ВОСТОЧНО-КАЗАХСТАНСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

А.В. Русакова

# ТЕХНОЛОГИИ И МЕТОДЫ АНАЛИЗА МЕДИКО – БИОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММЫ EXCEL

(теоретические сведения для решения практических задач)

Усть-Каменогорск, 2024

# Введение

Математико-статистическое описание данных медицинских исследований и оценка значимости различия величин, характеризующих эффективность проводимых профилактических, диагностических и лечебных мероприятий, являются основополагающими для доказательной медицины.

Не отягощая читателя излишней математикой, авторы учебно-методического пособия пытались решить две задачи:

1. Определить базовые понятия теории вероятностей и статистики и объяснить их смысл. Без этого невозможно осмысленно применять методы статистического анализа данных (гл. 1 и 2).

2. Рассмотреть этапы проведения анализа на конкретных примерах, используя табличный процессор Excel таким образом, чтобы любой студент-медик, аспирант или даже врач, взяв это учебно-методическое пособие и определив цель исследования, всегда мог самостоятельно получить нужный результат (гл. 3).

В издании рассмотрены далеко не все используемые сегодня на практике статистические методы. Мы ограничиваемся следующими: «Описательная статистика», «Элементы корреляционного анализа», «Оценка значимости различия признаков (статистические гипотезы и критерии проверки гипотез)». Именно они, как показывает знакомство с медицинской литературой, прежде всего необходимы при анализе полученных результатов. Ограничение так же связано с малым временем, отведенным программой на эту работу.

# Глава 1
# Основные понятия теории вероятностей. Случайные величины. Законы распределения случайных величин

## 1.1. Закономерность и случайность, случайная изменчивость в точных науках, биологии и медицине

Теория вероятностей — область математики, которая изучает закономерности в *случайных явлениях*. Случайное явление — это явление, которое при неоднократном воспроизведении одного и того же опыта может протекать каждый раз несколько по-иному.

Очевидно, что в природе нет ни одного явления, в котором не присутствовали бы в той или иной мере элементы случайности, но в различных ситуациях мы учитываем их по-разному. Так, в ряде практических задач ими можно пренебречь и рассматривать вместо реального явления его упрощенную схему — «модель», предполагая, что в данных условиях опыта оно протекает вполне определенным образом. При этом выделяются самые главные, решающие факторы, характеризующие явление. Именно такая схема изучения явлений чаще всего применяется в физике и технике.

Однако при решении многих задач многочисленные, тесно переплетающиеся между собой случайные факторы часто играют определяющую роль. Здесь на первый план выступает *случайная природа* явления, которой уже нельзя пренебречь. Это явление необходимо изучать именно с точки зрения закономерностей, присущих ему как случайному явлению.

Предмет изучения биологов и медиков — живой организм, зарождение, развитие и существование которого определяется очень многими и разнообразными, часто случайными внешними и внутренними условиями. Именно поэтому явления и события живого мира во многом тоже *случайны* по своей природе.

Элементы неопределенности, сложности, многопричинности, присущие случайным явлениям, обусловливают необходимость создания специальных математических методов для их изучения.

*Разработка таких методов, установление специфических закономерностей, свойственных случайным явлениям — главные задачи теории вероятностей.* Характерно, что эти закономерности выполняются лишь при массовости случайных явлений. Причем индивидуальные особенности отдельных случаев как бы взаимно погашаются, а усредненный результат для массы случайных явлений оказывается уже закономерным. В значительной

мере данное обстоятельство — причина широкого распространения вероятностных методов исследования в биологии и медицине.

Статистика возникла существенно раньше теории вероятностей. Еще в глубокой древности проводились переписи населения и велись земельные кадастры. Эти операции были связаны с наблюдениями и вычислениями. На протяжении веков статистика искала свой математический аппарат и нашла его в теории вероятностей. *В результате возник такой раздел математики, как математическая статистика, в котором устанавливаются закономерности случайных явлений на основании обработки статистических данных — результатов наблюдений и измерений.*

## 1.2. Вероятность случайного события

*Случайное событие — это всякое явление (факт), которое в результате опыта (испытания) может произойти или не произойти*. Случайные события часто обозначаются буквами A, B, C … и т. д.

Основной количественной характеристикой случайного события является его вероятность. Пусть $A$ — какое-то случайное событие. *Вероятность случайного события — это математическая величина, которая определяет возможность его появления.* Она обозначается $P(A)$. Рассмотрим два основных метода определения этой величины.

*Классическое* определение вероятности случайного события обычно базируется на результатах анализа умозрительных опытов (испытаний), суть которых определяется условием поставленной задачи. При этом вероятность случайного события $P(A)$ равна:

$$P(A) = \frac{m}{n}, \qquad (1)$$

где $m$ — число случаев, благоприятствующих появлению события $A$; $n$ — общее число равновозможных случаев.

***Пример.*** Лабораторная крыса помещена в лабиринт и должна выбрать один из пяти возможных путей, так как лишь один из них ведет к поощрению в виде пищи. В предположении равновозможности выбора пути определите вероятность выбора пути, ведущего к пище.

***Решение.*** По условию задачи, из пяти равновозможных случаев ($n = 5$) событию $A$ — «крыса находит пищу» — благоприятствует один из них, т. е. $m = 1$. Тогда

$$P(A) = P\text{ (крыса находит пищу)} = \frac{m}{n} = \frac{1}{5} = 0{,}2 = 20\ \%.$$

Перечислим свойства вероятности, следующие из ее классического определения:

1. Вероятность случайного события — величина безразмерная.

2. Вероятность случайного события всегда положительна и меньше единицы, т. е. $0 < P(A) < 1$.

3. Вероятность достоверного события, т. е. события которое в результате опыта обязательно произойдет ($m = n$), равна единице.

4. Вероятность невозможного события ($m = 0$) равна нулю.

5. Вероятность любого события — величина не отрицательная и не превышающая единицу: $0 \le P(A) \le 1$.

*Статистическое* определение вероятности случайного события применяется тогда, когда невозможно использовать классическое определение (1). Это часто имеет место в биологии и медицине. В таком случае вероятность $P(A)$ определяют путем обобщения результатов реально проведенных серий испытаний (опытов).

Введем понятие *относительной частоты появления случайного события*. Пусть была проведена серия испытаний, состоящая из $N$ опытов (число $N$ может быть выбрано заранее); интересующее нас событие $A$ произошло в $M$ из них ($M < N$).

Отношение числа опытов $M$, в которых это событие произошло, к общему числу проведенных опытов $N$ называют *относительной частотой появления случайного события A в данной серии опытов* — $P^*(A)$:

$$P^*(A) = \frac{M}{N} \qquad (2)$$

Именно эту величину используют для приближенной оценки статистической вероятности:

$$P(A) \approx P^*(A) = \frac{M}{N}. \qquad (3)$$

Чем больше $N$, тем точнее оценка, тем ближе значения $P^*(A)$ к $P(A)$. Точное значение статистической вероятности события определяется пределом этого отношения при $N \to \infty$:

$$P(A) = \lim_{N\to\infty}\left(\frac{M}{N}\right) \qquad (3а)$$

Например, в опытах по бросанию монеты относительная частота появления герба при 12 000 бросаний оказалась равной 0,5016, а в серии из 24 000 бросаний — 0,5005. В соответствии с формулой (3)

$$P(\text{появление герба}) = \frac{1}{2} = 0{,}5 = 50\ \%.$$

***Пример.*** При врачебном обследовании 500 человек у 5 нашли опухоль в легких (о. л.). Определите относительную частоту и вероятность этого заболевания.

***Решение.*** По условию задачи $M = 5$, $N = 500$, относительная частота $P^*(\text{о. л.}) = M/N = 5/500 = 0{,}01$. В этой задаче $N$ велико и можно с достаточной точностью считать, что $P(\text{о. л.}) = P^*(\text{о. л.}) = 0{,}01 = 1\ \%$.

Перечисленные ранее свойства вероятности случайного события сохраняются и при статистическом определении данной величины.

## 1.3. Случайные величины. Виды случайных величин

*Величина, которая принимает различные числовые значения под влиянием случайных обстоятельств, называется случайной величиной.* Примеры случайных величин: число больных на приеме у врача, точные размеры внутренних органов людей и т. д.

Различают дискретные и непрерывные случайные величины**.**

*Случайная величина называется дискретной, если она принимает только определенные, отделенные друг от друга значения, которые можно установить и перечислить.*

***Примеры*:**

1) число студентов в аудитории может быть только целым положительным числом: 0, 1, 2, 3, 4… 20…

2) число событий, происходящих за одинаковые промежутки времени: частота пульса, число вызовов скорой помощи за час, количество операций в месяц с летальным исходом и т. д.

*Случайная величина называется непрерывной, если она может принимать любые значения внутри некоторого интервала, который иногда имеет резко выраженные границы, а иногда и нет*[1]. К непрерывным случайным величинам относятся, например, масса тела и рост взрослых людей, масса и объем мозга, количественное содержание ферментов у здоровых людей, размеры форменных элементов крови, pH крови и т. п.

Если случайная величина зависит от времени, то можно говорить о случайном процессе.

Понятие случайной величины играет определяющую роль в современной теории вероятностей, разработавшей специальные приемы перехода от случайных событий к случайным величинам.

## 1.4. Закон распределения дискретной случайной величины

Чтобы дать полную характеристику дискретной случайной величины необходимо указать все ее значения и их вероятности.

*Соответствие между возможными значениями дискретной случайной величины и их вероятностями называется законом распределения этой величины.* Обозначим возможные значения случайной величины $X$

[1] В этом случае считают, что значения некоторой случайной величины $X$ могут лежать в интервале $(-\infty; \infty)$, т. е. на всей числовой оси.

через $x_i$, а соответствующие им вероятности через $p_i$[1]. Тогда закон распределения дискретной случайной величины можно задать тремя способами:

1. В виде таблицы, которая называется рядом распределения:

| $X$ | $x_1$ | $x_2$ | … | $x_i$ | … | $x_n$ |
|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | … | $p_i$ | … | $p_n$ |

При этом сумма всех вероятностей $p_i$ равна 1 (условие нормировки):

$$p_1 + p_2 + \ldots p_n = \sum_{i=1}^{n} P(x_i) = 1. \qquad (4)$$

2. Графически — в виде ломаной линии (рис. 1), которую принято называть многоугольником распределения:

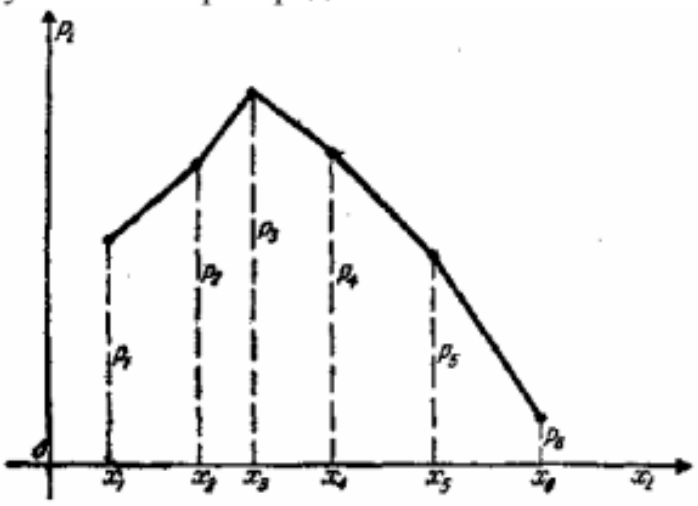


*Рис. 1*

3. Аналитически — в виде формулы. Например, если вероятность попадания в цель при одном выстреле равна $p$, то вероятность поражения цели 1 раз при $n$ выстрелах дается формулой $P(n) = n \cdot q^{n-1} \cdot p$, где $q = 1 - p$ — вероятность промаха при одном выстреле.

## 1.5. Закон распределения непрерывной случайной величины. Плотность распределения вероятностей

Для непрерывных случайных величин невозможно применить закон распределения в формах, приведенных выше, поскольку такая величина имеет бесчисленное («несчетное») множество возможных значений, сплошь заполняющих некоторый интервал. Поэтому составить таблицу, в которой были бы перечислены все ее возможные значения, или построить многоугольник распределения нельзя. Кроме того, вероятность какого-

[1] Обычно случайные величины обозначают большими буквами латинского алфавита, а их возможные значения и вероятности этих значений — малыми.

