_1, x_2, \ldots, x_n$ мы рассмотрим случайные величины $X_1, X_2, \ldots, X_n$, независимые и одинаково распределённые, как генеральная совокупность X, то получим случайную величину $f(X_1, X_2, \ldots, X_n)$, которая также называется выборочной статистикой или просто статистикой. В математической статистике случайные величины и их значения часто обозначаются одними и теми же маленькими буквами.

Рассмотрим два примера.

Пример 1. Выборочное среднее $\bar{x}$.

Случайную величину $\bar{x}$ – выборочное среднее – определяют по формуле

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Найдём математическое ожидание и дисперсию случайной величины $\bar{x}$. Обозначим математическое ожидание и дисперсию генеральной совокупности X через a и σ^2 соответственно. Таким образом, $M(X)=a$, $D(X)=\sigma^2$.

$$M(\bar{x}) = M(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n}\sum_{i=1}^{n} a = a.$$

Математическое ожидание выборочного среднего равно математическому ожиданию генеральной совокупности.

$$D(\bar{x}) = D(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n};$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Дисперсия выборочного среднего в n раз меньше дисперсии генеральной совокупности. При вычислении математического ожидания и

дисперсии мы воспользовались известными свойствами этих числовых характеристик.

Пример 2. Выборочная дисперсия S^2.

Случайная величина S^2 задаётся формулой

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2$$

Найдём математическое ожидание этой случайной величины:

$$M(S^2) = M(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2) = \frac{1}{n}\sum_{i=1}^{n} M(X_i^2) - M(\bar{x}^2) = M(X^2) - M(\bar{x}^2) +$$

$$+(a^2 - a^2) = D(X) - D(\bar{x}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Математическое ожидание случайной величины S^2 не равно дисперсии генеральной совокупности X. Чтобы получить равенство, рассматривают другую случайную величину. Она обозначается $\tilde{S}^2$, называется исправленной выборочной дисперсией и связана с выборочной дисперсией S^2 формулой

$$\tilde{S}^2 = \frac{n}{n-1} S^2.$$

$$M(\tilde{S}^2) = \frac{n}{n-1} M(S^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2.$$

7.2. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Допустим, что нам известен закон распределения генеральной совокупности. Но каждый закон распределения зависит от нескольких параметров. Например, плотность вероятности нормального закона зависит от параметров a и σ.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{-(x-a)^2}{2\sigma^2}}, \; a = M(x), \; \sigma^2 = D(x), \; -\infty < x < \infty.$$

В формулу для показательного закона входит параметр λ:

$$f(x) = \lambda \cdot e^{-\lambda x}, \; x \geq 0, \; \lambda = \frac{1}{M(x)}.$$

Равномерный закон распределения зависит от параметров a и b:

$$f(x) = \frac{1}{b-a}, \; a \leq x \leq b; \; M(x) = \frac{a+b}{2}, \; D(x) = \frac{(b-a)^2}{12}.$$

Вероятности значений, которые принимают дискретные случайные величины, также зависят от параметров. Например, вероятности значений

случайной величины, распределённой по закону Пуассона, зависят от параметра λ:

$$P(x=k)=\frac{\lambda^k}{k!}\cdot e^{-\lambda},\ k=0,1,2,\ldots,\ M(x)=\lambda.$$

Вероятность "успеха" p и число независимых экспериментов n – параметры биномиального закона распределения.

Вероятность $P_{n,k}$ того, что случайная величина, распределённая по биномиальному закону, примет значение, равное k, считается по формуле

$$P_{n,k}=C_n^k p^k (1-p)^{n-k},\ k=0,1,2,\ldots,n;\ M(x)=np;\ D(x)=npq.$$

Если закон распределения генеральной совокупности известен, а значения параметров, от которых этот закон зависит, неизвестны, возникает задача оценки значений этих параметров по имеющимся значениям $x_1, x_2, \ldots, x_n$ извлечённой из генеральной совокупности выборки. Точечные оценки параметров – это оценки с помощью числовых значений подходящих статистик. При этом можно оценивать не только параметры, непосредственно входящие в формулу для закона распределения, но и числовые характеристики генеральной совокупности – математическое ожидание, дисперсию, коэффициент корреляции и т. п. К точечным оценкам предъявляют три следующих требования:

1. Если статистика $f(x_1,x_2,\ldots,x_n)$ - оценка параметра a, то при $n\to\infty$ она должна сходиться по вероятности к числу a, т.е. $\forall\varepsilon>0$:

$$\lim_{n->\infty} P(|f(x_1,x_2,\ldots,x_n)-a|<\varepsilon)=1.$$

Такая оценка называется состоятельной.

2. Математическое ожидание статистики $f(x_1,x_2,\ldots,x_n)$ должно равняться числу a: $M(f(x_1,x_2,\ldots,x_n))=a$.

Такая оценка называется несмещённой.

3. Значения случайной величины $f(x_1,x_2,\ldots,x_n)$ должны быть достаточно близкими, другими словами, статистика $f(x_1,x_2,\ldots,x_n)$ должна иметь маленькую дисперсию.

Оценка, обладающая минимальной дисперсией, называется эффективной.

В качестве примера рассмотрим точечные оценки математического ожидания a и дисперсии σ^2 генеральной совокупности – выборочное среднее $\bar{x}$, выборочную дисперсию S^2, исправленную выборочную дисперсию $\tilde{S}^2$. Состоятельность выборочного среднего

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

вытекает из закона больших чисел.

Ранее мы показали, что $M(\bar{x}) = a$, поэтому $\bar{x}$ - несмещённая оценка математического ожидания *a.* Мы показали также, что

$$D(\bar{x}) = \frac{\sigma^2}{n}.$$

Можно доказать, что если генеральная совокупность X имеет нормальное распределение, то $\bar{x}$ - эффективная оценка параметра a.

Состоятельность статистики S^2 как оценки дисперсии

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

также вытекает из закона больших чисел. Но оценка S^2 – смещённая оценка, ведь

$$M(S^2) = \frac{n}{n-1}\sigma^2.$$

Несмещенной оценкой дисперсии σ^2 является исправленная выборочная дисперсия $\tilde{S}^2$.

Чтобы оценить некоторый параметр закона распределения генеральной совокупности X, нужно выразить его через теоретические моменты (математическое ожидание, дисперсию и т.п.), а затем подставить в полученную формулу значения соответствующих выборочных статистик (выборочного среднего, выборочной дисперсии и т. д.).

Например, оценками параметров a и σ^2 нормального распределения служат статистики $\bar{x}$ и $\tilde{S}^2$, так как $M(\bar{x})=a$, $D(\bar{x})=\sigma^2$. Оценкой параметров λ показательного закона является статистика $(1/\bar{x})$, т.к. $M(\bar{x})=1/\lambda$.Оценкой параметра λ закона Пуассона является статистика $\bar{x}$, т.к. $M(\bar{x})=\lambda$.

Такой метод оценки параметров называется методом моментов. Мы фактически пользовались им в процедуре проверки гипотезы о законе распределения генеральной совокупности по критерию Пирсона.

В заключение отметим, что поправочный множитель $n/(n-1)$, вводимый для статистики S^2, при больших n практически равен 1 и его нет смысла использовать.

7.3. О ТОЧНОСТИ И НАДЁЖНОСТИ ТОЧЕЧНЫХ ОЦЕНОК

Рассмотрим здесь только случай оценки математического ожидания a генеральной совокупности значением $\bar{x}$. Заменяя неизвестное значение a

числом $\bar{x}$, мы совершаем ошибку. Тогда случайная величина $|\bar{x}-a|$ – абсолютное значение ошибки. Если известен закон распределения случайной величины $\bar{x}$, можно найти вероятность

$$P(|\bar{x}-a|<\varepsilon)=\beta_\varepsilon$$

Число ε характеризует точность оценки, вероятность β_ε - её надёжность. Если для небольших ε вероятность β_ε достаточно велика, число $\bar{x}$ можно считать точной и надёжной оценкой математического ожидания *a*.

Когда генеральная совокупность имеет нормальное распределение, случайная величина $\bar{x}$ распределена нормально (сумма независимых нормально распределённых случайных величин). Если закон распределения генеральной совокупности отличен от нормального, но число *n* достаточно велико, случайную величину $\bar{x}$ можно считать приблизительно нормально распределённой в силу центральной предельной теоремы. Числовые характеристики $\bar{x}$ известны:

$$M(\bar{x})=a,\ D(\bar{x})=\frac{\sigma^2}{n},\ \sigma(\bar{x})=\frac{\sigma}{\sqrt{n}}.$$

Если дисперсия генеральной совокупности неизвестна, заменим её на значение исправленной выборочной дисперсии $\widetilde{S}^2$. Применяя известную формулу для нормального закона, получим

$$P=(|\bar{x}-a|<\varepsilon)=2\Phi\left(\frac{\varepsilon}{\sigma(\bar{x})}\right)=2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\approx 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\widetilde{S}}\right).$$

Пример. Из генеральной совокупности извлечена выборка объёма *n*=47. Найденное по выборке значение $\widetilde{S}$=2,35. Какова вероятность того, что точность ε оценки математического ожидания *a* генеральной совокупности не больше 0,3?

Решение. Нужно найти вероятность события

$$P\left(|\bar{x}-a|<0{,}3\right).$$

Имеем

$$P\left(|\bar{x}-a|<0{,}3\right)\approx 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\widetilde{S}}\right)=2\Phi\left(\frac{0{,}3\cdot\sqrt{47}}{2{,}35}\right)=2\Phi(0{,}875)\approx 2\cdot 0{,}31=0{,}62.$$

Найденная вероятность достаточно мала. Для того чтобы получить большую надёжность (при той же точности) или большую точность (при той же надёжности), нужно увеличить число *n*.

Пример. Каков должен быть минимальный объём выборки *n* для того, чтобы с надёжностью 0,98 точность оценки математического ожидания *a* с

помощью выборочного среднего $\bar{x}$ была 0,2, если среднее квадратичное отклонение σ генеральной совокупности равно 1,5?

Решение. Число n определяется из условия

$$P\left(\left|\bar{x}-a\right|<0{,}2\right)=0{,}98,$$

или

$$0{,}98=2\Phi\left(\frac{0{,}2\sqrt{n}}{\sigma}\right)=2\Phi\left(\frac{0{,}2\sqrt{n}}{1{,}5}\right).$$

Тогда

$$\Phi\left(\frac{0{,}2\sqrt{n}}{1{,}5}\right)=0{,}49\Rightarrow\frac{2\sqrt{n}}{15}=2{,}34\Rightarrow n\geq\left(\frac{2{,}34\cdot 15}{2}\right)^2=308.$$

Пример. Как изменится точность математического ожидания a из предыдущего примера, если объём выборки увеличить до 500, а надёжность оставить равной 0,98?

Решение. Из условия

$$P\left(\left|\bar{x}-a\right|<\varepsilon\right)=0{,}98$$

получаем, что

$$0{,}98=2\Phi\left(\frac{\varepsilon\sqrt{500}}{1{,}5}\right)\Rightarrow\frac{\varepsilon\sqrt{500}}{1{,}5}=2{,}34\Rightarrow\varepsilon=0{,}157.$$

Пример. Оценка вероятности p “успеха”.

Пусть проведено n независимых испытаний, в каждом из которых вероятность события A (“успеха”) равна p (следовательно, вероятность не-появления события A равна $q=1-p$). Если в n независимых испытаниях событие A появилось k раз, то насколько точно число k/n оценивает вероятность p?

Решение. В этом случае генеральная совокупность X имеет следующий закон распределения:

X_i	0	1
p_i	q	p

Случайная величина X равна единице, если событие A произошло, и равна нулю в противном случае.

$$M(X)=p,\ D(X)=pq,\ \sigma(X)=\sqrt{pq}.$$

Если число n достаточно велико, то случайная величина $\bar{x}=k/n$ - выборочное среднее - имеет приближённое нормальное распределение с параметрами:

$$M\left(\bar{x}\right)=p,\ D\left(\bar{x}\right)=\frac{pq}{n},\ \sigma\left(\bar{x}\right)=\sqrt{\frac{pq}{n}}.$$

Тогда точность и надёжность оценки числа p числом k/n определяют из равенства

$$P\left(\left|\bar{x}-p\right|<\varepsilon\right)=2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right)\approx 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\bar{x}(1-\bar{x})}}\right),$$

так как

$$S^2=\frac{1}{n}\left[0^2\cdot(n-k)+1^2\cdot k\right]-\bar{x}^2=\frac{k}{n}-\frac{k^2}{n^2}=\frac{k}{n}\left(1-\frac{k}{n}\right)=\bar{x}\left(1-\bar{x}\right),$$

$$S=\sqrt{\bar{x}(1-\bar{x})}.$$

Пример. Из большой партии некоторых изделий отобрано наугад для контроля 500 штук, причём 20 штук оказались бракованными. Найти вероятность того, что, приняв вероятность p изделию быть бракованным, равной 0,04, мы совершаем ошибку, не превосходящую 0,01. Сколько нужно отобрать изделий, чтобы с вероятностью 0,95 была совершена ошибка, не превосходящая 0,01?

Решение.

1. Здесь $\bar{x}=\frac{20}{500}=0{,}04;\quad n=500;\quad S^2=\bar{x}\left(1-\bar{x}\right)=0{,}04\cdot 0{,}96=0{,}0384;$

$S=\sqrt{0{,}0384}\approx 0{,}2;\quad \varepsilon=0{,}01.$

Положим $\sigma\left(\bar{x}\right)\approx\frac{S}{\sqrt{n}}=0{,}009.$

$$P\left(\left|\bar{x}-p\right|<0{,}01\right)\approx 2\Phi\left(\frac{0{,}01\cdot\sqrt{500}}{0{,}2}\right)=2\Phi\left(1{,}14\right)=0{,}746.$$

2. Здесь $\bar{x}=0{,}04;\ S=0{,}2;\ \varepsilon=0{,}01;\ \beta=0{,}95.$ Требуется найти объём выборки n.

$$\frac{\beta}{2}=0{,}475=\Phi\left(\frac{0{,}01\cdot\sqrt{n}}{0{,}2}\right).$$

По таблице функции Лапласа, зная её значение, равное числу 0,475, определяем аргумент

$$t=\frac{0{,}01\sqrt{n}}{0{,}2}=1{,}96,\ \text{отсюда}\ n\geq\left(\frac{1{,}96\cdot 0{,}2}{0{,}01}\right)^2=1537.$$

7.3.1. Ещё об определении нужного объёма выборки

Пользуясь формулой

$$P\left(\left|\bar{x}-a\right|<\varepsilon\right)=\beta=2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\approx 2\Phi\left(\frac{\varepsilon\sqrt{n}}{S}\right),$$

можно поставить три несложные задачи. Примеры решения этих задач уже были подробно разобраны выше. Здесь мы просто подведём итоги.

Задача 1. Зная объём выборки n и точность оценки математического ожидания a, найти надёжность β этой оценки:

$$\beta=2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\approx 2\Phi\left(\frac{\varepsilon\sqrt{n}}{S}\right).$$

В случае выборки из биномиально распределённой совокупности

$$\beta=2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\bar{x}\left(1-\bar{x}\right)}}\right).$$

Задача 2. Зная объём выборки n и надёжность β, оценить точность ε оценки математического ожидания a:

$$\varepsilon=\frac{t\sigma}{\sqrt{n}}\approx\frac{tS}{\sqrt{n}},$$

где число t определяется из таблицы функции Лапласа из условия

$$\Phi(t)=\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\frac{\beta}{2}.$$

В случае выборки из биномиально распределённой генеральной совокупности

$$\varepsilon\approx\frac{t\sqrt{\bar{x}\left(1-\bar{x}\right)}}{\sqrt{n}}.$$

Задача 3. Зная точность ε и надёжность β оценки математического ожидания a, определить минимальный объём n выборки, обеспечивающий заданные точность и надёжность:

$$n\geq\frac{t^2\sigma^2}{\varepsilon^2}\approx\frac{t^2S^2}{\varepsilon^2}.$$

В случае выборки из биномиально распределённой генеральной совокупности

$$n\geq\frac{t^2\bar{x}\left(1-\bar{x}\right)}{\varepsilon^2}.$$

Все эти формулы были выведены в предположении, что

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

где $X_1, X_2, ..., X_n$ - независимые и одинаково распределённые (как генеральная совокупность X) случайные величины. Такое предположение не всегда можно принять. На практике генеральная совокупность X - это N объектов, из которых отбирают для исследования n объектов. Если выборка бесповторная (один раз отобранный объект не возвращается назад), то дисперсия выборочного среднего $\bar{x}$ в общем случае зависит от чисел N и n.

Пусть генеральная совокупность X состоит из N чисел $x_1, x_2, ..., x_N$. Если вероятность выбора каждого числа (при извлечении одного числа) равна $1/N$, то математическое ожидание a и дисперсия σ^2 случайной величины x - выбранного числа - равны соответственно:

$$a = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - a^2.$$

Пусть теперь $\bar{x}$ - это среднее арифметическое n наудачу отобранных чисел, причём отбор бесповоротный. Нетрудно показать, что в этом случае

$$M(\bar{x}) = a; \; D(\bar{x}) = \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} \approx \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right); \; \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}.$$

На практике величину $\sqrt{1-n/N}$ заменяют на 1, если $n/N<0{,}05$, и рассчитывают в противном случае. Для отыскания необходимого объёма выборки, обеспечивающего заданную точность и надёжность, имеем:

$$\frac{\varepsilon\sqrt{n}}{\sigma\sqrt{1-\frac{n}{N}}} \geq t, \qquad n \geq \frac{\sigma^2 t^2 N}{\varepsilon^2 N + \sigma^2 t^2} \approx \frac{S^2 t^2 N}{\varepsilon^2 N + S^2 t^2}.$$

Если среди чисел $x_1, x_2, ..., x_N$ встречаются только нули и единицы, то

$$n \geq \frac{\bar{x}(1-\bar{x})t^2 N}{\varepsilon^2 N + \bar{x}(1-\bar{x})t^2}.$$

Напомним, что пользоваться указанными формулами можно, только если закон распределения выборочного среднего $\bar{x}$ можно хотя бы приближённо считать нормальным. Так почти всегда получается, когда $n>30$.

Пример. Фермер хочет оценить среднюю массу своих 5000 индеек с точностью до 0,5 фунта, чтобы как можно точнее определить доход от продажи этих индеек. Отобрав случайным образом 20 индеек, он нашёл, что их средняя масса составляет 9,25 фунта, а $\tilde{S}$=4,39 фунта. Теперь он в состоянии определить минимальный объём выборки n, позволяющий оценить среднюю массу индейки с точностью ε=0,5 и надёжностью β=0,95:

$$n \geq \frac{S^2 t^2 N}{\varepsilon^2 N + S^2 t^2} = \frac{4{,}39^2 \cdot 1{,}96^2 \cdot 5000}{0{,}5^2 \cdot 5000 + 4{,}39^2 \cdot 1{,}96^2} = 280 \text{ (штук)}.$$

Предварительная малая выборка потребовалась, чтобы оценить σ, иначе дальнейшие вычисления невозможны.

7.4. ПОНЯТИЕ ОБ ИНТЕРВАЛЬНЫХ ОЦЕНКАХ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Ещё один способ оценить известное значение параметра - указать интервал $(\varepsilon_1, \varepsilon_2)$ на числовой оси, про который известно, что он содержит это неизвестное значение с достаточно большой вероятностью $\beta, P(\varepsilon_1 < a < \varepsilon_2) = \beta.$

Вероятность β называется **доверительной вероятностью**, а интервал $(\varepsilon_1, \varepsilon_2)$ - **доверительным интервалом**. Для построения доверительных интервалов используют подходящим образом подобранные выборочные статистики.

7.4.1. Построение доверительного интервала для неизвестного математического ожидания *a* нормально распределённой генеральной совокупности, когда дисперсия σ^2 генеральной совокупности известна

Рассмотрим случайную величину $\bar{x}$ - выборочное среднее:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Так как генеральная совокупность X распределена по нормальному закону, $\bar{x}$ тоже имеет нормальное распределение, $M(\bar{x}) = a$, $D(\bar{x}) = \sigma^2/n$. График функции плотности вероятности случайной величины $\bar{x}$ симметричен относительно оси $x=a$.

Рассмотрим интервал $|\bar{x} - a| < \varepsilon$, или $a - \varepsilon < \bar{x} < a + \varepsilon$. Ширину 2ε этого интервала определим из условия

$$P(|\bar{x} - a| < \varepsilon) = \beta,$$

где β - заданная доверительная вероятность.

Как мы уже знаем, $\varepsilon = \frac{t\sigma}{\sqrt{n}}$, где число t находится по таблице функции Лапласа из условия $\Phi(t) = \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = \frac{\beta}{2}$.

