

Раздел 4. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

§ 1. Выборка и ее описание

Теория вероятностей и математическая статистика занимаются анализом закономерностей случайных массовых явлений. В теории вероятностей определяются вероятности тех или иных событий по известным вероятностям более простых событий, числовые характеристики СВ или вероятности, связанные с этими величинами, по известным законам распределения этих СВ. На практике для нахождения законов распределения СВ необходимо использовать экспериментальные данные.

Основной задачей математической статистики является разработка методов получения вероятностных характеристик случайных явлений на основе результатов наблюдений или эксперимента.

Математическая статистика опирается на теорию вероятностей и в свою очередь служит основой для разработки методов обработки и анализа статистических результатов в конкретных областях человеческой деятельности.

Понятия генеральной совокупности и выборки

Исходными понятиями математической статистики являются понятия генеральной и выборочной совокупностей.

Опр. 1. Выборка (случайная выборка, выборочная совокупность) – множество значений результатов наблюдений над одной и той же СВ при одних и тех же условиях. Элементы выборки называются **выборочными значениями**. Количество проведенных наблюдений называется **объемом выборки**.

Опр. 2. Генеральной совокупностью называется множество всех возможных наблюдений над СВ при данном комплексе условий.

В большинстве случаев генеральная совокупность бесконечна (можно производить сколь угодно много наблюдений).

При контроле качества данной партии товаров объем генеральной совокупности равен объему этой партии. Если обследо-

ние всей партии невозможно (например, обследование объекта связано с его уничтожением или требует больших материальных затрат), то о качестве партии судят по случайной выборке товаров из этой партии.

Назначение статистических методов в том, чтобы по выборке ограниченного объема сделать вывод о свойствах генеральной совокупности в целом.

Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем нас признаке генеральной совокупности, необходимо, чтобы объекты выборки «правильно» его представляли.

Опр. 3. Выборка называется *репрезентативной* (представительной), если она достаточно хорошо отражает изучаемые свойства генеральной совокупности.

Считается, что это требование выполняется, если объем выборки достаточно велик и все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку, т.е. при отборе сохраняется принцип случайности. Такую выборку называют *случайной выборкой*.

Опр. 4. Выборка называется *повторной*, если каждый выбранный элемент перед отбором следующего возвращается в генеральную совокупность. Если такого возвращения не происходит, выборка называется *бесповторной*.

Например, в задаче контроля качества, как правило, рассматриваются бесповторные выборки, а если производится несколько измерений некоторой величины, то выборка считается повторной.

Статистический ряд и его графическое изображение

Пусть имеется выборка объема n : $x_1; x_2; \dots; x_n$.

Опр. 5. *Вариационным рядом* выборки x_1, x_2, \dots, x_n называется способ ее записи, при котором ее элементы упорядочены (как правило, в порядке неубывания).

Пример 1. Дана выборка: 2; 4; 7; 3; 1; 1; 3; 2; 7; 3.

Запишем ее вариационный ряд: 1; 1; 2; 2; 3; 3; 3; 4; 7; 7. •

Опр. 6. Разность W между максимальным и минимальным элементами называется *размахом выборки*: $W = x_{\max} - x_{\min}$.

Как правило, некоторые выборочные значения могут совпадать, поэтому часто выборку представляют в виде статистического ряда.

Опр. 7. Пусть в выборке элемент x_i встречается n_i раз. Число n_i называется *частотой* выборочного значения x_i , а $\frac{n_i}{n}$ — *относительной частотой*.

Очевидно, что $\sum_{i=1}^k n_i = n$, где k — число различных элементов выборки.

Опр. 8. Последовательность пар $(x_i^*; n_i)$, где $x_1^*, x_2^*, \dots, x_k^*$ — различные выборочные значения, а n_1, n_2, \dots, n_k — соответствующие им частоты, называется *статистическим рядом*.

Обычно статистический ряд записывается в виде таблицы, первая строка которой содержит различные выборочные значения x_i^* , а вторая — их частоты n_i (или относительные частоты $\frac{n_i}{n}$, иногда и те, и другие):

x_i^*	x_1^*	x_2^*	...	x_k^*
n_i	n_1	n_2	...	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$...	$\frac{n_k}{n}$

При большом объеме (больше 30) выборки ее элементы объединяют в группы (разряды), представляя результаты опытов в виде *интервального (группированного) статистического ряда*. Для этого интервал, содержащий все элементы выборки, разбивают на k непересекающихся интервалов. Число интервалов выбирается произвольно и, как правило, $5 \div 10 \leq k \leq 20 \div 25$. Вычисления значительно упрощаются, если интервалы имеют одинаковую длину $h \approx \frac{W}{k}$. (В дальнейшем будет рассматриваться именно этот случай.) После того, как частичные интервалы выбраны, определяют частоты n_i — количество элементов выборки, попавших в i -й интервал (элемент, совпадающий с верхней границей интервала,

относится к последующему интервалу) и относительные частоты $\frac{n_i}{n}$. Полученные данные сводятся в таблицу:

$[x_i; x_{i-1})$	$[x_0; x_1)$	$[x_1; x_2)$	\dots	$[x_{k-1}; x_k]$
$x_i^* = \frac{x_{i-1} + x_i}{2}$	x_1^*	x_2^*	\dots	x_k^*
n_i	n_1	n_2	\dots	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	\dots	$\frac{n_k}{n}$

Для наглядного представления выборки используют полигон (для дискретных статистических рядов) и гистограмму (для интервальных статистических рядов) частот (или относительных частот).

Опр. 9. Полигоном частот называется ломаная с вершинами в точках $(x_i^*; n_i), 1 \leq i \leq k$ (см. рис. 27); **полигоном относительных частот** — ломаная линия с вершинами в точках $\left(x_i^*; \frac{n_i}{n}\right), 1 \leq i \leq k$.

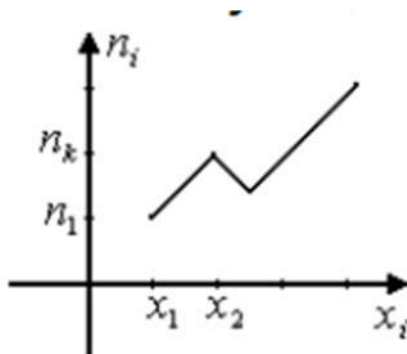


Рис. 27. Полигон частот

Опр. 10. Гистограммой относительных частот (частот) называют ступенчатую фигуру, составленную из прямоугольников, построенных на интервалах группировки так, что площадь

каждого прямоугольника равна соответствующей данному интервалу относительной частоте (частоте) (высоты прямоугольников равны $\frac{n_i}{nh}$ в случае гистограммы относительных частот и $\frac{n_i}{h}$ в случае гистограммы частот, см. рис. 28).

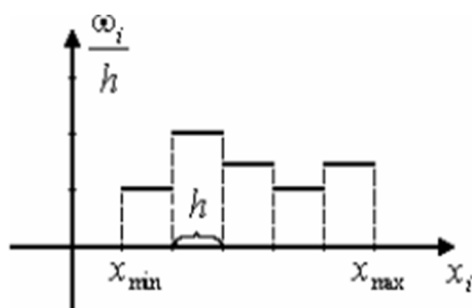


Рис. 28. Гистограмма относительных частот

Гистограмма относительных частот обладает тем свойством, что ее площадь равна 1. Площадь гистограммы частот равна объему выборки n .

При достаточно большом объеме выборки n и достаточно малых интервалах группировки h гистограмма относительных частот является хорошим приближением графика плотности распределения наблюдаемой случайной величины. Поэтому по виду гистограммы можно выдвинуть предположение (гипотезу) о распределении изучаемой случайной величины.

Эмпирическая функция распределения

Опр. 11. *Эмпирической функцией распределения* называется функция $F_n^*(x)$, определяющая для каждого значения x относительную частоту наблюдения значений, меньших x :

$$F_n^*(x) = \sum_{x_i^* < x} \frac{n_i}{n}.$$

Слово «эмпирический» означает «полученный по экспериментальным (опытным) данным», т.е. по выборке. По этой причине иногда употребляют термин «выборочная функция распределения».

Из определения эмпирической функции распределения видно, что она обладает такими же свойствами, как функция распределения дискретной случайной величины в теории вероятностей, а именно:

- 1) $0 \leq F_n^*(x) \leq 1$;
- 2) $F_n^*(x)$ – неубывающая функция;
- 3) $F_n^*(x)$ – непрерывная слева кусочно-постоянная функция;
- 4) если x_{\min} – наименьшее, а x_{\max} – наибольшее выборочные значения, то $F_n^*(x) = 0$ при $x \leq x_{\min}$ и $F_n^*(x) = 1$ при $x > x_{\max}$.

Пример 1 (продолжение). Для выборки 2; 4; 7; 3; 1; 1; 3; 2; 7; 3 запишем эмпирическую функцию распределения и построим ее график.

Объем выборки $n = 10$. Составим статистический ряд:

x_i^*	1	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	0,2	0,2	0,3	0,1	0,2

Запишем эмпирическую функцию распределения, накапливая относительные частоты:

$$F_n^*(x) = \begin{cases} 0 & \text{при } x \leq 1, \\ 0,2 & \text{при } 1 < x \leq 2, \\ 0,4 & \text{при } 2 < x \leq 3, \\ 0,7 & \text{при } 3 < x \leq 4, \\ 0,8 & \text{при } 4 < x \leq 7, \\ 1 & \text{при } x > 7. \end{cases}$$

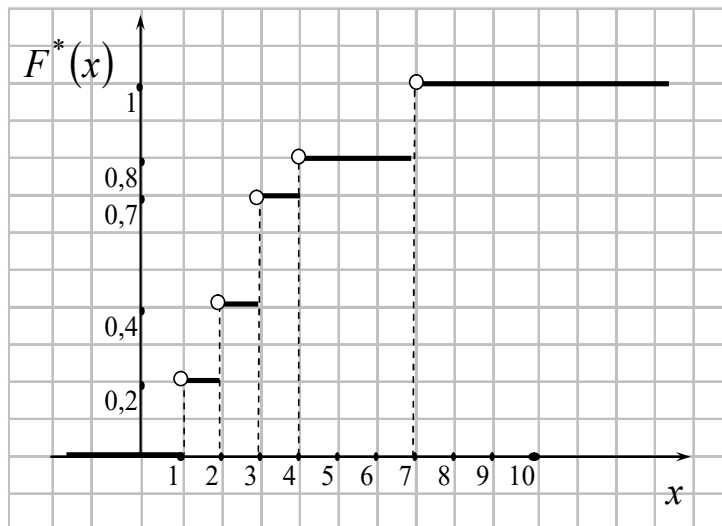


Рис. 29. График эмпирической функции распределения

График $F_n^*(x)$ представлен на рис. 29. •

Замечание. Если график $F_n^*(x)$ строится по интервальному статистическому ряду, то скачки происходят в точках, соответствующих серединам интервалов группировки.

Основное значение эмпирической функции распределения в том, что она используется в качестве оценки теоретической функции распределения $F(x) = P(\xi < x)$ наблюдаемой СВ ξ .

Пусть имеется выборка наблюдений над СВ ξ . Значение $F_n^*(x)$ эмпирической функции распределения в точке x равно относительной частоте наблюдения значений, меньших x , т. е. относительной частоте события $\{\xi < x\}$. Согласно закону больших чисел (в форме Я. Бернулли), при $n \rightarrow \infty$ относительная частота стремится к вероятности события, т. е., в данном случае, к $P(\xi < x) = F(x)$.

§ 2. Точечное оценивание параметров распределения. Свойства точечных оценок

Основные задачи статистической обработки одной выборки:

- 1) оценивание параметров распределения;
- 2) проверка статистических гипотез о виде или параметрах распределения.

Пусть имеется выборка объема n : $x_1; x_2; \dots; x_n$. Выборка представляет собой ряд наблюдений над одной и той же СВ. При ста-

статистическом анализе выборки в первую очередь стремятся оценить математическое ожидание и дисперсию. По результатам этого ограниченного числа наблюдений невозможно *вычислить* числовые характеристики наблюдаемой СВ, а можно только *оценить* их.

Опр. 1. Любая функция $\hat{\theta}_n = \hat{\theta}_n(x_1; x_2; \dots; x_n)$, зависящая от выборочных значений, называется **статистикой** или **выборочной функцией**.

Опр. 2. Точечной оценкой параметра θ называется любая статистика $\hat{\theta}_n$, предназначенная для оценки этого параметра и определяемая одним числом.