либо ее конкретного значения очень мала (близка к 0)[1]. Вместе с тем различные области (интервалы) возможных значений непрерывной случайной величины не равновероятны. Таким образом, и в данном случае действует некий закон распределения, хотя и не в прежнем смысле. Рассмотрим непрерывную случайную величину $X$, возможные значения которой сплошь заполняют некий интервал $(a, b)$[2]. Закон распределения вероятностей такой величины должен позволить найти вероятность попадания ее значения в любой заданный интервал $(x_1, x_2)$, лежащий внутри $(a, b)$ (рис. 2).

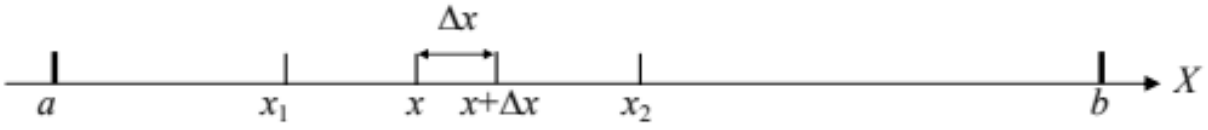



*Рис. 2*

Эту вероятность обозначают $P(x_1 < X < x_2)$, или $P(x_1 \le X \le x_2)$.

В теории вероятностей показано, что ее можно вычислить, введя величину, которая называется плотностью распределения вероятностей случайной величины $X$, или, короче, плотностью вероятности, плотностью распределения, обозначим ее $f(x)$. Для малого интервала $\Delta x$ значений $X$ (рис. 2) вероятность того, что случайная величина $X$ примет какое-то значение из этого интервала равна $\Delta P$, тогда

$$\Delta P = f(x)\cdot\Delta x, \text{ а } f(x) = \Delta P/\Delta x. \qquad (5)$$

Вероятность попадания значений величины $X$ в конечный интервал $(x_1, x_2)$ (рис. 2) определяется следующей формулой:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx. \qquad (6)$$

Графически вероятность $P(x_1 < X < x_2)$ равна площади криволинейной трапеции, ограниченной осью абсцисс, кривой $f(x)$ и прямыми $X = x_1$ и $X = x_2$ (рис. 3). Это следует из геометрического смысла определенного интеграла (6). Кривая $f(x)$ при этом называется *кривой распределения*.

Из (6) и рис. 3 следует, что если известна функция $f(x)$, то, изменяя пределы интегрирования, можно найти вероятность для любых интересующих нас интервалов. *Поэтому именно задание функции f(x) полностью определяет закон распределения для непрерывных случайных величин.*

---

[1] Приведем пример, поясняющий этот факт. Пусть случайная величина — уровень осадков, выпавших за год. Она может принимать любые значения из некоторого интервала. Однако вероятность того, что в заданный год этот уровень окажется точно равен 40 см, фактически равна 0.

[2] Иногда рассматривают интервал $(-\infty; +\infty)$.

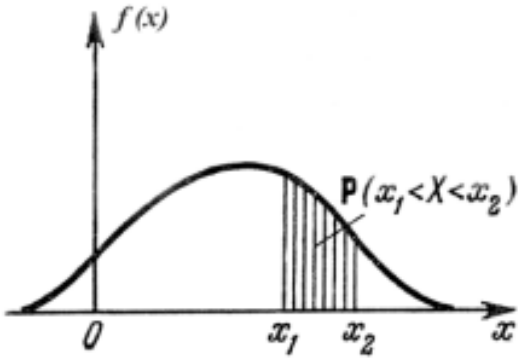


*Рис. 3*

Для плотности вероятности $f(x)$ должно выполняться условие нормировки в виде

$$\int_a^b f(x)dx = 1, \qquad (7)$$

если известно, что все значения $X$ лежат в интервале $(a, b)$, или в виде

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \qquad (8)$$

если границы интервала для значений $X$ точно не определены. Условия нормировки плотности вероятности (7) или (8) являются следствием того, что значения случайной величины $X$ достоверно лежат в пределах $(a, b)$ или $(-\infty, +\infty)$. Из (7) и (8) следует, что *площадь фигуры*, ограниченной кривой распределения и осью абсцисс, *всегда равна 1*.

## 1.6. Основные числовые характеристики случайных величин

Результаты, изложенные в подразд. 1.4 и 1.5, показывают, что полную характеристику дискретной и непрерывной случайных величин можно получить, зная законы их распределения. Однако во многих практически значимых ситуациях пользуются так называемыми числовыми характеристиками случайных величин. Главное назначение этих характеристик — выразить в сжатой форме наиболее существенные особенности распределения случайных величин. Важно, что данные параметры представляют собой конкретные (постоянные) значения, которые можно оценивать с помощью полученных в опытах данных. Этими оценками занимается «Описательная статистика».

В теории вероятностей и математической статистике используется достаточно много различных характеристик, но мы рассмотрим только наиболее употребляемые, не приводя формулы для их вычисления:

1. *Характеристики положения* — математическое ожидание, мода, медиана.

Они характеризуют положение случайной величины на числовой оси, т. е. указывают некоторое ориентировочное значение случайной величины, около которого группируются все другие ее возможные значения. Среди них важнейшую роль играет математическое ожидание $M(X)$.

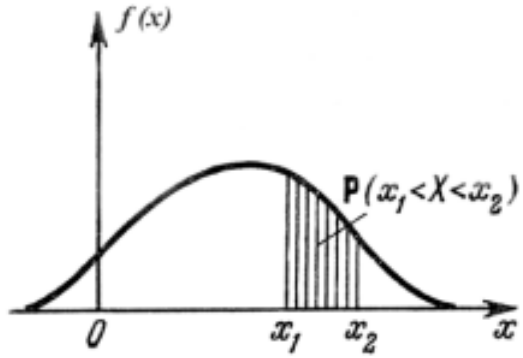


Рис. 3

Для плотности вероятности $f(x)$ должно выполняться условие нормировки в виде

$$\int_a^b f(x)dx = 1, \qquad (7)$$

если известно, что все значения $X$ лежат в интервале $(a, b)$, или в виде

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \qquad (8)$$

если границы интервала для значений $X$ точно не определены. Условия нормировки плотности вероятности (7) или (8) являются следствием того, что значения случайной величины $X$ достоверно лежат в пределах $(a, b)$ или $(-\infty, +\infty)$. Из (7) и (8) следует, что *площадь фигуры*, ограниченной кривой распределения и осью абсцисс, *всегда равна 1*.

## 1.6. Основные числовые характеристики случайных величин

Результаты, изложенные в подразд. 1.4 и 1.5, показывают, что полную характеристику дискретной и непрерывной случайных величин можно получить, зная законы их распределения. Однако во многих практически значимых ситуациях пользуются так называемыми числовыми характеристиками случайных величин. Главное назначение этих характеристик — выразить в сжатой форме наиболее существенные особенности распределения случайных величин. Важно, что данные параметры представляют собой конкретные (постоянные) значения, которые можно оценивать с помощью полученных в опытах данных. Этими оценками занимается «Описательная статистика».

В теории вероятностей и математической статистике используется достаточно много различных характеристик, но мы рассмотрим только наиболее употребляемые, не приводя формулы для их вычисления:

1. *Характеристики положения* — математическое ожидание, мода, медиана.

Они характеризуют положение случайной величины на числовой оси, т. е. указывают некоторое ориентировочное значение случайной величины, около которого группируются все другие ее возможные значения. Среди них важнейшую роль играет математическое ожидание $M(X)$.

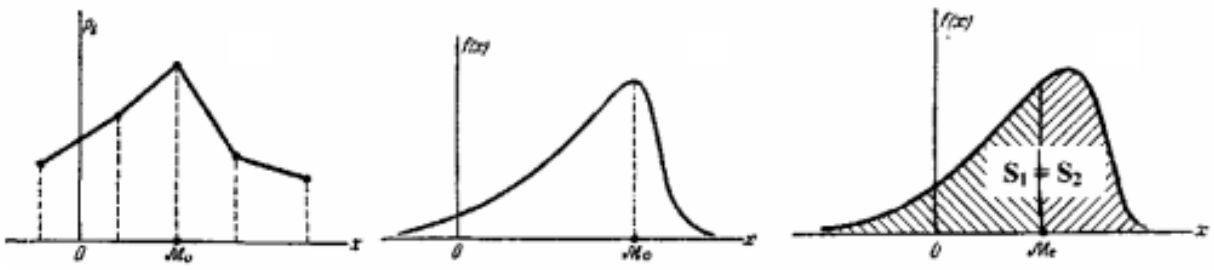
*Математическое ожидание* $M(X)$ случайной величины $X$ является вероятностным аналогом ее среднего арифметического $\overline{X}$: $M(X) = \overline{X}$ или $M(X) \approx \overline{X}$.

*Модой* $Mo(X)$ дискретной случайной величины называют ее наиболее вероятное значение (рис. 4, *а*), а непрерывной — значение $X$, при котором плотность вероятности максимальна (рис. 4, *б*).

*Медианой* ($Me$) случайной величины обычно пользуются только для непрерывных случайных величин, хотя формально ее можно определить и для дискретных $X$. Медианой $Me(X)$ случайной величины называют такое значение $X$, которое делит все распределение на две равновероятные части, т. е. вероятности $P(X < \text{Me})$ и $P(X > \text{Me})$ оказываются равными между собой:

$$P(X < Me) = P(X > Me) = \frac{1}{2}.$$

Графически медиана — это значение случайной величины, ордината которой делит площадь, ограниченную кривой распределения, пополам: $S_1 = S_2$ (рис. 4, *в*).




*а* *б* *в*

*Рис. 4*

Если $M(X)$, $Mo(X)$ и $Me(X)$ совпадают, то распределение случайной величины называют симметричным, в противном случае — асимметричным.

2. *Характеристики рассеяния* — это дисперсия и стандартное отклонение.

*Дисперсия* $D(X)$ случайной величины $X$ характеризует рассеяние, разбросанность значений случайной величины $X$ относительно ее математического ожидания. Само слово «дисперсия» означает «рассеяние».

Дисперсия $D(X)$ имеет размерность квадрата случайной величины. Это весьма неудобно при оценке разброса в физике, биологии, медицине. Поэтому обычно пользуются параметром, размерность которого совпадает с размерностью $X$. Это *стандартное отклонение* случайной величины $X$, которое обозначают $\sigma(X)$: $\sigma(X) = \sqrt{D(X)}$.

*3. Характеристики формы* — асимметрия и эксцесс.

*Асимметрия As* (коэффициент асимметрии) характеризует «скошенность» распределения. Если распределение симметрично относительно математического ожидания, коэффициент асимметрии равен нулю. На рис. 5 показаны два асимметричных распределения. Одно из них (кривая I) имеет положительную асимметрию ($As > 0$), другое (кривая II) — отрицательную ($As < 0$).

*Эксцесс Ex* (коэффициент эксцесса) используется для характеристики так называемой «крутости», т. е. островершинности или плосковершинности распределения. Для нормального распределения (см. подразд. 1.7) эксцесс равен 0. Кривые, более островершинные по сравнению с нормальной, обладают положительным эксцессом, кривые, более плосковершинные, — отрицательным эксцессом. На рис. 6 представлены: нормальное распределение (кривая I), распределение с положительным эксцессом (кривая II) и распределение с отрицательным эксцессом (кривая III).

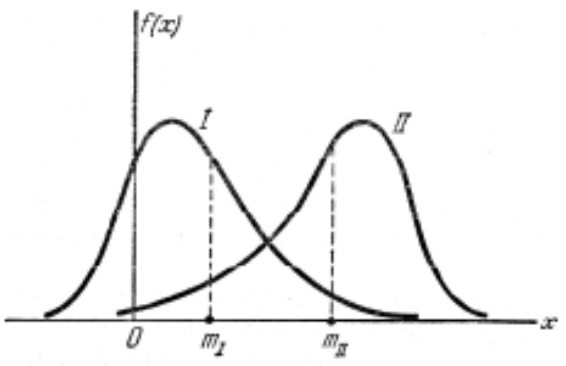



*Рис. 5*

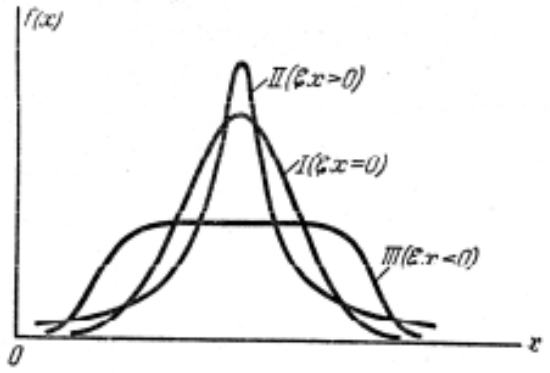



*Рис. 6*

Итак, математическое ожидание, мода, медиана, дисперсия, стандартное отклонение, асимметрия и эксцесс являются наиболее употребляемыми числовыми характеристиками случайных величин, каждая из которых выражает какое-нибудь характерное свойство их распределения.