Доверительный интервал $(\bar{x}-t\sigma/\sqrt{n},\ \bar{x}+t\sigma/\sqrt{n})$ содержит число a с вероятностью β.

Приведём значения t для наиболее часто встречающихся значений вероятности β.

β	0,9	0,95	0,99	0,9973	0,999
t	1,64	1,96	2,58	3,00	3,37

Пример. У ста случайно отобранных двадцатилетних юношей измерили рост. Оказалось, что средний рост $\bar{x}$=1,73 м, а исправленная выборочная дисперсия $\widetilde{S}^2$=0,00245 м^2. Построить доверительный интервал среднего всей совокупности, если β=0,99.

Решение. Подразумевается, что случайная величина X - рост юноши - имеет нормальное распределение с неизвестными нам параметрами a и σ Так как объём выборки n велик, то вместо неизвестного значения σ можно взять значение $\widetilde{S}=\sqrt{0{,}00245}=0{,}0495.$ Если β=0,99 => t=2,58. Отсюда:

$$\varepsilon=\frac{t\widetilde{S}}{\sqrt{n}}=\frac{2{,}58\cdot 0{,}0495}{10}\approx 0{,}013;\quad 1{,}717<a<1{,}743.$$

Подчеркнём, что условие $P(\varepsilon_1<a<\varepsilon_2)=\beta$ следует интерпретировать следующим образом. Случайный интервал $(\varepsilon_1,\varepsilon_2)$ со случайными границами $\varepsilon_1,\varepsilon_2$ с вероятностью β содержит неслучайное число a.

7.4.2. Построение доверительного интервала для неизвестной вероятности *p* “успеха”

Пусть в серии из n независимых испытаний “успех” произошел k раз. Требуется построить доверительный интервал, содержащий значения вероятности p появления “успеха” в каждом испытании с данной вероятностью β.

В этом случае генеральная совокупность X распределена по закону

x_i	0	1
p_i	$q=1-p$	p

$M(X)=p,\ D(X)=pq.$

Выборка, соответствующая k появлениям “успеха” в n независимых испытаниях, имеет вид

x_i	0	1
n_i	$n-k$	k

Выборочное среднее $\bar{x}=k/n$ можно считать приближённо нормально распределённым только в случае большего числа испытаний. Но если это условие выполнено, можно воспользоваться таблицей функции Лапласа,

подставив в формулу для ε вместо неизвестного среднего квадратического отклонения σ генеральной совокупности исправленное выборочное среднее квадратическое отклонение $\widetilde{S}$:

$$\widetilde{S} = \sqrt{\widetilde{S}^2} = \sqrt{\frac{n}{n-1} S^2} = \sqrt{\frac{n}{n-1}\left(\overline{x} - \overline{x}^2\right)}.$$

Причём если *n* велико, то подправлять *S* нет смысла.

Окончательно ε определяется по формуле

$$\varepsilon = \frac{tS}{\sqrt{n}} = \frac{t\sqrt{\overline{x}\left(1-\overline{x}\right)}}{\sqrt{n}},$$

где число *t* берётся из таблицы функции Лапласа из условия $\Phi(t) = \beta/2$.

Пример. В 100 бросаниях монеты герб выпал 64 раза. Построить доверительные интервалы для вероятности *p* выпадения герба в одном бросании с доверительными вероятностями $\beta_1 = 0{,}9;\ \beta_2 = 0{,}95;\ \beta_3 = 0{,}99$.

Можно ли считать монету правильной?

Решение. Здесь n=100; k=64; t_1=1,64; t_2=1,96; t_3=2,58; выборочное среднее $\overline{x} = 0{,}64$; выборочная дисперсия $S^2 = \overline{x}\left(1-\overline{x}\right) = 0{,}64 \cdot 0{,}36 = 0{,}2304$; выборочное среднее квадратическое отклонение $S = \sqrt{S^2} = 0{,}48$. Число опытов *n* велико, поправкой для дисперсии можно пренебречь.

$$\varepsilon_1 = \frac{t_1 S}{\sqrt{n}} = \frac{1{,}64 \cdot 0{,}48}{10} \approx 0{,}079,\ \ \varepsilon_2 = \frac{t_2 S}{\sqrt{n}} = \frac{1{,}96 \cdot 0{,}48}{10} \approx 0{,}094,$$

$$\varepsilon_3 = \frac{t_3 S}{\sqrt{n}} = \frac{2{,}58 \cdot 0{,}48}{10} \approx 0{,}124.$$

Границы доверительных интервалов для вероятности *p* таковы:

0,561<*p*<0,719, если β=0,9; 0,546<*p*<0,734, если β=0,95;

0,516<*p*<0,764, если β=0,95.

Ни один из этих интервалов не содержит числа 0,5. Монету следует признать неправильной, вероятность *p* выпадения герба больше 0,5.

7.4.3. Построение доверительного интервала для неизвестного математического ожидания нормально распределённой генеральной совокупности, когда дисперсия σ^2 генеральной совокупности неизвестна

Как уже было сказано выше, когда объём выборки n>30, при построении доверительного интервала для *a* можно пользоваться нормальным распределением, подставляя в формулу для ширины

интервала ε вместо неизвестного значения σ число $\widetilde{S}$, определяемое по выборке. Рассмотрим случай малых *n*.

В этой ситуации пользуются случайной величиной *T*, определяемой формулой

$$T = \frac{\bar{x} - a}{S}\sqrt{n-1} = \frac{\bar{x} - a}{\widetilde{S}}\sqrt{n}.$$

Подчеркнём, что $\bar{x}$ и *S* - это случайные величины, а *n* и *a* - числа.

Случайная величина *T* распределена по закону Стьюдента (Стьюдент - псевдоним английского статистика В. Госсета (1876-1937), одного из создателей теории проверки статистических гипотез).

График функции плотности вероятности случайной величины, распределённой по закону Стьюдента, симметричен относительно оси ординат. Функция плотности *f(t, r)* зависит от одного параметра *r,* который называется числом степеней свободы.

Случайная величина

$$T = \frac{\bar{x} - a}{S}\sqrt{n-1} = \frac{\bar{x} - a}{\widetilde{S}}\sqrt{n}$$

имеет число степеней свободы $r=n-1$.

Для распределения Стьюдента составлены специальные таблицы, по которым, зная число степеней свободы *r* и вероятность β события $\{T > t_\beta\}$, можно найти число t_β. Таблица распределения Стьюдента приведена в прил. 2.

Заключим случайную величину *T* в интервал, симметричный относительно нуля, и обозначим его границы через $-t_\beta$ и t_β.

$$P(|T| < t_\beta) = \beta.$$

Тогда вероятности событий $\{T > t_\beta\}$ и $\{T<t_\beta\}$ равны:

$$P(T>t_\beta) = P(T < t_\beta) = \frac{1-\beta}{2}.$$

Зная число степеней свободы $r=n-1$ и вероятность $(1-\beta)/2$, можно по таблице найти число t_β. Неравенство $|T| < t_\beta$ означает, что

$$\left|\frac{\bar{x} - a}{\widetilde{S}}\sqrt{n}\right| < t_\beta.$$

Раскрывая знак модуля, получаем, что

$$\bar{x} - \frac{t_\beta \widetilde{S}}{\sqrt{n}} < a < \bar{x} + \frac{t_\beta \widetilde{S}}{\sqrt{n}}, \quad \text{или} \quad \bar{x} - \frac{t_\beta S}{\sqrt{n-1}} < a < \bar{x} + \frac{t_\beta S}{\sqrt{n-1}}.$$

Мы построили доверительный интервал, содержащий число a с вероятностью β. Если генеральная совокупность конечна, состоит из N единиц, из неё извлекается выборка объёма n, причём $n>0{,}05N$; при вычислении границ доверительных интервалов для a следует ввести поправочный коэффициент, равный $\sqrt{1-n/N}$ (см. §5.3). Таким образом,

$$\varepsilon = \frac{t\widetilde{S}}{\sqrt{n}}\sqrt{1-\frac{n}{N}},$$

где число t определяется по таблице функции Лапласа при $n \geq 30$ и по таблице распределения Стьюдента при $n<30$.

Если закон распределения генеральной совокупности далёк от нормального, то выборками малого объёма лучше не пользоваться, иначе закон распределения выборочного среднего $\bar{x}$ также будет отличаться от нормального. Во всяком случае лучше брать n не менее 15-20 единиц.

7.4.4. Построение доверительного интервала для неизвестной дисперсии σ^2 нормально распределённой генеральной совокупности

Для построения такого доверительного интервала пользуются случайной величиной

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\widetilde{S}^2}{\sigma^2}.$$

Здесь σ и n — числа, S^2, $\widetilde{S}^2$ - случайные величины.

Случайная величина

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\widetilde{S}^2}{\sigma^2}$$

имеет распределение χ^2 с числом степеней свободы $r=n-1$. Заключим случайную величину χ^2 в интервал $\chi_1^2 < \chi^2 < \chi_2^2$

из условий: $P\left(\chi_1^2 < \chi^2 < \chi_2^2\right) = \beta$; $P\left(\chi^2 > \chi_2^2\right) = \frac{1-\beta}{2}$; $P\left(\chi^2 < \chi_1^2\right) = \frac{1-\beta}{2}$.

По таблице распределения χ^2, зная число степеней свободы $r=n-1$ и вероятности событий

$$P\left(\chi^2 > \chi_1^2\right) = 1 - \frac{1-\beta}{2} = \frac{1+\beta}{2},\ P\left(\chi^2 > \chi_2^2\right) = \frac{1-\beta}{2},$$

находим числа χ_1^2 и χ_2^2. В отличие от нормального распределения и распределения Стьюдента, распределение Хи-квадрат не симметрично, для определения границ доверительного интеравла нужно задать два условия.

Неравенство $\chi_1^2 < \frac{nS^2}{\sigma^2} < \chi_2^2$ можно записать в виде

$$\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}.$$

Пример. Построить доверительный интервал с доверительной вероятностью $\beta = 0{,}96$ для неизвестной дисперсии σ^2 нормально распределённой генеральной совокупности, если n=20, S^2=10.

Решение. r=n-1=19; $\frac{1-\beta}{2} = 0{,}02$; $\frac{1+\beta}{2} = 0{,}98$. По таблице распределения χ^2 находим числа χ_1^2 и χ_2^2: χ_1^2=8,6; χ_2^2=33,7.

$$\frac{20 \cdot 10}{33{,}7} < \sigma^2 < \frac{20 \cdot 10}{8{,}6}; \quad 5{,}93 < \sigma^2 < 23{,}26.$$

Извлекая квадратный корень из чисел 5,93 и 23,26, получаем границы доверительного интервала для σ: $2{,}44 < \sigma < 4{,}82$.

7.4.5. Построение доверительного интервала для разности математических ожиданий нормально распределённых генеральных совокупностей

В этом случае имеются две нормально распределённые генеральные совокупности X_1 и X_2 с параметрами a_1, σ_1^2 и a_2, σ_2^2 соответственно. Из первой совокупности извлекается выборка объёма n_1, из второй - объёма n_2. Требуется с заданной доверительной вероятностью β построить доверительный интервал для разности чисел (a_1-a_2).

Рассмотрим случай, когда числа σ_1^2 и σ_2^2 известны. Тогда случайная величина $\bar{x}_1$ - выборочное среднее для генеральной совокупности X_1 - имеет нормальное распределение с параметрами $M(\bar{x}_1) = a_1$, $D(\bar{x}_1) = \sigma_1^2/n_1$. Случайная величина $\bar{x}_2$ - выборочное среднее для генеральной совокупности X_2 - имеет нормальное распределение с параметрами $M(\bar{x}_2) = a_2$, $D(\bar{x}_1) = \sigma_2^2/n_2$. Случайная величина $x = \bar{x}_1 - \bar{x}_2$ - разность выборочных средних - имеет нормальное распределение (как разность нормально распределённых случайных величин) с параметрами

$$M(x) = M(\bar{x}_1 - \bar{x}_2) = a_1 - a_2; \quad D(x) = D(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Теперь можно заключить случайную величину x в интервал

$$|x - M(x)| < \varepsilon$$

и найти число ε, пользуясь таблицей функции Лапласа, из условия

$$P(|x - M(x)| < \varepsilon) = \beta.$$

Более подробно:

$$P(|x - M(x)| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma(x)}\right) = 2\Phi\left(\frac{\varepsilon}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = \beta.$$

Тогда по таблице функции Лапласа находим число

$$t = \varepsilon \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \Phi(t) = \beta/2,$$

интервал для разности (a_1-a_2) таков:

$$(\bar{x}_1 - \bar{x}_2) - t\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < a_1 - a_2 < (\bar{x}_1 - \bar{x}_2) + t\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Если значения σ_1 и σ_2 неизвестны, но объёмы выборок достаточно велики (n_1, n_2>30), то также пользуются описанной процедурой, подставляя вместо σ_1^2 и σ_2^2 исправленные выборочные дисперсии $\widetilde{S}_1^2$ и $\widetilde{S}_2^2$, определённые по выборкам.

Пример. Почва двух участков земли была тщательно проанализирована и оказалась одинаковой по составу. На этих участках была посеяна пшеница одного сорта. На участок *A* внесено удобрение, а на участок *B* нет. Через месяц со дня посева пшеницы с каждого участка была произведена случайная выборка 50 растений, измерялась их длина. Средние значения и несмещённые выборочные дисперсии, вычисленные по выборкам, оказались равными:

$$\bar{x}_1 = 323 \text{ мм};\quad \bar{x}_2 = 297 \text{ мм};\quad \widetilde{S}_1^2 = 441 \text{ мм}^2;\quad \widetilde{S}_2^2 = 529 \text{ мм}^2.$$

С доверительной вероятностью $\beta = 0{,}99$ построить доверительный интервал для разности средних (a_1-a_2). Оказывает ли удобрение влияние на рост растений?

Решение. Имеется в виду, что случайные величины X_1 и X_2 - длины растений на участках *A* и *B* соответственно - нормально распределены. Нужно построить доверительный интервал для разности (a_1-a_2) их математических ожиданий. Дисперсии σ_1^2 и σ_2^2 неизвестны, но ввиду больших объёмов выборок (n_1=n_2=50) можно воспользоваться

нормальным распределением, подставив вместо σ_1^2 и σ_2^2 числа $\widetilde{S}_1^2$ и $\widetilde{S}_2^2$ соответственно.

Так как $\beta = 0{,}99$, число t=2,58,

$$\varepsilon = t\sqrt{\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}} = 2{,}58\sqrt{\frac{441}{50} + \frac{529}{50}} = 11{,}36.$$

Доверительный интервал для разности (a_1-a_2) таков:

(323 - 297) - 11,36< a_1-a_2<(323 - 297) + 11,36; 14,64<a_1-a_2<37,36.

Интервал не содержит нуля. Удобрение способствует росту растений.

Если выборки небольшие, но генеральные совокупности X_1 и X_2 имеют одну и ту же дисперсию σ^2, вместо числа

$$\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

рассматривают случайную величину

$$\widetilde{S}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

где $\widetilde{S}$ - объединённая несмещённая оценка дисперсии σ^2.

Ранее была выведена формула, позволяющая найти выборочную дисперсию для объединения двух выборок:

$$S = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2},$$

где S_1^2, S_2^2 - выборочные дисперсии, определённые по первой и второй выборкам соответственно. Как точечная оценка дисперсии σ^2 статистика S^2 смещена. Действительно,

$$M\left(S^2\right) = \frac{n_1 M\left(S_1^2\right) + n_2 M\left(S_2^2\right)}{n_1 + n_2} = \frac{n_1 \dfrac{n_1 - 1}{n_1}\sigma^2 + n_2 \dfrac{n_2 - 1}{n_2}\sigma^2}{n_1 + n_2} =$$

$$= \frac{\left(n_1 + n_2 - 2\right)}{n_1 + n_2}\sigma^2 \neq \sigma^2.$$

Несмещённой объединённой оценкой дисперсии σ^2 является статистика

$$\widetilde{S}^2 = \frac{(n_1 - 1)\widetilde{S}_1^2 + (n_2 - 1)\widetilde{S}_2^2}{n_1 + n_2 - 2} = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

Тогда случайная величина

$$T = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(a_1 - a_2\right)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

имеет распределение Стьюдента с числом степеней свободы $r = n_1 + n_2 - 2.$

Эту случайную величину можно заключить в интервал $|T| < t_\beta$ из условия

$$P\left(|T| < t_\beta\right) = \beta.$$

Число t_β находится по таблице распределения Стьюдента. Тогда интервал для разности (a_1-a_2) таков:

$$\left(\bar{x}_1 - \bar{x}_2\right) - t_\beta \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < a_1 - a_2 < \left(\bar{x}_1 - \bar{x}_2\right) +$$

$$+ t_\beta \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Пример. Химик делает шесть измерений концентрации серной кислоты и обнаруживает, что средняя концентрация $\bar{x}_1 = 9{,}234$, S_1=0,12. Проводя опыты с кислотой из другой бутылки, он делает одиннадцать измерений и получает, что средняя концентрация $\bar{x}_2 = 8{,}86,$ а S_2=0,21. Найти границы доверительного интервала для разности средних величин концентрации кислоты в двух бутылках при $\beta = 0{,}99.$ Были ли наполнены бутылки одной и той же кислотой?

Решение. Предполагается, что показания прибора, измеряющего концентрацию кислоты, можно считать значениями нормально распределённой случайной величины.

Число степеней свободы $r = 6 + 11 - 2 = 15;$ $(1\text{-}\beta)/2$=0,005. Тогда $t_\beta = 2{,}95.$ Объединённая несмещённая оценка дисперсии:

$$\widetilde{S} = \sqrt{\frac{6 \cdot 0{,}12^2 + 11 \cdot 0{,}21^2}{15}} = 0{,}195;$$

$$t_\beta \widetilde{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2{,}95 \cdot 0{,}195 \cdot \sqrt{\frac{1}{6} + \frac{1}{11}} = 0{,}292;$$

$\bar{x}_1 - \bar{x}_2 = 9{,}234 - 8{,}86 = 0{,}374;$ 0,374 - 0,292<a_1-a_2<0,374+0,292;

$$0{,}08 < a_1\text{-}a_2 < 0{,}67.$$

Интервал не содержит нуля, концентрация кислоты в бутылках разная.

7.5. ЗАДАЧИ

1. Пусть $x_1, x_2, \ldots, x_n$ - выборка из генеральной совокупности X с известным математическим ожиданием a и неизвестной дисперсией σ^2. Показать, что несмещённой оценкой для σ^2 будет статистика

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2.$$

2. В результате проведения n независимых экспериментов в одних и тех же условиях событие A произошло k раз. Показать, что относительная частота k/n появления события A будет несмещённой и состоятельной оценкой вероятности p события A в одном эксперименте. Определить такое значение p, при котором дисперсия этой оценки будет максимальна.

3. Пусть генеральная совокупность X имеет равномерное распределение на интервале $(a, a+1)$, причём a неизвестно. Из генеральной совокупности извлечена выборка $x_1, x_2, \ldots, x_n$ объёма n. Для оценки параметра a можно использовать статистики:

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} X_i - 0{,}5 \text{ и } a_2 = \max(x_i) - \frac{n}{n+1}.$$

Показать, что оценки эти не смещены. Найти дисперсию случайных величин a_1 и a_2.

4. Из генеральной совокупности X, равномерно распределённой на интервале (a, b), извлечена выборка объёма n: $x_1, x_2, \ldots, x_n$. Длина интервала $h=b-a$ известна, но середина интервала $C=(a+b)/2$ неизвестна. В качестве оценки середины интервала предлагается случайная величина

$$\overline{C} = \frac{\max(x_i) + \min(x_i)}{2}.$$

Доказать несмещённость этой оценки.

5. Из равномерно распределённой на интервале (a, b) генеральной совокупности X извлечена выборка объёма n: $x_1, x_2, \ldots, x_n$. В качестве оценки длины интервала $(b-a)$ предлагается случайная величина

$$\overline{h} = \max(x_i) - \min(x_i).$$

Доказать смещённость этой оценки.

6. Из нормально распределённой генеральной совокупности с параметрами $a = 18{,}1$ и $\sigma = 2{,}3$ извлечена выборка объёма 9. Найти вероятности следующих событий:

$$\text{а)}\ \{\overline{x} < 16\}, \quad \text{б)}\ \{15 < \overline{x} < 17\}, \quad \text{в)}\ \{16 < \overline{x} < 19\}.$$

Как изменится решение для выборки объёма n=36?

7. Некий кандидат набрал на выборах 45% голосов. Из всей совокупности избирателей случайным образом отобрали две группы

людей. Оценить вероятность того, что разность между долями голосов, поданных за этого кандидата в каждой из групп, окажется больше 0,05.