Подчеркнем, что точечная оценка практически никогда не совпадает с истинным значением параметра, она может только оценивать его с большей или меньшей точностью.

Для любого параметра можно предложить разные оценки. Так, в качестве оценки для математического ожидания можно использовать первый элемент выборки x_1 , среднее арифметическое наибольшего и наименьшего элементов выборки, среднее арифметическое всех элементов выборки и т. д.

Задача статистического оценивания параметров заключается в том, чтобы из всего множества оценок выбрать в некотором смысле наилучшую. Это означает, что распределение СВ $\hat{\theta}_n(x_1; x_2; \dots; x_n)$ должно концентрироваться около истинного значения параметра θ .

Замечание. Если, имея выборку $x_1; x_2; \dots; x_n$ значений некоторой СВ, повторно провести n независимых наблюдений над этой СВ, то новая выборка $x'_1; x'_2; \dots; x'_n$, вообще говоря, не будет совпадать с первоначальной. Поэтому выборочные значения можно рассматривать как СВ.

Основное предположение математической статистики: выборочные значения $x_1; x_2; \dots; x_n$ являются независимыми в совокупности одинаково распределенными СВ. Следовательно, любая оценка $\hat{\theta}_n(x_1; x_2; \dots; x_n)$ также является СВ.

Основные характеристики точечных оценок

Качество точечной оценки характеризуется следующими основными свойствами.

Опр. 3. Оценка $\hat{\theta}$ называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру: $M\hat{\theta} = \theta$.

Требование несмещенности гарантирует отсутствие систематических ошибок при оценивании. Оно особенно важно при малом числе наблюдений (в случае выборок объема не более 30).

Опр. 4. Оценка $\hat{\theta}_n$ называется *состоятельной*, если при увеличении объема выборки n оценка $\hat{\theta}_n$ сходится по вероятности к θ : $\hat{\theta}_n \xrightarrow{P} \theta$ при $n \rightarrow \infty$.

Это свойство означает, что при большом объеме выборки практически достоверно, что $\hat{\theta}_n \approx \theta$. Чем больше объем выборки, тем более точные оценки можно получить.

Опр. 5. Пусть $\hat{\theta}_1$ и $\hat{\theta}_2$ – две различные *несмещенные* оценки параметра. Если для их дисперсий выполняется условие $D\hat{\theta}_1 < D\hat{\theta}_2$, то говорят, что оценка $\hat{\theta}_1$ более эффективна, чем оценка $\hat{\theta}_2$. Оценка с наименьшей дисперсией называется *эффективной*.

Это означает, что распределение эффективной оценки наиболее тесно сконцентрировано около истинного значения параметра.

Замечание. Не всегда можно найти оценки, которые имели бы все указанные свойства.

Точечные оценки для математического ожидания и дисперсии

Для негруппированной выборки	Для группированного статистического ряда
Выборочное среднее	
$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i^* n_i$
Выборочная дисперсия	

$D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ $D_B = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2$	$D_B = \frac{1}{n} \sum_{i=1}^k (x_i^* - \bar{x})^2 n_i$ $D_B = \frac{1}{n} \sum_{i=1}^k (x_i^*)^2 n_i - (\bar{x})^2$
<p style="text-align: center;">Несмещенная оценка дисперсии</p> $s^2 = \frac{n}{n-1} D_B$	
$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ $s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right)$	$s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i^* - \bar{x})^2 n_i$ $s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i^*)^2 n_i - \frac{n}{n-1} (\bar{x})^2$

Выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (среднее арифметическое элементов выборки) характеризует центр распределения (рассеивания) изучаемой СВ.

Выборочная дисперсия D_B характеризует степень разброса (рассеяния) выборочных значений относительно среднего.

Свойства выборочных среднего и дисперсии как оценок для математического ожидания и дисперсии

Утв. 1. Выборочное среднее является *несмещенной* и *состоятельной*, а в случае выборки из нормального распределения и *эффективной* оценкой для математического ожидания наблюдаемой СВ.

Докажем несмещенность выборочного среднего как оценки для математического ожидания.

Пусть x_1, x_2, \dots, x_n – некоторая выборка, т. е. x_1, x_2, \dots, x_n – независимые СВ, имеющие одинаковое распределение. Обозначим $Mx_i = a$. Надо показать, что $M\bar{x} = a$. По свойствам математического ожидания,

$$M\bar{x} = M\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n Mx_i = \frac{1}{n} \sum_{i=1}^n a = \frac{1}{n} na = a.$$

Состоятельность выборочного среднего как оценки математического ожидания следует из закона больших чисел. <

Утв. 2. Выборочная дисперсия $D_{\text{в}}$ является *состоятельной*, но *смещенной* оценкой дисперсии изучаемой СВ:

$$MD_{\text{в}} = \frac{n-1}{n} \sigma^2.$$

Докажем смещенность выборочной дисперсии как оценки для дисперсии. Пусть выборочные значения x_1, x_2, \dots, x_n – независимые СВ, имеющие одинаковое распределение с $Mx_i = a$, $Dx_i = \sigma^2$.

Покажем сперва, что для любого постоянного c справедливо равенство

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - c)^2 - n(\bar{x} - c)^2.$$

Действительно,

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n ((x_i - c) - (\bar{x} - c))^2 = \\ &= \sum_{i=1}^n ((x_i - c)^2 - 2(x_i - c)(\bar{x} - c) + (\bar{x} - c)^2) = \\ &= \sum_{i=1}^n (x_i - c)^2 - 2(\bar{x} - c) \sum_{i=1}^n (x_i - c) + \sum_{i=1}^n (\bar{x} - c)^2 = \\ &= \sum_{i=1}^n (x_i - c)^2 - 2(\bar{x} - c) \left(\sum_{i=1}^n x_i - nc \right) + n(\bar{x} - c)^2 = \\ &= \sum_{i=1}^n (x_i - c)^2 - 2(\bar{x} - c)(n\bar{x} - nc) + n(\bar{x} - c)^2 = \\ &= \sum_{i=1}^n (x_i - c)^2 - 2n(\bar{x} - c)^2 + n(\bar{x} - c)^2 = \sum_{i=1}^n (x_i - c)^2 - n(\bar{x} - c)^2. \end{aligned}$$

Таким образом, $D_{\text{в}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2 - (\bar{x} - a)^2.$

Используя определение дисперсии СВ, а также свойства математического ожидания и дисперсии, имеем:

$$MD_{\text{в}} = \frac{1}{n} \sum_{i=1}^n M(x_i - a)^2 - M(\bar{x} - a)^2 = \frac{1}{n} \sum_{i=1}^n Dx_i - D\bar{x} =$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n Dx_i - D\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n Dx_i - \frac{1}{n^2} \sum_{i=1}^n Dx_i = \\
&= \frac{1}{n} n\sigma^2 - \frac{1}{n^2} n\sigma^2 = \sigma^2 \left(1 - \frac{1}{n}\right) = \frac{n-1}{n} \sigma^2. \triangleleft
\end{aligned}$$

Итак, выборочная дисперсия D_v является *смещенной* оценкой дисперсии изучаемой СВ (дает заниженное значение). В связи с этим вместо нее вводится другая статистика – **исправленная выборочная дисперсия** $s^2 = \frac{n}{n-1} D_v$, которая является *несмещенной оценкой дисперсии*.

Утв. 3. Исправленная дисперсия s^2 является *несмещенной состоятельной* оценкой дисперсии. В случае выборки из нормального распределения s^2 является также *асимптотически эффективной* оценкой дисперсии.

§ 3. Интервальное оценивание параметров распределения

Точечные оценки не дают информации о степени близости оценки к истинному значению оцениваемого параметра. Чтобы получить информацию о точности и надежности оценки, используют интервальные оценки.

Опр. 1. *Интервальной оценкой (доверительным интервалом)* параметра θ называется интервал, границы которого $\hat{\theta}_1 = \hat{\theta}_1(x_1; x_2; \dots; x_n)$ и $\hat{\theta}_2 = \hat{\theta}_2(x_1; x_2; \dots; x_n)$ являются функциями выборочных значений и который с заданной вероятностью γ накрывает истинное значение оцениваемого параметра θ :

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = \gamma.$$

Интервал $(\hat{\theta}_1; \hat{\theta}_2)$ называется **доверительным интервалом**; число γ – **доверительной вероятностью** или **надежностью** интервальной оценки; значение $\alpha = 1 - \gamma$ – **уровнем значимости**.

Величина доверительного интервала существенно зависит от объема выборки (уменьшается с ростом n , т. е. *чем больше объем выборки, тем более точную оценку можно получить*) и от доверительной вероятности γ (величина доверительного интервала увеличивается с приближением γ к 1, т. е. *чем более надежный вывод*

мы хотим получить, тем меньшую точность мы можем гарантировать).

Выбор доверительной вероятности определяется конкретными условиями. Обычно используются значения 0,90; 0,95; 0,99; 0,9973, т. е. такие, чтобы получить интервал, который с большой вероятностью накроет истинное значение оцениваемого параметра.

Приведем примеры доверительных интервалов для параметров нормального распределения. Для этого нам нужно знать, каким распределениям подчиняются статистики \bar{x} и s^2 . (Напомним, что запись $\xi \sim \mathcal{N}(a; \sigma)$ означает, что СВ ξ имеет нормальное распределение с $M\xi = a$, $D\xi = \sigma^2$. Тот факт, что СВ ξ имеет закон распределения \mathcal{P} , будем символически обозначать $\xi \sim \mathcal{P}$.)

Построение доверительного интервала для математического ожидания в случае выборки из нормального распределения с известной дисперсией σ^2

Утв. 1. Пусть имеется выборка объема n из нормального распределения с математическим ожиданием a и дисперсией σ^2 , т. е. $x_1, x_2, \dots, x_n \sim \mathcal{N}(a; \sigma)$. Тогда статистика \bar{x} распределена по нормальному закону с параметрами a и $\frac{\sigma}{\sqrt{n}}$, а статистика $\frac{\bar{x} - a}{\sigma / \sqrt{n}}$ имеет стандартное нормальное распределение:

$$\bar{x} \sim \mathcal{N}\left(a; \frac{\sigma}{\sqrt{n}}\right); \quad \frac{\bar{x} - a}{\sigma / \sqrt{n}} \sim \mathcal{N}(0; 1).$$

Отметим, что

$$M\bar{x} = M \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n Mx_i = \frac{1}{n} \sum_{i=1}^n a = \frac{1}{n} na = a;$$

$$D\bar{x} = D \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n^2} \sum_{i=1}^n Dx_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Это означает, в частности, что \bar{x} является более точной, чем одиночное наблюдение, оценкой для математического ожидания,

поскольку чем меньше дисперсия, т. е. разброс значений, тем точнее оценка.

Утв. 2. Доверительный интервал для математического ожидания a в случае выборки из нормального распределения с известной дисперсией σ^2 определяется соотношением

$$P\left(\bar{x} - u_{\alpha} \frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \quad (1)$$

где n – объем выборки; \bar{x} – выборочное среднее; α – уровень значимости; u_{α} – квантиль нормального распределения уровня α , т. е. такое число, что для СВ $\xi \sim \mathcal{N}(0; 1)$, имеющей стандартное нормальное распределение, $P(|\xi| \geq u_{\alpha}) = \alpha$.

Квантиль u_{α} определяется по таблице функции Лапласа из соотношения $\Phi(u_{\alpha}) = \frac{1 - \alpha}{2}$.

Формула (1) означает, что при достаточно большом количестве выборок одного и того же объема n примерно в $100(1 - \alpha)\%$ выборок интервал $\left(\bar{x} - u_{\alpha} \frac{\sigma}{\sqrt{n}}; \bar{x} + u_{\alpha} \frac{\sigma}{\sqrt{n}}\right)$ покрывает истинное значение математического ожидания a .

Доказательство. Из утверждения 1 следует, что

$$P\left(\left|\frac{\bar{x} - a}{\sigma / \sqrt{n}}\right| < u_{\alpha}\right) = 1 - \alpha;$$

$$P\left(-u_{\alpha} < \frac{a - \bar{x}}{\sigma / \sqrt{n}} < u_{\alpha}\right) = 1 - \alpha;$$

$$P\left(-u_{\alpha} \frac{\sigma}{\sqrt{n}} < a - \bar{x} < u_{\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha;$$

$$P\left(\bar{x} - u_{\alpha} \frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \triangleleft$$

Выборочные распределения

С нормальным распределением связаны также следующие три наиболее часто используемые в статистике распределения: χ^2 -распределение, t -распределение Стьюдента и F -распределение Фишера.