## 1.7. Нормальный закон распределения случайных величин

Нормальный закон распределения (закон Гаусса) играет исключительно важную роль в теории вероятностей. Во-первых, это наиболее часто встречающийся на практике закон распределения непрерывных случайных величин. Во-вторых, он является предельным законом в том смысле, что к нему при определенных условиях приближаются другие законы распределения.

Нормальный закон распределения характеризуется следующей формулой для плотности вероятности:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{[x-M(x)]^2}{2\sigma^2}}, \qquad (9)$$

где $x$ — текущие значения случайной величины $X$; $M(X)$ и $\sigma$ — ее математическое ожидание и стандартное отклонение. Из (9) видно, что если случайная величина распределена по нормальному закону, то достаточно знать только два числовых параметра — $M(X)$ и $\sigma$ — чтобы полностью знать закон ее распределения.

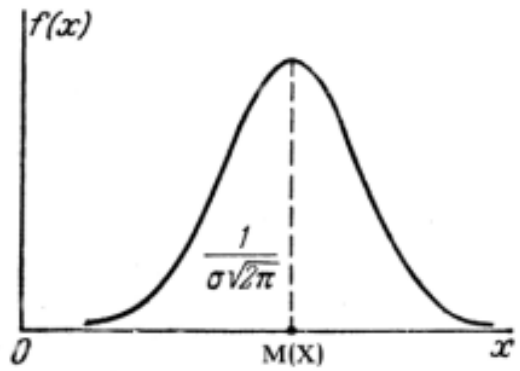



*Рис. 7*

График функции (9) называется нормальной кривой распределения (кривой Гаусса). Он имеет симметричный вид относительно прямой $x = M(X)$ (рис. 7).

Величина M(X) называется также центром рассеяния. Стандартное отклонение $\sigma$ характеризует ширину кривой распределения.

При изменении значения $M(X)$ в (9) нормальная кривая не меняется по форме, но сдвигается вдоль оси абсцисс. С возрастанием $\sigma$ максимальное значение $f(x)$ убывает, а сама кривая, становясь более пологой, растягивается вдоль оси абсцисс, при уменьшении $\sigma$ кривая вытягивается вверх, одновременно сжимаясь с боков. Вид кривой распределения при разных значениях $\sigma$: ($\sigma_3 < \sigma_2 < \sigma_1$) показан на рис. 8.

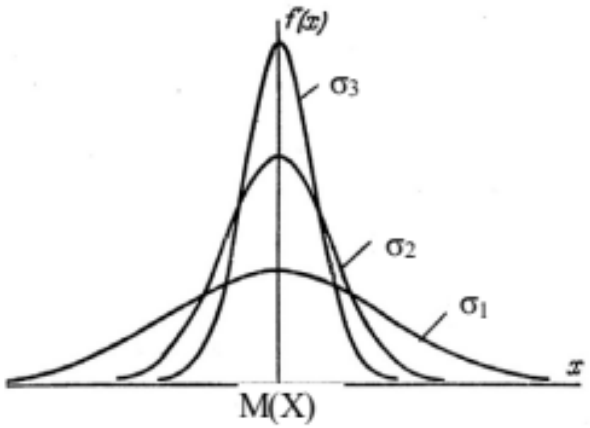



*Рис. 8*

Естественно, что при любых значениях $M(X)$ и $\sigma$ площадь, ограниченная нормальной кривой и осью $X$, остается равной 1 (условие нормировки):

$$\int_a^b f(x)dx = 1, \text{ или } \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Нормальное распределение симметрично, поэтому среднее, мода и медиана равны друг другу: $M(X) = Mo(X) = Me(X)$, асимметрия $As = 0$, эксцесс $Ex = 0$.

Вероятность попадания значений случайной величины $X$ в интервал $(x_1, x_2)$, т. е. $P(x_1 < X < x_2)$, равна:

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{[x-M(X)]^2}{2\sigma^2}} dx. \qquad (10)$$

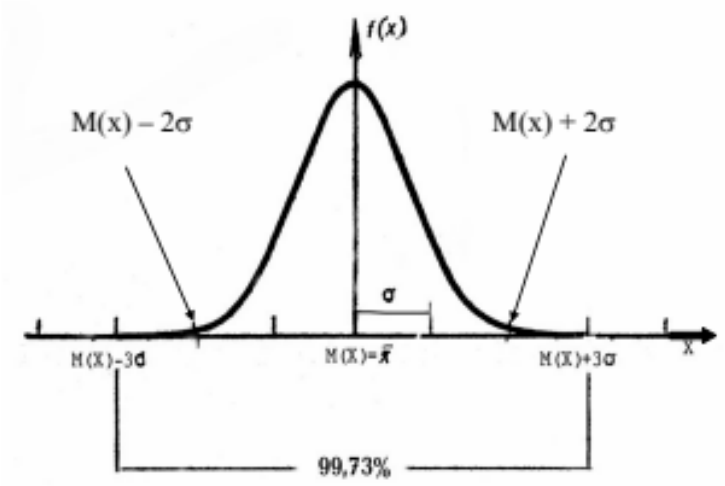



*Рис. 9*

На практике часто приходится вычислять вероятности попадания значений нормально распределенной случайной величины на участки, симметричные относительно $M(X)$. В частности, рассмотрим следующую, важную в прикладном отношении задачу. Отложим от $M(X)$ вправо и влево отрезки, равные σ, 2σ и 3σ (рис. 9), и проанализируем результат вычисления вероятности попадания значений $X$ в соответствующие интервалы:

$$P(M(X) - \sigma < X < M(X) + \sigma) = 0{,}6827 = 68{,}27\ \%. \qquad (11)$$

$$P(M(X) - 2\sigma < X < M(X) + 2\sigma) = 0{,}9545 = 95{,}45\ \%. \qquad (12)$$

$$P(M(X) - 3\sigma < X < M(X) + 3\sigma) = 0{,}9973 = 99{,}73\ \%. \qquad (13)$$

Из (13) следует: практически достоверно, что значения нормально распределенной случайной величины $X$ с параметрами $M(X)$ и σ лежат в интервале $M(X) \pm 3\sigma$. Иначе говоря, зная $M(X) = \overline{X}$ и σ, можно указать интервал, в который с вероятностью $P = 99{,}73\ \%$ попадают значения данной случайной величины. Такой способ оценки диапазона возможных значений $X$ известен как «правило трех сигм».

***Пример.*** Известно, что для здорового человека pH крови является нормально распределенной величиной со средним значением (математическим ожиданием) 7,4 и стандартным отклонением 0,2. Определите диапазон значений этого параметра.

***Решение.*** Для ответа на этот вопрос воспользуемся «правилом трех сигм». С вероятностью равной 99,73 % можно утверждать, что диапазон значений pH для здорового человека составляет 6,8–8.

# Глава 2
# Элементы математической статистики

## 2.1. Предмет и задачи математической статистики. Генеральная и выборочная совокупность

Предмет математической статистики — это разработка методов получения, описания и анализа статистических данных, определенных в результате исследования массовых случайных явлений. Статистические данные часто можно рассматривать как совокупность экспериментальных результатов, которые представляют собой набор возможных значений случайных величин (роста, массы тела, длительности пребывания больного на койке, содержания сахара в крови и т. д.).

*Фундаментальными понятиями* математической статистики являются генеральная и выборочная совокупности (выборка). Существуют разные подходы к пониманию смысла этих величин. Мы определяем их так. *Генеральная совокупность — это множество подлежащих статистическому изучению однородных объектов, которые характеризуются определенными качественными или количественными признаками.* Например, конечная и реально существующая генеральная совокупность — конкретно выбранная популяция: все жители Беларуси в фиксированный момент времени или только все мужчины, или женщины, или дети. Следующий пример: бесконечная и реально существующая генеральная совокупность — множество чисел, лежащих между 0 и 1.

Чтобы изучить генеральную совокупность по какому-либо из ее количественных признаков $X$ (острота зрения, показатели анализа крови и т. д.), нужно определить закон распределения данного признака и основные характеристики этого распределения, например, математическое ожидание и дисперсию. Для этого следовало бы изучить все ее объекты и затем обработать полученный массив данных методами теории вероятностей. Однако на практике провести сплошное обследование объектов генеральной совокупности часто физически невозможно и экономически невыгодно. Поэтому обычно исследуется только часть объектов, так называемая выборка.

*Совокупность «n» объектов, отобранных из интересующей нас генеральной совокупности для конкретного статистического исследования, называется выборочной совокупностью, или выборкой.*

Исследование выборки дает некоторое приближенное, оценочное значение интересующего нас параметра, принимающего различные значения для разных выборок. Поэтому главная цель выборочного метода, основного в математической статистике, — по вычисленной характери-

стике выборки как можно точнее определить соответствующую характеристику генеральной совокупности. Это возможно лишь в том случае, когда отобранная для работы часть объектов репрезентативна целому, т. е. типична, обладает теми же основными чертами, что и все целое. Иначе говоря, выборка должна быть представительной, т. е. по возможности полнее «представлять» свою генеральную совокупность. Это одно из важнейших требований, предъявляемых к выборке, несоблюдение которого ведет к грубым ошибкам и обесценивает результаты исследования. Например, если при изучении заболеваемости населения республики (генеральная совокупность) ишемической болезнью сердца в качестве выборки будет взята группа студентов, то результаты окажутся ошибочными, поскольку свойства выборки не будут соответствовать свойствам генеральной совокупности, то же будет, когда в качестве выборки взяты только пациенты кардиологического диспансера. Репрезентативность выборки обеспечивается ее достаточным объемом и определенными правилами ее формирования, которые в данном издании не рассматриваются.

Из многочисленных **задач**, решаемых математической статистикой, выделим следующие:

1. Определение статистических характеристик выборки (методы описательной статистики).
2. Определение параметров генеральной совокупности по данным выборки: точечные оценки и доверительные интервалы для параметров распределения.
3. Проверка статистических гипотез.
4. Исследование статистической связи между двумя признаками выборочной совокупности (элементы корреляционного анализа).

В данной главе излагаются общие подходы к решению этих задач. Конкретные примеры разобраны, главным образом, в гл. 3.

## 2.2. Статистическое распределение выборки

Обычно необходимо знать распределение признака $X$ в генеральной совокупности, но реально исследуется лишь некоторая выборка из нее.

В серии экспериментов, проводимых с выборкой, величина $X$ принимает определенные значения. Значения, записанные для всех элементов выборки в том порядке, в котором они были получены в опытах, представляют собой *простой статистический ряд:* $x_1, x_2, x_3 \ldots x_n$. Каждое значение $X$ в полученном числовом ряду называют *вариантой*. Полученные данные и подлежат статистической обработке, статистическому анализу.

Первый шаг при обработке этого материала — наведение в нем определенного порядка, ведущего к получению статистического распределе-

ния выборки. Здесь возможны два основных способа: создание вариационного или интервального ряда.

Рассмотрим *вариационный ряд*. Пусть некоторая выборка исследуется по количественному признаку $X$, который представляет собой дискретную случайную величину. В имеющемся у нас простом статистическом ряду варианта $x_1$ встречается (повторяется) $m_1$ раз, $x_2$ — $m_2$ раза, … $x_к$ — $m_к$ раз, при этом $\sum_{i=1}^{k} m_i = n$ , т. е. равна объему выборки. Далее по данным простого статистического ряда строится статистическое распределение (в медицинской литературе — *вариационный ряд*), которое удобно представить в виде таблицы, включающей в себя:

1) различные по значению варианты $x_i$, расположенные в определенной, ранжированной[1], заранее выбранной последовательности (обычно в порядке возрастания);

2) $m_i$ — частоты вариант, т. е. числа наблюдений (повторений) варианты $x_i$ в простом статистическом ряду;

3) $p_i = m_i / n$ — относительные частоты вариант, т. е. отношения частот $m_i$ к объему выборки $n$; они являются выборочными (эмпирическими) оценками вероятностей появления значений $x_i$ (см. 1.2).