8. У случайно отобранных 525 студентов мужского пола измерили рост. Средний рост оказался равным 181 см, выборочное среднее квадратическое отклонение роста S=3,35 см. Какова точность оценки среднего роста с надёжностью 0,95?

9. При испытании на крепость случайно отобранных 400 отрезков одиночной нити были получены следующие результаты:

Крепость, г	Число испытанных образцов	Крепость, г	Число испытанных образцов
105-125	8	205-225	120
125-145	24	225-245	35
145-165	40	245-265	20
165-185	56	265-285	7
185-205	84	285-305	6

1. С вероятностью 0,95 определить среднюю крепость нити во всей партии.
2. С какой вероятностью можно утверждать, что разность между средней крепостью пряжи в выборочной и генеральной совокупностях не превысит 3,5 г?
3. Сколько нужно отобрать для испытаний образцов одиночной нити, чтобы с вероятностью 0,99 утверждать, что разность между средней крепостью пряжи в выборочной и генеральной совокупностях не превысит 3г?
4. С вероятностью 0,95 определить гарантийные пределы, в которых находится доля пряжи во всей совокупности с крепостью, превышающей 250 г.
5. Как изменить объём выборки, чтобы с вероятностью 0,9 можно было гарантировать отклонение выборочной доли от генеральной, не превышающее 0,02?
6. С какой вероятностью можно утверждать, что выборочная доля будет отличаться от генеральной не более чем на 0,025?

10. Для определения средней дальности пробега автомобилей наугад отобрали 100 путёвок. Получены следующие данные:

Дальность пробега автомобилей, км	Число путёвок	Дальность пробега автомобилей, км	Число путёвок
20-50	3	170-200	13
50-80	7	200-230	6
80-110	14	230-260	4
110-140	28	260-290	1
140-170	24	—	—

1. С вероятностью 0,98 определить пределы, в которых находится средний пробег машин базы.

2. Как изменить гарантийную вероятность, чтобы предельную ошибку средней дальности пробега уменьшить на 20%?

3. Сколько нужно отобрать путёвок, чтобы с вероятностью 0,99 гарантировать, что средняя дальность пробега всех машин автобазы не вышла за пределы 130,6-150,6 км?

4. С вероятностью 0,98 определить пределы, в которых находится доля всех машин автобазы с дальностью пробега, превышающей 200 км.

5. Какова вероятность того, что предельная доля ошибки машин, дальность пробега которых превышает 200 км, не превзойдёт 0,05?

6. Сколько нужно отобрать путёвок, чтобы с вероятностью 0,99 гарантировать отклонение выборочной доли от генеральной, не превышающее 0,06?

11. Средняя масса пакетов, расфасованных на автомате, равна 1 кг при среднем квадратическом отклонении 3 г. Сколько нужно отобрать пакетов, чтобы с вероятностью 0,95 гарантировать отклонение средней массы отобранных пакетов от 1 кг, не превышающее 0,1%?

12. Для определения среднего количества деловой древесины в одном дереве подвергли выборочному обследованию 1000 деревьев, растущих в большом лесу. Были получены следующие данные:

Количество деловой древесины в одном дереве, $м^3$	0,5	1	1,5
Количество деревьев, шт.	208	484	308

Определить вероятность того, что генеральная средняя отличается от выборочного среднего не более чем на 0,03 $м^3$.

13. Произведено пять независимых равноточных измерений для определения заряда электрона. Получены следующие результаты (кулоны): $1{,}594 \cdot 10^{-19}$; $1{,}597 \cdot 10^{-19}$; $1{,}598 \cdot 10^{-19}$; $1{,}593 \cdot 10^{-19}$; $1{,}590 \cdot 10^{-19}$. Найти доверительные границы для величины заряда электрона, если $\beta = 0{,}99$.

14. Средний возраст глав семей для случайной выборки из 3704 семейств равен 51,07 года, а выборочное среднее квадратическое отклонение - 19,98 года. Построить доверительные интервалы для среднего возраста всех глав семейств при $\beta_1 = 0{,}95$ и $\beta_2 = 0{,}99$.

15. Построить доверительные интервалы для математического ожидания *a* с доверительными вероятностями $\beta_1 = 0{,}9$ и $\beta_2 = 0{,}99$ в каждом из следующих случаев:

Измерение	$\bar{x}$	n	σ(известно)
Число слов в предложении	27	225	7
Длина предплечья	18,1 единицы	100	0,82
Диаметр мускульной мышцы	17,1 единицы	625	3,4

В задачах 16 - 19 найти доверительные интервалы для математического ожидания с доверительными вероятностями $\beta_1 = 0{,}9;\ \beta_2 = 0{,}95;\ \beta_3 = 0{,}99$.

16. Содержание углерода в килограмме чугуна, если $\bar{x} = 28$ г, n=16, $S = 4$ г.

17. Диаметры шести шаров в шарикоподшипнике: 2,01; 1,99; 2,00; 2,00; 2,01; 1,98 мм.

18. Увеличение частоты пульса солдат после проверки физических данных: 10,13, 6, 8, 12, 8, 7, 10, 12, 14 ударов.

19. Процентное содержание витамина *C* в выборке витаминных драже: 14,3; 15,2; 16,3; 14,8; 12,9.

Для каждой ситуации, описанной в задачах 16 – 19, были проведены повторные выборки. Используя объединённые выборочные оценки, снова построить доверительные интервалы для математического ожидания с теми же доверительными вероятностями.

20. Содержание углерода в килограмме чугуна, если $n = 9$, $\bar{x} = 28{,}8$ г, S^2=20г^2.

21. Диаметр шаров в шарикоподшипнике: $\bar{x} = 2$ см; $S^2 = 6{,}4 \cdot 10^{-5}$ см^2; n=8.

22. Увеличение частоты пульса: $\bar{x}$=9 ударов; S^2=4; n=11.

23. Процентное содержание витамина *C*: $\bar{x}$=14; S^2=0,25; n=7.

24. С производственной линии, производящей сигареты, было отобрано 900 сигарет, 45 из них оказались бракованными. Оценить долю дефектных сигарет во всей совокупности и найти для неё доверительные границы, если $\beta = 0{,}9$. Какой объём выборки с производственной линии нужно взять, чтобы со степенью доверия 99,7% утверждать, что ошибка оценивания не превосходит 0,01?

25. В 10000 сеансах игры с автоматом выигрыш появился 4000 раз. Найти доверительный интервал для вероятности выигрыша, если $\beta = 0{,}95$. Сколько сеансов игры следует провести, чтобы с вероятностью 0,99 вероятность выигрыша отличалась от частоты не более чем на 0,01?

26. 4709 семей в случайной выборке, включающей 7148 семей, зарегистрировали расходы на алкогольные напитки. Каковы границы доверительного интервала для доли всех семей, имевших расходы на алкогольные напитки в период опроса, если $\beta = 0{,}999$?

27. Случайная величина *Z* равна разности двух независимых нормальных случайных величин *X* и *Y*. Выборочные оценки для *X* и *Y* определяли по результатам n_1=16 и n_2=36 наблюдений соответственно. С доверительной вероятностью 0,95 найти доверительный интервал для математического ожидания *Z*, если $\bar{x} = 10,\ \bar{y} = 4,\ \sigma(X) = 1,\ \sigma(Y) = 2$.

28. Были произведены выборки из трёх генеральных совокупностей. Каждый раз вычислялись выборочные средние $\bar{x}$ и исправленные выборочные средние квадратические отклонения $\widetilde{S}$. После изменений условий экспериментов выборки были повторены и снова были найдены те же числовые характеристики. Результаты представлены в таблице. По её данным построить доверительные интервалы для средних совокупностей $\beta_1 = 0{,}9$, $\beta_2 = 0{,}95$; с доверительными вероятностями $\beta_3 = 0{,}99$ и $\beta_4 = 0{,}997$ построить доверительные интервалы для разности средних. Можно ли считать, что расхождения между средними объясняются только случайностью выборок?

Измерения	$\bar{x}$	$\widetilde{S}$	n
Скорость чтения, слов/мин	110	12	100
Скорость чтения после упражнений, слов/мин	130	14	81
Урожай зерна, удобрение *A,* ц/га	43	5	30
Урожай зерна, удобрение *B,* ц/га	40	4,5	45
Добыча парафина из торфа, %: растворитель *A*	5,3	2,1	40
Добыча парафина из торфа, %: растворитель *B*	6,4	2,4	50

29. Выборка лампочек сорта *A* исследовалась на продолжительность горения. Время непрерывного свечения выбранных лампочек (условные единицы): 21, 32, 28, 14, 30, 27, 30. Проверка лампочек сорта *B* дала такие результаты: 28, 29, 34, 18, 30. С доверительной вероятностью 0,99 построить доверительный интервал для разности средних продолжительностей работ лампочек двух сортов.

30. Покрышки, произведённые на заводе *A*, исследовались на износ (км пройденного пути). Оказалось, что средний пробег равен 25000 км, а S_1=1200 км. В тех же условиях испытывались покрышки, изготовленные на заводе *B*. Для них средний пробег оказался равен 23500 км, S_2=1000 км. Найти доверительный интервал для разности средних пробегов, $\beta = 0{,}99$. Каждый раз испытывалось 4 покрышки.

В задачах 31 - 35 требуется построить доверительные интервалы для дисперсии σ^2 и среднего квадратического отклонения σ нормально распределенной генеральной совокупности. Положить $\beta_1 = 0{,}9$; $\beta_2 = 0{,}95$; $\beta_3 = 0{,}98$.

31. Обратиться к данным задачи 16.

32. Обратиться к данным задачи 17.

33. Обратиться к данным задачи 18.

34. Обратиться к данным задачи 19.

35. Результаты 10 независимых измерений длины стержня (мм): 23, 24, 23, 25, 25, 26, 26, 25, 24, 25.

36. Требуется оценить средние еженедельные расходы на питание студентов некоторого университета. Эта оценка используется для определения размера дотаций нуждающимся студентам.

1. Выборку какого объёма следует взять для получения доверительного интервала шириной 0,4 долл. при доверительной вероятности $\beta = 0{,}95$? Из опыта известно, что S=5 долл.
2. Оказалось, что выборочное среднее равно 70 долл. Указать пределы доверительного интервала для средних расходов на питание.
3. Сколько примерно расходует студент на питание в течение учебного года (35 недель)?

37. Известно, что S^2=4. При каком n левая граница доверительного интервала для σ^2 отличается от S^2 на 1,8? Положить $\beta = 0{,}95$.

38. На некотором предприятии работает 2000 человек. Дирекция хочет оценить долю рабочих, которые в понедельник опоздали на работу более чем на 5 минут.

1. Выборку какого объёма нужно взять, чтобы доверительный интервал имел ширину не более 4%, если β=0,95? (Дирекции известно, что количество опоздавших не больше 30%).
2. Была взята случайная выборка того объёма, который был определён в предыдущем пункте. Оказалось, что $\bar{x} = 0{,}18$. Оценить общее число рабочих, опоздавших на работу в понедельник.

39. 800 студентам задали следующий вопрос: купили бы вы в период с сентября по июль в магазине нашего университетского городка хотя бы одну пару обуви? Число положительных ответов оказалось 100.

1. Построить доверительный интервал для доли студентов, сделавших такую покупку, если $\beta = 0{,}99$.
2. Оценить общее число покупателей-студентов, если всего в университете учится 20000 студентов.
3. Предположим, что студенты, попавшие в выборку, ответили также, сколько пар обуви они купили и за какую цену. Сто студентов, каждый из которых купил хотя бы одну пару обуви, купили в общей сложности 120 пар по средней цене $\bar{x} = 45$ долл. за пару, причём S=6,3 долл. Построить доверительный интервал для средней стоимости пары обуви, приобретённой всеми студентами, $\beta = 0{,}95$.
4. Оценить общую сумму денег, истраченных на обувь студентами в магазинах этого городка.

8. ПОНЯТИЕ О ПРОВЕРКЕ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

О сколько нам открытий чудных
Готовят просвещенья дух,
И опыт, сын ошибок трудных,
И гений, парадоксов друг,
И случай, бог изобретатель.
А. С. Пушкин

8.1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

8.1.1. Что такое статистическая гипотеза

Под статистической гипотезой мы будем понимать либо предположение о законе распределения генеральной совокупности (закон неизвестен), либо предположение о значениях (неизвестных) параметров известного закона распределения.

В соответствии со сказанным статистические гипотезы делятся на непараметрические, если в них высказывается предложение о виде закона распределения, и параметрические, если в них говорится о значениях параметров известного закона распределения.

В главе 1 мы уже рассматривали процедуру проверки непараметрических гипотез по критерию Пирсона.

Параметрические гипотезы рассматриваются попарно. Гипотезы, образующие пару, взаимно исключают друг друга; называются они нулевой и альтернативой. Процедура проверки применяется к нулевой гипотезе H_0. Если в результате проверки оказывается целесообразным отвергнуть гипотезу H_0, то принимается альтернативная гипотеза H_a.

Нулевая гипотеза - это простая гипотеза, в ней говорится о конкретных значениях параметров. Альтернативная гипотеза сложная, в ней подразумевается бесконечно много возможностей. Рассмотрим несколько примеров нулевых и альтернативных гипотез.

Нулевая гипотеза: математическое ожидание a генеральной совокупности равно числу a_0. Коротко это записывается так: H_0: $a = a_0$. Возможные для этого случая альтернативные гипотезы:

1. H_a: $a \neq a_0$. 2. H_a: $a > a_0$. 3. H_a: $a < a_0$.

Нулевая гипотеза: математические ожидания a_1 и a_2 двух генеральных совокупностей равны. Коротко это записывается так H_0: $a_1 = a_2$. Возможные альтернативные гипотезы:

1. H_a: $a_1 \neq a_2$. 2. H_a: $a_1 > a_2$. 3. H_a: $a_1 < a_2$.

8.1.2. О процедуре проверки нулевой гипотезы

Всякую гипотезу (не только статистическую) желательно обосновать. Вряд ли можно дать строгое определение такому понятию, как обоснование. В статистике под обоснованием гипотезы понимается следующая процедура.

1. Подбирается случайная величина K (выборочная статистика), закон распределения которой известен в предположении справедливости нулевой гипотезы H_0.

2. Область значений случайной величины K разбивается на два непересекающихся подмножества. Вероятность того, что случайная величина K примет значение из первого множества (если только справедлива гипотеза H_0), велика. Вероятность того, что случайная величина K примет значение из второго множества, мала. Первое множество значений называется областью принятия гипотезы H_0, второе – областью отвержения гипотезы H_a или критической областью. Вероятность попадания в критическую область называется уровнем значимости и обозначается буквой α. Тогда вероятность попадания в область принятия гипотезы H_0 равна $(1 - \alpha)$.

3. По выборке, извлеченной из генеральной совокупности, вычисляется экспериментальное значение случайной величины K – число $K_{эксп}$. Если это число принадлежит критической области, то гипотеза H_0 отвергается, как противоречащая опытным данным. Справедливой считается альтернативная гипотеза H_a. Если число $K_{эксп}$ принадлежит области принятия гипотезы H_0, эта гипотеза считается согласующейся с опытными данными.

Случайная величина K называется статистическим критерием или тестом.

В зависимости от условия эксперимента критическую область можно выбрать двусторонней, левосторонней и правосторонней. На рис. 8.1 – 8.3 проиллюстрированы введенные определения. На каждом из этих рисунков показан график функции плотности вероятности случайной величины K при условии справедливости гипотезы H_0, обозначена эта функция $f(K / H_0)$.

8.1.3. Ошибки, допускаемые при проверке статистических гипотез

Всякое заключение об истинности или ложности статистической гипотезы достаточно условно. Выборка конечного объёма не может отражать всю генеральную совокупность. Если экспериментальное значение

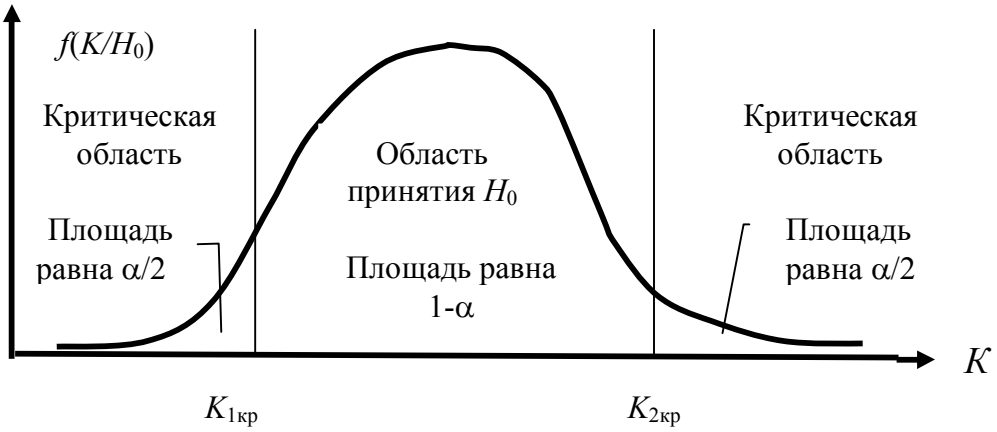


Рис. 8.1

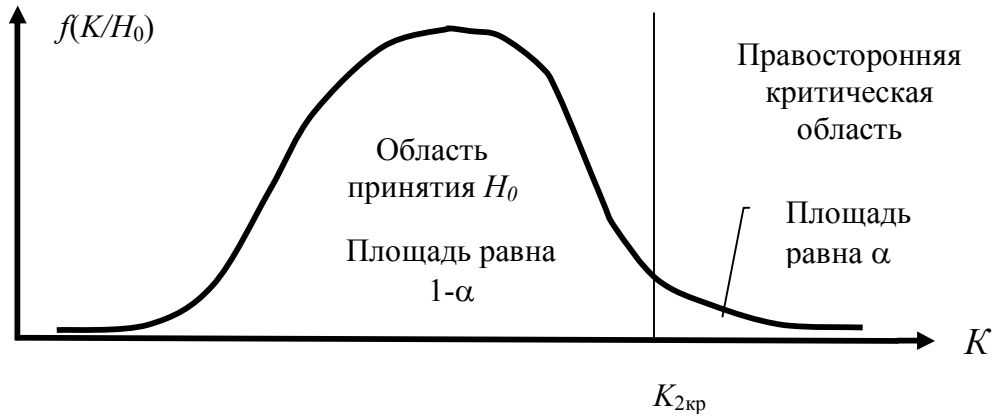


Рис. 8.2

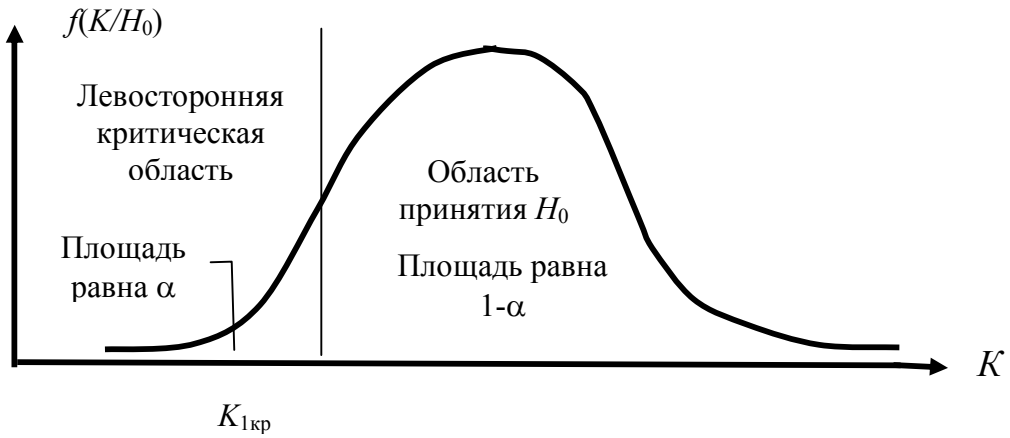


Рис. 8.3

статистического критерия, вычисленное по этой выборке, попало в область принятия гипотезы H_0, это ещё не означает, что H_0 обязательно верна. Если число $K_{эксп}$ попало в критическую область, это не гарантия справедливости

альтернативной гипотезы.

Таким образом, применяя процедуру проверки некоторой статистической гипотезы, можно совершить одну из двух ошибок.

Ошибкой первого рода называется ошибка отклонения верной гипотезы H_0. Вероятность этой ошибки равна вероятности попадания случайной величины K в критическую область, т.е. равна числу α. Ошибкой второго рода называется ошибка принятия ложной гипотезы H_0. Чтобы найти вероятность ошибки второго рода, нужно знать закон распределения случайной величины K в предположении справедливости альтернативной гипотезы H_a. Тогда искомая вероятность равна вероятности попадания случайной величины K в область принятия гипотезы H_0. Это число обозначается буквой β. Мощностью критерия K называется вероятность $(1 - \beta)$ несовершения ошибки второго рода.