Опр. 2. *Распределением χ^2 с k степенями свободы* называется распределение суммы квадратов k независимых СВ, распределенных по нормальному закону с параметрами 0 и 1, т. е. если $\xi_1, \xi_2, \dots, \xi_k \sim \mathcal{N}(0; 1)$, то $\eta = \xi_1^2 + \xi_2^2 + \dots + \xi_k^2 \sim \chi_k^2$.

Плотность распределения χ^2 с k степенями свободы имеет вид

$$f_{\chi_k^2}(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & \text{при } x > 0, \\ 0 & \text{при } x \leq 0. \end{cases}$$

Здесь $\Gamma\left(\frac{k}{2}\right)$ — значение *гамма-функции* которая определяется для $\alpha > 0$

формулой $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$ и обладает свойствами: 1) $\Gamma(1) = \Gamma(2) = 1$;

2) $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$; 3) $\Gamma(\alpha) = (\alpha - 1)!$, если α — натуральное;

4) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

На рис. 30 изображены кривые плотностей распределения χ^2 при числе степеней свободы $k = 1, 2, 3, 5$.

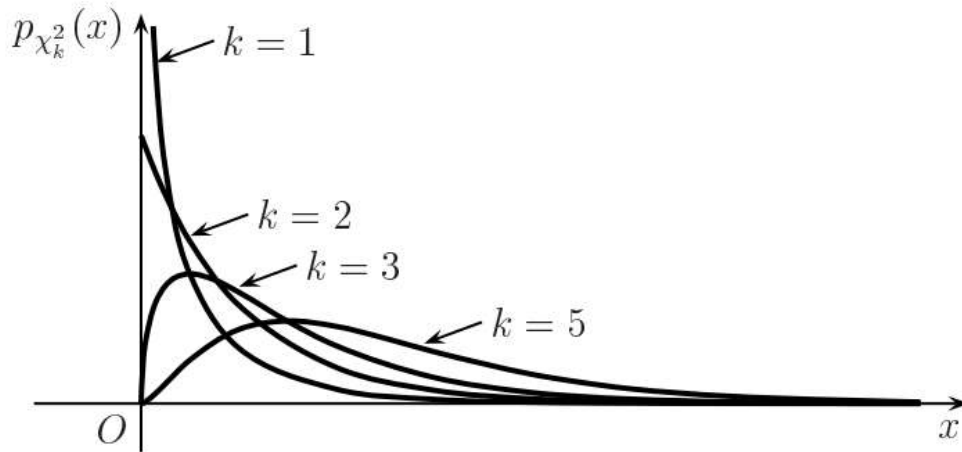


Рис. 30. Графики плотности распределения χ^2 с числом степеней свободы $k = 1, 2, 3, 5$

В частном случае при $k = 2$ распределение χ_k^2 совпадает с показательным (экспоненциальным) распределением.

С ростом k распределение χ_k^2 приближается к нормальному.

Считается, что при $k > 30$ оно практически не отличается от нормального.

Утв. 3. В случае выборки объема n из нормального распределения с *известным* математическим ожиданием статистика $\frac{nD_B}{\sigma^2}$ имеет χ^2 -распределение с n степенями свободы:

$$\frac{nD_B}{\sigma^2} \sim \chi_n^2;$$

в случае выборки объема n из нормального распределения с *неизвестным* математическим ожиданием статистика $\frac{(n-1)s^2}{\sigma^2}$ имеет χ^2 -распределение с числом степеней свободы $n - 1$:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Доказательство. Докажем утверждение для случая известного математического ожидания. Пусть имеется выборка объема n из нормального рас-

пределения, т. е. СВ x_1, x_2, \dots, x_n независимы и $x_i \sim \mathcal{N}(a; \sigma)$. Тогда в силу свойств нормального распределения СВ $\frac{x_i - a}{\sigma} \sim \mathcal{N}(0; 1)$.

Покажем, что в случае, когда математическое ожидание известно, статистика $\frac{nD_B}{\sigma^2}$ представляет собой сумму квадратов этих СВ. В этом случае для расчета D_B используется известное значение математического ожидания:
 $D_B = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$, поэтому

$$\frac{nD_B}{\sigma^2} = \frac{n \frac{1}{n} \sum_{i=1}^n (x_i - a)^2}{\sigma^2} = \frac{\sum_{i=1}^n (x_i - a)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - a}{\sigma} \right)^2,$$

что в силу определения 2 доказывает, что $\frac{nD_B}{\sigma^2} \sim \chi_n^2$. \triangleleft

Упражнение 1. Доказать утверждение 3 для случая неизвестного математического ожидания.

Замечание. В общем случае число степеней свободы, соответствующее той или иной оценке дисперсии, определяется как количество независимых наблюдений, по которым вычисляется данная оценка дисперсии, минус число параметров, которые оцениваются по этой выборке, кроме дисперсии. В случае, когда математическое ожидание неизвестно, одна степень свободы «расходуется» на вычисление оценки \bar{x} .

Опр. 3. Распределением Стьюдента с k степенями свободы называется распределение СВ $t = \frac{\xi}{\sqrt{\eta/k}}$ $\sim t_k$, где СВ $\xi \sim \mathcal{N}(0; 1)$ и $\eta \sim \chi_k^2$ независимы.

[WWWBIKISПРАВКАWWW](http://www.bikispravka.ru)



Уильям Сили Госсет
(англ. *William Sealy Gosset*)
(1876–1937)

британский химик и статистик, работавший на пивоваренном заводе «Гиннесс» (Arthur Guinness Son & Co). Один из основоположников теории статистических оценок и проверки гипотез.

«Гиннесс» был передовым предприятием, ориентированным на использование новейших достижений науки для принятия экономических и технических решений, благодаря чему Госсет имел полную свободу в проведении научных исследований. В частности, для контроля качества продукции необходимо было понять, как с точки зрения математики можно обосновать достоверность выводов, полученных при исследовании малой выборки, и правомерность применения этих выводов к выборке большой. В результате этих исследований Госсет разработал математическое обоснование «закона ошибок» для малых статистических выборок.

Госсет более известен под своим псевдонимом Студент (Student), поскольку по условиям контракта с корпорацией «Гиннесс» не имел права открыто публиковать результаты своих исследований (таким способом охранялась коммерческая тайна, «ноу-хау» в виде вероятностно-статистических методов, разработанных Госсетом).