Итак, для дискретной величины $X$ вариационный ряд — статистическое распределение выборки — имеет следующий вид (табл. 1).

*Таблица 1*

| Варианта $x_i$ ($x_1 < x_2 < x_3 \ldots < x_k$) | $x_1$ | $x_2$ | $x_3$ | … | $x_k$ | Контроль |
|---|---|---|---|---|---|---|
| Частота $m_i$ | $m_1$ | $m_2$ | $m_3$ | … | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |
| Относительная частота $p_i^* = \frac{m_i}{n}$ | $\frac{m_1}{n}$ | $\frac{m_2}{n}$ | $\frac{m_3}{n}$ | … | $\frac{m_k}{n}$ | $\sum_{i=1}^{k} \frac{m_i}{n} = 1$ |

Напомним, что под распределением дискретной случайной величины в теории вероятностей понимается соответствие между возможными значениями случайной величины и их вероятностями; в математической статистике — соответствие между наблюдаемыми вариантами $x_i$ и их частотами или относительными частотами.

*Интервальный ряд* удобен тогда, когда количественный признак $X$, характеризующий выборку, непрерывен, т. е. может принимать любые значения в некотором интервале. В этом случае статистическое распределение выборки (интервальный ряд) строится следующим образом. Область изменения признака ($x_{макс} - x_{мин}$) разбивают на несколько интер-

[1] В математической статистике ранжированным рядом часто называется последовательность всех полученных в эксперименте вариант, записанных в порядке возрастания.

валов обычно равной ширины. Число интервалов $k$, как правило, не менее 5 и не более 25 и приближенно определяется следующими эмпирическими формулами:

$$k \approx \sqrt{n}\ , \text{ или } k \approx 1 + 3{,}32\ \lg n,\ k \approx 5\lg n \qquad (14)$$

где $n$ — объем выборки.

Если ширина интервалов одинакова, то она равна:

$$\Delta x = h = \frac{x_{\text{макс}} - x_{\text{мин}}}{k} \qquad (15)$$

Затем вычисляют границы интервалов: $x_{\text{мин}} = x_0$, $x_1 = x_0 + h$, $x_2 = x_1 + h$, $x_3 = x_2 + h$,…, $x_{\text{макс}} = x_k$. Поскольку некоторые варианты могут являться границей двух соседних интервалов, то, во избежание недоразумений, придерживаются следующего правила: к интервалу $(a, b)$ относят варианты, удовлетворяющие неравенству $a \le x < b$.

Затем для каждого интервала подсчитывают частоты $m_i$ и (или) относительные частоты $p_i = m_i/n$ попадания вариант в данный интервал (эмпирические оценки вероятности попадания значений $X$ в выбранный интервал). Нередко используют также плотность относительной частоты:

$$\frac{m_i}{n\,\Delta x} = \frac{m_i}{n\,h}.$$

Данную величину можно считать выборочной (эмпирической) оценкой плотности вероятности.

Рассмотренное выборочное распределение непрерывной случайной величины $X$ — интервальный ряд — обычно представляется в виде таблицы, имеющей следующий вид (табл. 2).

*Таблица 2*

| Интервал | $x_0 - x_1$ | $x_1 - x_2$ | … | $x_{k-1} - x_k$ | Контроль |
|---|---|---|---|---|---|
| Частота $m_i$ | $m_1$ | $m_2$ | … | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |
| Относительная частота $p^*_i = m_i/n$ | $m_1/n$ | $m_2/n$ | … | $m_k/n$ | $\sum_{i=1}^{k} \frac{m_i}{n} = 1$ |
| Плотность относительной частоты $p^*_i/\Delta x = m_i/n\,\Delta x$ | $m_1/n\,\Delta x$ | $m_2/n\,\Delta x$ | … | $m_k/n\,\Delta x$ | $\sum_{i=1}^{k} \frac{m_i}{n\cdot\Delta x} = \frac{1}{\Delta x}$ |

Обобщим изложенный выше материал:

1. Если выборка исследуется по количественному признаку $X$, который представляет собой дискретную случайную величину, то статистическим распределением выборки является вариационным статистический ряд — полученные разные значения признака, записанные в упорядоченном виде с указанием их частот и относительных частот.

валов обычно равной ширины. Число интервалов $k$, как правило, не менее 5 и не более 25 и приближенно определяется следующими эмпирическими формулами:

$$k \approx \sqrt{n}\ , \text{ или } k \approx 1 + 3{,}32\ \lg n,\ k \approx 5\lg n \qquad (14)$$

где $n$ — объем выборки.

Если ширина интервалов одинакова, то она равна:

$$\Delta x = h = \frac{x_{\text{макс}} - x_{\text{мин}}}{k} \qquad (15)$$

Затем вычисляют границы интервалов: $x_{\text{мин}} = x_0$, $x_1 = x_0 + h$, $x_2 = x_1 + h$, $x_3 = x_2 + h$,…, $x_{\text{макс}} = x_k$. Поскольку некоторые варианты могут являться границей двух соседних интервалов, то, во избежание недоразумений, придерживаются следующего правила: к интервалу $(a, b)$ относят варианты, удовлетворяющие неравенству $a \le x < b$.

Затем для каждого интервала подсчитывают частоты $m_i$ и (или) относительные частоты $p_i = m_i/n$ попадания вариант в данный интервал (эмпирические оценки вероятности попадания значений $X$ в выбранный интервал). Нередко используют также плотность относительной частоты:

$$\frac{m_i}{n\,\Delta x} = \frac{m_i}{n\,h}.$$

Данную величину можно считать выборочной (эмпирической) оценкой плотности вероятности.

Рассмотренное выборочное распределение непрерывной случайной величины $X$ — интервальный ряд — обычно представляется в виде таблицы, имеющей следующий вид (табл. 2).

*Таблица 2*

| Интервал | $x_0 - x_1$ | $x_1 - x_2$ | … | $x_{k-1} - x_k$ | Контроль |
|---|---|---|---|---|---|
| Частота $m_i$ | $m_1$ | $m_2$ | … | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |
| Относительная частота $p^*_i = m_i/n$ | $m_1/n$ | $m_2/n$ | … | $m_k/n$ | $\sum_{i=1}^{k} \frac{m_i}{n} = 1$ |
| Плотность относительной частоты $p^*_i/\Delta x = m_i/n\,\Delta x$ | $m_1/n\,\Delta x$ | $m_2/n\,\Delta x$ | … | $m_k/n\,\Delta x$ | $\sum_{i=1}^{k} \frac{m_i}{n\cdot\Delta x} = \frac{1}{\Delta x}$ |

Обобщим изложенный выше материал:

1. Если выборка исследуется по количественному признаку $X$, который представляет собой дискретную случайную величину, то статистическим распределением выборки является вариационным статистический ряд — полученные разные значения признака, записанные в упорядоченном виде с указанием их частот и относительных частот.

2. Если выборка исследуется по количественному признаку $X$, который представляет собой непрерывную случайную величину, то статистическим распределением выборки является интервальный статистический ряд. Он включает в себя интервалы вариант, частоты попадания вариант в эти интервалы, относительные частоты, при необходимости — плотности относительных частот для этих интервалов.

## 2.3. Графическое представление статистических распределений выборок

Для получения наглядного представления о распределении выборок строят соответствующие графики, в частности, полигон частот или гистограмму распределения.

Вариационный ряд часто изображают графически в виде **полигона частот** или **полигона относительных частот**.

Для построения полигона частот на оси абсцисс откладывают варианты $x_i$, а на оси ординат — соответствующие им частоты $m_i$. Точки $(x_i; m_i)$ соединяют отрезками прямых. *Полигоном частот называют ломаную линию, отрезки которой соединяют точки* $(x_1, m_1)$; $(x_2, m_2)$... $(x_к, m_к)$.

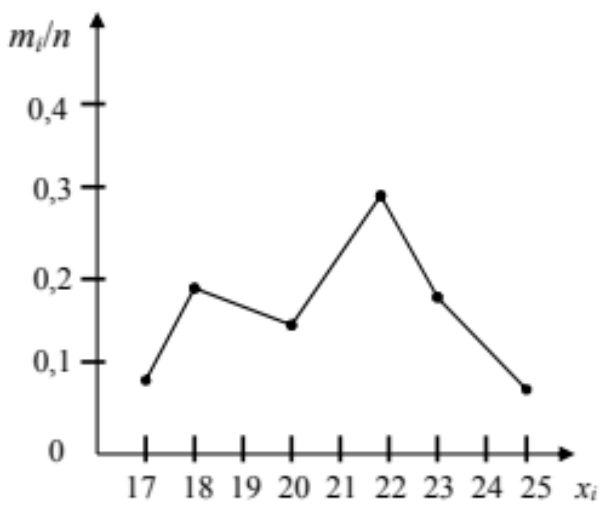



*Рис. 10*

*Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки* $(x_1, \frac{m_1}{n})$; $(x_2, \frac{m_2}{n})$; $(x_к, \frac{m_k}{n})$. На рис. 10 показан полигон относительных частот, построенный по данным выборки для некоторой случайной величины.

Для непрерывной случайной величины обычно строят **гистограммы**.

*Гистограммой называют диаграмму, состоящую из вертикальных прямоугольников, основаниями которых являются интервалы длиной* $\Delta x = h$, *а высоты равны* $m_i$ (частоте), $m_i/n$[1] (относительной частоте), $m_i/n$ %, *отношению* $\frac{m_i}{\Delta x}$ или $\frac{m_i}{\Delta x n}$ для соответствующих интервалов.

[1] Эти варианты позволяют сравнить гистограммы, построенные на одних и тех же интервалах, но для различных выборок из той же генеральной совокупности.

В случае, когда строят прямоугольники высотой $\frac{m_i}{\Delta x}$, площадь каждого из них равна количеству вариант в $i$-м интервале, т. е. площадь гистограммы равна сумме частот для всех интервалов, иначе говоря, равна объему выборки. Такую гистограмму называют *гистограммой частот*.

Если строят прямоугольники высотой $\frac{m_i}{\Delta xn}$, то площадь каждого $i$-го прямоугольника является оценкой вероятности попадания значений $x$ в выбранный интервал. В этом случае площадь гистограммы равна единице, а гистограмма называется *гистограммой относительных частот* (рис. 11, здесь анализируемый показатель — масса тела новорожденного).

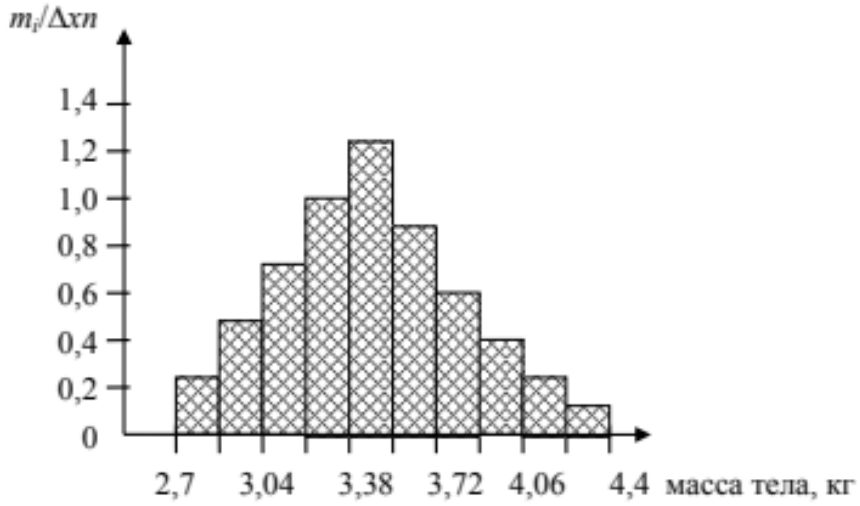



*Рис. 11*

Важно, что гистограммы можно использовать для оценки закона распределения признака в генеральной совокупности (в популяции). Соединяя средние точки верхних оснований прямоугольников гистограммы относительных частот плавной линией, можно по данным выборки получить примерный вид графика зависимости плотности вероятности $f$ от $x$.