Хороший статистический критерий тот, для которого минимальны вероятности совершения ошибок как первого, так и второго рода. Но мы не будем рассматривать теорию построения критериев максимальной мощности, ограничимся только обсуждением так называемых критериев значимости, применяя которые заранее фиксируют вероятность совершения ошибки первого рода (уровень значимости α).

Во многих случаях экспериментатору достаточно быть уверенным в том, что выдвинутая им нулевая гипотеза не противоречит опытным данным.

8.2. ПРОВЕРКА ПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ ПО КРИТЕРИЯМ ЗНАЧИМОСТИ

В дальнейшем ограничимся только случаем нормально распределенной генеральной совокупности или нескольких нормально распределенных генеральных совокупностей.

8.2.1. Проверка гипотезы о значении математического ожидания

8.2.1.1. Случай, когда дисперсия σ^2 генеральной совокупности известна

Задача ставится следующим образом. Из нормально распределенной генеральной совокупности с неизвестным математическим ожиданием a и известной дисперсией σ^2 извлечена выборка объёма n. Выдвигается нулевая гипотеза, что значение математического ожидания равно числу a_0. Требуется подтвердить или опровергнуть эту гипотезу. Для проверки гипотезы используется случайная величина

$$u = \frac{(\overline{x} - a_0)}{\sigma}\sqrt{n},$$

где $\overline{x}$ - выборочное среднее.

Если нулевая гипотеза верна, случайная величина u имеет нормальное распределение с математическим ожиданием, равным нулю, и дисперсией, равной единице (так называемое стандартное нормальное распределение). Альтернативная гипотеза, область принятия нулевой гипотезы и критическая область выбираются, исходя из условий эксперимента,

Та же случайная величина используется и тогда, когда дисперсия σ^2 неизвестна, но объём выборки n достаточно велик, $n \geq 30$. Вместо числа σ в формулу для u подставляется значение выборочного среднего квадратического отклонения S, найденного по выборке.

Пример. На станке-автомате должны изготовляться детали с номинальным контролируемым размером a_0 = 12 мм. Известно, что распределение контролируемого размера является нормальным. Были измерены размеры 36 случайно отобранных деталей. Среднее значение контролируемого размера оказалось равным $\overline{x}$ = 11,7 мм, выборочное среднее квадратическое отклонение S оказалось равным 0,5 мм. Можно ли считать, что станок-автомат изготавливает детали уменьшенного размера и, следовательно, требует наладки?

Решение. Проверяется нулевая гипотеза H_0: a = 12. Дисперсию генеральной совокупности будем считать равной 0,5. Если верна гипотеза H_0, значения выборочного среднего $\overline{x}$ не должны сильно отличаться от 12, соответственно значения случайной величины

$$u = \frac{(\overline{x} - 12)}{0{,}5/\sqrt{n}}$$

не должны сильно отличаться от нуля. Если же верна гипотеза H_a, то значения $\overline{x}$ следует ожидать меньше 12, соответственно значения u должны быть существенно меньше нуля. Критическая область – левосторонняя (рис. 8.4).

Положим уровень значимости α равным 0,05. Теперь можно найти $u_{кр}$, т.е. придать точный смысл высказыванию "значения u должны быть существенно меньше нуля".

По определению критической области $P(u < u_{кр}) = \alpha = 0{,}05$.

С другой стороны, если гипотеза H_0 справедлива, то $P(u < u_{кр}) = \Phi(u_{кр}) - \Phi(-\infty) = \Phi(u_{кр}) + 0{,}5$, тогда $\Phi(u_{кр}) = -0{,}45$, $u_{кр} = -1{,}65$.

Найдем $u_{эксп}$.

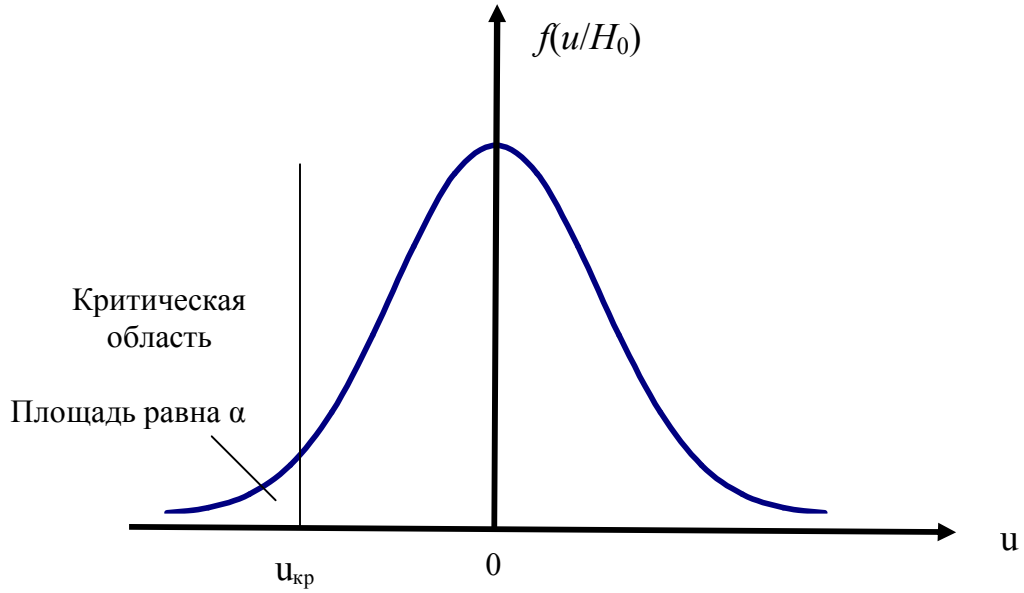


Рис. 8.4

$$u_{эксп} = \frac{(11{,}7 - 12)}{0{,}5/6} = -3{,}6 < -1{,}65\,.$$

Экспериментальное значение критерия u попало в критическую область, $u_{эксп} < u_{кр}$. Нулевую гипотезу следует отклонить в пользу альтернативной, станок нуждается в наладке.

8.2.1.2. Проверка гипотезы о значении вероятности "успеха"

Произведено n испытаний по схеме Бернулли, "успех" произошел k раз. Требуется проверить нулевую гипотезу, что вероятность появления "успеха" в каждом испытании равна числу P_0.

Если число испытаний $n \geq 50$, для проверки нулевой гипотезы используют случайную величину

$$u = \frac{(\overline{x} - P_0)}{\sqrt{\frac{P_0(1-P_0)}{n}}},$$

где $\overline{x} = k / n$ – выборочное среднее. Если верна нулевая гипотеза, то случайную величину u можно считать распределённой по стандартному нормальному закону.

Пример. Игральную кость подбросили 600 раз, шестёрка появилась

125 раз. Можно ли утверждать, что кость правильная? Принять α = 0,05.

Решение. Требуется проверить нулевую гипотезу $H_0 : p = 1/6$. Если кость правильная, то математическое ожидание числа выпадений шестерки в шестистах бросаниях равно 1/6 * 600 = 100.

Так как 125 > 100, в качестве альтернативной гипотезы примем гипотезу $H_a : p > 1/6$. Критическая область - правосторонняя (рис. 8.5).

Число $u_{кр}$ находится из условия $P(u > u_{кр}) = \alpha = 0,05$. С другой стороны, $P(u > u_{кр}) = \Phi(\infty) - \Phi(u_{кр}) = 0,5 - \Phi(u_{кр})$.

Отсюда $\Phi(u_{кр}) = 0,45 \Rightarrow u_{кр}=1,65$. Найдём число $u_{эксп}$.

$$u_{эксп} = \frac{125/600 - 1/6}{\sqrt{\frac{(1/6)*(5/6)}{600}}} = 2,73 > 1,65.$$

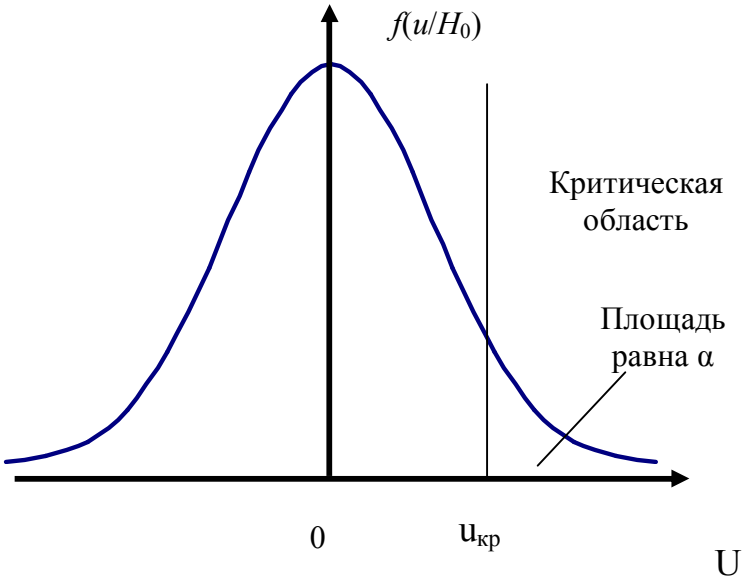


Рис. 8.5

Экспериментальное значение критерия больше критического, нулевая гипотеза отвергается в пользу альтернативной.

8.2.1.3. Проверка гипотезы о значении математического ожидания, когда дисперсия генеральной совокупности неизвестна

Когда объём выборки мал ($n < 30$), а дисперсия генеральной совокупности неизвестна, для проверки нулевой гипотезы $H_0 : a = a_0$ используют статистику

$$T - \frac{\bar{x} - a_0}{s}\sqrt{n-1} = \frac{\bar{x} - a_0}{\tilde{s}}\sqrt{n},$$

где $\overline{x}$ – выборочное среднее; S – выборочное среднее квадратическое отклонение; $\widetilde{S}$ – исправленное выборочное среднее квадратическое отклонение.

Если нулевая гипотеза верна, случайная величина T распределена по закону Стьюдента с $n - 1$ степенями свободы.

Пример. По паспортным данным автомобильного двигателя, расход топлива на 100 км пробега составляет 10 л. В результате изменения конструкции двигателя ожидается, что расход топлива уменьшится. Для проверки были испытаны 25 случайно отобранных автомобилей с модернизированным двигателем. Средний расход топлива на 100 км пробега оказался равным $\overline{x}$ = 9,6 л. Исправленное выборочное среднее квадратическое отклонение расхода топлива равно 2 л. Можно ли утверждать, что изменение конструкции двигателя не повлияло на расход топлива? Положить α = 0,05. Считать, что выборка произведена из нормальной генеральной совокупности.

Решение. Проверяется нулевая гипотеза H_0: $a = 10$ против альтернативной гипотезы H_a : $a < 10$. Если нулевая гипотеза верна, то случайная величина

$$T = \frac{\overline{x} - 10}{\widetilde{S}}\sqrt{25}$$

имеет распределение Стьюдента с числом степеней свободы $n = 24$. По таблице распределения Стьюдента, полагая α = 0,05 и $r = 24$, находим, что критическое значение в случае правосторонней критической области $t_{кр}$ = l,71. Так как график функции плотности вероятности распределения Стьюдента симметричен относительно нуля, для левосторонней критической области $T_{кр}$ = -1,71. Вычислим $T_{эксп}$.

$$T_{эксп} = \frac{9{,}6 - 10}{2}\sqrt{25} = -1.$$

$T_{эксп} > T_{кр}$, нет оснований считать, что новая конструкция двигателя позволила уменьшить расход топлива.

Пример. В условиях предыдущего примера определить такое максимальное значение выборочного среднего, что при α = 0,05 нулевая гипотеза будет отвергнута.

Решение. Значение $\overline{x}$ определяется из условия

$$T_{эксп} = \frac{\overline{x} - 10}{2}\sqrt{25} \le -1{,}71.$$

Отсюда $\overline{x} \le 9{,}316$.

Пример. Пусть теперь критическая область задается условием $\overline{x} < 9{,}47$. Какова вероятность ошибки первого рода, если нулевая

гипотеза H_0 справедлива? (Полагаем, что $\widetilde{S} = 2$).

Решение. Вероятность ошибки первого рода – это вероятность того, что случайная величина

$$T = \frac{\overline{x} - 10}{2}\sqrt{25},$$

имеющая распределение Стьюдента с числом степеней свободы $r = 24$, примет значение меньше чем (9,47 – 10) * 2,5 = -1,32.

По таблице распределения Стьюдента находим, что α = 0,1. Это означает, что принятый критерий примерно 10% автомобилей, имеющих расход топлива 10 л на 100 км пробега, отнесет к автомобилям, имеющим меньший расход топлива.

Пример. Группе школьников младших классов был дан стандартный тест на проверку скорости чтения. Затем со школьниками был проведен специальный курс занятий, после чего детям был предложен второй тест. Скорость чтения оценивалась в баллах, результаты представлены в таблице.

Школьник	1	2	3	4	5	6	7	8	9	10
Результаты первого тестирования	7	5	6	8	3	8	7	8	2	4
Результаты второго тестирования	6	7	9	8	6	7	7	9	4	7

Можно ли сказать, что занятия по увеличению скорости чтения эффективны? Предполагается, что оба теста эквивалентны по трудности, а оценки, полученные школьниками, можно считать нормально распределенными.

Решение. Рассмотрим выборку, варианты которой – разности баллов, полученных каждым школьником до и после занятий. Эта выборка такова:

Школьник	1	2	3	4	5	6	7	8	9	10
Разности	-1	2	3	0	3	-1	0	1	2	3

Предположим, что упражнения не влияют на скорость чтения (нулевая гипотеза). Тогда полученную выборку следует считать выборкой из нормально распределённой генеральной совокупности со средним значением $a = 0$. Если же упражнения увеличивают скорость чтения (альтернативная гипотеза), то математическое ожидание a должно быть больше нуля. Значит, для проверки нулевой гипотезы нужно использовать статистику

$$T = \frac{\overline{x}}{S}\sqrt{9},$$

которая, если верна нулевая гипотеза (a = 0), имеет распределение Стьюдента с числом степеней свободы r = 9. Найдём значения $\overline{x}$ и S.

$\overline{x}$ = (–1 + 2 + 3 + 0 + 3 – 1 + 0 – 1 + 2 + 3) / 10=1,2;

S^2 = (1 + 4 + 9 + 9 + 1 + 1 + 4 +9) / 10 – 1,2^2 = 3,8 – 1,44 = 2,36;

$S = \sqrt{2,36} = 1,54$. $T_{эксп} = \frac{\overline{x}}{S}\sqrt{9} = \frac{1,2}{1,54} * 3 \approx 2,34$.

Положим α = 0,05, тогда $T_{кр}$ = 1,83.

Так как $T_{эксп} > T_{кр}$, следует признать, что средняя скорость чтения у детей увеличилась.

8.2.2. Проверка гипотезы о равенстве математических ожиданий двух генеральных совокупностей

Задача ставится так. Имеются две нормально распределенные генеральные совокупности с параметрами a_1 и σ_1^2, a_2 и σ_2^2 соответственно. Из первой генеральной совокупности извлечена выборка объема n_1, из второй – объёма n_2. Требуется проверить нулевую гипотезу H_0: $a_1 = a_2$ о равенстве средних значений этих генеральных совокупностей.

8.2.2.1. Случай, когда дисперсии σ_1^2 и σ_2^2 считаются известными

В этом случае для проверки нулевой гипотезы применяют случайную величину

$$u = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

где $\overline{x}_1$ и $\overline{x}_2$ – выборочные средние.

Если верна нулевая гипотеза, статистика u имеет нормальное распределение со средним, равным нулю, и дисперсией, равной единице. Дисперсии считаются известными, когда n_1, $n_2 \geq 30$, тогда $\sigma_1^2 \approx S_1^2$, $\sigma_2^2 \approx S_2^2$.

Эту же статистику используют, проверяя гипотезу о равенстве вероятностей "успеха". Объёмы выборок n_1, n_2 должны быть достаточно велики, чтобы биномиальное распределение можно было приближенно считать нормальным.

Пример. Двое рабочих на одинаковых станках изготовляют одинаковые детали. Есть ли значимая разница между долями выпускаемого ими брака? Была собрана следующая информация: $n_1 = 200$, $k_1 = 12$; $n_2 = 200$, $k_2 = 18$. Положим $\alpha = 0{,}05$, тогда $\overline{x}_1 = 0{,}06$; $\overline{x}_2 = 0{,}09$; $\sigma_1^2 \approx S_1^2 = \overline{x}_1 (1 - \overline{x}_1) = 0{,}0564$, $\sigma_2^2 \approx S_2^2 = \overline{x}_2 (1 - \overline{x}_2) = 0{,}0819$;

$u_{\text{эксп}} = -1{,}14$.

Если альтернативная гипотеза имеет вид H_a: $P_1 < P_2$, то $u_{\text{кр}} = -1{,}65 < u_{\text{эксп}}$. Нельзя утверждать, что второй рабочий в среднем делает больше брака, чем первый.

8.2.2.2. Случай, когда σ_1^2 и σ_2^2 неизвестны,
но известно, что $\sigma_1^2 = \sigma_2^2$

Для проверки нулевой гипотезы используют случайную величину

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1 - 1)\widetilde{S}_1^2 + (n_2 - 1)\widetilde{S}_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}},$$

где S_1^2, S_2^2 – выборочные дисперсии; $\widetilde{S}_1^2$, $\widetilde{S}_2^2$ – исправленные выборочные дисперсии.

Если нулевая гипотеза верна, то случайная величина T имеет распределение Стьюдента с числом степеней свободы $r = n_1 + n_2 - 2$.

8.2.3. Проверка гипотезы о значении дисперсии

Из нормально распределённой генеральной совокупности извлечена выборка объема n. На основании этой выборки требуется проверить нулевую гипотезу, что дисперсия генеральной совокупности равна числу σ_0^2.

Для проверки гипотезы используется случайная величина

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{(n-1)\widetilde{S}^2}{\sigma_0^2},$$

где $\widetilde{S}^2$ – выборочная дисперсия.

Если нулевая гипотеза верна, то случайная величина χ^2 имеет распределение χ^2 с числом степеней свободы $r = n - 1$.

Пример. Точность наладки станка-автомата, производящего некоторые детали, характеризуется дисперсией длины деталей. Если эта величина больше 400 мкм^2, станок останавливают для наладки. Выборочная

дисперсия длины 15 случайно отобранных деталей из продукции станка оказалась равной $S^2 = 680$ мкм^2. Нужно ли проводить наладку станка, если: а) уровень значимости $\alpha = 0{,}01$; б) уровень значимости $\alpha = 0{,}1$?

Решение. Требуется проверить нулевую гипотезу H_0: $\sigma^2 = 400$ мкм^2 против альтернативной H_a: $\sigma^2 > 400$ мкм^2 . Величина уровня значимости α определяет ширину критической области. Чем больше α, тем шире критическая область (рис.8.6).

По таблице распределения χ^2, положив $\alpha = 0{,}1$ и $r = 14$, находим $\chi^2_{кр} = 21{,}06$. Если положить $\alpha = 0{,}01$, то $\chi^2_{кр} = 29{,}14$.

Вычислим $\chi^2_{эксп}$:

$$\chi^2_{эксп} = \frac{15 * 680}{400} = 25{,}5.$$

Если принять $\alpha = 0{,}01$, нулевую гипотезу можно считать не противоречащей опытным данным. При $\alpha = 0{,}1$ нулевая гипотеза отвергается.

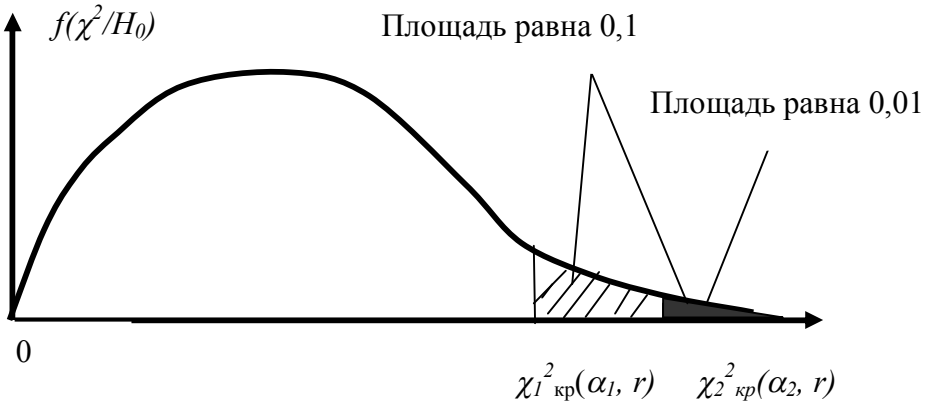


Рис. 8.6

8.2.4. Проверка гипотезы о равенстве дисперсий двух генеральных совокупностей

Имеются две генеральные совокупности, дисперсии которых σ_1^2 и σ_2^2 неизвестны. Из этих генеральных совокупностей извлечены выборки объёмов n_1 и n_2 соответственно. По этим выборкам вычислены исправленные выборочные дисперсии $\widetilde{S}_1^{\,2}$ и $\widetilde{S}_2^{\,2}$. Для определенности будем полагать, что $\widetilde{S}_1^{\,2} > \widetilde{S}_2^{\,2}$. Требуется проверить нулевую гипотезу H_0: $\sigma_1^2 = \sigma_2^2$ о равенстве дисперсий против альтернативной H_a: $\sigma_1^2 > \sigma_2^2$.