Первым, кто понял значение работ Госсета по оценке параметров малой выборки, был английский статистик Р. Э. Фишер (1890–1962), считавший, что Госсет совершил «логическую революцию» в математической статистике.

~~~~~

Плотность распределения Стьюдента с  $k$  степенями свободы имеет вид

$$f_{t_k}(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k} \Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}};$$

ее график представлен на рис. 31.

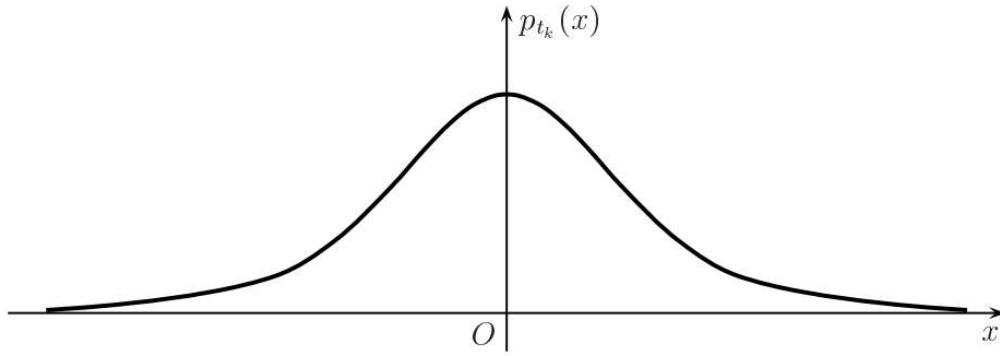


Рис. 31. График плотности распределения Стьюдента

При  $k \rightarrow \infty$  график приближается к графику плотности нормального распределения. Практически уже при  $k > 30$  можно считать  $t$ -распределение приближенно нормальным.

**Утв. 4.** В случае выборки объема  $n$  из нормального распределения с *неизвестной* дисперсией статистика  $\frac{\bar{x} - a}{\sqrt{s^2 / n}}$  имеет распределение Стьюдента с числом степеней свободы  $n - 1$ :

$$\frac{\bar{x} - a}{\sqrt{s^2 / n}} \sim t_{n-1}.$$

*Упражнение 2.* Сравнить с формулировкой утверждения 1.

*Доказательство.* Пусть имеется выборка объема  $n$  из нормального распределения, т. е. СВ  $x_1, x_2, \dots, x_n$  независимы и  $x_i \sim \mathcal{N}(a; \sigma)$ . Тогда в силу утверждений 1 и 3 СВ  $\xi = \frac{\bar{x} - a}{\sigma / \sqrt{n}} \sim \mathcal{N}(0; 1)$ , СВ  $\eta = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ , причем можно доказать также, что эти СВ независимы.

Представив рассматриваемую статистику в виде

$$\frac{\bar{x} - a}{\sqrt{s^2 / n}} = \frac{\frac{\bar{x} - a}{\sigma / \sqrt{n}}}{\frac{\sqrt{s^2 / n}}{\sigma / \sqrt{n}}} = \frac{\frac{\bar{x} - a}{\sigma / \sqrt{n}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\xi}{\sqrt{\frac{\eta}{n-1}}},$$

согласно определению 3 заключаем, что она подчиняется распределению Стьюдента с числом степеней свободы  $n - 1$ .  $\triangleleft$

**Опр. 4.** *F-распределением Фишера с числами степеней свободы  $k_1$  и  $k_2$*  называется распределение СВ

$$F = \frac{\eta_1 / k_1}{\eta_2 / k_2} \sim F_{k_1; k_2},$$

где СВ  $\eta_1 \sim \chi_{k_1}^2$  и  $\eta_2 \sim \chi_{k_2}^2$  независимы.

WWWИКИСПРАВКАWWWWWWWWWWWWWWW



**Рональд Эйлер Фишер**  
(англ. *Ronald Aylmer Fisher*)  
(1890–1962)

английский статистик, биолог-эволюционист и генетик, «отец современной статистики», один из основоположников математической генетики, по образованию математик и физик-теоретик.

С его именем связаны многие понятия математической статистики, он построил теорию точечных и интервальных статистических оценок, внес существенный вклад в создание современной теории проверки статистических гипотез, положил начало использованию статистических процедур при планировании

научного эксперимента. Фундаментальный в теории вероятностей термин «дисперсия» также был введен Фишером в 1916 г.

Большинство методов Фишера имеют общий характер и применяются в естественных науках, в экономике и в других областях деятельности. Его книга «Статистические методы для исследователей», опубликованная в 1925 г., переиздавалась в течение 50 лет.

$F$ -распределение исследовалось и было названо в честь Фишера его учеником Джорджем Снедекором (1881–1974), хотя сам Фишер рассматривал ве-

личину  $\zeta = \frac{1}{2} \ln F$ , распределение которой сейчас называют  $z$ -

распределением Фишера. В современной статистической практике предпочитают использовать  $F$ -распределение, имеющее более простые свойства.

WWWWWWWWWWWWWWW

Плотность распределения Фишера с  $k_1$  и  $k_2$  степенями свободы имеет вид

$$f_{F_{k_1; k_2}}(x) = \begin{cases} \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1} \left(1 + \frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}} & \text{при } x > 0, \\ 0 & \text{при } x \leq 0. \end{cases}$$

На рис. 32 представлены кривые плотностей распределения Фишера в зависимости от значений  $k_1$  и  $k_2$ .

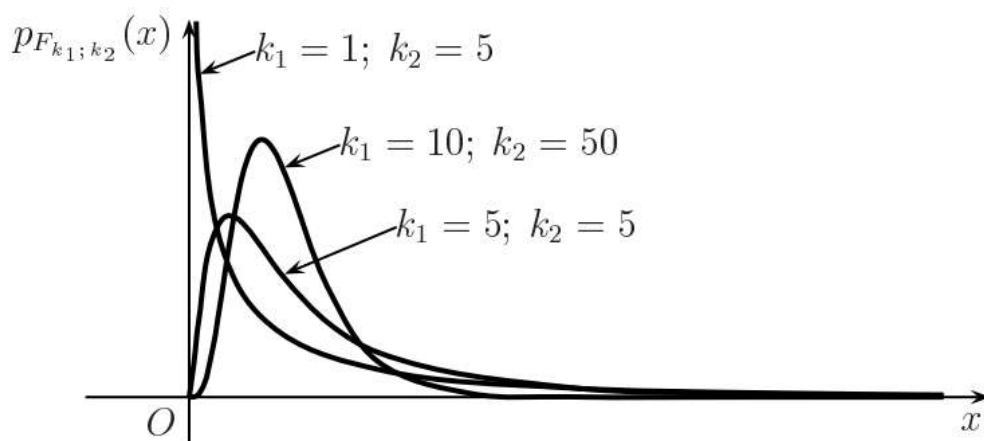


Рис. 32. Графики плотностей распределения Фишера в зависимости от значений  $k_1$  и  $k_2$

*Замечание.* Существуют таблицы распределений  $\chi^2$ , Стьюдента, Фишера. Однако при использовании статистических таблиц необходимо обращать пристальное внимание на то, какие величины затабулированы и какие нужны при вычислениях.

### Построение доверительного интервала для математического ожидания в случае выборки из нормального распределения с неизвестной дисперсией

**Утв. 5.** Доверительный интервал для математического ожидания  $a$  в случае выборки из нормального распределения с неизвестной дисперсией  $\sigma^2$  определяется соотношением

$$P\left(\bar{x} - t_{\alpha; n-1} \frac{s}{\sqrt{n}} < a < \bar{x} + t_{\alpha; n-1} \frac{s}{\sqrt{n}}\right) = 1 - \alpha,$$

где  $n$  – объем выборки;  $\bar{x}$  – выборочное среднее;  $s^2$  – несмещенная оценка дисперсии;  $\alpha$  – уровень значимости;  $t_{\alpha; n-1}$  – *квантиль* уровня  $\alpha$  распределения Стьюдента с числом степеней свободы  $k = n - 1$ , т. е. такое число, что для СВ  $\xi$ , имеющей распределение Стьюдента с числом степеней свободы  $k = n - 1$ , имеет место  $P(|\xi| \geq t_{\alpha; n-1}) = \alpha$ .

*Упражнение 3.* Доказать аналогично доказательству утверждения 2.

Квантиль  $t_{\alpha; n-1}$  определяется по таблице распределения Стьюдента.

При малых выборках ( $n < 30$ ) распределение Стьюдента дает не вполне определенные результаты (широкий доверительный интервал). Это объясняется тем, что малая выборка содержит малую информацию об интересующем нас признаке. С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормальному.

### **Построение доверительного интервала для дисперсии в случае выборки из нормального распределения с неизвестным математическим ожиданием**

**Утв. 3.** Доверительный интервал для дисперсии  $\sigma^2$  в случае выборки из нормального распределения с *неизвестным* математическим ожиданием  $a$  определяется соотношением

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2; n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2; n-1}}\right) = 1 - \alpha,$$

где  $n$  – объем выборки;  $s^2$  – несмещенная оценка дисперсии;  $\alpha$  – уровень значимости;  $\chi^2_{\alpha/2; n-1}$  и  $\chi^2_{1-\alpha/2; n-1}$  – *квантили* распределения  $\chi^2$  с числом степеней свободы  $k = n - 1$ , определяемые соотношением  $P(\xi \geq \chi^2_{\alpha; n-1}) = \alpha$  для СВ  $\xi$ , имеющей распределение  $\chi^2$  с числом степеней свободы  $k = n - 1$ .

Квантили  $\chi^2_{\alpha/2; n-1}$  и  $\chi^2_{1-\alpha/2; n-1}$  определяются по таблице распределения  $\chi^2$ .

*Доказательство.* Из утверждения 3 следует, что для выборки из нормального распределения с неизвестным математическим ожиданием

$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ . В силу несимметричности графика плотности распределения  $\chi^2$  для построения доверительного интервала будут использованы две квантили  $\chi_{\alpha/2; n-1}^2$  и  $\chi_{1-\alpha/2; n-1}^2$  (см. рис. 33), такие, что для СВ  $\xi \sim \chi_{n-1}^2$  имеют место соотношения  $P(\xi \geq \chi_{\alpha/2; n-1}^2) = \frac{\alpha}{2}$  и  $P(\xi \leq \chi_{1-\alpha/2; n-1}^2) = \frac{\alpha}{2}$ .

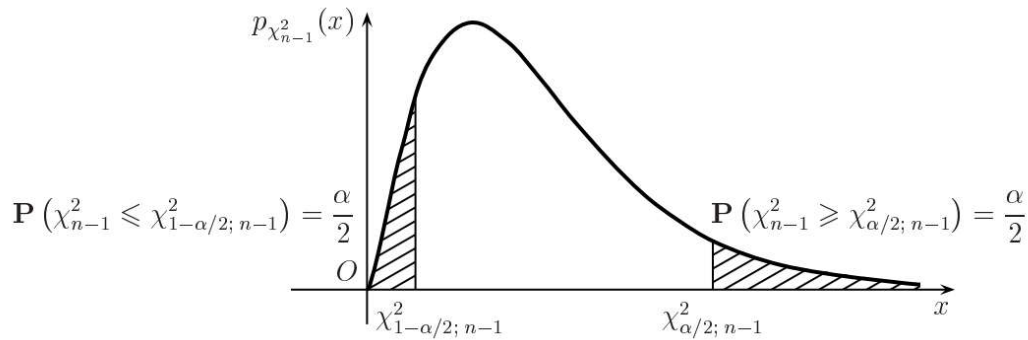


Рис. 33. К построению доверительного интервала для дисперсии в случае выборки из нормального распределения с неизвестным математическим ожиданием

Тогда

$$P\left(\chi_{1-\alpha/2; n-1}^2 < \frac{(n-1)s^2}{\sigma^2} < \chi_{\alpha/2; n-1}^2\right) = 1 - \alpha;$$

$$P\left(\frac{1}{\chi_{\alpha/2; n-1}^2} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi_{1-\alpha/2; n-1}^2}\right) = 1 - \alpha;$$

$$P\left(\frac{(n-1)s^2}{\chi_{\alpha/2; n-1}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2; n-1}^2}\right) = 1 - \alpha. \triangleleft$$

#### § 4. Проверка статистических гипотез

Основные задачи математической статистики разделяют на две категории, тесно связанные между собой, но отличающиеся постановкой задач: оценивание параметров и проверка статистических гипотез. Основной задачей оценивания параметров является

ся получение по выборке оценок, наилучших в том или ином смысле. При проверке гипотез задача ставится иначе: требуется по выборке принять или отвергнуть некоторое предположение о распределении генеральной совокупности, из которой извлечена выборка.

**Опр. 1.** *Статистической гипотезой* называется любое предположение о виде (*непараметрическая гипотеза*) или параметрах (*параметрическая гипотеза*) неизвестного распределения.

**Опр. 2.** Статистическая гипотеза называется *простой*, если она полностью определяет функцию распределения. В противном случае гипотеза называется *сложной*.

**Пример 1.** Предположим, что введен новый способ производства некоторого товара. Для определения качества товара измеряется некоторая его характеристика  $\xi \sim \mathcal{N}(a_0; \sigma_0)$ , где  $a_0, \sigma_0$  известны. Если необходимо выяснить, как новый способ производства влияет на качество товара, можно выдвинуть, например, такие гипотезы:

$H_1 : a = a_0, \sigma = \sigma_0$ , т. е. распределение СВ  $\xi$  не изменилось после изменения процесса производства;

$H_2 : a > a_0, \sigma = \sigma_0$ , т. е. увеличилось среднее значение показателя качества;

$H_3 : a = a_0, \sigma < \sigma_0$ , т. е. разброс значений показателя качества стал меньше.

Гипотеза  $H_1$  является простой, а гипотезы  $H_2$  и  $H_3$  – сложными. •

**Опр. 3.** Проверяемую гипотезу обычно называют *нулевой* и обозначают  $H_0$ . Наряду с нулевой рассматривают *альтернативную*, или *конкурирующую*, гипотезу  $H_a$  (или  $H_1$ , или  $\bar{H}$ ).

Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки гипотез.

**Пример 2.** Например,

$$\begin{array}{ll} 1) & \begin{array}{l} H_0 : \theta = \theta_0, \\ \bar{H}_1 : \theta \neq \theta_0; \end{array} \\ 2) & \begin{array}{l} H_0 : \theta = \theta_0, \\ \bar{H}_2 : \theta > \theta_0. \end{array} \end{array}$$



В первом случае нужно определить, можно ли считать, что значение параметра  $\theta$  равно заданному значению  $\theta_0$ . Ясно, что значение параметра  $\theta$ , оцененное по выборке, практически всегда будет отличаться от заданного значения  $\theta_0$ . Вопрос в том, каким должно быть это отличие, чтобы нулевую гипотезу можно было считать верной и при каком отличии нужно отвергнуть нулевую гипотезу и принять альтернативную.

Во втором случае говорят, что рассматривается *односторонняя* альтернатива. Такая задача возникает, если заранее известно, что значение параметра  $\theta$  не может быть меньше, чем  $\theta_0$ . •

**Опр. 4.** Правило, которое позволяет по выборке принять или отвергнуть проверяемую гипотезу, называется **критерием проверки статистической гипотезы (статистическим критерием)**.

*Замечание.* Статистическими методами *нельзя доказать* правильность гипотезы. Критерий проверки статистической гипотезы позволяет отбросить гипотезу как неправильную, но не позволяет доказать, что она верна, т. е. статистические критерии указывают лишь на отсутствие опровержения со стороны имеющихся экспериментальных данных. Если по результатам проверки статистическая гипотеза принимается, то говорят, что она *согласуется с выборочными данными* или что она *не противоречит результатам наблюдений*.

Статистический критерий обычно основывается на некоторой статистике  $\hat{\theta}_n$ , для которой известно ее точное или приближенное распределение. Множество всех возможных значений этой статистики разбивается на два непересекающихся подмножества:  $S$  – **область принятия нулевой гипотезы** и  $W$  – область отклонения нулевой гипотезы.  $W$  называется **критической областью**.

В задачах проверки гипотез возможны следующие четыре ситуации.

| Проверяемая гипотеза $H_0$ : | $H_0$ принимается –     | $H_0$ отвергается –     |
|------------------------------|-------------------------|-------------------------|
| объективно верна             | правильное решение      | <b>ошибка 1-го рода</b> |
| объективно неверна           | <b>ошибка 2-го рода</b> | правильное решение      |