## 2.4. Методы описательной статистики

Это методы описания выборок, исследуемых по количественному признаку $X$, с помощью их различных числовых характеристик.

Преимущество данных методов заключается в следующем. Несколько простых и достаточно информативных статистических показателей, если они известны, во-первых, избавляют нас от просмотра сотен, а порой и тысяч значений вариант, а во-вторых, позволяют получить более или менее точную оценку характеристик распределения признака в генеральной совокупности.

Описывающие выборку показатели разбиваются на несколько групп; в своем большинстве они имеют аналоги в виде числовых характеристик случайных величин в теории вероятностей.

**Показатели положения** описывают положение вариант выборки на числовой оси. Сюда относят:

а) минимальную и максимальную варианту;

б) выборочное среднее арифметическое значение (выборочное среднее), выборочные моду и медиану. Они определяют «центральную» точку распределения выборки — наиболее значимую для поставленной задачи варианту.

*Выборочным средним* называется величина

$$\bar{x}_{\text{в}} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (16)$$

где $x_i$ — $i$-я варианта, полученная в опыте с $i$-м элементом выборки; $n$ — объем выборки.

Выборочное среднее является той точкой, сумма отклонений значений $X$ от которой равна нулю. Это единственная точка, которая обладает данным свойством, оно выделяет ее среди всех других.

*Выборочная мода* $Mo_в$ — варианта, которая чаще всего встречается в исследуемой выборке, т. е. имеет наибольшую частоту.

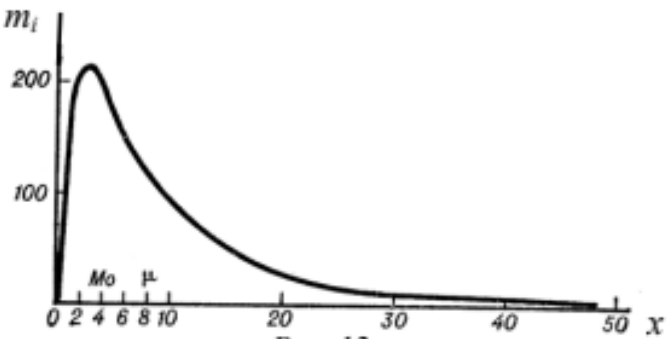


*Рис. 12*

***Пример 1.*** На рис. 12 приведено предполагаемое распределение по возрасту заболевших дифтерией (на 10 тыс. населения соответствующего возраста), которое явно не соответствует нормальному. Очевидно, что знание среднего возраста заболевших ($\bar{x}_{\text{в}} \approx 7{,}8$ года) в этом случае менее важно, чем знание возраста, в котором чаще всего возникает заболевание и который представляет собой моду ($Mo_{\text{в}} \approx 4$ года). Именно этот показатель указывает, где должны быть сосредоточены главные профилактические меры: в школах или дошкольных учреждениях.

Если выборочное распределение имеет несколько мод, то говорят, что оно мультимодально. Это служит индикатором того, что выборка не является однородной, и данные, возможно, порождены несколькими «наложенными» распределениями.

*Выборочная медиана* $Me_в$ — варианта, которая делит ранжированный статистический ряд (см. сноску на стр. 17) на две равные части по числу попадающих в них вариант.

***Пример 2.*** Дан статистический ряд: 1; 2; 3; 3; 4; 4; 5; 5; 6; 8; 9; $n = 11$. Варианта, разделяющая этот ряд на две равные по количеству вариант части, занимает в ряду 6 место и равна 4, т. е. $\text{Me}_\text{в} = 4$.

**Показатели разброса** описывают степень разброса данных относительно своего центра. Здесь обычно используются:

а) *стандартное отклонение* $S$ и *выборочная дисперсия* $D_в = S^2$ [1], характеризующие рассеяние вариант вокруг их среднего выборочного значения $\bar{x}_\text{в}$:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_\text{в})^2}{n-1}}. \qquad (17)$$

Именно эти параметры часто используются как меры изменчивости некоторого исследуемого показателя (случайной величины $X$). Чем больше $D_\text{в}$ и $S$, тем сильнее разбросаны значения $X$ относительно среднего;

б) *размах выборки* — разность между максимальной и минимальной вариантами: $x_\text{макс} - x_\text{мин}$;

в) *коэффициент вариации*:

$$\nu = \frac{S}{\bar{x}_\text{в}} \cdot 100\ \%. \qquad (18)$$

Применяется для сравнения величин рассеяния двух вариационных рядов: тот из них имеет большее рассеяние, у которого коэффициент вариации больше.

Отметим так же выборочный эксцесс ($Ex_\text{в}$) и выборочную асимметрию $As_\text{в}$ (их смысл — см. подразд. 1.6).

Для определения **закона распределения** исследуемой величины в генеральной совокупности, который нужно знать для проведения последующего анализа, можно использовать *гистограммы* и *полигон частот*. О них шла речь в предыдущем подразделе. О законе распределения также можно судить по выборочным числовым характеристикам случайной величины.

Заметим, что большинство методов статистического анализа данных разработано для случайных величин, распределенных по нормальному закону. Распределение исследуемой величины в генеральной совокупности можно рассматривать как близкое к нормальному, если:

1. Выборочные $\bar{x}_\text{в}$, $Me_\text{в}$, $Mo_\text{в}$ равны или незначительно отличаются друг от друга.

2. Минимальное и максимальное значения $X$ ($x_\text{макс}$ и $x_\text{мин}$) примерно равноудалены от $\bar{x}_\text{в}$.

[1] Точнее $S^2$ называется «исправленная выборочная дисперсия».

3. Выборочные $Ex_в$ и $As_в$ близки к нулю.

Можно так же проверить выполнение следующих условий:

1. Примерно 99,7 % отклонений значений исследуемого показателя $X$ от среднего по модулю меньше $3S$.

2. Примерно 95,5 % отклонений значений исследуемого показателя $X$ от среднего по модулю меньше $2S$.

3. Примерно 68,3 % отклонений значений исследуемого показателя $X$ от среднего по модулю меньше $S$.

Такой подход является следствием расчета вероятностей попадания значений нормально распределенного исследуемого показателя в интервалы: $M(x) \pm \sigma$, $M(x) \pm 2\sigma$, $M(x) \pm 3\sigma$ (см. подразд. 1.7).

## 2.5. Оценка параметров генеральной совокупности по ее выборке. Точечная и интервальная оценки

Напомним, что главная цель любого статистического исследования — установить закон распределения и получить значения характеристик изучаемого признака генеральной совокупности путем анализа выборки. Иначе говоря, надо определить генеральную среднюю $\overline{x}_г = M(X)$, генеральные дисперсию $D_г(X)$, стандартное отклонение $\sigma_г$, генеральную моду $Mo_г$, медиану $Me_г$ и другие характеристики генеральной совокупности путем статистического исследования выборки.

**Точечная оценка характеристик генеральной совокупности** — наиболее простой, но не очень достоверный способ. При данном способе в качестве оценок характеристик генеральной совокупности используются соответствующие числовые характеристики выборки. Например, в качестве генерального среднего используется выборочное среднее, в качестве генеральной дисперсии — выборочная дисперсия и т. д. Такие оценки и называются точечными. Их недостаток состоит в том, что не ясно, насколько сильно они отличаются от истинных значений параметров генеральной совокупности. Ошибка может быть особенно большой в случае малых выборок.

**Интервальная оценка параметров генеральной совокупности** более достоверна. В этом случае определяется интервал, в который с заданной вероятностью попадает истинное значение исследуемого признака. Такой интервал называется *доверительным интервалом*, а вероятность того, что истинное значение оцениваемой величины находится внутри этого интервала — *доверительной вероятностью*, или *надежностью*. *В медицинской литературе для этой величины используется термин «вероятность безошибочного прогноза»*. Обозначим ее $\gamma$. Значения $\gamma$ задаются заранее (обычно в медико-биологических исследованиях выбирают

значения $\gamma = 0{,}95 = 95\ \%$ или $\gamma = 0{,}99 = 99\ \%$), после чего находят соответствующий доверительный интервал. *Иногда вместо доверительной вероятности используется величина* $\alpha = 1 - \gamma$, *которая называется уровнем значимости.*

Для построения надежных интервальных оценок необходимо знать закон, по которому оцениваемый случайный признак распределен в генеральной совокупности.

Рассмотрим, вначале для малых выборок ($n < 30$), как строится интервальная оценка генеральной средней $\bar{x}_г = M(X)$ признака, который в генеральной совокупности распределен по нормальному закону. В этом случае интервальной оценкой (с доверительной вероятностью $\gamma$) генеральной средней (математического ожидания, $\bar{x}_г = M(X)$) количественного признака $X$ по выборочной средней $\bar{x}_в$ при неизвестном $\sigma_г$ является доверительный интервал

$$\bar{x}_в - \delta < M(X) < \bar{x}_в + \delta, \qquad (19)$$

или, в другой форме записи:

$$M(X) = \bar{x}_в \pm \delta, \qquad (20)$$

где $\delta = t_{\gamma,n}(S/\sqrt{n})$ — полуширина доверительного интервала — предельная ошибка выборки, характеризующая точность оценки; $n$ — объем выборки; $S$ — выборочное стандартное отклонение; $S/\sqrt{n} = S_{\bar{x}_в}$ — стандартная ошибка выборочного среднего (в медицинской и биологической литературе эта величина иногда обозначается буквой $m$ и называется ошибкой репрезентативности), $t_{\gamma,n}$ — коэффициент Стьюдента (его значения определяются либо по соответствующим таблицам, либо содержатся в программных статистических пакетах обработки данных).

Анализ формулы (19) показывает, что:

а) чем больше доверительная вероятность $\gamma$, тем больше коэффициент $t_{\gamma,n}$ и шире доверительный интервал;

б) чем больше объем выборки $n$, тем у́же доверительный интервал.

При большой выборке ($n > 30$) полуширину доверительного интервала $\delta$ определяют по соотношениям:

$$\delta \approx 1{,}96S/\sqrt{n} \text{ при } \gamma = 95\ \% \text{ или } \delta \approx 2{,}6S/\sqrt{n} \text{ при } \gamma = 99\ \%. \qquad (21)$$

Рассматривая интервальную оценку $M(X)$, обратим внимание на следующие обстоятельства. Иногда экспериментаторы приводят результат в виде:

$$M(X) = \bar{x}_в \pm S/\sqrt{n} = \bar{x}_в \pm m.$$

По существу, такая запись содержит указание доверительного интервала при $t_{\gamma,n} = 1$. Рассмотрим, на примере, к чему это может привести.

***Пример.*** Произведено 31 измерение какой-то величины $n = 31$. Необходимо определить доверительный интервал с доверительной вероятностью 90 % для математического ожидания $M(X)$ этой величины.

Обработка полученных данных дает: $\bar{x}_в = 58{,}29$ ед., $S = 5{,}58$ ед., $t_{0{,}9;31} = 1{,}7$. Тогда $M(X) = \bar{x}_в \pm t_{\gamma,n} \cdot (S/\sqrt{n}) = (58{,}29 \pm 1{,}7)$ ед. Если указать результат в виде $M(X) = \bar{x}_в \pm S/\sqrt{n}$ ($t_{\gamma,n} = 1$), то в нашем примере $M(X) = 58{,}29 \pm 1$. По таблицам значений коэффициента Стьюдента легко определить, что при $t_{\gamma,31} = 1$ доверительная вероятность $\gamma$ равна лишь 68 %. Это резко снижает доверие к полученному результату (с 90 % при правильном расчете до 68 %) несмотря на сужение доверительного интервала.

Из формул (19), (21) понятно, как при заданной доверительной вероятности и объему выборки получить оценку $\bar{x}_г = M(X)$.

Поставим обратную, практически значимую задачу. По заданной точности оценки $\delta$, т. е. по заданной полуширине доверительного интервала, определим необходимый объем выборки, обеспечивающий нужное $\delta$. Эта задача решается особенно просто в случае больших выборок ($n > 30$). Здесь, например, при доверительной вероятности 95 % $\delta = 1{,}96S/\sqrt{n}$. Тогда из (21) следует, что необходимый объем выборки равен: $n \geq (1{,}96)^2 S^2/\delta^2$.