Для проверки нулевой гипотезы используют статистику

$$F = \widetilde{S}_1^{\,2} / \widetilde{S}_2^{\,2}.$$

Если нулевая гипотеза верна, то случайная величина F распределена по

так называемому закону Фишера. График функции плотности вероятности распределения Фишера показан на рис. 8.7. Распределение Фишера зависит от двух параметров, которые обозначаются r_1 и r_2 и называются числом степеней свободы большей и меньшей дисперсии соответственно. В нашем случае $r_1 = n_1 - 1$, $r_2 = n_2 - 1$. Зная r_1 и r_2, по таблице распределения Фишера можно найти число $F_{кр}$ и сравнить его с найденным по выборкам.

Числа в таблице Фишера больше 1, поэтому критическая область всегда правосторонняя, а при вычислении экспериментального значения F большую дисперсию делят на меньшую, чтобы $F_{кр}$ получилось больше 1. (Р. Фишер (1890-1968) – выдающийся английский статистик, создатель дисперсионного анализа. Труды Р. Фишера способствовали развитию теории проверки статистических гипотез и многомерного статистического анализа).

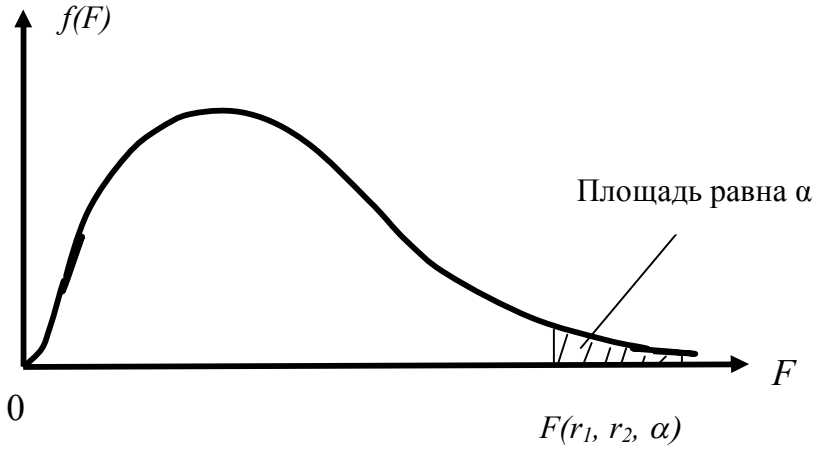


Рис. 8.7

8.2.5. Проверка гипотезы о значении коэффициента корреляции ρ

Из двумерной нормально распределенной генеральной совокупности извлечена двумерная выборка объёма n. Требуется проверить гипотезу о том, что коэффициент корреляции между составляющими двумерной генеральной совокупности равен числу ρ_0.

Для проверки гипотезы H_0: $\rho = \rho_0$ используют следующую статистику (преобразование Фишера): $z = 0{,}5 \ln\left(\frac{1+r}{1-r}\right)$, где r –выборочный коэффициент корреляции.

Если нулевая гипотеза верна, то случайная величина z имеет распределение, близкое к нормальному. Параметры этого распределения таковы:

$$M(z) = 0{,}5 \ln\left(\frac{1+\rho_0}{1-\rho_0}\right); \quad D(z) = 1/(n-3).$$

Чем больше объём выборки n, тем ближе распределение статистики z к нормальному распределению. Практическая рекомендация – $n \geq 10$.

Пример. Пусть $n = 40$; $r = 0{,}8$; $\alpha = 0{,}05$. Проверить нулевую гипотезу H_0: $\rho = 0{,}6$ против альтернативной H_a: $\rho > 0{,}6$. При каких значениях n нулевая гипотеза будет принята? При каких значениях r нулевая гипотеза будет отвергнута?

Решение.

1. Вычислим $z_{эксп}$: $\quad z = 0{,}5 \ln\left(\frac{1+0{,}8}{1-0{,}8}\right) \approx 1{,}1$.

Если нулевая гипотеза справедлива, то случайная величина z имеет такие математическое ожидание и дисперсию:

$$M(z) = 0{,}5 \ln\left(\frac{1+0{,}6}{1-0{,}6}\right) \approx 0{,}69; \quad D(z) = \frac{1}{40-3} \approx 0{,}027.$$

Тогда $\sigma(z) \approx 0{,}16$.

Критическая область в нашем случае – правосторонняя, число $z_{кр}$ определяется из условия $P(z_{кр} < z < \infty) = 0{,}05$.

Но $P(z_{кр} < z < \infty) = \Phi(\infty) - \Phi\left(\frac{z_{кр} - M(z)}{\sigma(z)}\right) = 1{,}65 - \Phi\left(\frac{z_{кр} - 0{,}69}{0{,}16}\right)$.

Отсюда $\Phi\left(\frac{z_{кр} - 0{,}69}{0{,}16}\right) = 0{,}45 \Rightarrow \frac{z_{кр} - 0{,}69}{0{,}16} = 1{,}65 \Rightarrow z_{кр} = 0{,}954$, $z_{кр} < 1{,}1$.

Следует считать, что коэффициент корреляции ρ больше, чем 0,6.

2. Чтобы нулевая гипотеза была принята, должно быть $z_{кр} > z_{эксп} = 1{,}1$.

$$z_{кр} = M(z) + 1{,}65\sigma(z) = 0{,}69 + \frac{1{,}65}{\sqrt{n-3}}.$$

Решая неравенство $0{,}69 + \frac{1{,}65}{\sqrt{n-3}} > 1{,}1$, получаем, что $10 \leq n < 19$. Левая граница ($n \leq 10$) указана, чтобы подчеркнуть, что гипотезу не проверяют при малых n.

3. Чтобы нулевая гипотеза была отвергнута, должно быть $z_{кр} = 0{,}954 < z_{эксп}$.

Но $z_{эксп} = 0{,}5 \ln\left(\frac{1+r}{1-r}\right)$.

Решая неравенство $0{,}5 \ln\left(\frac{1+r}{1-r}\right) > 0{,}954$, получаем, что $0{,}74 \le r \le 1$. Правая граница ($r \le 1$) указана, чтобы подчеркнуть, что значения выборочного коэффициента корреляции всегда не превосходят числа 1.

Преобразование Фишера используют также для проверки нулевой гипотезы о равенстве двух коэффициентов корреляции. Нулевая гипотеза H_0: $\rho_1 = \rho_2$ проверяется при помощи статистики

$$z = \left[0{,}5 \ln\left(\frac{1+r_1}{1-r_1}\right) - 0{,}5 \ln\left(\frac{1+r_2}{1-r_2}\right) \right],$$

которую, в случае справедливости нулевой гипотезы, можно считать нормально распределенной со средним $M(z) = 0$ и дисперсией

$$D(z) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}, \quad (n_1, n_2 \ge 10).$$

8.3. ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Мы рассмотрим процедуры проверки следующих непараметрических гипотез: о законе распределения генеральной совокупности, об извлечении двух выборок из одной и той же генеральной совокупности, о независимости двух случайных величин.

8.3.1. Проверка гипотезы о законе распределения генеральной совокупности по критерию Колмогорова ─ Смирнова (λ - критерию)

Критерий Колмогорова ─ Смирнова используют только в случае непрерывно распределенных генеральных совокупностей. Кроме того, объём выборки n должен быть достаточно большим ($n \ge 50$), а значения параметров закона распределения должны быть оценены заранее (например, по параллельной выборке). Если определять параметры закона по той же выборке, по которой вычисляется экспериментальное значение критерия, он покажет хорошее согласование с теоретическим распределением, даже если это распределение подобрано неточно. Поэтому, когда все-таки приходится использовать только одну выборку,

уровень значимости α берется достаточно большим (0,1 – 0,2), чтобы критическая область была широкой.

С помощью критерия Колмогорова ─ Смирнова проверяют, совпадает ли функция распределения генеральной совокупности с предлагаемой теоретической функцией $F(x)$. Для этого сравнивают значения функции $F(x)$ с вычисленными по выборке относительными накопленными частотами $\nu_{xi}^{нак}$, где x_i – границы интервалов, по которым сгруппирована выборка. Если закон распределения генеральной совокупности указан правильно, то случайная величина

$\lambda = \max_i \left| F(x_i) - \nu_{x_i}^{нак} \right| \sqrt{n}$ при $n \to \infty$ имеет функцию распределения

$$K(\lambda_0) = P(\lambda < \lambda_0) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda_0^2} .$$

Зададим уровень значимости α. Из соотношения $P(\lambda \geq \lambda_{кр}) = 1 - P(\lambda < \lambda_{кр}) = 1 - K(\lambda_{кр})$ можно найти число $\lambda_{кр}$. Вот некоторые значения $\lambda_{кр}$:

α	0,2	0,1	0,05	0,02	0,01
$\lambda_{кр}$	1,073	1,224	1,358	1,520	1,627

Сравним число $\lambda_{эксп}$, определенное по выборке, с числом $\lambda_{кр}$, определённым по заданному значению уровня значимости α. Если $\lambda_{эксп} > \lambda_{кр}$, проверяемая гипотеза отклоняется, если $\lambda_{эксп} < \lambda_{кр}$, то считается, что указанная функция распределения генеральной совокупности согласуется с опытными данными.

Пример. Вернемся к параграфу 6.3.1. Проверим гипотезу о нормальном законе распределения отклонений диаметра вала от номинального размера. Функция распределения нормального закона с параметрами α и σ такова:

$$F = 0{,}5 + \Phi\left(\frac{x-a}{\sigma}\right).$$

Определим теперь $\lambda_{эксп}$. Вычисления сведем в табл. 8.1. Максимальное значение разности $| F(x_i) - \nu_{xi}^{нак} |$ выделено стрелкой. Напомним, что значения a и σ были оценены по выборке и равны -0,03 и 0,05 соответственно.

$$\lambda_{эксп} = 0{,}042 * \sqrt{200} = 0{,}6.$$

Положим $\alpha = 0{,}1$, тогда $\lambda_{кр} = 1{,}224$. Так как $\lambda_{кр} > \lambda_{эксп}$, гипотезу о нормальном распределении отклонений диаметра вала от номинального размера можно считать не противоречащей опытным данным.

Таблица 8.1

x_i	$\frac{x-a}{\sigma}$	$\Phi\left(\frac{x-a}{\sigma}\right)$	$F(x_i)$	$\nu_{xi}^{нак}$	$\lvert F(x_i) - \nu_{xi}^{нак} \rvert$
-0,15	-2,4	-0,492	0,008	0	0,008
-0,13	-2	-0,477	0,023	0,015	0,008
-0,11	-1,6	-0,445	0,055	0,055	0,000
-0,09	-1,2	-0,387	0,113	0,110	0,003
-0,07	-0,8	-0,288	0,212	0,210	0,002
-0,05	-0,4	-,0155	0,345	0,345	0,000
-0,03	0	0,000	0,500	0,525	0,025
-0,01	0,4	0,155	0,655	0,670	0,015
0,01	0,8	0,288	0,788	0,760	0,028
0,03	1,2	0,387	0,887	0,845	0,042←
0,05	1,6	0,445	0,945	0,930	0,015
0,07	2	0,477	0,977	0,970	0,007
0,09	2,4	0,492	0,992	0,990	0,002
0,11	2,8	0,497	0,997	0,995	0,002
0,13	3,2	0,499	0,999	1,000	0,001

8.3.2. Проверка гипотезы об извлечении двух выборок из одной и той же генеральной совокупности

Пусть имеются две выборки объёмов n_1 и n_2 соответственно. Требуется проверить нулевую гипотезу о том, что обе выборки извлечены из одной и той же генеральной совокупности.

8.3.2.1. Проверка по λ - критерию

Предположим, что $n_1, n_2 \geq 50$ и обе выборки сгруппированы по одним и тем же интервалам. Проверяется утверждение, что функции распределения $F_1(x)$ и $F_2(x)$ генеральных совокупностей, из которых извлечены выборки, тождественны. Для этого используют выборочную статистику

$$\lambda = \max_i \left| \nu_{x_i}^{1нак} - \nu_{x_i}^{2нак} \right| \sqrt{n},$$

где $n = n_1 n_2 / (n_1 + n_2)$; $\nu_{xi}^{1\ нак}$, $\nu_{xi}^{2\ нак}$ – относительные накопленные частоты, вычисленные для границы x_i по первой и второй выборкам соответственно.

Если нулевая гипотеза верна, то при $n \to \infty$ выборочная статистика λ имеет распределение Колмогорова.

8.3.2.2. Проверка по критерию Вилкоксона

Обозначим варианты первой выборки через x_i, $i = 1, 2, ..., n_i$; варианты второй выборки – через y_j, $j = 1, 2, ..., n_j$. Для проверки нулевой гипотезы о равенстве законов распределения генеральных совокупностей, из которых извлечены выборки, применяется следующая процедура. Варианты выборок располагают в общую неубывающую последовательность, например,

$$x_1 \quad y_1 \quad y_2 \quad x_2 \quad y_3 \quad x_3 \quad x_4 \quad y_4 \quad x_5 \quad x_6 \quad y_5,$$

где $x_1, \ldots, x_6$ – варианты первой выборки; $y_1, ..., y_5$ – второй.

Говорят, что варианты x_i, y_j образуют инверсию, если $x_i > y_j$. В нашем примере варианта x_1 не образует ни одной инверсии с элементами второй выборки; варианта x_2 образует две инверсии (с числами y_1, y_2); варианты x_3, x_4 образуют по три инверсии каждая (с y_1, y_2, y_3); варианты x_5 и x_6 образуют по четыре инверсии (с y_1, y_2, y_3, y_4). Общее число инверсий

$$u = 0 + 2 + 3 + 3 + 4 + 4 = 16.$$

Можно показать, что при n_1, $n_2 \geq 10$ и справедливости нулевой гипотезы случайная величина u – число инверсий – распределена приближенно нормально с параметрами

$$M(u) = 0{,}5\, n_1 n_2; \quad D(u) = n_1 n_2 (n_1 + n_2 + 1) / 12.$$

Критическую область в данном случае разумно выбрать двустороннюю, так как и слишком маленькое, и слишком большое число инверсий равно свидетельствуют о несовпадении законов распределения генеральных совокупностей.

Пример. Можно ли утверждать, что данные выборки извлечены из одной генеральной совокупности?

1-я выборка	50	41	48	60	46	60	51	42	62	54	42	46
2-я выборка	38	40	47	51	63	50	63	57	59	51	-	-

Принять $\alpha = 0{,}1$.

Решение. Воспользуемся критерием Вилкоксона. Сольем две наши выборки в одну, расположив варианты в порядке возрастания. Элементы второй выборки подчеркнем:

38, 40, 42, 42, 42, 46, 46, 47, 48, 50, 50, 51, 51, 51, 54, 57, 59, 60, 60, 62, 63, 63.

Одинаковые варианты из разных выборок будем чередовать. Число инверсий

$$u_{\text{эксп}} = 2 + 2 + 2 + 2 + 2 + 3 + 4 + 5 + 6 + 8 + 8 + 8 = 52.$$

Найдем $u_{кр}$. Если H_0 верна, то $M(u) = 0{,}5n_1n_2 = 0{,}5*10*12 = 60$; $D(u) = n_1n_2*$ $*(n_1 + n_2 + 1)/12 = 10*12*23/12 = 230$; $\sigma(u) = \sqrt{230} = 15{,}17$.

Критическая область и область принятия нулевой гипотезы показаны на рис.8.8.

Границы критической области определяются из условия

$$P(\,|\,u - M(u)\,|\,< \varepsilon\,) = 1\text{-}\alpha = 0{,}9,$$

где ε – половина ширины области принятия нулевой гипотезы.

Отсюда: $2\Phi\left(\dfrac{\varepsilon}{\sigma(u)}\right) = 2\Phi\left(\dfrac{\varepsilon}{15{,}17}\right) = 0{,}9$; $\varepsilon = 1{,}65*15{,}17 = 25$;

$$u^1_{кр} = 60 - 25 = 35; \qquad u^2_{кр} = 60 + 25 = 85.$$

Экспериментальное значение критерия u, равное 52, лежит внутри интервала $(u^1_{кр}, u^2_{кр})$. Можно считать, что обе выборки извлечены из одной и той же генеральной совокупности.

Заметим, что вследствие того, что критерий Вилкоксона универсален, его можно применять для любых законов распределения генеральных совокупностей, он одновременно достаточно груб. Точнее говоря, велика вероятность ошибки второго рода.

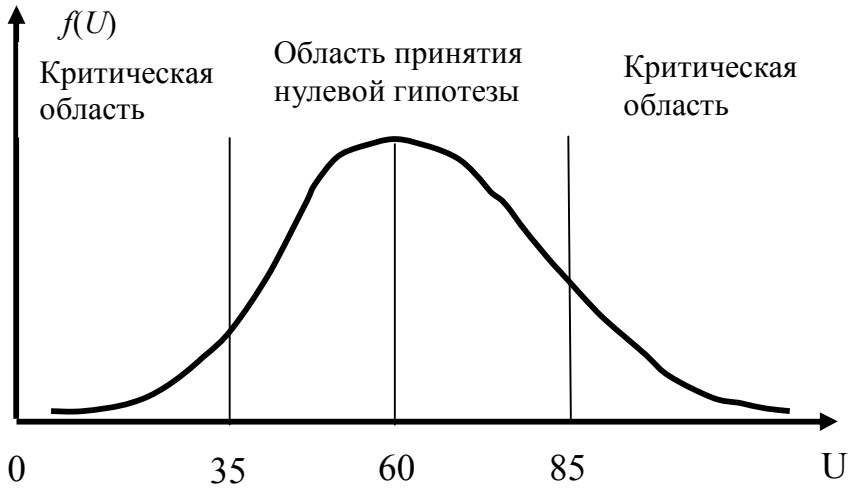


Рис. 8.8

Пользуясь критерием Вилкоксона, можно принять нулевую гипотезу даже в случае очевидно разных генеральных совокупностей.

Пример. Пусть выборки таковы:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-	-	-
y	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8

Решение. Ясно, что они извлечены из разных генеральных совокупностей. Посмотрим, что даст применение критерия Вилкоксона.

$$u_{эксп} = 15*6 = 90;\ M(u) = 12*15/2 = 90;$$
$$D(u) = 12*15*28/12 = 420;\ \sigma(u) = 20{,}5.$$

Так как $u_{эксп}$ просто совпадает с математическим ожиданием $M(u)$, ширина области принятия гипотезы H_0 не имеет значения. Экспериментальное значение попадает в самую середину этой области, а нулевую гипотезу следует считать верной, хотя она очевидным образом неверна.

8.3.2.3. Критерий знаков

Этот критерий применяют в случае попарно связанных выборок. Такая ситуация возникает, например, когда у n объектов измеряется некоторый параметр двумя приборами. Тогда для 1-го объекта имеем два результата: x_i – показание первого прибора; y_i – показание второго прибора, i= 1,2,…,n. Нужно проверить нулевую гипотезу о тождественности законов распределения случайных величин X и Y – ошибок измерения – при использовании первого и второго приборов.

Пусть значения x_i, y_i извлечены из одной генеральной совокупности. Если эта генеральная совокупность распределена по непрерывному закону, то

$$P(x_i > y_i) = P(x_i < y_i) = 0{,}5; \quad P(x_i = y_i) = 0,\ i = 1, 2, 3, \ldots, n.$$

Событие $\{ x_i > y_i \}$ обозначим знаком «+»; событие $\{ x_i < y_i \}$ – знаком «-». В силу сделанных предположений случайная величина z - число появлений знака «+» в n независимых испытаниях - имеет биномиальное распределение, причем вероятность появления «успеха» $P = 0{,}5$.