**Опр. 5.** Вероятность ошибки 1-го рода, т. е. вероятность отвергнуть нулевую гипотезу, когда она верна, называется **уровнем значимости** статистического критерия и обозначается  $\alpha$  :

$$P(H_0 \text{ отвергается} | H_0 \text{ верна}) = P(\hat{\theta}_n \in W | H_0 \text{ верна}) = \alpha.$$

Вероятность ошибки 2-го рода, т. е. вероятность ошибочно принять нулевую гипотезу, обозначается  $\beta$  :

$$P(H_0 \text{ принимается} | H_0 \text{ не верна}) = P(\hat{\theta}_n \in S | H_0 \text{ не верна}) = \beta.$$

Пользуясь терминологией статистического контроля качества продукции, можно сказать, что  $\alpha$  – это риск поставщика (забраковка партии, удовлетворяющей стандарту), а  $\beta$  – риск потребителя (принятие партии, не удовлетворяющей стандарту).

**Опр. 6. Мощностью критерия** называется вероятность отклонить проверяемую гипотезу  $H_0$ , когда она неверна. Эта вероятность равна

$$P(H_0 \text{ отвергается} | H_0 \text{ не верна}) = 1 - \beta.$$

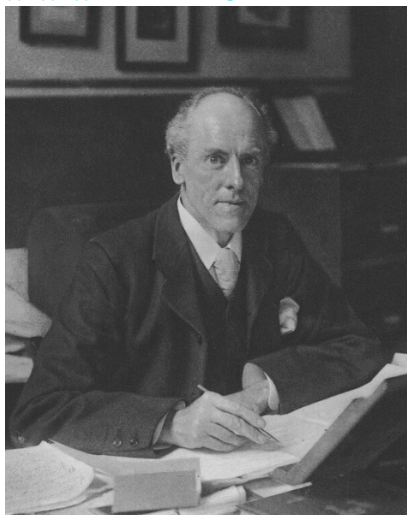
Ясно, что при построении статистических критериев желательно, чтобы вероятности ошибок обоих родов были как можно меньше. Однако это требование противоречивое. Невозможно одновременно уменьшить обе ошибки. Реально поступают следующим образом: задают уровень значимости  $\alpha$  (как правило, равный 0,05; 0,01 или 0,1), а затем выбирают статистический критерий так, чтобы ошибка 2-го рода была наименьшей.

Выбор критической области, а следовательно, критерий проверки гипотезы, определяется:

- 1) выбором нулевой и альтернативной гипотез;
- 2) заданием уровня значимости  $\alpha$ ;
- 3) выбором статистики, на которой основан критерий.

**Опр. 7.** Статистические критерии, с помощью которых проверяются гипотезы о значениях параметров распределения или о соотношениях между ними, в предположении, что тип распределения известен, называются **критериями значимости** или **параметрическими критериями**.

Наиболее известными критериями согласия являются критерий  $\chi^2$  Пирсона и критерий Колмогорова.



английский математик и биолог, основатель знаменитой английской школы биометрики. Внес существенный вклад в распространение методов статистического анализа в биологии и психологии. Основные идеи Пирсона были опубликованы в серии из 19 статей под общим названием «Математический вклад в теорию эволюции». Пирсон считается одним из отцов современной статистики.

Проверяемая гипотеза представляет собой предположение о распределении наблюдаемой СВ и является простой (конкретно указывает предполагаемое распределение):

$H_0$ : функция распределения наблюдаемой СВ совпадает с  $F(x)$ ;

$\bar{H}$ : функция распределения наблюдаемой СВ не совпадает с  $F(x)$ .

Критерий согласия  $\chi^2$  Пирсона основан на сравнении эмпирических и теоретических частот попадания СВ в рассматриваемые группы (интервалы):

$n_i$  – эмпирическая частота наблюдения значений из интервала  $[x_{i-1}; x_i)$ ;

$np_i = n P(\xi \in [x_{i-1}; x_i)) = n(F(x_i) - F(x_{i-1}))$  – теоретическое значение соответствующей частоты.

Рассмотрим статистику

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

*Упражнение.* Показать, что контроль вычислений можно осуществить по формуле  $\chi_{\text{расч}}^2 = \sum_{i=1}^k \frac{n_i^2}{np_i} - n$ .

Для вычисления статистики  $\chi_{\text{расч}}^2$  нужно знать сгруппированный статистический ряд и теоретическую функцию распределения  $F(x)$  для расчета вероятностей  $p_i$ .

При этом теоретическое распределение  $F(x)$  может зависеть от одного или нескольких параметров. Пусть  $r$  – число неизвестных параметров теоретического распределения. В этом случае вместо значений параметров используются их оценки.

*Замечание.* Оценки параметров рассчитываются по сгруппированному статистическому ряду *до объединения групп*.

**Т 1.** Если теоретическая функция распределения зависит от  $r$  параметров и оценки этих параметров обладают свойствами асимптотической нормальности и асимптотической эффективности, то независимо от вида теоретической функции распределения  $F(x)$  в пределе (при  $n \rightarrow \infty$ ) статистика  $\chi_{\text{расч}}^2$  имеет распределение  $\chi^2$  с числом степеней свободы  $k - r - 1$ , где  $n_i$  – объем выборки;  $k$  – число интервалов группировки;  $r$  – количество параметров теоретической функции распределения, оцениваемых по данной выборке.

Таким образом, **критерий согласия  $\chi^2$  Пирсона** заключается в следующем: *если  $\chi^2_{\text{расч}} < \chi^2_{\alpha; k-r-1}$ , где  $\chi^2_{\alpha; k-r-1}$  определяется по таблице квантилей распределения  $\chi^2$ , то гипотеза  $H_0$  принимается (признается непротиворечащей экспериментальным данным; нет оснований отвергнуть гипотезу  $H_0$ ) на уровне значимости  $\alpha$ , а если  $\chi^2_{\text{расч}} \geq \chi^2_{\alpha; k-r-1}$ , то гипотеза  $H_0$  отвергается (не согласуется с данными эксперимента).*

Основное достоинство критерия согласия  $\chi^2$  Пирсона – его универсальность, т. е. применимость для любого закона распределения, в том числе с неизвестными параметрами. Основной недостаток – необходимость большого объема выборки (не менее 60–100 наблюдений) и произвольность группировки, влияющая на величину  $\chi^2_{\text{расч}}$ .

### Критерий согласия $\lambda$ Колмогорова

Критерий Колмогорова ( $\lambda$ -критерий) применяется для проверки гипотез о распределениях только *непрерывных* СВ. Он основан на сравнении эмпирической  $F_n^*(x)$  и гипотетической (теоретической)  $F(x)$  функций распределения.

Кроме того, при применении критерия Колмогорова предполагается, что значения параметров теоретического распределения известны. Это обстоятельство существенно сужает область практического применения критерия Колмогорова. Тем не менее, он часто используется. При этом неизвестные параметры теоретического распределения следует оценивать по выборкам большого объема, параллельных исследуемой. Если же эти параметры оцениваются по исследуемой выборке, то критерий Колмогорова оказывается менее точным в том смысле, что фактический уровень значимости оказывается меньше заданного уровня значимости, и в результате в ряде случаев повышается риск принять нулевую гипотезу  $H_0$  как правильную, когда на самом деле она противоречит опытным данным. В силу этого рекомендуется при использовании критерия Колмогорова в случае неизвестных параметров распределения задавать несколько больший уровень значимости:  $\alpha = 0,1 \div 0,2$ .

Применение критерия Колмогорова основано на использовании статистики

$$D_n = \max_x |F_n^*(x) - F(x)|.$$

Доказано, что в случае, когда значения параметров теоретического распределения известны, при любой теоретической функции распределения  $F(x)$  имеет место соотношение

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n < \lambda) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 \lambda^2} = K(\lambda).$$

При заданном уровне значимости  $\alpha$  по соответствующей таблице находят значение  $\lambda_\alpha$  такое, что  $K(\lambda_\alpha) = 1 - \alpha$ . Если  $\sqrt{n}D_n < \lambda_\alpha$ , то гипотеза  $H_0$  принимается (признается непротиворечащей экспериментальным данным).

В случае выборок малого объема  $n$  используют специальные таблицы для статистики  $D_n$ , учитывающие объем выборки  $n$ .

В случае, когда параметры теоретического распределения неизвестны, распределение статистики  $D_n$  зависит, вообще говоря, от вида теоретического распределения. В настоящее время составлены таблицы распределений статистики  $D_n$  для многих семейств распределений (нормального, показательного и др.), а также разработаны некоторые модификации статистики  $D_n$ .

## О применении критериев согласия

При выдвижении гипотезы о распределении наблюдаемой СВ выбор теоретического распределения должен базироваться в первую очередь на понимании механизма изучаемого явления. Если же этот механизм неизвестен или недостаточно изучен, предварительный выбор распределения основывается на опыте аналогичных исследований и виде гистограммы, построенной по выборке. Достоинство использования гистограммы для этой цели – простота применения и наглядность; недостаток – гистограмма может напоминать несколько распределений одновременно.

В качестве приближенного критерия для предварительного выбора закона распределения могут быть использованы **выборочные коэффициенты асимметрии и эксцесса**:

$$\hat{A} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{nD_B^{3/2}}, \quad \hat{E} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{nD_B^2} - 3.$$

Коэффициенты асимметрии и эксцесса характеризуют форму (симметричность и крутость) графика плотности распределения. Если плотность рас-

распределения симметрична, то коэффициент асимметрии равен нулю. Если плотность имеет длинный «правый хвост», то  $\hat{A} > 0$ , а если длинный «левый хвост», то  $\hat{A} < 0$ . Если распределение «сконцентрировано» вокруг среднего больше, чем нормальное (плотность имеет «острую» вершину), то  $\hat{E} > 0$ , а если меньше – то  $\hat{E} < 0$ .

Для нормального распределения эти параметры равны нулю. Поэтому если для изучаемого распределения асимметрия и эксцесс имеют небольшие значения, то можно предположить близость этого распределения к нормальному.

После предварительного выбора распределения проводится проверка выдвинутой гипотезы по одному из критериев согласия. Кроме критериев  $\chi^2$  Пирсона и  $\lambda$  Колмогорова, существует ряд других, например, критерий  $\omega^2$  Мизеса – Смирнова, критерий Шапиро – Уилка и т. д. Существует также специальный критерий для проверки гипотезы нормальности по совокупности достаточно большого числа выборок малого объема.

Выбор того или иного критерия согласия зависит от дальнейших целей исследования. Как правило, знания о типе распределения нужны для того, чтобы на их основе сделать по выборочным данным те или иные выводы. Нередко оказывается, что для справедливости этих выводов особенно важны лишь некоторые свойства теоретического закона распределения. Именно эти свойства и должны быть проверены в первую очередь. Например, при применении критерия Стьюдента к выборкам, отличающимся от нормальных, результаты будут близки к правильным, если выборки достаточно велики и коэффициенты асимметрии и эксцесса такие же, как у нормального закона.

## Критерии значимости

Напомним, что статистические критерии, с помощью которых проверяются гипотезы о значениях параметров распределения или о соотношениях между ними в предположении, что тип распределения известен, называются **критериями значимости** или **параметрическими критериями**.

Пусть по выборке объема  $n$  получена некоторая оценка  $\hat{\theta}$  для параметра  $\theta$  теоретического распределения и есть основания полагать, что истинное значение параметра  $\theta$  есть  $\theta_0$ . Тогда проверяется нулевая гипотеза  $H_0: \theta = \theta_0$  в сравнении с альтернативой  $\bar{H}: \theta \neq \theta_0$ .

Если даже нулевая гипотеза  $H_0$  верна, выборочное значение  $\hat{\theta}$  обычно не совпадает точно с  $\theta_0$ , поэтому возникает вопрос,

насколько значимо это отличие, как сильно  $\hat{\theta}$  должно отличаться от  $\theta_0$ , чтобы можно было достаточно обоснованно отвергнуть гипотезу  $H_0$ .

Если известно распределение оценки  $\hat{\theta}$  в предположении, что гипотеза  $H_0$  справедлива, то можно найти такой интервал  $(\theta_1; \theta_2)$ , что

$$P(\theta_1 < \hat{\theta} < \theta_2) = 1 - \alpha$$

для некоторого заданного уровня значимости  $\alpha$ .

Интервал  $S = (\theta_1; \theta_2)$  является областью принятия гипотезы (при уровне значимости  $\alpha$ ). В отличие от доверительного интервала  $(\hat{\theta}_1; \hat{\theta}_2)$ , границы которого определяются по результатам наблюдений, границы области принятия гипотезы задаются до проведения эксперимента и определяются выбором  $H_0, \bar{H}, \alpha$ .

Рассмотрим критерии значимости, предназначенные для проверки гипотез о значениях параметров *в случае выборок из нормального распределения*, которое на практике встречается наиболее часто. Нормальное распределение имеет два параметра: математическое ожидание  $a$  и дисперсию  $\sigma^2$ , которые оцениваются с помощью выборочного среднего и выборочной дисперсии (исправленной выборочной дисперсии) соответственно.