Определим доверительный интервал для $D$ и $\sigma$, предполагая, что случайная величина в генеральной совокупности опять же распределена по нормальному закону, а ее математическое ожидание неизвестно. Тогда при заданной доверительной вероятности $\gamma$ или уровне значимости $\alpha$, он определяется следующими соотношениями:

а) для дисперсии:

$$\frac{(n-1)S^2}{\chi^2_{n-1,(1-\gamma)/2}} < D_г < \frac{(n-1)S^2}{\chi^2_{n-1,(1+\gamma)/2}}; \qquad (22)$$

б) для стандартного отклонения:

$$S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,(1-\gamma)/2}}} < \sigma_г < S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,(1+\gamma)/2}}}. \qquad (23)$$

В этих формулах значения $\chi^2$ (хи-квадрат) определяются либо по соответствующим таблицам, либо с помощью встроенных функций программ обработки статистических данных.

Если вместо доверительной вероятности $\gamma$ использовать уровень значимости $\alpha$ и учесть, что $\alpha = 1 - \gamma$, то формулы для расчета доверительных интервалов принимают следующий вид:

а) для дисперсии: $\dfrac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} < D < \dfrac{(n-1)S^2}{\chi^2_{n-1,(1-\alpha/2)}}$

б) для стандартного отклонения: $S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,\alpha/2}}} < \sigma < S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,(1+\alpha/2)}}}$.

Подобные интервальные оценки с заданной надежностью даются и в тех случаях, когда рассматриваемый случайный признак распределен в генеральной совокупности не по нормальному, а по другим законам. Соответствующие формулы для вычисления будут другими.

## 2.6. Понятие о статистических гипотезах и критериях проверки гипотез

Во многих случаях требуется на основе экспериментальных данных решить, справедливо ли некоторое утверждение. Например, верно ли, что два набора данных (2 выборки) происходят из одного источника (из одной генеральной совокупности), или что *A* лучший стрелок, чем *B*, или что данное лекарство лучше другого при лечении определенного заболевания? При ответе на подобные вопросы, во-первых, хотелось бы принять наиболее обоснованное решение, во-вторых, оценить вероятность ошибочности этого решения.

Рассмотрение таких задач в строгой математической постановке приводит к понятию **статистической гипотезы**.

В обычном языке понятие «гипотеза» означает предположение. В математической статистике *гипотеза — это: а) предположение о виде неизвестного закона распределения исследуемой экспериментально случайной величины (непараметрическая гипотеза); б) предположение о значениях характеристик (параметров) известного распределения (параметрическая гипотеза).*

***Примеры статистических гипотез.*** Делаются предположения:

1. Данный признак в генеральной совокупности распределен по нормальному закону (непараметрическая гипотеза).

2. Дисперсии двух совокупностей, распределенных по нормальному закону, равны между собой (параметрическая гипотеза).

Эти предположения подлежат проверке.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей. Если выдвинутая гипотеза отвергнута, то имеет место противоречащая гипотеза.

*Выдвинутую гипотезу $H_0$ называют нулевой* (*основной*). *Конкурирующую* (*альтернативную) гипотезу, которая несовместна с нулевой, обозначают* $H_1$. Например, если $H_0$ состоит в предположении, что математическое ожидание $M(X)$ нормального распределения равно 2, то конкурирующая гипотеза, в частности, может состоять в предположении, что $M(X) \neq 2$. Коротко это записывают так:

$$H_0: M(X) = 2;$$
$$H_1: M(X) \neq 2.$$

Заключение о справедливости нулевой или альтернативной гипотезы всегда делается на основании анализа выборки определенного объема.

Если для исследуемого явления сформулирована та или иная гипотеза, то надо найти правило, которое позволяло бы по имеющимся статистическим данным (по выборке) принять решение о соответствии либо несоответствии выдвинутой гипотезы этим данными. Это правило называется *статистическим критерием (иногда просто критерием, статистикой) проверки гипотезы*.

Для проверки непараметрических гипотез существуют *критерии согласия*, которые должны подтвердить или опровергнуть правильность выбора закона распределения. В данном случае чаще других используются *критерий* $\chi^2$ (*хи-квадрат*) *Пирсона и критерий Колмогорова–Смирнова*. Первый из них более универсален, так как приемлем для случайных величин любого типа (дискретных и непрерывных), второй — только для непрерывных случайных величин. В данном издании эти критерии лишь упоминаются. При необходимости ими можно воспользоваться, работая со специальными статистическими пакетами (Biostat, Statistica и т. д.) или с соответствующей литературой, например [4, 5, 9].

Параметрические гипотезы проверяются с помощью *параметрических критериев значимости*, разработанных, прежде всего, для случайных величин, распределенных по нормальному закону.

Создание критериев потребовало разработки достаточно сложной теории. Рассмотрим лишь основные ее идеи и продемонстрируем на примерах работу соответствующих правил.

Обычно проверяется нулевая гипотеза ($H_0$). Тогда статистическим критерием проверки $H_0$ называют случайную величину (статистику) $K$, точное или приближенное распределение которой известно. В конкретных задачах $K$ имеет и свое конкретное обозначение. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве статистики $K$ используют отношение выборочных дисперсий (критерий $F$ Фишера–Снедекора, иногда просто критерий Фишера):

$$K = F = S_1^2/S_2^2 \ (S_1 > S_2).$$

Так как входящие в критерий величины рассчитывают по данным определенных выборок, то вычисленное значение $K$ называют *наблюдаемым значением критерия* $K_{\text{набл.}}$ Например, если по выборкам $S_1^2 = 20$, а $S_2^2 = 5$, то $K_{\text{набл.}} = F_{\text{набл.}} = 20/5 = 4$.

После выбора определенного критерия множество всех его возможных значений разделяют на две области:

1. *Критическая область — совокупность значений критерия К, при которых нулевую гипотезу ($H_0$) отвергают.*

2. *Область принятия гипотезы (область допустимых значений) — совокупность значений критерия К, при которых нет оснований в рамках данного критерия отвергнуть гипотезу $H_0$.*

Таким образом, основной принцип проверки гипотезы $H_0$ достаточно прост: если вычисленное по выборке значение критерия $K_{набл.}$ принадлежит критической области, то гипотезу $H_0$ отвергают; если оно принадлежит области принятия гипотезы, то $H_0$ можно принять.

Указанные области (интервалы) разделяются *критическими точками* (границами) $к_{кр.}$, при этом различают одностороннюю (право- или левостороннюю) и двустороннюю критические области.

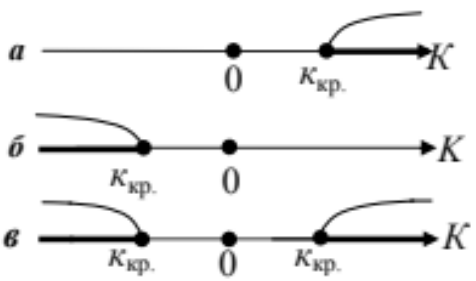



*Рис. 13*

*Правосторонней* называют критическую область, определяемую неравенством $K > к_{кр.}$, где $к_{кр.} > 0$ (рис. 13, *а*)

*Левосторонней* называют критическую область, определяемую неравенством $K < к_{кр.}$, где $к_{кр.} < 0$ (рис. 13, *б*).

*Двусторонней* называют критическую область, которая определяется неравенствами $K < к_1$, $K > к_2$, где $к_2 > к_1$. В частности, если критические точки симметричны относительно нуля, т. е. распределение критерия симметрично относительно нуля, то двусторонняя критическая область определяется неравенствами (в предположении, что $к_{кр.} > 0$): $K < -к_{кр.}$, $K > к_{кр.}$ или $|K| > к_{кр.}$ (рис 13, *в*).

Каким же образом можно отыскать критическую точку $к_{кр.}$ на оси значений *K*? Обычно задают вероятность отклонения гипотезы $H_0$, когда она верна. Эта вероятность определяется выбранным уровнем значимости α, уже введенным нами ранее (см. подразд. 2.5), и здесь она называется *уровнем значимости критерия*. Обычно α = 0,05, 0,01 или 0,001. Если, например, принят уровень значимости равный 0,05, то это означает, что в 5 случаях из 100 мы рискуем допустить ошибку — отвергнуть правильную гипотезу.

В статистике, принимая решение по результатам проверки гипотезы, можно допустить ошибки различного характера. *Ошибка первого рода* состоит в том, что с вероятностью α отклоняется правильная гипотеза $H_0$. *Ошибка второго рода* состоит в том, что будет принята гипотеза $H_0$, в то время как она не верна. Вероятность ошибки второго рода обозначают β. Величину 1 – β называют *мощностью критерия*. Фактически мощность критерия определяется вероятностью не допустить ошибку второго рода ($\beta \to 0$). Чем ближе мощность критерия к единице, тем более эффективен критерий. Многие статистические критерии получены путем нахождения

наиболее мощного критерия при заданных предположениях об основной и альтернативной гипотезах.

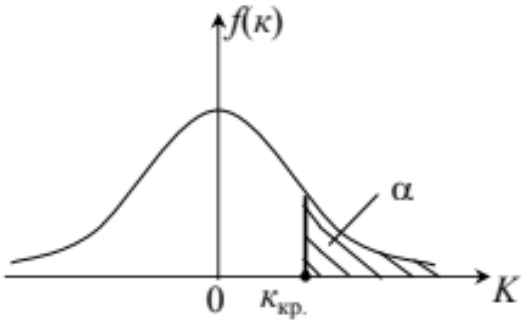


*Рис. 14*

Вернемся к вероятности α. Она численно равна площади под кривой распределения критерия $K$, которая соответствует критической области (например, заштрихованная область на рис. 14).

Для каждого критерия существуют соответствующие таблицы (см. например (4, 5, 9)), по которым при заданном уровне значимости α находят $к_{кр.}$, соответствующие расчетные функции имеются в программах обработки статистических данных, в том числе в табличном процессоре Excel.

*Итак, если вычисленное по выборке $K_{набл.}$ попадает в критическую область, нулевая гипотеза $H_0$ отвергается, если нет, то нет оснований ее отвергнуть.*

В современных статистических пакетах обычно сравниваются не только $K_{набл.}$ и $к_{кр.}$, но и заданный уровень значимости α и вероятность того, что, например, для правосторонней критической области $K > K_{набл.}$. Обозначим эту вероятность $P$, в нашем примере она равна площади под кривой распределения критерия, расположенной справа от $K_{набл.}$ (рис. 15, *а*, *б*).

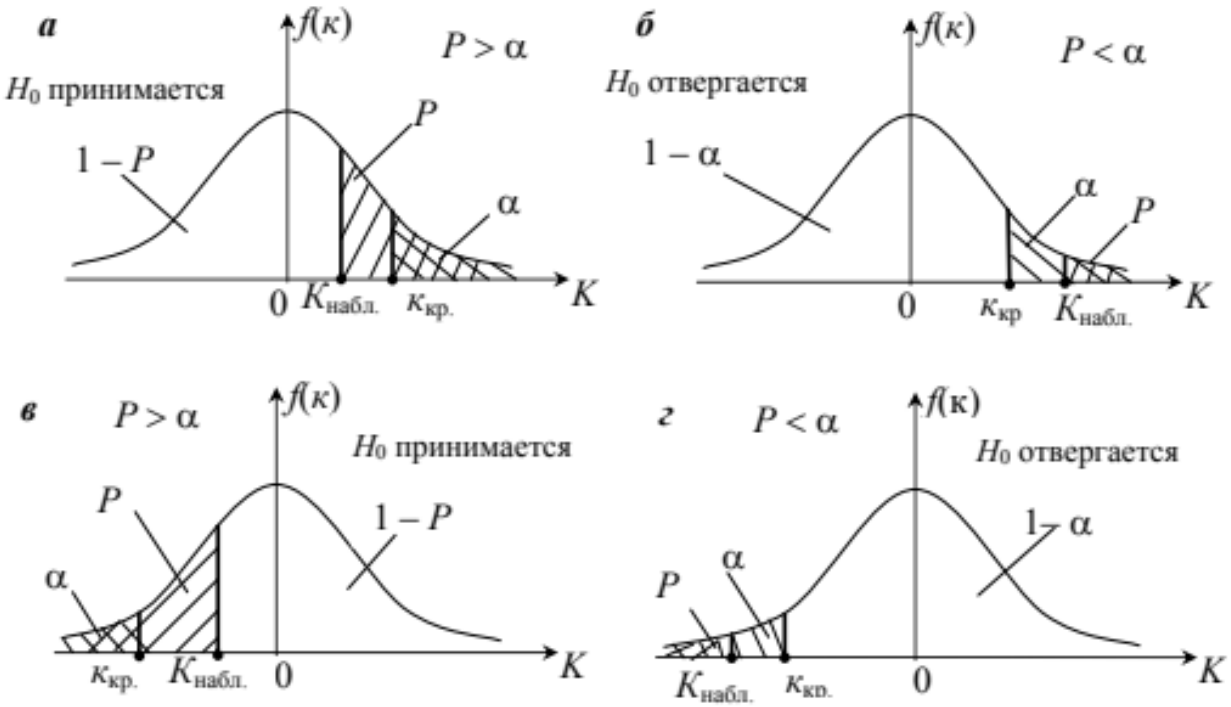


*Рис. 15*

Если вероятность $P$ оказывается больше заданного уровня значимости $\alpha$ ($P > \alpha$), то гипотеза $H_0$ принимается (рис 15, *а*), в противном случае — не принимается (рис 15, *б*, $P < \alpha$). Рис. 15, *в*, *г* иллюстрирует такой