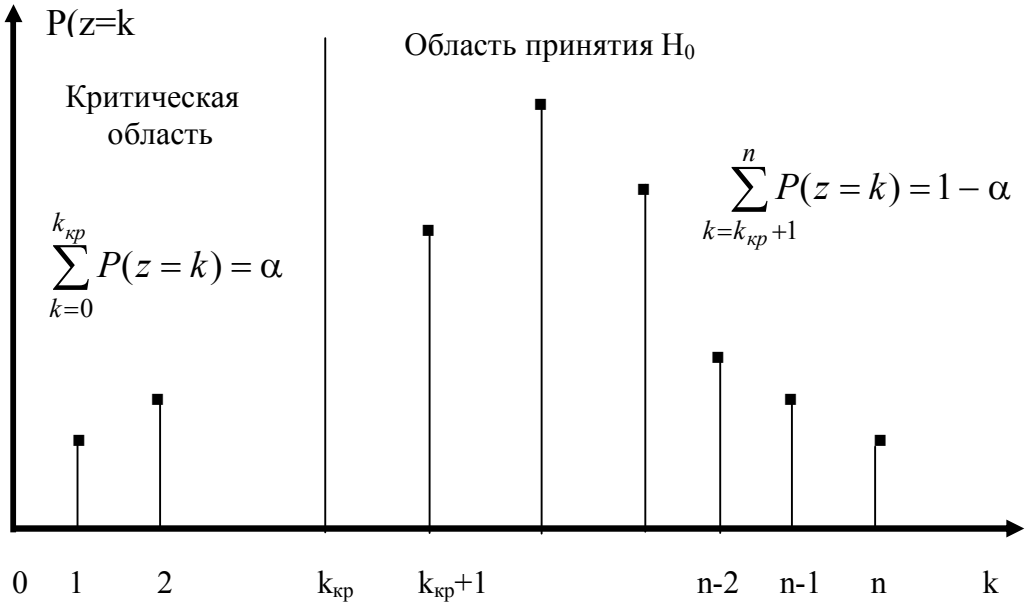
Задача сводится к проверке нулевой гипотезы H_0: $P = 0{,}5$ против одной из альтернативных ($H_a^1 : P < 0{,}5$; $H_a^2 : P > 0{,}5$; $H_a^3 : P \neq 0{,}5$).Если верна нулевая гипотеза, то случайная величина z принимает значения 0, 1, ... , n с вероятностями $P(z = k) = C^k_n (0{,}5)^n,\ k = 0, 1, \ldots n$.

Критическая область и область принятия гипотезы H_0 в случае альтернативной гипотезы P < 0,5 показаны на рис. 8.9.

Замечание. Вследствие ошибок округления может возникнуть ситуация, когда $x_i = y_i$. Такие пары просто исключаются из рассмотрения, соответственно уменьшается объём выборок.

Пример. Предполагается, что один из двух приборов, определяющих скорость автомобиля, систематически завышает её. Для проверки этого положения определили скорость 10 автомобилей, причем скорость каждого фиксировалась одновременно двумя приборами. Получены следующие данные:

X_i, км\ч	70	85	63	54	65	80	75	95	52	55
Y_i, км\ч	72	86	62	55	63	80	78	90	53	57



H_a: P< 0,5.

Рис. 8.9

Завышает ли второй прибор значения скорости? Принять α = 0,1.

Решение. Применим критерий знаков, считая, что показания приборов не зависят друг от друга. Так как один раз показания приборов совпали, этот случай не рассматривается. Объём выборки $n = 9$, причем показания первого прибора три раза ($k = 3$) были больше показаний второго и 6 раз оказались меньше. Проверяется нулевая гипотеза $H_0 : P = 0{,}5$ против альтернативной $H_a : P < 0{,}5$. В предложении справедливости H_0 вычислим несколько первых вероятностей $P(z = k)$ для $k = 0, 1, 2, \ldots$:

$$P(z = 0) = C^0_9(0{,}5)^9 = 0{,}002; \qquad P(z = 1) = C^1_9(0{,}5)^9 = 0{,}018;$$

$$P(z = 2) = C^2_9(0{,}5)^9 = 0{,}070; \qquad P(z = 3) = C^3_9(0{,}5)^9 = 0{,}164.$$

Таким образом, $\sum_{k=0}^{2} P(z = k) = 0{,}09$, $\sum_{k=0}^{3} P(z = k) = 0{,}25$. Критическим значением следует признать число $k = 2$. Так как экспериментальное значение статистики z равно 3, гипотеза H_0 не противоречит результатам наблюдений. Различия в показаниях приборов вызваны случайными ошибками.

Мощность критерия знаков, так же как и критерия Вилкоксона, не велика. Вычислим, например, вероятность ошибки второго рода в предположении, что второй прибор всё-таки завышает истинное значение скорости. Пусть вероятность события $\{x_i > y_i\}$ равна 0,4. Если объём выборки $n = 9$, то неверная нулевая гипотеза $H_0 : P = 0{,}5$ будет принята, если показания первого прибора не менее трёх раз превзойдут показания второго. Вероятность этого события:

$P(z \geq 3) = 1 - (\ P(z < 3) = 1 - (P(z = 2) + P(z = 1) + P(z = 0))$.

$P(z = 2) = C^2_9 * (0{,}4)^2 * (0{,}6)^7 = 0{,}16$; $P(z = 1) = C^1_9 * (0{,}4)^1 * (0{,}6)^8 = 0{,}06$.

$P(z = 0) = C^0_9 * (0{,}4)^0 * (0{,}6)^9 = 0{,}01$.

Тогда $P(z \geq 3) = 1 - 0{,}23 = 0{,}77$.

В 77 % случаев критерий знаков "ошибается", считая, что различие в показаниях приборов случайно.

8.3.3. Проверка гипотезы о независимости двух дискретных случайных величин

Пусть X и Y – две дискретные случайные величины, причём X принимает k разных значений $x_1, x_2, \ldots, x_k$ с вероятностями $p_1, p_2, \ldots, p_k$ соответственно, а Y принимает l различных значений $y_1, y_2, \ldots, y_l$ с вероятностями $q_1, q_2, \ldots, q_l$ соответственно.

Случайные величины X и Y называются независимыми тогда и только тогда, когда справедливо соотношение

$$P(X = x_i,\ Y = y_j) = p_i q_j,\ i = l, 2, \ldots, k;\ j = 1, 2, \ldots, l.$$

Требуется описать процедуру проверки нулевой гипотезы о независимости случайных величин X и Y.

Далее мы будем достаточно широко трактовать понятие "значения" случайной величины. Как и в современных алгоритмических языках, под значением мы будем понимать не только число, но и, например, символьную строку вида "да", "нет", "одобряю" и т.п. У случайной величины, принимающей подобные "значения", нет, конечно, числовых характеристик, как нет и функции распределения. Но для нас важно только наличие "закона распределения": перечня "значений" и соответствующих им вероятностей.

Опишем выборку, на основании которой осуществляется проверка. Итог каждого эксперимента - пара (x_i, y_j), где x_i – значение случайной величины X, которое она приняла в результате этого эксперимента; y_j – значение, принятое случайной величиной Y. Выборка объёма n состоит из

n таких пар. Если у случайной величины X k разных значений, а у случайной величины Y l разных значений, всего возможно $k * l$ разных сочетаний вида (x_i, y_j). Обозначим частоту каждого такого сочетания через n_{ij} . Одновременно обозначим через n_i частоту значения x_i (сколько раз в n экспериментах случайная величина X приняла значение x_i), через m_j − частоту значения y_j, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, l$.

Ясно, что $\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n$; $\sum_{i=1}^{k} n_i = n$; $\sum_{j=1}^{l} m_j = n$;

$\sum_{j=1}^{l} n_{ij} = n_i$; $\sum_{i=1}^{k} n_{ij} = m_j$.

Результаты n экспериментов можно представить в виде так называемой таблицы сопряженности признаков размера $k \times l$.

	y_1	y_2	……	y_l	$\sum_{j=1}^{l} n_{ij} = n_i$
x_1	n_{11}	n_{12}	……	n_{1l}	n_1
x_2	n_{21}	n_{22}	……	n_{2l}	n_2
……	……	……	……	……	……
x_k	n_{k1}	n_{k2}	……	n_{kl}	n_k
$\sum_{i=1}^{k} n_{ij} = m_j$	m_1	m_2	……	m_l	$\sum_{i=1}^{k} n_i = n$; $\sum_{j=1}^{l} m_j = n$;

Если гипотеза H_0 верна, вероятность каждой пары (x_i, y_j) равна произведению $p_i q_j$, а математическое ожидание числа появлений пары (x_i, y_j) в n независимых экспериментах равно произведению np_iq_j.

Тогда случайную величину

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - np_i q_j\right)^2}{np_i q_j}$$

(при условии, что H_0 верна, а все математические ожидания $np_iq_j \geq 4$, $i = 1, 2, ..., k$, $j = 1, 2, ..., l$) можно считать распределенной по закону χ^2 с $(k - 1)(l - 1)$ степенями свободы. Зная уровень значимости α и число степеней свободы, можно найти $\chi_{кр}^2$ и сравнить его с числом $\chi_{эксп}^2$, определённым по выборке. Если $\chi_{кр}^2 > \chi_{эксп}^2$, гипотеза H_0 о независимости случайных величин принимается, иначе H_0 отклоняется.

Несколько замечаний.

1. Вероятности p_i, q_j обычно неизвестны. Они оцениваются по выборке.

В качестве значения p_i берется число n_i / n, $i = 1, 2,..., k$, вместо q_j берется число m_j / n, $j = 1, 2, ..., l$.

2. Если числа $np_iq_j < 4$, то соответствующие строки и столбцы должны быть объединены с соседними строками и столбцами.

3. Если $(k – 1)(l – 1) \geq 8$ и $n \geq 40$, то минимально допустимое значение ожидаемых частот может быть равным единице.

4. Формулу, по которой вычисляется $\chi_{эксп}^2$, можно упростить. Если вероятности p_i, q_j оцениваются по выборке, то $p_i = n_i / n$, $q_j = m_j / n$, тогда

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - n\frac{n_i}{n}\frac{m_j}{n}\right)^2}{n\frac{n_i}{n}\frac{m_j}{n}} = n\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{n_{ij}^2}{n_i m_j} - \frac{2n_{ij}n_i m_j}{n n_i m_j} + \frac{n_i^2 m_j^2}{n^2 n_i m_j}\right) =$$

$$= \left(n\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_i m_j}\right) - n, \text{ так как } \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n; \qquad \sum_{i=1}^{k}\sum_{j=1}^{l} n_i m_j = n^2.$$

Пример. Утверждается, что результат действия лекарства зависит от способа его применения. Проверить это утверждение при $\alpha = 0{,}05$ по следующим данным:

Результат	Способ применения		
	A	B	C
Неблагоприятный	11	17	16
Благоприятный	20	23	19

Решение. Вычислим экспериментальное значение критерия χ^2.

$n = 11 + 17 + 16 + 20 + 23 + 19 = 106$; $n_1 = 11 + 17 + 16 = 44$;

$n_2 = 20 + 23 + 19 = 62$; $m_1 = 11 + 20 = 31$; $m_2 = 7 + 23 = 40$; $m_3 = 16+19 = 35$.

В соответствии с выведенной формулой

$$\chi^2 = 106\left(\frac{11^2}{44\cdot 31} + \frac{17^2}{44\cdot 40} + \frac{16^2}{44\cdot 35} + \frac{20^2}{62\cdot 31} + \frac{23^2}{62\cdot 40} + \frac{19^2}{62\cdot 35}\right) - 106 = 0{,}73.$$

Число степеней свободы $r = (2 – 1)(3 – 1) = 2$.

Если $\alpha = 0{,}05$, то $\chi_{кр}^2 = 6 > \chi_{эксп}^2$, нулевая гипотеза не отвергается, результат действия лекарства не зависит от способа его применения.

8.4. РАНГОВАЯ КОРРЕЛЯЦИЯ

Пусть из двумерной генеральной совокупности извлечена выборка (x_i, y_i) объёма n. Упорядочим по возрастанию или убыванию варианты x_i. Каждому значению x_i, $i = 1, 2, ..., n$, поставим в соответствие номер этого значения в упорядоченной последовательности. Этот номер называется рангом варианты x_i. Аналогично ранжируем варианты y_i. Таким образом, каждой паре (x_i, y_i) соответствует пара рангов её элементов. Обозначим эту пару рангов также (x_i, y_i).

Пример. Измерения длины головы (x_i) и длины грудного плавника (y_i) у 10 окуней дали такие результаты (мм):

Таблица 8.2

x_i	66	61	67	73	51	59	48	47	45	44
y_i	38	31	36	43	29	35	28	25	26	23

Определить ранги элементов этой выборки. Решение не требует комментариев. Выборка рангов такова (табл. 8.3)

Таблица 8.3

x_i	8	7	9	10	5	6	4	3	2	1
y_i	9	6	8	10	5	7	4	2	3	1

8.4.1. Коэффициент ранговой корреляции Спирмена

Вычислим теперь коэффициент корреляции по выборке рангов. В этом случае он называется выборочным коэффициентом ранговой корреляции Спирмена (Ч.Спирмен − английский психолог (1863-1945)) и обозначается r_s. Формулу для вычислений

$$r_s = \frac{1/n \sum x_i y_i - \bar{x} \cdot \bar{y}}{S_x S_y}$$

можно упростить. Воспользуемся формулами для суммы первых степеней и квадратов первых n натуральных чисел.

$$1+2+3+...+n = \frac{n(n+1)}{2}; \qquad 1^2+2^2+3^2+...+n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Отсюда:

$$\bar{x} = \bar{y} = \frac{1}{n}\sum_{1}^{n} i = \frac{n-1}{2}; \quad S_x{}^2 = S_y{}^2 = \frac{1}{n}\sum_{1}^{n} i^2 - \left(\frac{n-1}{2}\right)^2 = \frac{n^2-1}{12};$$

$$S_x S_y = \frac{n^2-1}{12}.$$

Далее обозначим через d_i разность $x_i - y_i$.

Так как $\sum_{i=1}^{n}(x_i - y_i)^2 = \sum_{i=1}^{n} x_i{}^2 + \sum_{i=1}^{n} y_i{}^2 - 2\sum_{i=1}^{n} x_i y_i$, то в нашем случае получаем

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y} = \frac{(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i{}^2 - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} - \frac{1}{2n}\sum_{i=1}^{n} d_i{}^2.$$

Окончательно $r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i{}^2$.

Найдём значение r_s для нашего примера. Разности рангов d_i таковы:

d_i	-1	1	1	0	0	-1	0	1	-1	0

$$\sum_{i=1}^{6} d_i{}^2 = 6; \qquad r_s = 1 - \frac{6\cdot 6}{10\cdot 99} = 0{,}964.$$

Мы получили число, очень близкое к единице. Следует считать, что длина головы и длина грудного плавника тесно связаны между собой.

Проверку на значимость выборочного коэффициента ранговой корреляции можно произвести строго. Гипотеза H_0: $|\rho_s| = 0$ при альтернативной гипотезе H_a : $|\rho_s| > 0$ и при объеме выборки $n \geq 10$ проверяется по значению случайной величины

$$T_{эксп} = |r_s|\sqrt{\frac{n-2}{1-r_s{}^2}}.$$

Если гипотеза H_0 верна, эта статистика имеет распределение Стьюдента с $n - 2$ степенями свободы. Закон распределения исходной двумерной генеральной совокупности не имеет значения, хотя предполагается , что составляющие X и Y генеральной совокупности – непрерывные случайные величины.

В нашем случае

$$T_{эксп} = |0{,}964| \sqrt{\frac{8}{1-0{,}964^2}} = 10{,}25.$$

Если положить $\alpha = 0{,}05$, то $t_{кр} = 1{,}86$ (число степеней свободы $r = 8$, а критическая область - правосторонняя), $t_{кр} < 10{,}25$, гипотеза H_0 отвергается.

При помощи статистики T можно проверять нулевую гипотезу о равенстве нулю коэффициента корреляции ρ двумерной нормально распределённой генеральной совокупности.

8.4.2. Связанные ранги

На практике часты случаи, когда несколько значений x_i (y_i) исходной выборки одинаковы, им нужно приписывать одинаковые ранги. Говорят, что несколько подряд идущих одинаковых значений x_i (y_i) образуют связку, называются такие элементы связанными. Каждый из связанных элементов получает ранг, равный среднему арифметическому рангов, которые имели бы элементы связки, если бы они были различны.

Одинаковые ранги называются связанными рангами (табл. 8.4).

Таблица 8.4

x_i	10	12	10	12	12	15	17
Ранг x_i	1,5	4	1,5	4	4	6	7
y_i	2	4	2	3	7	2	9
Ранг y_i	2	5	2	4	6	2	7

Формула для вычисления коэффициента корреляции Спирмена при наличии связанных рангов становится громоздкой и здесь не приводится. Практика показывает, что использование обычной формулы для r_s, без поправки на связанные ранги, обеспечивает достаточную точность вычислений.

8.4.3. Коэффициент ранговой корреляции Кендэла

М. Кендэл (английский статистик с мировым именем) предложил другой коэффициент ранговой корреляции τ. Он вычисляется по двумерной выборке рангов следующим образом. Столбцы $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ переставляются так, чтобы ранги x_i образовали возрастающую

последовательность 1, 2, …, n. Теперь $x_i = i$. Для каждого ранга y_i обозначим через p_i число рангов $y_к > y_i$, причем $k > i$; через q_i обозначим число рангов $y_к < y_i$, причём $k > i$.

Пусть $P = \sum_i p_i$; $Q = \sum_i q_i$; $S = P - Q$.

Коэффициент τ вычисляется по одной из эквивалентных формул:

$$\tau = \frac{2S}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1.$$

(Нетрудно показать, что $P + Q = \frac{n(n-1)}{2}$).

Число τ лежит в пределах $-1 \le \tau \le 1$, причем $\tau = 1$, если $x_i = y_i$ i=1, 2, …, n; $\tau = -1$, если $x_i + y_i = n + 1$, $i = 1, 2, \ldots, n$.

Пример вычисления τ для рангов из табл. 8.3 приведён в табл. 8.5

Таблица 8.5

№ п\п	Ранги		p_i	q_i
	x_i	y_i		
10	1	1	9	0
9	2	3	7	1
8	3	2	7	0
7	4	4	6	0
5	5	5	5	0
6	6	7	3	1
2	7	6	3	0
1	8	9	1	1
3	9	8	1	0
4	10	10	0	0
Сумма	55	55	$P = 42$	$Q = 3$

$$\tau = \frac{2 \times 39}{90} = 1 - \frac{3 \cdot 4}{90} = \frac{42 \cdot 4}{90} - 1 = 0{,}87.$$

Коэффициенты Спирмена и Кендэла никак не связаны между собой. Обычно $r_s > \tau$, но сравнение этих коэффициентов не дает никакой дополнительной информации о связи между рангами.

Значимость коэффициента ранговой корреляции Кендэла (H_0: в генеральной совокупности $\tau = 0$) при $n \ge 10$ проверяется при помощи статистики

$$\lambda = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}},$$

которую можно считать приближённо нормально распределённой со

средним $M(\lambda) = 0$ и дисперсией $D(\lambda) = 1$ (если верна нулевая гипотеза). Для нашего примера при H_a: $\tau > 0$ и $\alpha = 0{,}05$ по таблице функции Лапласа находим, что $\lambda_{кр} = 1{,}65$. Между тем

$$\lambda_{эксп} = \frac{0{,}867}{\sqrt{\frac{2(2*10+5)}{9*10(10-1)}}} = 3{,}49.$$

Нулевую гипотезу следует отвергнуть.

8.4.4. Коэффициент конкордации Кендэла

Пусть m экспертов независимо один от другого ранжируют n элементов по некоторому признаку. Каждый получает свою выборку рангов, всего выборок m. Обозначим через R_{ij} ранг, приписанный i-му элементу j-м экспертом, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Чтобы оценить, насколько хорошо мнения экспертов согласуются между собой, М.Кендэл определил следующий коэффициент конкордации (согласованности) W:

$$W = \frac{12\sum_{i-1}^{n} D_i^2}{m^2(n^3-n)}, \quad \text{где } D_i = \sum_{j=1}^{m} R_{ij} - \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} R_{ij}}{n}, \; i = 1,2,\ldots, n \text{ — сумма}$$

рангов, приписанных всеми экспертами i-му элементу, минус среднее значение этих сумм рангов.

При наличии связанных рангов коэффициент W вычисляется по формуле

$$W = \frac{12\sum_{i-1}^{n} D_i^2}{m^2(n^3-n) - mB}, \quad \text{где} \quad B = \sum_{k=1}^{z}(B_k^3 - B_k),$$

z - число связок рангов, $B_к$ - число связанных рангов в k-й связке, $k = 1,2, \ldots, z$.

Коэффициент W принимает значения в интервале $0 \le W \le 1$, причем $W = 1$, когда мнения экспертов полностью совпадают, $W = 0$, когда мнения экспертов полностью рассогласованы (никакой элемент не имеет двух одинаковых рангов).