Выборочное среднее является оценкой для среднего значения измеряемой величины и может служить оценкой того или иного показателя качества. Дисперсия характеризует разброс экспериментальных значений, а следовательно, служит мерой точности. Например, если произведено несколько измерений одной и той же величины, то дисперсия может характеризовать точность прибора, метода измерения и т. д.

### **1. Проверка гипотезы о равенстве математического ожидания нормального распределения заданному значению.**

*Нулевая гипотеза  $H_0$  :  $a = a_0$ .*

*Альтернативная гипотеза  $\bar{H}$  :  $a \neq a_0$ .*

Требуется по выборке объема  $n$  проверить гипотезу  $H_0$  при заданном уровне значимости  $\alpha$ . При этом предполагается, что выборка взята из нормально распределенной генеральной совокупности.



Если дисперсия  $\sigma^2$  известна, то утверждение 1 §3 гласит, что при справедливости гипотезы  $H_0$  имеет место  $\frac{\bar{x} - a_0}{\sigma / \sqrt{n}} \sim \mathcal{N}(0; 1)$ . Следовательно, критерий принятия гипотезы может быть выбран из условия

$$P\left(\left|\frac{\bar{x} - a}{\sigma / \sqrt{n}}\right| < u_\alpha\right) = 1 - \alpha.$$

Таким образом, *если дисперсия  $\sigma^2$  известна, то гипотеза  $H_0$  принимается* (т. е. согласуется с результатами наблюдений) *при условии, что*

$$u_{\text{расч}} = \frac{|\bar{x} - a_0|}{\sqrt{\sigma^2 / n}} < u_{\text{табл}} = u_\alpha, \quad (1)$$

где квантиль  $u_\alpha$  удовлетворяет соотношению  $\Phi(u_\alpha) = \frac{1 - \alpha}{2}$ .

*Если дисперсия  $\sigma^2$  неизвестна, то гипотеза  $H_0$  принимается при*

$$t_{\text{расч}} = \frac{|\bar{x} - a_0|}{\sqrt{s^2 / n}} < t_{\text{табл}} = t_{\alpha; n-1}, \quad (2)$$

где квантиль  $t_{\alpha; n-1}$  определяется по таблице распределения Стьюдента.

## **2. Проверка гипотезы о равенстве заданному значению дисперсии нормального распределения.**

*Нулевая гипотеза  $H_0: \sigma^2 = \sigma_0^2$ .*

*Альтернативная гипотеза  $\bar{H}: \sigma^2 \neq \sigma_0^2$ .*

*Гипотеза  $H_0$  при заданном уровне значимости  $\alpha$  принимается, если*

$$\chi_{1-\alpha/2; n-1}^2 < \chi_{\text{расч}}^2 = \frac{(n-1)s^2}{\sigma_0^2} < \chi_{\alpha/2; n-1}^2, \quad (3)$$

где квантили  $\chi_{1-\alpha/2; n-1}^2$  и  $\chi_{\alpha/2; n-1}^2$  определяются по таблице распределения  $\chi^2$ .

**3. Сравнение двух дисперсий нормально распределенных признаков.** Такая задача возникает, если требуется сравнить точность приборов, инструментов, методов измерения. Лучшим будет тот прибор, инструмент, метод, который дает меньший разброс результатов, т. е. меньшую дисперсию.

*Нулевая гипотеза*  $H_0 : \sigma_1^2 = \sigma_2^2$ .

*Альтернативная гипотеза*  $\bar{H} : \sigma_1^2 \neq \sigma_2^2$ .

Пусть для первой дисперсии по выборке объема  $n_1$  найдена несмещенная оценка  $s_1^2$ , для второй – по выборке объема  $n_2$  оценка  $s_2^2$ .

В случае двух независимых выборок из нормального распределения, согласно утверждению 3 §3 и определению  $F$ -распределения Фишера, отношение  $\frac{s_1^2}{s_2^2}$  имеет распределение Фишера с числами степеней свободы  $f_1 = n_1 - 1$  и  $f_2 = n_2 - 1$ . Следовательно, критерий принятия гипотезы может быть выбран из условия

$$P\left(F_{1-\alpha/2; f_1; f_2} < \frac{s_1^2}{s_2^2} < F_{\alpha/2; f_1; f_2}\right) = 1 - \alpha.$$

Для квантилей распределения Фишера имеет место соотношение

$$F_{1-\alpha; f_1; f_2} = \frac{1}{F_{\alpha; f_2; f_1}}.$$

Поэтому

$$F_{1-\alpha/2; f_1; f_2} < \frac{s_1^2}{s_2^2} \Leftrightarrow \frac{1}{F_{\alpha/2; f_2; f_1}} < \frac{s_1^2}{s_2^2} \Leftrightarrow \frac{s_2^2}{s_1^2} < F_{\alpha/2; f_2; f_1}.$$

Это позволяет сформулировать критерий проверки гипотезы  $H_0$  следующим образом.

*Гипотеза  $H_0$  при заданном уровне значимости  $\alpha$  принимается, если*

$$F_{\text{расч}} = \frac{s_{\text{max}}^2}{s_{\text{min}}^2} < F_{\text{табл}} = F_{\alpha/2; f_1; f_2}. \quad (4)$$

Здесь  $F_{\text{расч}}$  равно отношению *большой* несмещенной оценки дисперсии к *меньшей*, квантиль  $F_{\alpha/2; f_1; f_2}$  определяется по таблице распределения Фишера, причем  $f_1$  и  $f_2$  – числа степеней свободы *соответственно* числителя и знаменателя, т. е. *большой* и *меньшей* оценок дисперсий.

**Опр. 9.** Если гипотеза о равенстве дисперсий принимается, то эти дисперсии считаются *однородными*. (Термин «однородные» в статистике означает «являющиеся оценкой одного и того же параметра».)

*Замечание 1.* Критерий Фишера (4) может использоваться также для проверки гипотезы о равенстве дисперсии заданному значению  $\sigma_0^2$ . В этом случае число степеней свободы известной дисперсии принимается равным  $\infty: f_{\sigma_0^2} = \infty$ .

*Замечание 2.* Критерий Фишера (4) может применяться также для проверки гипотезы о равенстве нескольких дисперсий нормально распределенных признаков. В этом случае проверяют гипотезу о равенстве наибольшей и наименьшей из сравниваемых дисперсий. Если они признаются однородными, то можно принять гипотезу о равенстве всех сравниваемых дисперсий.

#### **4. Сравнение двух средних в случае независимых нормально распределенных признаков.**

*Нулевая гипотеза*  $H_0: a_1 = a_2$ .

*Альтернативная гипотеза*  $\bar{H}: a_1 \neq a_2$ .

Требуется по выборкам объемов  $n_1$  и  $n_2$  проверить гипотезу  $H_0$  при заданном уровне значимости  $\alpha$ .

*1 случай.* Если дисперсии  $\sigma_1^2$  и  $\sigma_2^2$  известны, то гипотеза  $H_0$  принимается при условии, что

$$u_{\text{расч}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < u_{\text{табл}} = u_{\alpha}, \quad (5)$$

где квантиль  $u_{\alpha}$  удовлетворяет соотношению  $\Phi(u_{\alpha}) = \frac{1-\alpha}{2}$ .

2 случай. Если дисперсии  $\sigma_1^2$  и  $\sigma_2^2$  не известны, но на основании проверки соответствующей гипотезы по критерию Фишера признаны *однородными*, то гипотеза  $H_0$  принимается при

$$t_{\text{расч}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} < t_{\text{табл}} = t_{\alpha; f}, \quad (6)$$

где общая средневзвешенная дисперсия  $s^2$  вычисляется по формуле

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

и имеет число степеней свободы  $f = n_1 + n_2 - 2$ , значение  $t_{\alpha; f}$  определяется по таблице квантилей распределения Стьюдента.

3 случай. Если дисперсии  $\sigma_1^2$  и  $\sigma_2^2$  не известны и на основании проверки по критерию Фишера признаны *неоднородными*, то проверка также проводится по критерию Стьюдента, однако этот критерий является приближенным. В этом случае гипотеза  $H_0$  принимается, если

$$t_{\text{расч}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < t_{\text{табл}} = t_{\alpha; f}, \quad (7)$$

где квантиль  $t_{\alpha; f}$  определяется по таблице распределения Стьюдента при

$$f \approx \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left( \frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left( \frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}.$$

Отметим, что при сравнении двух средних в случае неизвестных дисперсий возникает необходимость проверки двух различ-

ных гипотез по одним и тем же данным. Сперва проверяют гипотезу о равенстве дисперсий, а затем гипотезу о равенстве средних.

**5. Сравнение нескольких средних в случае независимых нормально распределенных признаков.** Для сравнения нескольких средних в случае независимых нормально распределенных признаков используется специальная статистическая процедура, которая называется дисперсионным анализом. Однако можно сделать вывод и на основании критерия Стьюдента, проверив гипотезу о равенстве наибольшего и наименьшего средних.

Опишем процедуру *однофакторного дисперсионного анализа*.

Пусть имеется  $N$  независимых выборок объемов  $n_1, n_2, \dots, n_N$  соответственно и задан уровень значимости  $\alpha$ . Обозначим через  $\bar{x}_i, s_i^2$  несмещенные оценки математического ожидания и дисперсии, полученные по  $i$ -й выборке,  $f_i = n_i - 1$ .

*Нулевая гипотеза*  $H_0 : a_1 = a_2 = \dots = a_N$ .

*Альтернативная гипотеза*  $\bar{H}$ : не все эти математические ожидания равны между собой.

Условием применимости метода дисперсионного анализа является, помимо нормальности выборок, однородность дисперсий. Следовательно, как и в случае двух выборок, процедуре сравнения средних должно предшествовать сравнение дисперсий.

Идея однофакторного дисперсионного анализа заключается в разбиении общей дисперсии, которая получается при объединении всех наблюдений в одну выборку, на два независимых слагаемых – *факторную (межгрупповую)* дисперсию  $s_{\text{факт}}^2$ , порождаемую различием между группами (выборками), и *остаточную (внутригрупповую)* дисперсию  $s_{\text{ост}}^2$ , обусловленную случайными помехами и неучтенными факторами:  $s_{\text{общ}}^2 = s_{\text{факт}}^2 + s_{\text{ост}}^2$ . Дисперсионный анализ был первоначально предложен Р. Фишером и определен им как метод «отделения дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам».

Межгрупповая дисперсия рассчитывается по формуле

$$s_{\text{факт}}^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x}_i - \bar{\bar{x}})^2 n_i,$$

где  $\bar{\bar{x}}$  – выборочное среднее, рассчитанное по объединенной выборке; число степеней межгрупповой дисперсии равно  $f_{\text{факт}} = N - 1$ . Остаточная дисперсия представляет собой взвешенное среднее оценок дисперсий и рассчитывается по формуле

$$s_{\text{ост}}^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \dots + f_N s_N^2}{f_1 + f_2 + \dots + f_N}, \quad f_{\text{ост}} = f_1 + f_2 + \dots + f_N.$$

Гипотеза  $H_0$  при заданном уровне значимости  $\alpha$  принимается (не противоречит экспериментальным данным), если

$$F_{\text{расч}} = \frac{s_{\text{факт}}^2}{s_{\text{ост}}^2} < F_{\text{табл}} = F_{\alpha; f_{\text{факт}}; f_{\text{ост}}},$$

где  $F_{\alpha; f_{\text{факт}}; f_{\text{ост}}}$  определяется по таблице квантилей распределения Фишера.

**6. Сравнение двух средних в случае зависимых нормально распределенных признаков.** Такая задача возникает, если две выборки взаимосвязаны. Например, проводятся измерения одних и тех же величин на одних и тех же объектах двумя разными методами и требуется определить, одинаковы ли результаты использования двух методов измерения. Либо если проводятся измерения какой-то характеристики для одних и тех же объектов до и после некоторого воздействия и требуется определить, влияет ли это воздействие на значение характеристики.

В этом случае имеются две выборки одинакового объема  $n$ :

$$\begin{aligned} x_{11}, \quad x_{12}, \quad \dots, \quad x_{1n}; \\ x_{21}, \quad x_{22}, \quad \dots, \quad x_{2n}. \end{aligned}$$

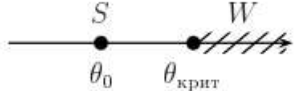
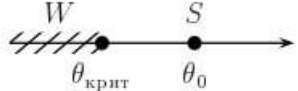
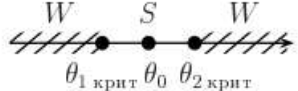
Поскольку значения в каждой паре  $x_{1i}, x_{2i}$  связаны (например, измерены на одном и том же объекте), то получим новую выборку с элементами  $\Delta x_i = x_{1i} - x_{2i}$ .

Задача сводится к проверке гипотезы о равенстве нулю среднего значения новой выборки, т. е.  $H_0: a_{\Delta x} = 0$ . Эта проверка проводится по критерию (2).

### Двусторонние и односторонние критические области

Иногда возникает необходимость сравнения гипотезы  $H_0: \theta = \theta_0$  с **односторонней** альтернативой  $\bar{H}_1: \theta > \theta_0$  или  $\bar{H}_2: \theta < \theta_0$ . Например, если известно, что неравенство  $\theta < \theta_0$  невозможно, то в качестве альтернативной рассматривается гипотеза  $\bar{H}: \theta > \theta_0$ .

Вид критической области  $W$  и области  $S$  принятия гипотезы зависит от вида альтернативной гипотезы.