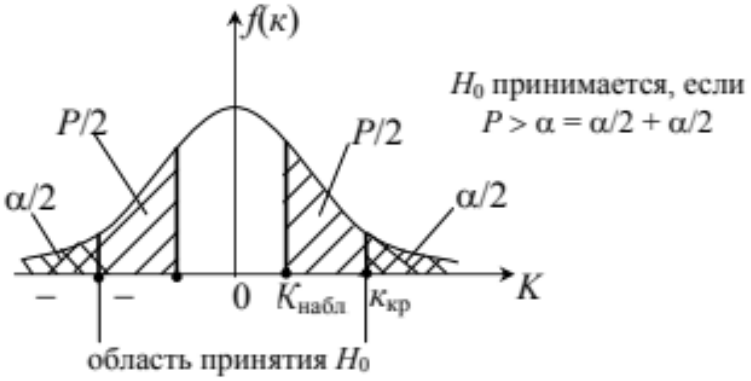



*Рис. 16*

подход для левосторонней критической области. Этот подход также верен для двусторонней симметричной критической области (рис. 16).

Отметим, что в данном учебно-методическом пособии, рассматривая примеры использования различных критериев, мы ограничиваемся проверкой гипотез, связанных с параметрами нормального распределения.

## 2.7. Примеры различных критериев и правила работы с ними

**I. *Работа с одной выборкой.***

*Проверка гипотезы о значении математического ожидания случайной величины, распределенной по нормальному закону (одновыборочный t-критерий Стьюдента).*

Постановка задачи: пусть получена выборка $x_1$, $x_2$, $x_3$… $x_n$ из нормальной генеральной совокупности случайной величины $X$. Математическое ожидание (среднее) генеральной совокупности неизвестно, но есть основание предполагать, что $M(X)$ равно некоторому конкретному числу $A$, например, какому-то стандартному значению.

Требуется проверить гипотезу $H_0$: $M(X) = A$, исходя из полученной выборки.

Для проверки $H_0$ необходимо учитывать дисперсию, т. е. степень разброса значений $X$ относительно среднего. В связи с этим рассматриваются 2 случая: дисперсия генеральной совокупности известна (большие выборки) и дисперсия генеральной совокупности неизвестна (малые выборки, $n < 30$). Выберем второй случай, как наиболее практически значимый.

Здесь в качестве критерия (статистики) используют одновыборочный критерий Стьюдента с $(n-1)$ степенями свободы[1]:

$$t = \frac{\bar{x}_в - A}{S}\sqrt{n}, \qquad (26)$$

где $\bar{x}_в$ и $S$ — выборочные среднее и стандартное отклонение, $n$ — объем выборки.

Отметим, что при решении соответствующих задач с использованием программы Excel формула (26) вводится с клавиатуры.

Подставляя в (26) заданные $A$ и $n$ и найденные по выборке числовые значения $\bar{x}_в$ и $S$, вычисляют наблюдаемое (расчетное) значение критерия $t_{набл.}$

При получении конкретного вывода о принятии либо нет нулевой гипотезы $H_0$ придерживаются следующих правил, которые определяются видом конкурирующей гипотезы.

1. Если конкурирующая (альтернативная) гипотеза $H_1$: $M(X) \neq A$ (двухсторонняя критическая область), то по заданному уровню значимости α и числу степеней свободы $m = n - 1$ находят критическую точку $t_{двухст.\ кр.}(\alpha, m)$. Если $|t_{набл.}| < t_{двухст.\ кр.}$ — нет оснований отвергать нулевую гипотезу, $H_0$ принимается с заданным уровнем значимости. Если $|t_{набл.}| > t_{двухст.\ кр.}$ — $H_0$ отвергают.

При известной вероятности $P$ нулевую гипотезу $(H_0)$ принимают при $P > \alpha$ и не принимают, если $P < \alpha$.

2. Если конкурирующая гипотеза $H_1$: $M(X) > A$ (правосторонняя критическая область), то по заданным α и $m$ находят $t_{прав.\ кр.}(\alpha, m)$.

При $t_{набл.} < t_{прав.\ кр.}$ $H_0$ принимают, в противном случае — отвергают и принимается конкурирующая гипотеза.

При известной вероятности $P$ нулевую гипотезу $(H_0)$ принимают при $P > \alpha$ и не принимают, если $P < \alpha$.

3. Если конкурирующая гипотеза $H_1$: $M(X) < A$ (левосторонняя критическая область), то поступают следующим образом. Вначале находят «вспомогательную» критическую точку $t_{правост.\ кр.}(\alpha, m)$ и полагают границу левосторонней критической области $t_{левост.кр.} = -t_{правост.\ кр.}$. Если $t_{набл.} > -t_{прав.\ кр.}$, нет оснований отвергать $H_0$, если $t_{набл.} < -t_{правост.\ кр.}$ $H_0$ отвергают. Так же как в предыдущих случаях $H_0$ принимают при $P > \alpha$.

Следует отметить, что данный критерий позволяет провести сравнение выборочной средней с предполагаемой генеральной средней нормальной совокупности. Его часто называют *критерием сравнения выборочной средней с гипотетической генеральной средней*. Такая возможность важна

[1] Степени свободы — специальные параметры (характеристики распределения), используемые при работе со статистическими гипотезами.

при решении многих прикладных задач, в том числе возникающих в медицинской промышленности.

Если $H_0$ принимается, т. е. $M(X) = A$, то выборочная средняя $\bar{x}_в$ также незначимо отличается от гипотетической генеральной средней $A$, а если справедлива гипотеза $H_1$, то различие этих величин значимо.

***Пример.*** Проектный размер изделий, изготавливаемых станком-автоматом, $A = 35$ мм. Предприятием получен станок, требует ли он корректировки в конкретных условиях работы, если измерения 20 случайно отобранных изделий дали следующие результаты: $\bar{x}_в = 35{,}07$ мм, $S = 0{,}16$ мм, $t_{набл.} = 1{,}96$?

Проверим при $\alpha = 0{,}05$ основную гипотезу $H_0$: $M(X) = 35$ при $H_1$: $M(X) \neq 35$.

Найденное $t_{двухст.\ кр.}(0{,}05;\ 19) = 2{,}09$ и так как $t_{набл.} < t_{двухст.\ кр.}$, то нет оснований при уровне значимости 0,05 отвергать $H_0$: $M(X) = 35$ мм, так что отличие $\bar{x}_в$ от $A = 35$ мм тоже незначимо. Станок обеспечивает проектный размер изделий и не требует корректировки своей работы.

В заключение отметим, что рассмотренный критерий нечувствителен к умеренным отклонениям от предположения о нормальности распределения.

Важно так же следующее: для проверки гипотезы $H_0$: $M(X) = A$ против любой из возможных альтернатив $H_1$ можно использовать доверительный интервал. Мы отвергаем $H_0$ с уровнем значимости $\alpha$, если $A$ лежит вне определенного с доверительной вероятностью $\gamma = 1 - \alpha$ доверительного интервала.

***Пример.*** Средний рост младенцев в нормально распределенной популяции новорожденных составляет 51,35 см ($A = 51{,}35$ см). По данным выборки (здесь мы ее не приводим), предоставленной одним из родильных домов, средний рост новорожденных мальчиков $\bar{x}_в = 51{,}8$ см, выборочная дисперсия $S^2 = 2{,}1$ см$^2$, объем выборки $n = 25$.

Можно предположить, что средний рост $M(X)$ в популяции новорожденных мальчиков больше чем 51,35 см. Чтобы подтвердить либо опровергнуть это предположение при уровне значимости $\alpha = 0{,}05$, проверим гипотезу $H_0$: $M(X) = 51{,}35$ см против альтернативы $H_1$: $M(X) > 51{,}35$ см (правосторонняя критическая область). Вычислим $t_{набл.}$ по (26):

$$t_{набл.} = \frac{51{,}8 - 51{,}35}{1{,}45} \cdot 5 = 1{,}55$$

и найдем $t_{правост.\ кр.}$; $t_{правост.\ кр.} = 1{,}71$.

Так как $t_{набл.} < t_{правост.\ кр.}$, с заданным уровнем значимости $\alpha = 0{,}05$ принимается $H_0$. Это подтверждается также соотношением между $P$ и $\alpha$: вычисленное $P = 0{,}067$, значит $P > \alpha$.

при решении многих прикладных задач, в том числе возникающих в медицинской промышленности.

Если $H_0$ принимается, т. е. $M(X) = A$, то выборочная средняя $\bar{x}_в$ также незначимо отличается от гипотетической генеральной средней $A$, а если справедлива гипотеза $H_1$, то различие этих величин значимо.

***Пример.*** Проектный размер изделий, изготавливаемых станком-автоматом, $A = 35$ мм. Предприятием получен станок, требует ли он корректировки в конкретных условиях работы, если измерения 20 случайно отобранных изделий дали следующие результаты: $\bar{x}_в = 35{,}07$ мм, $S = 0{,}16$ мм, $t_{набл.} = 1{,}96$?

Проверим при $\alpha = 0{,}05$ основную гипотезу $H_0$: $M(X) = 35$ при $H_1$: $M(X) \neq 35$.

Найденное $t_{двухст.\ кр.}(0{,}05;\ 19) = 2{,}09$ и так как $t_{набл.} < t_{двухст.\ кр.}$, то нет оснований при уровне значимости 0,05 отвергать $H_0$: $M(X) = 35$ мм, так что отличие $\bar{x}_в$ от $A = 35$ мм тоже незначимо. Станок обеспечивает проектный размер изделий и не требует корректировки своей работы.

В заключение отметим, что рассмотренный критерий нечувствителен к умеренным отклонениям от предположения о нормальности распределения.

Важно так же следующее: для проверки гипотезы $H_0$: $M(X) = A$ против любой из возможных альтернатив $H_1$ можно использовать доверительный интервал. Мы отвергаем $H_0$ с уровнем значимости $\alpha$, если $A$ лежит вне определенного с доверительной вероятностью $\gamma = 1 - \alpha$ доверительного интервала.

***Пример.*** Средний рост младенцев в нормально распределенной популяции новорожденных составляет 51,35 см ($A = 51{,}35$ см). По данным выборки (здесь мы ее не приводим), предоставленной одним из родильных домов, средний рост новорожденных мальчиков $\bar{x}_в = 51{,}8$ см, выборочная дисперсия $S^2 = 2{,}1$ см$^2$, объем выборки $n = 25$.

Можно предположить, что средний рост $M(X)$ в популяции новорожденных мальчиков больше чем 51,35 см. Чтобы подтвердить либо опровергнуть это предположение при уровне значимости $\alpha = 0{,}05$, проверим гипотезу $H_0$: $M(X) = 51{,}35$ см против альтернативы $H_1$: $M(X) > 51{,}35$ см (правосторонняя критическая область). Вычислим $t_{набл.}$ по (26):

$$t_{набл.} = \frac{51{,}8 - 51{,}35}{1{,}45} \cdot 5 = 1{,}55$$

и найдем $t_{правост.\ кр.}$; $t_{правост.\ кр.} = 1{,}71$.

Так как $t_{набл.} < t_{правост.\ кр.}$, с заданным уровнем значимости $\alpha = 0{,}05$ принимается $H_0$. Это подтверждается также соотношением между $P$ и $\alpha$: вычисленное $P = 0{,}067$, значит $P > \alpha$.