Пример. Три эксперта оценивают качество шести изделий ($m = 3$, $n = 6$).Результаты оценки и дальнейшие расчёты приведены в табл. 8.6.

$$\frac{\sum_{j=1}^{3}\sum_{i}^{6} R_{ij}}{6} = \frac{63}{6} = 10{,}5; \; B = (2^3 - 2) + (3^3 - 3) = 30,$$

$$W = \frac{12 * 133}{3^2(6^3 - 6) - 3 * 30} = 0{,}887 .$$

Значение $W = 0{,}887$ близко к 1, мнения экспертов хорошо согласуются. Для проверки значимости коэффициента конкордации W используют статистику

$$\chi^2 = m(n-1)W = \frac{12\sum_{i=1}^{n} D_i^2}{mn(n+1) - B/(n-1)} .$$

Если справедлива нулевая гипотеза, статистика χ^2 имеет распределение χ^2 с $r = n$ - 1 степенями свободы.

Таблица 8.6

Изделие i	Эксперт j			Сумма рангов		
	1	2	3	$\sum_{j=1}^{3} R_{ij}$	D_i	D_i^2
1	1	2	1	4	-6,5	42,25
2	2	1	3	6	-4,5	20,25
3	4	4,5	3	11,5	1,0	1,00
4	5	4,5	6	15,5	5,0	25,00
5	3	3	3	9	-1,5	2,25
6	6	6	5	17	6,5	42,25
Сумма	21	21	21	63	—	133,00

В нашем случае $\chi^2_{эксп} = 3(6 - 1) \times 0{,}887 = 13{,}3$; $r = 5$, положим $\alpha = 0{,}05$, тогда $\chi^2_{кр} = 11{,}1 < \chi^2_{эксп}$. Нулевая гипотеза отвергается, оценки экспертов можно считать согласованными.

8.5. ЗАДАЧИ

1. Большая партия изделий может содержать некоторую долю дефектных. Поставщик утверждает, что эта доля составляет 5%; покупатель предполагает, что доля дефектных изделий больше. Из партии случайным образом отбираются и проверяются 10 изделий; партия принимается, если при проверке обнаружено не более одного дефектного изделия. В противном случае партия возвращается поставщику. Описать ситуацию в терминах теории проверки статистических гипотез и ответить на следующие вопросы:

а) Каковы нулевая и альтернативная гипотезы?

б) Каков закон распределения теста, применяемого для проверки нулевой гипотезы?

в) Каковы критическая область и область принятия нулевой гипотезы?

г) В чем заключается ошибка первого рода и какова её вероятность?

д) Если доля дефектных изделий в партии на самом деле равна 10%, то какова вероятность ошибки второго рода?

2. В аттракционе на ярмарке аттракционист предъявляет колоду из 10 карт, среди которых, как он утверждает, два туза (нулевая гипотеза). Игрок извлекает одну карту из тщательно перетасованной колоды. Если извлекается туз, игрок получает назад свои деньги плюс призовую сумму. Некто полагает, что в колоде на самом деле тузов не более одного. Для проверки он наблюдает результаты 10 последовательных игр.

а) Описать закон распределения теста, использованного для проверки нулевой гипотезы.

б) Описать область принятия нулевой гипотезы и критическую область, если $\alpha = 0{,}1$.

в) Какова вероятность ошибки второго рода, если в колоде на самом деле всего один туз?

3. Школьная администрация обеспокоена тем обстоятельством, что, когда детям на завтрак подаётся шпинат, они плохо его едят. Продавец замороженного шпината утверждает, что 80 % детей будут есть этот шпинат, который, однако, несколько дороже. Чтобы проверить это предположение, 10 случайно отобранным детям предлагают попробовать новый сорт шпината.

а) Каковы нулевая и альтернативная гипотезы?

б) Каков закон распределения теста, использованного для проверки нулевой гипотезы?

в) Выбрать уровень значимости α и построить область принятия нулевой гипотезы и критическую область. Если 7 детей съедят шпинат, будет ли принята H_0?

г) Найти вероятность ошибки второго рода, если в среднем только половине школьников нравится замороженный шпинат?

4. Бросают 5 монет. Некто подозревает, что в этой пятёрке есть монеты с гербом с двух сторон (двугербовые).

Известно, что монета может быть либо правильной, либо двугербовой. Предлагается следующая процедура проверки нулевой гипотезы о том, что все монеты правильные: нулевая гипотеза принимается, если при одновременном бросании этих пяти монет менее четырёх выпадут гербом кверху. Найти вероятность ошибки первого рода. Какова вероятность ошибки второго рода, если из пяти монет две двугербовые?

5. Ответить на вопросы задачи 4, если процедура проверки нулевой гипотезы такова: из данных пяти монет случайным образом выбирается одна и затем подбрасывается три раза. Нулевая гипотеза отвергается, если герб выпадает все три раза.

6. В эксперименте с дегустированием кофе каждому испытуемому предлагается попробовать кофе в 10 парах чашек. В каждой паре одна чашка содержит кофе первого сорта, одна – второго. Испытуемый должен определить, в какой чашке находится кофе первого сорта. Если он правильно определяет сорт кофе не менее 8 раз из 10, его приглашают на работу в качестве дегустатора кофе. Каковы шансы получить работу случайно угадывающему человеку? Какой должна быть нулевая гипотеза относительно вероятности P правильно угадать сорт кофе, чтобы уровень значимости предложенного критерия не превосходил 0,05?

7. В урне содержатся неотличимые на ощупь черные и белые шары. Предполагается, что число чёрных шаров равно числу белых. Эта гипотеза принимается, если при извлечении 50 шаров (с возвращением) число черных будет в пределах от 20 до 30.

Какова вероятность ошибки первого рода? Какова вероятность ошибки второго рода, если вероятность появления чёрного шара равна 1/3?

8. Из продукции автомата, обрабатывающего болты с номинальным значением контролируемого размера a = 40 мм, была взята выборка болтов объёма n = 36. Выборочное среднее контролируемого размера $\overline{x}$ = 40,2 мм. Результаты предыдущих измерений дают основание предполагать, что действительные размеры болтов имеют нормальное распределение с дисперсией σ^2 = 1 мм^2. Можно ли по результатам приведенного выборочного обследования утверждать, что контролируемый размер не имеет положительного смещения по отношению к номинальному размеру? Принять α = 0,01.

9. Пусть в условиях задачи 8 партия болтов бракуется, если выборочное среднее контролируемого размера будет больше 40,1 мм. Выполнить следующие задания:

а) Найти вероятность ошибки первого рода, если решение принимается по выборке объёма n = 36.

б) Какова вероятность ошибки второго рода, если a = 40,3мм. Объём выборки n = 36.

10. Решить задачу 9, если партия болтов бракуется при выполнении одного из неравенств: $\overline{x} > 40,1$ мм и $\overline{x} < 39,9$ мм, где $\overline{x}$ – выборочное среднее контролируемого размера.

11. Длительное наблюдение за производственным процессом изготовления химической смеси привело к заключению, что величину выпуска можно считать нормально распределённой случайной величиной со средним значением 250 единиц в единицу времени и средним квадратическим отклонением 50 единиц. Производственный процесс был модифицирован, причём модификация не изменила величину среднего квадратического отклонения. При уровне значимости α = 0,05 проверяется нулевая гипотеза о том, что среднее значение выпуска продукции не

изменилось против альтернативной, что выпуск продукции увеличился. Выполнить следующие задания:

а) Найти критическую область при 25 наблюдениях.

б) Известно, что число наблюдений n = 25, а модификация процесса привела к увеличению среднего выпуска продукции до 260 единиц в единицу времени. Найти вероятность отвергнуть нулевую гипотезу.

в) Найти такое число наблюдений, которое необходимо взять, чтобы вероятность отвергнуть нулевую гипотезу при условии, что произошло увеличение среднего выпуска продукции до 260 единиц, была бы не меньше 0,9.

12. Торговый контролер, в задачу которого входит контроль за правильной массой отдельных товаров, отобрал 10 однофунтовых пакетов с кофе и взвесил содержимое каждого из них. Он получил такие результаты (фунты): 0,94; 0,95; 0,92; 1,02; 0,97; 0,95; 1,02; 0,96; 0,92; 0,97. Если окажется, что средняя масса пакета меньше 1, то контролер должен сделать замечание магазину. Будет ли сделано замечание? Положить α = 0,05.

13. Торговый инспектор отобрал в магазине 10 однофунтовых пакетов с кофе и взвесил их содержимое. Результаты (фунты) таковы: 0,89; 0,90; 0,87; 0,97; 0,92; 0,90; 0,97; 0,91; 0,87; 0,92. При уровне значимости $\alpha = 0{,}05$ проверить нулевые гипотезы H_0: a = 1 и H_{0-}: a = 0,95 при альтернативных H_a: $a < 1$ и H_a: $a < 0{,}95$ соответственно (a - средняя масса пакета с кофе).С доверительной вероятностью β = 0,95 оценить, сколько фунтов кофе понадобится владельцу магазина, чтобы наполнить 1000 пакетов, если он будет продолжать делать это так же, как и раньше.

14. Количество бракованных изделий в партии не должно превышать 5%. Из этой партии было случайным образом отобрано 100 изделий, 6 штук оказались бракованными. Можно ли считать, что процент брака превосходит допустимый, если $\alpha = 0{,}01$?

15. При проведении телепатического опыта индуктор независимо от предшествующих проб выбирает один из двух предметов и думает о нем, а реципиент пытается угадать мысли индуктора. Опыт был повторён 100 раз, причем было получено 60 правильных ответов. Можно ли приписать полученный результат чисто случайному совпадению?

16. В случайной выборке, состоящей из 900 темноволосых людей, 150 человек имеют голубые глаза. Доля голубоглазых людей среди всего населения равна 0,25. Проверить гипотезу о том, что доля темноволосых людей с голубыми глазами меньше, чем доля голубоглазых среди всего населения. Положить $\alpha = 0{,}05$.

17. Известно, что 5% всех застраховавших свою жизнь умирает по достижении 60 лет. В группе из 1000 человек этого возраста, работающих

в строительстве, умерло 68. Проверить гипотезу о том, что застрахованные люди, работающие в строительстве, умирают чаще в 60 лет, чем все остальные застрахованные,

18. Фирма, производящая новый вид лекарства от гриппа, объявила, что оно излечивает от болезни менее чем за 4 дня и его эффективность равна 95%. В случайной выборке из 300 человек, заболевших гриппом и принимающих это лекарство, 272 поправились менее чем за 4 дня. Было ли заявление фирмы состоятельным? Какова вероятность ошибки 2-го рода, если вероятность выздороветь менее чем за 4 дня на самом деле равна 0,9?

19. Пусть уровень значимости $\alpha = 0{,}05$; H_0: $p = 0{,}1$; H_a: $p > 0{,}1$. Найти вероятность ошибки 2-го рода, если на самом деле: а) $p = 0{,}15$; б) $p = 0{,}12$. Объём выборки $n = 100$.

20. Пусть $N_1 = N_2 = 200$; H_0: $p_1 = p_2$; H_a: $p_1 \neq p_2$; $\alpha = 0{,}05$; $\bar{x}_1 = 0{,}4$; $\bar{x}_2 = 0{,}6$. Будет ли принята нулевая гипотеза? Каким окажется ответ, если $\bar{x}_1 = 0{,}48$; $\bar{x}_2 = 0{,}52$?

21. Пусть H_0: $p_1 = p_2$; H_a: $p_1 \neq p_2$; $\alpha = 0{,}05$; $\bar{x}_1 = 0{,}2$; $\bar{x}_2 = 0{,}25$. Найти $n_1 = n_2$, чтобы разница между выборочными средними позволила отвергнуть нулевую гипотезу.

22. На двух аналитических весах, в одном и том же порядке, взвешены 10 проб химических веществ и получены следующие результаты взвешивания (мг):

x_i	25	30	28	50	20	40	32	36	42	38
y_i	28	31	26	52	24	36	33	35	45	40

При уровне значимости 0,01 установить, значимо или незначимо различаются результаты взвешиваний в предположении, что они распределены нормально.

23. Физическая подготовка 9 спортсменов была проверена при поступлении в спортивную школу, а затем после недели тренировок. Итоги проверки в баллах оказались следующими (в первой строке указано число баллов, полученных каждым спортсменом при поступлении в школу, во второй − после обучения):

x_i	76	71	57	49	70	69	26	65	59
y_i	81	85	52	52	70	63	33	83	62

При уровне значимости 0,05 установить, значимо или незначимо улучшилась физическая подготовка спортсменов. Считать, что балл, полученный спортсменом, нормально распределенная случайная величина.

24. Новую технику изготовления резиновых трубок проверяют следующим образом. Пару трубок (одна из которых изготовлена по старому методу, а вторая – по новому) обрабатывают различными видами кислот. Данные о сроке службы трубок (годы) следующие:

Тип кислоты	1	2	3	4	5	6	7	8	9	10
Срок службы трубок, изготовленных по новому методу	1,7	4,6	3,7	3,9	2,8	3,1	2,4	4,2	3,6	3,3
Срок службы трубок, изготовленных по старому методу	1,2	4,4	3,1	4,0	2,4	2,7	2,6	3,7	3,5	3,0

Улучшает ли новая технология качество изготовления резиновых трубок? Предполагается, что срок службы резиновых трубок – нормально распределенная случайная величина. Как ответить на поставленный вопрос, если закон распределения срока службы неизвестен? Положить $\alpha = 0{,}05$.

25. Для проведения сельскохозяйственного эксперимента было выбрано небольшое поле из 25 участков. Каждый участок был разделён на 2 равные части; случайным образом выбиралась одна половина и засеивалась первым сортом пшеницы, другая часть засеивалась вторым сортом. Урожайность измерили и записали разности для каждого участка (урожайность пшеницы второго сорта минус урожайность пшеницы первого сорта).Выборочное среднее разностей оказалось равным $3{,}5\times10^3$ $м^3/км^2$, а выборочное среднее квадратическое отклонение равно $0{,}4\times10^3$ $м^3/км^2$. Даёт ли второй сорт пшеницы существенно большую урожайность, чем первый? Положить $\alpha = 0{,}05$ и считать, что урожайность пшеницы − нормально распределенная случайная величина.

26. Исследовалась растяжимость некоторого вида резины после химической обработки. Было отобрано и пронумеровано шесть мотков резины. Каждый отобранный моток был разделён пополам, одна половина подвергалась химической обработке, другая – нет. Затем было измерено растяжение (%) двенадцати кусков резины. Результаты представлены в таблице.

Вид резины	Исследуемый кусок					
	1	2	3	4	5	6
Обработанная	18,1	17,3	19,1	18,4	17,2	16,7
Необработанная	16,3	17,0	18,4	17,6	17,0	16,0

Увеличила ли химическая обработка растяжимость резины? Положить $\alpha = 0{,}05$. Какая случайная величина предполагается нормально распределенной?

27. Приводимая таблица дает среднюю длину яйца кукушки вообще и среднюю длину яйца кукушки, положенного в гнездо определённого вида птиц. Существует ли действительно разница в длине или расхождение носит чисто случайный характер?

Отложенные яйца	Число наблюдений	Средняя длина, мм	Выборочная дисперсия, мм^2
Вообще	1572	22,3	0,9216
1-й вид	91	21,9	0,6241
2-й вид	115	22,4	0,5776
3-й вид	58	22,6	0,7396

28. Таблица дает результаты ряда наблюдений над скорлупой крабов, живущих в глубокой и мелкой воде.

Можно ли приписать получившуюся разницу влиянию среды или ее следует отнести за счет случайности выборочного исследования?

Крабы	Число наблюдений	Средняя длина скорлупы, см	Выборочная дисперсия, см^2
Мелководные	35	8,41	0,0016
Глубоководные	41	8,59	0,0025

29. Автомобилестроительная компания покупает у двух сталелитейных компаний заготовки обойм для шарикоподшипников. Пользуясь приведенными выборочными данными, проверить гипотезу о равенстве средних масс заготовок, $\alpha = 0{,}05$, считая одинаковыми дисперсии масс. Массы даны в граммах.

Сталелитейная компания А	41,6	41,7	41,8	42,2	41,2	40,9	41,9	41,5	41,7	41,8
Сталелитейная компания Б	40,5	41,1	40,9	41,4	41,7	41,8	41,1	40,7	41,2	41,4

Определить $n_1 = n_2$, для которого разница в средней массе в 0,2 г будет обнаружена при уровне значимости $\alpha = 0{,}01$.

30. Для проверки эффекта нового вида лекарства были отобраны две случайные группы заболевших гриппом людей по 60 человек в каждой. Одной группе давали таблетки, не содержащие никакого лекарства, другой группе давали по внешнему виду такие же таблетки, но содержащие новый вид лекарства.

Больные из первой группы выздоровели в среднем через 7 дней, из второй − в среднем через 6 дней. Среднее квадратическое отклонение продолжительности болезни (независимо от того, принимал больной лекарство или нет) можно положить равным двум дням. Ускоряет ли новое лекарство выздоровление?

31. В школе случайным образом отобрали 40 мальчиков и 40 девочек. Каждый из отобранных учеников получил один и тот же тест для оценки умственных способностей. Результаты теста таковы:

Показатель	Мальчики	Девочки
Средний балл	71	76
Выборочная дисперсия	25	36

Проверить гипотезу о том, что девочки в среднем более способны, чем мальчики.

32. Пусть H_0: $a_1 = a_2$; H_a: $a_1 \neq a_2$; $\alpha = 0{,}05$; $\bar{x}_1 = 1{,}5$; $\bar{x}_2 = 1{,}2$; $n_1 = 20$; σ_1=0,2; σ_2=0,1. При каких значениях n_2 нулевая гипотеза будет принята? Отвергнута?

33. В результате длительного хронометража времени сборки узла различными сборщиками установлено, что дисперсия этого времени равна 2 мин^2. Результаты 20 наблюдений за работой новичка таковы:

Время сборки одного узла, мин	56	58	60	62	64
Частота	1	4	10	3	2

Можно ли, при уровне значимости $\alpha = 0{,}05$, считать, что новичок работает ритмично? Распределение какой случайной величины предполагается здесь нормальным?

34. Предлагается определенная процедура проверки коэффициента трения шины по мокрому асфальту. Утверждается, что дисперсия результатов измерений этого коэффициента равна 0,1. Выборочная дисперсия, вычисленная по результатам 25 измерений коэффициента трения, оказалась равной 0,2. Справедливо ли утверждение авторов процедуры? Положить $\alpha = 0{,}05$.

35. При каком значении S^2 можно будет считать, что σ^2 существенно меньше, чем $\sigma_0^2 = 0{,}0015$, если $\alpha = 0{,}05$, $n = 15$?

36. Пусть $\sigma_0^2 = 10$. Каким должен быть объём выборки, чтобы отвергнуть нулевую гипотезу при $S^2 = 15$, $\alpha = 0{,}05$? При $S^2 = 5$, $\alpha = 0{,}05$?

37. Исследование длины и ширины 121 черепа, найденных в Верхнем Египте и относимых к расе, живущей за 8000 лет до нашей эры, показало, что выборочные средние квадратические отклонения длины и ширины черепа равны 5,7 и 4,6 мм соответственно. Те же величины, выведенные на основании обследования 1000 европейцев, оказались равными 6,2 и 5,1 мм. Предполагая, что законы распределения длины и ширины черепа нормальные, выяснить, можно ли считать расхождение случайным. Принять $\alpha = 0{,}05$.

38. Чтобы выяснить, оказывает ли влияние на прочность бетона способ его изготовления, выполнили небольшой эксперимент. Из данной партии сырья взяли 6 выборок, причем выборки были сделаны однородными, насколько это возможно. Затем 6 выборок разделили случайным образом на две группы из трех выборок каждая, и из каждой выборки был сделан пробный куб, причем выборки из второй группы подверглись особой обработке. После 28-дневной выдержки шести пробных кубов определили их сопротивление на сжатие. Результаты таковы:

Предел прочности бетона 1, МПа	29	31,1	28,4
Предел прочности бетона 2, МПа	30,9	31,8	31,8

Проверить гипотезу, что бетон обеих групп одинаково прочен, $\alpha = 0{,}05$.

39. Автомобиль имеет четыре покрышки, каждая наполовину сделана

из резины сорта А, наполовину – из резины сорта В. Можно считать, что все колеса автомобиля находятся в одинаковых условиях. После 10000 км пути покрышки были исследованы на износ, результаты исследования (баллы) приводятся в таблице.

Марка резины	Покрышка			
	1	2	3	4
А	32	40	36	35
В	25	28	27	26

Проверить гипотезу о том, что резина В изнашивается быстрее, чем резина А, $\alpha = 0{,}05$.

В задачах 40 – 53 предполагается, что выборки получены из двумерных нормальных генеральных совокупностей.