|                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $H_0 : \theta = \theta_0$<br>$\bar{H} : \theta > \theta_0$<br>$W = \{\hat{\theta}_n > \theta_{\text{крит}}\}$  $P(\hat{\theta}_n > \theta_{\text{крит}}   H_0) = \alpha$<br>правосторонняя<br>критическая область | $H_0 : \theta = \theta_0$<br>$\bar{H} : \theta < \theta_0$<br>$W = \{\hat{\theta}_n < \theta_{\text{крит}}\}$  $P(\hat{\theta}_n < \theta_{\text{крит}}   H_0) = \alpha$<br>левосторонняя<br>критическая область | $H_0 : \theta = \theta_0$<br>$\bar{H} : \theta \neq \theta_0$<br>$W = \left\{ \hat{\theta}_n < \theta_{1 \text{ крит}} \right.$<br>или $\left. \hat{\theta}_n > \theta_{2 \text{ крит}} \right\}$  $P(\hat{\theta}_n < \theta_{1 \text{ крит}}   H_0) =$<br>$= P(\hat{\theta}_n > \theta_{2 \text{ крит}}   H_0) = \frac{\alpha}{2}$<br>двусторонняя<br>критическая область |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Таким образом, в зависимости от вида альтернативной гипотезы  $\bar{H}$  выбирают *правостороннюю, левостороннюю или двустороннюю* критическую область.

Например, при проверке гипотезы  $H_0 : a = a_0$  против альтернативы  $\bar{H}_1 : a > a_0$  требуется выяснить, соответствует ли выборочное среднее значение норме или превосходит ее. Пусть дисперсия  $\sigma^2$  известна. Оценкой для параметра  $a$  является  $\bar{x}$ . Ясно, что если  $\bar{x} < a_0$ , то гипотезу  $H_0$  следует предпочесть альтернативе  $\bar{H}_1$ . Если же  $\bar{x} > a_0$ , то гипотеза  $H_0$  принимается на уровне значимости  $\alpha$ , если выполняется условие (1) с  $u_{\text{табл}} = u_{2\alpha}$ , т. е. табличное значение определяется для удвоенного уровня значимости.

Аналогично с удвоенным уровнем значимости определяются табличные значения при использовании критериев (2), (5), (6), (7) в случае односторонних альтернатив.

Критерий (3) используется следующим образом. В случае альтернативы  $\bar{H}_1 : \sigma^2 > \sigma_0^2$  гипотеза  $H_0 : \sigma^2 = \sigma_0^2$  при заданном уровне значимости  $\alpha$  принимается, если

$$\chi_{\text{расч}}^2 = \frac{(n-1)s^2}{\sigma_0^2} < \chi_{\alpha; n-1}^2;$$

в случае альтернативы  $\bar{H}_2: \sigma^2 < \sigma_0^2$  гипотеза  $H_0$  принимается, если

$$\chi_{\text{расч}}^2 = \frac{(n-1)s^2}{\sigma_0^2} > \chi_{1-\alpha; n-1}^2.$$

*Замечание.* Приведенные критерии значимости являются наилучшими в указанных ситуациях, так как обеспечивают максимальную мощность.

### Условия применимости критериев значимости

Сделаем теперь несколько замечаний о границах применимости указанных критериев значимости. Все рассмотренные выше критерии проверки гипотез о средних и дисперсиях предназначены для случая нормально распределенных совокупностей. Критерии сравнения дисперсий (3)–(4) весьма чувствительны к отклонениям распределений от нормального. В то же время критерии сравнения средних (1), (2), (5)–(7) устойчивы к умеренным отклонениям распределений от нормального; критерий (6) может использоваться при умеренном отклонении от выполнения требования о равенстве дисперсий, если объемы выборок приблизительно равны. В случае невыполнения этих условий для сравнения средних двух выборок используют критерий Манна-Уитни, который не требует предположения о нормальности распределения.

## § 5. Элементы регрессионного и корреляционного анализа

Пусть на основании экспериментальных данных (по выборке объема  $n$  связанных пар наблюдений  $(x_i, y_i)$ ) изучается связь между двумя величинами.

Две СВ могут быть:

- 1) независимыми;
- 2) связаны функциональной зависимостью (каждому значению одной из них соответствует строго определенное значение другой);
- 3) связаны статистической зависимостью.

**Опр. 1. Статистической (стохастической, вероятностной)** называется такая зависимость между СВ, при которой каждому значению одной из них соответствует множество возможных значений другой и изменение значения одной величины влечет изменение *распределения* другой, в частности, может изменяться *среднее значение* другой.



**Пример 1.** Статистической является зависимость урожайности некоторой культуры от количества вносимых удобрений или количества осадков; зависимость спроса на товар от его цены; надежности автомобиля от его возраста и т. д. •

Статистическая зависимость возникает из-за того, что на зависимую переменную влияют какие-то неучтенные или неконтролируемые факторы.

При изучении статистической зависимости обычно ограничиваются исследованием усредненной зависимости: как в среднем будет изменяться значение одной величины при изменении другой. Такая зависимость называется *регрессионной*.

**Опр. 2.** *Регрессионная зависимость* между двумя СВ – это функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Основным методом исследования статистических зависимостей является *корреляционно-регрессионный* анализ.

*WWWИКИСПРАВКАWWW*

Понятия корреляции и регрессии появились во второй половине XIX в. в работах К. Пирсона и Ф. Гальтона (1822–1911, английский психолог и антрополог, первым ввел в биологию и психологию математические методы). Слово «корреляция» восходит к лат. correlatio – «соотношение, взаимосвязь»; «регрессия» происходит от лат. regressio – «движение назад». Термин «регрессия» ввел Ф. Гальтон, который, изучая зависимость между ростом отцов и сыновей, обнаружил явление «регрессии к среднему»: у детей, родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней величине.

*WWW*

**Основными задачами корреляционного анализа** являются выявление связи между наблюдаемыми СВ и оценка тесноты этой связи.

**Основными задачами регрессионного анализа** являются установление *формы зависимости* между наблюдаемыми величинами и определение по экспериментальным данным уравнения зависимости, которое называют **выборочным (эмпирическим) уравнением регрессии**, а также прогнозирование с помощью уравнения регрессии среднего значения зависимой переменной при заданном значении независимой переменной.

Вид эмпирической функции регрессии определяют исходя из: 1) соображений о физической сущности исследуемой зависимости; 2) опыта предыдущих исследований; 3) характера расположения точек на **корреляционном поле**, которое получается, если отметить на плоскости все точки с координатами  $(x_i, y_i)$ , соответствующие наблюдениям.

Наибольший интерес представляет линейное эмпирическое уравнение регрессии  $\hat{y} = b_0 + b_1x$ , так как: 1) это наиболее простой случай для расчетов и анализа; 2) при нормальном распределении функция регрессии является линейной.

### Выборочный коэффициент корреляции и его свойства

Пусть на основании экспериментальных данных изучается связь между двумя величинами. Тогда выборка объема  $n$  представляет собой  $n$  пар значений  $(x_i, y_i)$ .

Количественной мерой *линейной связи* между двумя наблюдаемыми величинами служит выборочный коэффициент корреляции.

**Опр. 3. Выборочный коэффициент корреляции** между наблюдаемыми величинами  $x$  и  $y$  определяется соотношением

$$r_{x,y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{D_B(x)D_B(y)}}, \quad (1)$$

где  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ,  $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$  – выборочные средние величин  $x$ ,  $y$  и произведения  $xy$  соответственно;  $D_B(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2$ ,  $D_B(y) = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2$  – выборочные дисперсии величин  $x$  и  $y$ .

**Свойства выборочного коэффициента корреляции.**

1.  $-1 \leq r_{x,y} \leq 1$ .
2. Если наблюдаемые величины  $x$  и  $y$  независимы, то  $r_{x,y} \approx 0$ .
3. Если  $|r_{x,y}| = 1$  (или близок к 1), то наблюдаемые величины  $x$  и  $y$  связаны линейной зависимостью, т. е.  $y = b_0 + b_1x$ .

4. Если  $r_{x,y} > 0$ , то с ростом значений одной величины значения другой также в основном возрастают; если  $r_{x,y} < 0$ , то с ростом значений одной величины значения другой, наоборот, убывают.

На рис. 34 приведены примеры корреляционных полей и соответствующих им коэффициентов корреляции.

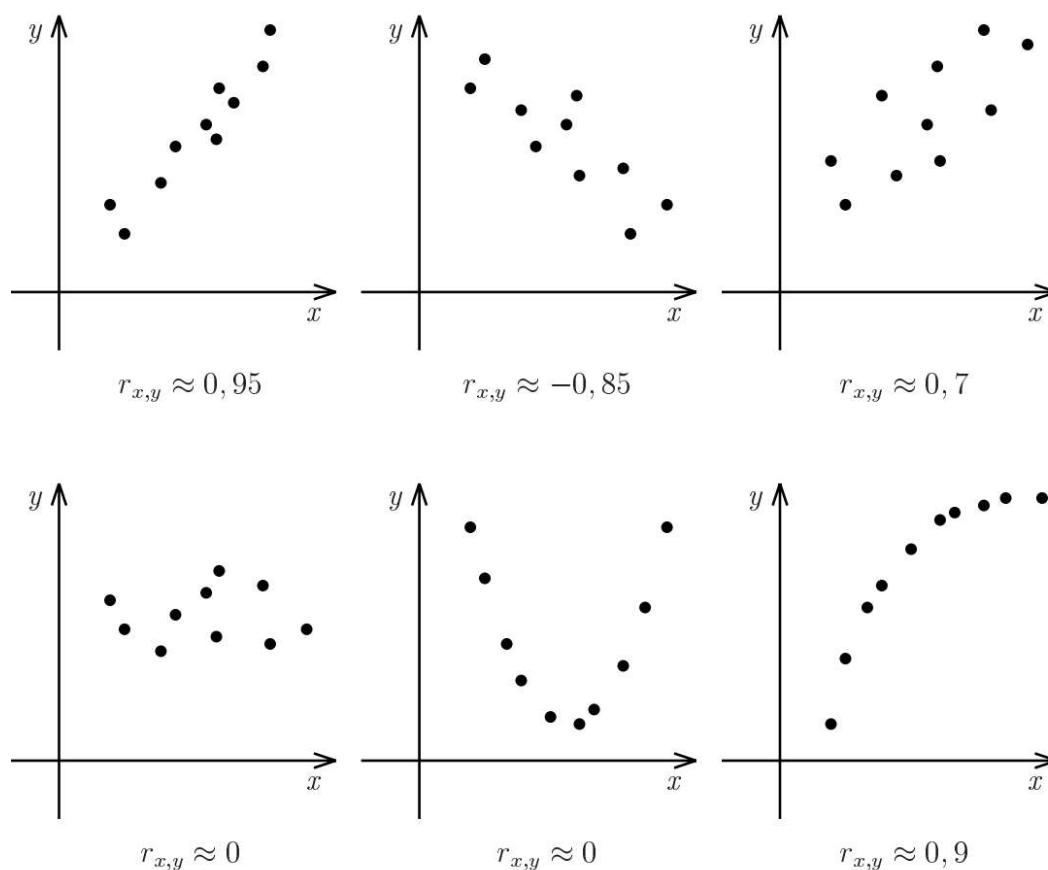


Рис. 34. Примеры корреляционных полей

Для первого корреляционного поля на рис. 8 выборочный коэффициент корреляции  $r_{x,y} \approx 0,95$ , т. е. близок к 1, экспериментальные точки расположены вдоль некоторой прямой.

Как можно видеть по второй и третьей диаграммам рис. 34, если корреляционное поле менее сконцентрировано около прямой линии, коэффициент корреляции уменьшается (по модулю). В практических задачах даже значение  $r_{x,y} \approx 0,7$  может быть признано довольно высоким, свидетельствующим о линейной зависимости между величинами.

На четвертом корреляционном поле точки разбросаны, зависимость между увеличением значений одной величины и увеличением или уменьше-

нием значений второй не прослеживается, коэффициент корреляции близок к 0.

Пятая иллюстрация показывает, что равенство 0 коэффициента корреляции не означает независимость между наблюдаемыми величинами.

Отметим, что коэффициент корреляции является мерой именно линейной зависимости. На пятом и шестом рисунках точки образуют не прямые, а кривые линии, т. е. зависимость между переменными явно присутствует, но отличается от линейной, и коэффициент корреляции не равен 1 по абсолютной величине.

## Проверка значимости коэффициента корреляции

**Проверка значимости коэффициента корреляции** – это проверка гипотезы о том, что коэффициент корреляции значимо отличается от нуля. Так как выборка произведена случайно, нельзя утверждать, что если выборочный коэффициент корреляции  $r_{x;y} \neq 0$ , то и коэффициент корреляции генеральной совокупности  $r_{\xi;\eta} \neq 0$ . Возможно, отличие  $r_{x;y}$  от нуля вызвано только случайными искажениями наблюдаемых значений.

Если выборка из нормального распределения, то проверка производится по *критерию Стьюдента*: если

$$t_{\text{расч}} = |r_{x;y}| \sqrt{\frac{n-2}{1-r_{x;y}^2}} > t_{\text{табл}} = t_{\alpha; n-2},$$

где  $t_{\alpha; n-2}$  – квантиль уровня  $\alpha$  распределения Стьюдента с числом степеней свободы  $k = n - 2$  (определяется по таблице), то при заданном уровне значимости  $\alpha$  (допускается, что вывод может быть ошибочным с небольшой вероятностью  $\alpha$ ) коэффициент корреляции считается значимо отличающимся от нуля, а следовательно, связь между величинами  $x, y$  признается статистически значимой.

Подчеркнем, что *коэффициент корреляции является мерой именно линейной зависимости*. В случае нелинейной зависимости связь между величиной коэффициента корреляции и близостью точек корреляционного поля к некоторой линии не прослеживается. Поэтому в практических задачах при выборе вида эмпирической функции регрессии обязательно учитывают характер расположения точек на корреляционном поле.

## Определение коэффициентов уравнения линейной регрессии методом наименьших квадратов

Пусть имеется выборка объема  $n$  наблюдений над двумя величинами  $x$  и  $y$ :  $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$ , и принята гипотеза о линейной зависимости между  $y$  и  $x$ .

Для определения коэффициентов линейного эмпирического уравнения регрессии

$$\hat{y} = b_0 + b_1 x$$

используется **метод наименьших квадратов (МНК)**. Суть этого метода в том, что коэффициенты  $b_0$  и  $b_1$  выбирают так, чтобы сумма квадратов отклонений наблюдаемых значений  $y_i$  от предсказываемых по уравнению  $\hat{y}_i = b_0 + b_1 x_i$  была минимальной (см. рис. 35). Таким образом, минимизируется функция

$$Q(b_0; b_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \rightarrow \min_{b_0, b_1}.$$

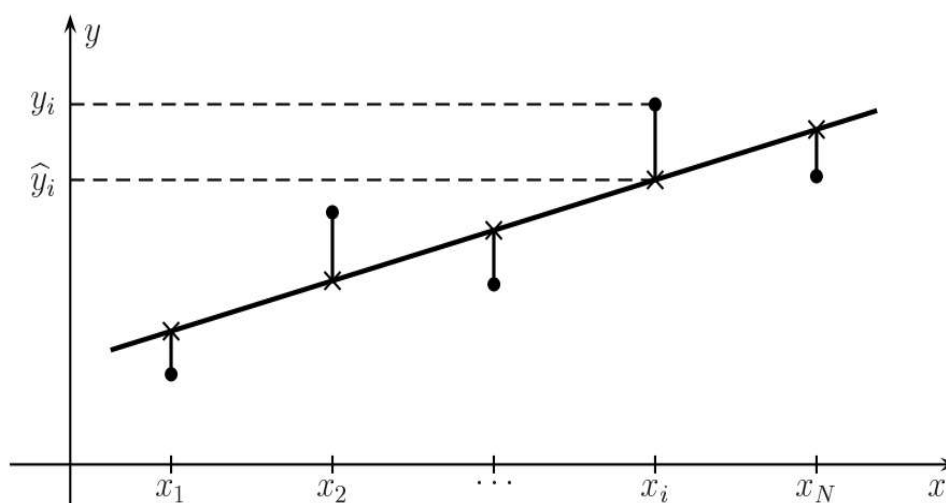


Рис. 35. Геометрическая интерпретация МНК

**WWWИКИСПРАВКАWWWWWWWWWWWW**

МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным и широко используется для статистических выводов в различных областях науки и техники.

МНК был разработан К. Гауссом в 1794–95 гг. и, независимо от него,

А. Лежандром в 1805–06 гг., название метода было предложено А. Лежандром. Первоначально МНК использовался как вычислительная процедура для обработки результатов астрономических и геодезических наблюдений. Строгое математическое обоснование и установление границ содержательной применимости этого метода даны А. А. Марковым (1856–1922, русский математик, академик, внесший большой вклад в теорию вероятностей, математический анализ и теорию чисел) и А. Н. Колмогоровым.