40. Выборочный коэффициент корреляции r, вычисленный по выборке объема $n = 33$, равен 0,41. Проверить нулевую гипотезу H_0: $\rho = 0$ против альтернативных : а) H_a: $\rho \neq 0$; б) H_a: $\rho > 0$. Принять $\alpha = 0{,}05$.

41. Пусть H_0: $\rho = 0{,}5$; H_a: $\rho \neq 0{,}5$; $\alpha = 0{,}05$; $n = 33$; $r = 0{,}41$. Проверить нулевую гипотезу. При каких значениях n нулевая гипотеза будет отвергнута? Принята? При каких значениях r нулевая гипотеза будет отвергнута? Принята?

Ответить на вопросы задачи 41 для следующих данных:

42. $r = 0{,}62$; $n = 52$; H_0: $\rho = 0{,}5$; H_a: $\rho > 0{,}5$.

43. $r = 0{,}88$; $n = 103$; H_0: $\rho = 0{,}95$; H_a: $\rho < 0{,}95$.

44. $r = -0{,}36$; $n = 12$; H_0: $\rho = 0$; H_a: $\rho < 0$.

45. $r = -0{,}71$; $n = 124$; H_0: $\rho = -0{,}6$; H_a: $\rho < -0{,}6$.

46. $r = 0{,}10$; $n = 228$; H_0: $\rho = 0$; H_a: $\rho \neq 0$.

47. $r = -0{,}13$; $n = 24$; H_0: $\rho = 0$; H_a: $\rho < 0$.

48. По двум выборкам объёмов $n_1 = 28$ и $n_2 = 39$ вычислены выборочные коэффициенты корреляции $r_1 = 0{,}71$ и $r_2 = 0{,}85$ соответственно. Можно ли утверждать, что коэффициенты корреляции ρ_1 и ρ_2 двумерных генеральных совокупностей различны? Для каких значений r_2 можно считать, что гипотеза H_0: $\rho_1 = \rho_2$ не противоречит опытным данным? Для каких значений n_1 нулевая гипотеза будет отвергнута? Положить $\alpha = 0{,}05$; H_a: $\rho_1 \neq \rho_2$.

Решить задачу 48 для следующих данных:

49. $r_1 = 0{,}62$; $n_1 = 52$; $r_2 = 0{,}88$; $n_2 = 60$; H_a: $\rho_1 < \rho_2$.

50. $r_1 = 0{,}36$; $n_1 = 28$; $r_2 = 0{,}41$; $n_2 = 33$; H_a: $\rho_1 \neq \rho_2$.

51. $r_1 = 0{,}71$; $n_1 = 19$; $r_2 = 0{,}62$; $n_2 = 28$; H_a: $\rho_1 > \rho_2$.

52. $r_1 = -0{,}88$; $n_1 = 103$; $r_2 = -0{,}71$; $n_2 = 84$; H_a: $\rho_1 \neq \rho_2$.

53. $r_1 = -0{,}41$; $n_1 = 12$; $r_2 = -0{,}62$; $n_2 = 19$; H_a: $\rho_1 > \rho_2$.

54. Проверить, используя критерий Колмогорова − Смирнова, гипотезу об извлечении двух выборок из одной генеральной совокупности.

Положить α = 0,1. Из продукции, изготовленной на двух станках, извлечены две выборки по 50 изделий. Результаты измерения одного из диаметров изделия (мм) таковы:

Станок 1	Станок 2	Станок 1	Станок 2	Станок 1	Станок 2
51,29	51,50	51,18	51,39	51,56	51,55
35	35	66	46	56	10
33	69	35	42	42	44
54	60	50	39	29	19
24	54	50	16	31	24
42	42	54	51	30	31
47	54	48	50	12	51
54	55	36	50	28	46
24	33	50	48	51	39
36	56	42	53	39	39
58	68	56	25	15	30
70	39	56	48	42	30
47	42	48	36	36	42
50	15	42	53	28	55
26	48	56	23	30	44
47	46	34	55	48	24
05	42	36	51		

55. Измерялась длина (мм) тела личинок щелкуна, обитающих в посевах озимой ржи и проса. Результаты таковы:

В посевах ржи	7,0	10,0	14,0	15,0	12,5	16,5	12,0	11,5	14,0	10,5	15,0
В посевах проса	11,0	12,0	16,0	13,5	18,0	15,5	16,0	17,0	11,0	18,0	14,5

Создается впечатление, что личинки щелкунов, обитающие в просе, в среднем крупнее личинок, обитающих во ржи. Проверить это предположение при α = 0,05.

56. Проверить предположение о том, что лечебный препарат не меняет состав крови (в частности, числа лейкоцитов), если препарат испытывался на 10 особях, а последующий анализ крови дал такие результаты: 0,97; 1,05; 1,09; 0,88; 1,01; 1,14; 1,03; 1,07; 0,94; 1,02 (числа выражают отношение числа лейкоцитов в опыте к числу лейкоцитов в норме).

57. Отношение зрителей к включению одной из телепередач в программе выразилось нижеследующими данными.

Можно ли считать, что отношение к данной передаче не зависит от пола зрителя? Принять α = 0,1.

Отношение к передаче	Положительное	Безразличное	Отрицательное
Мужчины	14	24	2
Женщины	29	36	15

58. При переписи населения Англии и Уэльса в 1901 г. были зарегистрированы (с точностью до тысячи) 15729000 мужчин и 16799000 женщин; 3497 мужчин и 3072 женщины были зарегистрированы как

глухонемые от рождения. Проверить гипотезу о том, что глухонемота не связана с полом. Положить $\alpha = 0{,}05$.

59. В статье о взаимосвязи между односторонним развитием глаз (измеренным по астигматизму, остроте зрения и т.д.) и односторонним развитием рук были приведены следующие данные:

Односторонность в развитии рук (по испытанию в поднимании тяжести)	Глазная односторонность по общему астигматизму			
	Левоглазие	Обоеглазие	Праволазие	Итого
Леворукие (левши)	34	62	28	124
Обоерукие	27	28	20	75
Праворукие	57	105	52	214
Итого	118	195	100	413

Можно ли на основании этих данных сделать вывод о том, что одностороннее развитие рук связано с односторонним развитием глаз?

60. 1000 человек классифицировали по признаку дальтонизма. Есть ли зависимость между способностью различать цвета и полом человека?

Способность различать цвета	Мужчины	Женщины
Дальтоники	38	6
Недальтоники	442	514

61. Результаты опроса общественного мнения с точки зрения поддержки 4-х кандидатов избирателями южных и северных районов некоторой страны таковы:

Районы	Кандидаты				Всего
	1	2	3	4	
Север	200	156	128	116	600
Юг	100	104	92	104	400
Всего	300	260	220	220	1.000

Имеется ли существенное различие в степени поддержки кандидатов избирателями каждого из регионов? Положить $\alpha = 0{,}05$.

62. Универмаг решил проанализировать сроки погашения кредита для различных категорий своих клиентов. Выборка, включающая $n = 1200$ платежей, дала следующие результаты:

Время	Рабочие	Священники	Служащие	Всего
До 30 суток	380	220	120	720
30-90 суток	220	200	60	480
Всего	600	420	180	1.200

Есть ли существенная разница между отдельными категориями покупателей с точки зрения сроков погашения кредита? Положить $\alpha = 0{,}05$.

В задачах 63 – 66 требуется найти выборочные коэффициенты корреляции Спирмена и Кендэла и проверить их на значимость.

63

Ранги	Студент								
	1	2	3	4	5	6	7	8	9
Оценки по математике	9	3	1	4	2	8	5	6	7
Оценки по химии	6	7	3	2	1	8	5	4	9

64

Ранги	Девушка (конкурс красоты)											
	1	2	3	4	5	6	7	8	9	10	11	12
Умение держать себя, манеры	3	11	4	10	1	8	9	2	12	6	7	5
Красота	4	11	1	12	6	2	10	5	9	7	8	3

65

Ранги	Бегун									
	1	2	3	4	5	6	7	8	9	10
Рост	1	2	3	4	5	6	7	8	9	10
Быстрота бега	5	6	10	7	9	4	3	1	8	2

66

Ранги	Цветной диск												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Порядок оттенков	1	2	3	4	5	6	7	8	9	10	11	12	13
Порядок оттенков, указанный испытуемым	7	4	2	3	1	10	6	8	9	5	11	13	12

67. Десять участниц конкурса красоты были ранжированы тремя судьями. Результаты приведены в таблице. С помощью коэффициента ранговой корреляции определить пару судей, оценки которых наиболее близки. Найти коэффициент конкордации.

Девушки	1	2	3	4	5	6	7	8	9	10
Судья 1	1	6	5	10	3	2	4	9	7	8
Судья 2	3	5	8	4	7	10	2	1	6	9
Судья 3	6	4	9	8	1	2	3	10	5	7

68. Предположим, что между переменным x и y существует линейная зависимость $y = ax+b$, $a > 0$. Показать, что в этом случае $r_s = \tau = 1$.

69. Пусть сумма рангов каждого объекта равна $n+1$. Показать, что тогда $r_s = \tau = -1$.

Приложение 1

Нормальное распределение

Таблица значений функции $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}\,dz$

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
0,00	0,0000	0,45	0,1736	0,90	0,3159	1,35	0,4115
0,01	0,0040	0,46	0,1772	0,91	0,3186	1,36	0,4131
0,02	0,0080	0,47	0,1808	0,92	0,3212	1,37	0,4147
0,03	0,0120	0,48	0,1844	0,93	0,3238	1,38	0.4162
0,04	0,0160	0,49	0,1879	0,94	0,3264	1,39	0,4177
0,05	0,0199	0,50	0,1915	0,95	0,3289	1,40	0,4192
0,06	0,0239	0,51	0,1950	0,96	0,3315	1,41	0,4207
0,07	0,0279	0,52	0,1985	0,97	0,3340	1,42	0,4222
0,08	0,0319	0,53	0,2019	0,98	0,3365	1,43	0,4236
0,09	0,0359	0,54	0,2054	0,99	0,3389	1,44	0,4251
0,10	0,0398	0,55	0,2088	1,00	0,3413	1,45	0,4265
0,11	0,0438	0,56	0,2123	1,01	0,3458	1,46	0,4279
0,12	0,0478	0,57	0,2157	1,02	0,3461	1,47	0,4292
0,13	0,0517	0,58	0,2190	1,03	0,3485	1.48	0,4306
0,14	0,0557	0,59	0,2224	1,04	0,3508	1,49	0,4319
0,15	0,0596	0,60	0,2257	1,05	0,3531	1,50	0,4332
0,16	0,0636	0,61	0,2291	1,06	0,3554	1,51	0,4345
0,17	0,0675	0,62	0,2324	1,07	0,3577	1,52	0,4357
0,18	0,0714	0,63	0,2357	1,08	0,3599	1,53	0,4370
0,19	0,0753	0,64	0,2389	1,09	0,3621	1,54	0,4382
0,20	0,0793	0,65	0,2422	1,10	0,3643	1,55	0,4394
0,21	0,0832	0,66	0,2454	1,11	0,3665	1,56	0,4406
0,22	0,0871	0,67	0,2486	1,12	0,3686	1,57	0,4418
0.23	0,0910	0,68	0,2517	1,13	0,3708	1,58	0,4429
0,24	0,0948	0,69	0.2549	1,14	0,3729	1,59	0,4441
0,25	0,0987	0,70	0,2580	1,15	0,3749	1,60	0,4452
0,26	0,1026	0,71	0,2611	1,16	0,3770	1,61	0,4463
0,27	0,1064	0,72	0,2642	1,17	0,3790	1,62	0,4474
0,28	0.1103	0,73	0,2373	1,18	0,3810	1,63	0,4484
0,29	0,1141	0,74	0,2703	1,19	0,3830	1,64	0,4495
0,30	0,1179	0,75	0,2734	1,20	0,3849	1,65	0,4505
0,31	0,l217	0,76	0,2764	1,21	0,3869	1,66	0,4515
0,32	0,1255	0,77	0,2794	1,22	0,3883	1,67	0,4525
0.33	0,1293	0,78	0,2823	1,23	0,3907	1,68	0,4535
0,34	0,1331	0,79	0,2852	1,24	0,3925	1,69	0,4545
0,35	0,1368	0,80	0,2881	1,25	0.3944	1,70	0,4554
0,36	0,1406	0,81	0,2910	1,26	0,3962	1,71	0,4564
0,37	0,1443	0,82	0,2939	1,27	0,3980	1,72	0,4573

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
0,38	0,1480	0,83	0,2967	1,28	0.3997	1,73	0,4582
0,39	0,1517	0,84	0,2995	1,29	0,4015	1,74	0,4591
0,40	0,1554	0,85	0,3023	1,30	0,4032	1,75	0,4599
0,41	0,1591	0,86	0,3051	1,31	0,4049	1,76	0,4608
0,42	0,1628	0,87	0,3078	1,32	0,4066	1,77	0,4616
0,43	0,1604	0,88	0,3106	1,33	0,4082	1,78	0,4625
0,44	0,1700	0,89	0,3133	1,34	0,4099	1,79	0,4633
1,80	0,4641	2,00	0,4772	2,40	0,4918	2,80	0,4974
1,81	0,4649	2,02	0,4783	2,42	0,4922	2,62	0,4370
1,82	0,4656	2,04	0,4793	2,44	0,4927	2,84	0,4977
1,83	0,4664	2,06	0,4803	2,46	0,4931	2,86	0,4979
1,84	0,4671	2,08	0,4812	2,48	0,4934	2,88	0,4980
1,85	0,1678	2,10	0,4821	2,50	0.4938	2,90	0,4981
1,86	0,4686	2,12	0,4830	2,52	0,4941	2,92	0,4982
1,87	0,4693	2,14	0,4838	2,54	0,4945	2,94	0,4984
1,88	0,4699	2,16	0,4846	2.56	0,4948	2,96	0,4985
1,89	0,4706	2,18	0,4854	2.58	0,4951	2,98	0,4986
1,90	0,4713	2,20	0,4861	2,60	0,4953	3,00	0,49865
1,91	0,4719	2,22	0,4868	2,62	0,4956	3,20	0,49931
1,92	0,4726	2,24	0,4875	2,64	0,4959	3,40	0,49966
1,93	0,4732	2,26	0,4881	2,66	0,4961	3,00	0,499841
1,94	0,4738	2,28	0,4887	2,68	0,4963	3,80	0,499928
1,95	0,4744	2,30	0,4893	2,70	0,4965	4,00	0,499968
1,96	0,4750	2,32	0,4898	2,72	0,4967	4,50	0,499997
1,97	0,4756	2,34	0,4904	2,74	0,4969	5,00	0,499997
1,98	0,4761	2,36	0,4909	2,76	0,4971		
1,99	0,4767	2,38	0,4913	2,78	0,4973		

Приложение 2

Распределение Стьюдента

Значения $t_{\alpha,r}$ удовлетворяют условию $P(t \geq t_{\alpha,r}) = \alpha$.

В таблице приведены значения квантилей $t_{\alpha,r}$ в зависимости от числа степеней свободы r и вероятности α.

α / r	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,603	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	5,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959

r \ α	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0.535	0,865	1,337	1.746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2.898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,887	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518'	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Приложение 3

χ^2 - распределение

В таблице приведены значения квантилей χ^2 в зависимости от числа степеней свободы r и вероятности α.

$$P(\chi^2 \geq \chi^2_{кр}) = \alpha$$

r \ α	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 4

Распределение Фишера

В таблице приведены значения квантилей F_{α,r_1,r_2} в зависимости от

числа степеней свободы r_1 и r_2 для $\alpha = 0{,}05$.

$$P\left(F \geq F_{\alpha, r_1, r_2}\right) = 0{,}05.$$

r_2 \ r_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,512	18,999	19,163	19,248	19,298	19,329	19,371	19,414	19,453	19,496
3	10,129	9,552	9,276	9,118	9,014	8,941	8,844	8,744	8,638	8,527
4	7,710	6,945	6,591	6,388	6,257	6,164	6,041	5,912	5,774	5,628
5	6,607	5,786	5,410	5,192	5,050	4,950	4,818	4,678	4,527	4,365
6	5,987	5,143	4,756	4,534	4,388	4,284	4,147	4,000	3,841	3,669
7	5,591	4,737	4,347	4,121	3,972	3,866	3,725	3,574	3,410	3,230
8	5,317	4,459	4,067	3,838	3,688	3,580	3,438	3,284	3,116	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,073	2,900	2,707
10	4,965	4,103	3,807	3,478	3,326	3,217	3,072	2,913	2,737	2,538
11	4,844	3,982	3,587	3,357	3,204	3,094	2,948	2,788	2,609	2,405
12	4,747	3,885	3,490	3,259	3,106	2,999	2,848	2,686	2,505	2,296
13	4,667	3,805	3,410	3,197	3,025	2,915	2,767	2,604	2,420	2,207
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,534	2,349	2,131
15	4,543	3,683	3,287	3,056	2,901	2,790	2,641	2,475	2,288	2,066
16	4,494	3,634	3,239	3,007	2,853	2,741	2,591	2,424	2,235	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,381	2,190	1,961
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,342	2,150	1,917
19	4,381	3,522	3,127	2,895	2,740	2,629	2,477	2,308	2,114	1,878
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,278	2,083	1,843
21	4,325	3,467	3,027	2,840	2,685	2,573	2,421	2,250	2,054	1,812
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,226	2,028	1,783
23	4,279	3,422	3,028	2,795	2,640	2.528	2,375	2,203	2,005	1,757
24	4,260	3,304	3,009	2,777	2,621	2,508	2,355	2,183	1,984	1,733
25	4,224	3,385	2,991	2,759	2,603	2,490	2,337	2,165	1,965	1,711
26	4,225	3,369	2,975	2,743	2,587	2,474	2,321	2,148	1,947	1,691
27	4,210	3,354	2,961	2,728	2,572	2,459	2,305	2,132	1,930	1,672
28	4.196	3,340	2,947	2,714	2,558	2,445	2,292	2,118	1,915	1,654
29	4,183	3,328	2,934	2,702	2,545	2,432	2,278	2,104	1,901	1,638
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,092	1,887	1,622
40	4,085	3,232	2,839	2,606	2,449	2,336	2,180	2,004	1,793	1,509
60	4,001	3,151	2,758	2,525	2,368	2,2.54	2,097	1,918	1,700	1,389
120	3,920	3,072	2,680	2,447	2,290	2,175	2,106	1,834	1,608	1,254
∞	3,841	2,996	2,605	2,372	2,214	2,098	1,938	1,752	1,517	1,000

Библиографический список

1. *Агапов П.И.* Задачник по теории вероятностей. - М.: Высшая школа, 1986.
2. *Андрухаев Х.М.* Сборник задач по теории вероятностей. - М.: Просвещение, 1985.
3. *Вентцель Е.С.* Теория вероятностей и её инженерные приложения /Е.С.Вентцель, Л.А.Овчаров. –М.: Наука, 1998.
4. *Герасимович А.И.* Математическая статистика. - Минск: Вышейшая школа, 1983.
5. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. - М.: Высшая школа, 1979.
6. *Емельянов Г.В.* Задачник по теории вероятностей и математической статистике /Г.В.Емельянов, В.П.Скитович.- Л.: Изд-во ЛГУ, 1967.
7. *Кендэл М.* Временные ряды. - М.: ФиС, 1981.
8. *Кокс Д.* Прикладная статистика. Принципы и примеры /Д.Кокс, Э.Снелл. - М.: Мир, 1984.
9. *Колкот Э.* Проверка значимости. - М.: Статистика, 1978.
10. *Мельник М.* Основы прикладной статистики. - М.: Энергоатомиздат, 1983.
11. *Мешалкин Л.Д.* Сборник задач по теории вероятностей. - М.: Изд-во МГУ, 1963.
12. *Нейман Ю.* Вводный курс теории вероятностей и математической статистики. - М.: Наука, 1968.
13. *Палий И.А.* Задачи по математической статистике/ СибАДИ. - Омск, 1991.
14. *Палий И.А.* Введение в теорию вероятностей/ СибАДИ. - Омск, 1993.
15. Сборник задач по математике для втузов. Специальные курсы / *Э.А. Вуколов*, *А.В.Ефимов, В.Н. Земсков* и др.; Под ред. *А.В. Ефимова*. - М.: Наука, 1984.
16. Теория статистики: Учебник / Под ред. проф. *Р.А. Шлоймовой*. — М.: ФиС, 1996.
17. *Тернер Д.* Вероятность, статистика и исследование операций. - М.: Статистика, 1976.
18. *Тюрин Ю.Н.* Статистический анализ данных на компьютере /Ю.Н.Тюрин, А.А.Макаров. - М.: ИНФРА, 1998.
19. *Ферстер Э.* Методы корреляционного и регрессионного анализа /Э.Ферстер, Б.Ренц. -М.: ФиС, 1983.