~~~~~

Необходимым условием существования минимума данной функции двух переменных является равенство нулю ее частных производных по неизвестным параметрам b_0, b_1 :

$$\frac{\partial Q}{\partial b_0} = 0, \frac{\partial Q}{\partial b_1} = 0.$$

Вычислим частные производные:

$$\begin{cases} \frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i), \\ \frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i. \end{cases}$$

Приравняв частные производные к 0, сократим оба уравнения на -2 и упростим:

$$\begin{cases} \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0, \\ \sum_{i=1}^n (x_i y_i - b_0 x_i - b_1 x_i^2) = 0; \end{cases} \quad \begin{cases} \sum_{i=1}^n y_i - n b_0 - b_1 \sum_{i=1}^n x_i = 0, \\ \sum_{i=1}^n x_i y_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 = 0. \end{cases}$$

Итак, значения параметров b_0 и b_1 находят из системы, которая называется **системой нормальных уравнений** метода наименьших квадратов:

$$\boxed{\begin{cases} n b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}} \quad (2)$$

Метод наименьших квадратов широко применяется при статистической обработке результатов измерений.

Пример 2. При контроле качества пищевых продуктов для определения концентрации тех или иных веществ находят эмпирическое линейное уравнение зависимости оптической плотности градуировочного раствора от концентрации. Имеются данные для определения концентрации фосфора в мясных изделиях.

Концентрация раствора, мг/кг	0,02	0,04	0,06	0,08	0,10
Оптическая плотность раствора	0,035	0,070	0,150	0,140	0,175

По имеющимся данным требуется:

- 1) построить корреляционное поле;
- 2) найти выборочный коэффициент корреляции и проверить его значимость при $\alpha = 0,05$;
- 3) определить коэффициенты линейного эмпирического уравнения регрессии, построить прямую на корреляционном поле.

Решение. 1) Построим корреляционное поле, отмечая по оси Ox данные концентрации, а по оси Oy – соответствующие им значения оптической плотности (рис. 36).

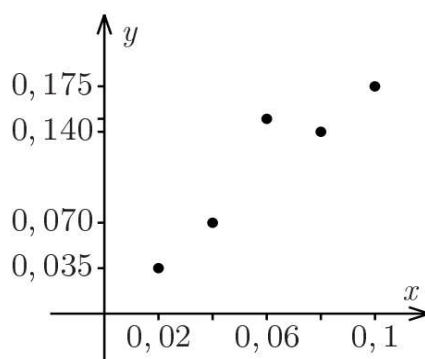


Рис. 36. Корреляционное поле

2) По виду корреляционного поля можно предположить, что выборочный коэффициент корреляции положителен и значительно отличается от 0. Рассчитаем его по формуле (1):

$$r_{x;y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{D_B(x)D_B(y)}},$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$; $D_B(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2$,
 $D_B(y) = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2$.

Число опытов $n = 5$. Для расчета необходимых сумм удобно составить вспомогательную таблицу.

	x_i – концен- трация	y_i – оптическая плотность	$x_i y_i$	x_i^2	y_i^2
1	0,02	0,035	0,0007	0,0004	0,001225
2	0,04	0,070	0,0028	0,0016	0,004900
3	0,06	0,150	0,0090	0,0036	0,022500
4	0,08	0,140	0,0112	0,0064	0,019600
5	0,10	0,175	0,0175	0,0100	0,030625
Σ	0,30	0,570	0,0412	0,0220	0,078850

Тогда $\bar{x} = \frac{0,30}{5} = 0,06$; $\bar{y} = \frac{0,570}{5} = 0,114$;

$$\overline{xy} = \frac{0,0412}{5} = 0,00824;$$

$$D_B(x) = \frac{0,022}{5} - (0,06)^2 = 0,0008;$$

$$D_B(y) = \frac{0,07885}{5} - (0,114)^2 = 0,002774.$$

Подставляя в (1), получим выборочный коэффициент корреляции

$$r_{x,y} = \frac{0,00824 - 0,06 \cdot 0,114}{\sqrt{0,0008 \cdot 0,002774}} \approx 0,94.$$

Поскольку

$$t_{\text{расч}} = 0,94 \cdot \sqrt{\frac{5-2}{1-0,94^2}} \approx 4,76 > t_{\text{табл}} = t_{0,05; 5-2} = 3,18,$$

то можно считать при уровне значимости $\alpha = 0,05$, что величины x и y связаны линейной зависимостью, т. е. $y = b_0 + b_1 x$.

3) Для того, чтобы найти коэффициенты b_0 и b_1 линейного эмпирического уравнения регрессии $\hat{y} = b_0 + b_1x$ по МНК, составим систему нормальных уравнений по формуле (2):

$$\begin{cases} 5b_0 + 0,3b_1 = 0,57, \\ 0,3b_0 + 0,022b_1 = 0,0412. \end{cases}$$

Решим систему по формулам Крамера:

$$b_0 = \frac{\begin{vmatrix} 0,57 & 0,3 \\ 0,0412 & 0,022 \end{vmatrix}}{\begin{vmatrix} 5 & 0,3 \\ 0,3 & 0,022 \end{vmatrix}} = \frac{0,57 \cdot 0,022 - 0,3 \cdot 0,0412}{5 \cdot 0,022 - 0,3 \cdot 0,3} = 0,009;$$

$$b_1 = \frac{\begin{vmatrix} 5 & 0,57 \\ 0,3 & 0,0412 \end{vmatrix}}{\begin{vmatrix} 5 & 0,3 \\ 0,3 & 0,022 \end{vmatrix}} = \frac{5 \cdot 0,0412 - 0,3 \cdot 0,57}{5 \cdot 0,022 - 0,3 \cdot 0,3} = 1,75.$$

Итак, эмпирическое линейное уравнение регрессии имеет вид $\hat{y} = 0,009 + 1,75x$.

Построим прямую на корреляционном поле (рис. 37).

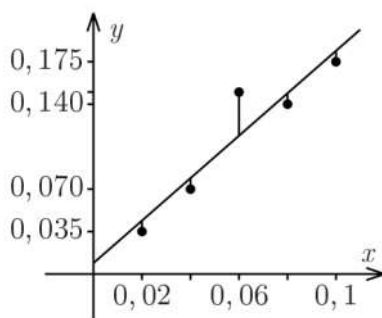


Рис. 37. Отклонение точек корреляционного поля от прямой регрессии

Согласно МНК, построенная прямая приближает экспериментальные данные наилучшим образом в том смысле, что будет наименьшей сумма квадратов длин вертикальных отрезков, показанных на рисунке. •

Криволинейная регрессия

Зависимость между двумя наблюдаемыми величинами далеко не всегда можно выразить линейной функцией. Иногда видно, что точки корреляционного поля образуют некоторую кривую. При выборе вида эмпирической функции регрессии необходимо учитывать теоретические сведения и опыт предыдущих аналогичных исследований.

Как правило, до начала исследования должен быть определен вид эмпирической функции регрессии с точностью до нескольких параметров, значения которых оцениваются по результатам эксперимента. В том случае, если функция регрессии линейна по параметрам или может быть сведена к таковой с помощью замены переменных, для определения оценок параметров используют МНК.

Например, степенная зависимость вида $y = ax^b$ может быть сведена к линейной с помощью логарифмирования:

$$\ln y = \ln a + \ln x^b \Rightarrow \ln y = \ln a + b \ln x.$$

Если ввести новые переменные $Y = \ln y$, $X = \ln x$, исходная зависимость сведется к линейной $Y = b_0 + b_1 X$, коэффициенты которой могут быть найдены по МНК. Тогда коэффициенты искомой зависимости определяются из соотношений $a = e^{b_0}$, $b = b_1$.

Логарифмируя уравнение $y = a e^{bx}$, получим

$$\ln y = \ln a + \ln e^{bx} \Rightarrow \ln y = \ln a + bx.$$

Следовательно, такая зависимость линеаризуется с помощью замены $Y = \ln y$, $X = x$.

Логарифмическая зависимость $y = a + b \ln x$ является линейной относительно переменных $Y = y$ и $X = \ln x$, а гиперболическая

зависимость вида $y = a + \frac{b}{x}$ — линейной относительно $Y = y$ и $X = \frac{1}{x}$.

Гиперболическая зависимость вида $y = \frac{1}{a + bx}$ может быть преобразована к виду $\frac{1}{y} = a + bx$, поэтому она сводится к линейной зависимости относительно переменных $Y = \frac{1}{y}$, $X = x$.

Аналогично можно показать, что гиперболическая зависимость вида $y = \frac{x}{a + bx}$ сводится к линейной посредством замены

$$Y = \frac{x}{y}, X = x.$$

Для проверки того, удачно ли выбран вид зависимости, следует построить новое корреляционное поле на плоскости OXY . Если вид зависимости y от x подобран правильно, то точки $(X_i; Y_i)$ будут располагаться вдоль прямой.