Вывод: средний рост новорожденных мальчиков $M(X)$ и $\bar{x}_в$ незначимо отличаются от среднего роста новорожденных младенцев.

Рассчитаем по данным выборки доверительный интервал для $M(X)$, коэффициент Стьюдента $t_{0,95;25} = 2,1$. Расчет дает: полуширина интервала $\delta = 0,61$ см, а 51,19 см $< M(X) <$ 52,41 см.

Вывод: этот расчет подтверждает справедливость $H_0$: с вероятностью 95 % полученный доверительный интервал содержит и средний рост новорожденных А = 51,35 см.

**II. *Работа с двумя выборками.***

Для работы с рассмотренными ниже критериями в программе Excel имеется инструмент «Анализ данных».

1. *Проверка гипотезы о равенстве математических ожиданий (средних) двух нормальных генеральных совокупностей при неизвестных, но одинаковых дисперсиях*[1] *(малые независимые выборки*[2]*, двухвыборочный t-критерий Стьюдента).*

Постановка задачи: получены 2 независимые выборки из нормальных генеральных совокупностей случайных величин $X$ и $Y$, их объемы $n_1$ и $n_2$. По этим выборкам найдены $\bar{x}_в$, $\bar{y}_в$, $S_x^2$ и $S_y^2$

Требуется проверить гипотезу $H_0$: $M(X) = M(Y)$.

Здесь в качестве критерия (статистики) используется двухвыборочный критерий Стьюдента:

$$t = \frac{(\bar{x}_в - \bar{y}_в)}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \qquad (28)$$

При $n_1 = n_2 = n$

$$t = \frac{(\bar{x}_в - \bar{y}_в)}{\sqrt{S_x^2 + S_y^2}} \sqrt{n} \qquad (29)$$

Дальнейшие действия стандартны. Необходимо вычислить $t_{набл.}$ с помощью формулы (28) или (29) и конкретных характеристик выборок. Если полученное значение $t_{набл.}$ принадлежит критической области, то $H_0$ отвергается и принимается конкурирующая гипотеза $H_1$: $M(X)$ и $M(У)$, а следовательно, $\bar{x}_в$ и $\bar{y}_в$ различаются значимо, т. е. их различие вызвано принципиальными причинами. Если $t_{набл.}$ оказывается в области принятия нулевой гипотезы, то $M(X) = M(У)$ и различие выборочных средних незначимо и обусловлено случайными факторами.

---

[1] Равенство дисперсий можно проверить используя критерий Фишера–Снедекора, который будет рассмотрен ниже.

[2] Независимые выборки — выборки, полученные для разных объектов, связанных определенным исследованием.

Правила, позволяющие сделать вывод о справедливости нулевой гипотезы, зависят от конкурирующей гипотезы $H_1$, от уровня значимости α и числа степеней свободы $m = n_1 + n_2 - 2$ или $m = 2(n - 1)$ (при $n_1 = n_2 = n$). Они не отличаются от рассмотренных в первом случае работы с одной выборкой.

Если рассчитывается вероятность $P$, то $H_0$ принимается при $P > \alpha$.

***Пример.*** Сравнительное исследование концентрации свинца в крови (в мг/100 г) группы рабочих аккумуляторного завода $X$ (подвергавшихся профессиональному воздействию) и группы рабочих текстильной фабрики $Y$ (не подвергавшихся профессиональному воздействию), привело к следующим результатам:

$\bar{x}_в = 0{,}08157$ мг/100 г, $S_x = 0{,}0067$ мг/100 г, $S_x^2 = 4{,}489 \cdot 10^{-5}$, $n = 7$

$\bar{y}_в = 0{,}03943$ мг/100 г, $S_y = 0{,}00355$ мг/100 г, $S_y^2 = 1{,}26 \cdot 10^{-5}$, $n = 7$.

Число степеней свободы $m = 12$.

Предполагается, что $D(X) = D(Y)$ и исследуемый показатель в генеральной совокупности распределен по нормальному закону.

При $\alpha = 0{,}05$ проверяется $H_0$: $M(X) = M(Y)$ против альтернативы $H_1$: $M(X) \neq M(Y)$. В соответствии с вышеприведенными числовыми данными $t_{набл.} = 19{,}6$, $t_{двухст.\ кр.} = 2{,}18$.

Так как $t_{набл.} > t_{двухст.\ кр.}$, нулевая гипотеза отвергается с заданным уровнем значимости.

То же подтверждает расчет $P$, $P < 0{,}05$.

Вывод: условия работы значимо влияют на содержание свинца в крови рабочих.

2. *Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей (F-критерий Фишера–Снедекора).*

Постановка задачи: пусть генеральные совокупности величин $X$ и $Y$ распределены по нормальному закону. По независимым выборкам объемами $n_1$ и $n_2$, извлеченным из этих совокупностей, найдены выборочные дисперсии $S_x^2$ и $S_y^2$. По этим дисперсиям при заданном уровне значимости α требуется проверить нулевую гипотезу, при этом генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0: D(X) = D(Y).$$

Если окажется, что гипотеза $H_0$ справедлива, т. е. генеральные дисперсии одинаковы, то различие выборочных дисперсий незначимо и объясняется случайными причинами. Если $H_0$ будет отвергнута, т. е. если генеральные дисперсии неодинаковы, то различие выборочных дисперсий значимо. Оно не может быть объяснено случайными причинами, а является следствием различия самих генеральных дисперсий.

В качестве критерия (статистики) проверки нулевой гипотезы о равенстве генеральных дисперсий принимают величину:

$$F = \frac{S_б^2}{S_м^2} \text{ (критерий } F \text{ Фишера–Снедекора)}, \qquad (30)$$

где $S_м$ и $S_б$ — соответственно меньшее и большее значение выборочных дисперсий.

При использовании критерия (30) критическая область, как всегда, определяется видом конкурирующей гипотезы. Рассмотрим два случая.

1. Нулевая гипотеза $H_0$: $D(X) = D(Y)$.

Конкурирующая гипотеза $H_1$: $D(X) \neq D(Y)$. Здесь критическая область *двусторонняя*.

2. Если есть основание предполагать, что одна из дисперсий обязательно не меньше другой, например, $D(X) \geq D(Y)$, тогда нулевая гипотеза $H_0$: $D(X) = D(Y)$, а конкурирующая гипотеза $H_1$: $D(X) > D(Y)$. В данном случае критическая область *правосторонняя*. Если $H_1$: $D(X) < D(Y)$, критическая *область левосторонняя*.

При решении конкретных задач придерживаются следующих правил:

1. Чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0$: $D(X) = D(Y)$ при конкурирующей гипотезе $H_1$: $D(X) > D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей выборочной дисперсии к меньшей): $F_{\text{набл.}} = \frac{S_б^2}{S_м^2} \geq 1$

Далее по заданному α и числам степеней свободы $m_1 = n_1 - 1$ и $m_2 = n_2 - 1$ ($n_1$ — объем выборки, для которой получена большая выборочная дисперсия) находят критическую точку $F_{\text{кр.}}(a, m_1, m_2)$. Если $F_{\text{набл.}} < F_{\text{кр.}}$ — нет оснований отвергать $H_0$ и она принимается. Если же $F_{\text{набл.}} > F_{\text{кр}}$, то $H_0$ отвергают и полагают, что $D(X) > D(Y)$.

При работе со статистическими пакетами гипотезу $H_0$ принимают, если $P > \alpha$, в противном случае — нет.

2. При конкурирующей гипотезе $H_1$: $D(X) \neq D(Y)$ критическую точку $F_{\text{кр.}}(\alpha/2, m_1, m_2)$ ищут по уровню значимости α/2. Если $F_{\text{набл.}} < F_{\text{кр.}}$ — нет оснований отвергать $H_0$. Если $F_{\text{набл.}} > F_{\text{кр.}}$ — $H_0$ отвергают.

При известной вероятности $P$ гипотезу $H_0$ принимают, если $P > \alpha = \alpha/2 + \alpha/2$.

***Пример.*** Условие задачи и соответствующие данные приведены в примере, который иллюстрирует работу двухвыборочного критерия Стьюдента. Воспользуемся ими в данном случае.

Сформулируем вопрос: при уровне значимости $\alpha = 0{,}05$ проверить $H_0$: $D(X) = D(Y)$, при $H_1$: $D(X) \neq D(Y)$.

По данным задачи, $F_{\text{набл.}} = \frac{S_x^2}{S_y^2} = 3{,}56$. Так как критическая область двусторонняя, то при отыскании критической точки следует брать уро-

вень значимости в два раза меньший заданного, то есть $\alpha/2 = 0{,}025$, тогда $F_{кр.} = (0{,}025; 6; 6) = 5{,}82$.

Так как $F_{набл.} < F_{кр.}$, $H_0$ принимается — $D(X) = D(Y)$.

3. *Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки[1]).*

Постановка задачи: пусть генеральные совокупности $X$ и $Y$ распределены нормально, причем их дисперсии неизвестны. Требуется, используя зависимые выборки, при уровне значимости $\alpha$, проверить основную гипотезу $H_0$: $M(X) = M(Y)$ при альтернативе $H_1$: $M(X) \neq M(Y)$ или $H_1$: $M(X) > M(Y)$, или $H_1$: $M(X) < M(Y)$.

В этом случае используется парный двухвыборочный t-критерий Стьюдента с $m = n - 1$ степенями свободы, который имеет вид:

$$t = \frac{\overline{d}}{S_d} \cdot \sqrt{n}, \qquad (31)$$

где $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, $d_i = x_i - y_i$ $(i = 1 \ldots n)$, $S_d = \sqrt{\sum_{i=1}^{n} \frac{(d_i - \overline{d})^2}{n-1}}$, $n$ — объем выборки.

Далее решение задачи находится тривиально: сравниваем $t_{набл.}$, полученное по данным выборки из (31), и $t_{кр.}$ при разных вариантах критических областей.

Например, если $H_1$: $M(X) \neq M(Y)$ и $|t_{набл.}| < t_{двухст.\ кр.}$, $H_0$ принимается, в противном случае принимается $H_1$.

При возможности вычисления $P$, $H_0$ принимается при $P > \alpha$.

***Пример.*** Рассмотрим популяцию, состоящую из критически больных пациентов с циркуляторным шоком. Была получена выборка из 108 пациентов и у каждого из них измерялось $X$ — венозное pH и $Y$ — артериальное pH.

Из клинического опыта известно, что для здоровых людей среднее венозное pH меньше, чем артериальное. Проверим, выполняется ли это соотношение для популяции больных с указанной выше патологией, используя парный t-критерий Стьюдента, примем $H_0$: $M(X) = M(Y)$, а $H_1$: $M(X) < M(Y)$.

По литературным данным [1], $\overline{d} = -0{,}04$, $S_d = 0{,}1533$, $\sqrt{n} = \sqrt{108} = 10{,}39$. Используя (31), получим, что $t_{набл.} = -2{,}71$, $t_{левост.\ кр.} = -1{,}66$ на уровне значимости $\alpha = 0{,}05$. Так как $t_{набл.} < t_{левост.\ кр.}$, то $H_0$ отвергается и принимается $H_1$.

---

[1] Зависимые выборки — выборки, полученные для одних и тех же объектов, связанных определенным исследованием.

Вывод: в популяции критически больных пациентов среднее венозное pH и среднее артериальное pH значимо отличаются друг от друга, причем $(pH)_{вен.} < (pH)_{арт.}$. Это неравенство подтверждается медицинскими фактами.

## 2.8. Основы корреляционного анализа

Одной из задач анализа данных является установление зависимости (связи) между признаками — случайными величинами (частота пульса, артериальное давление, показатель анализа крови и т. д.). Эта задача решается методами корреляционного[1] анализа.

Пусть $X$ и $Y$ — случайные величины. *Зависимость их друг от друга (если она существует) называется корреляционной зависимостью.* Она может быть установлена *качественно* — по виду диаграммы рассеяния (корреляционного поля), и *количественно* — путем вычисления коэффициента корреляции. При установлении корреляционной зависимости экспериментально для каждого обследованного объекта получают соответствующие пары значений величин $X$ и $Y$ (например, роста и массы тела людей определенного пола и возраста, числа эритроцитов и содержания гемоглобина в анализе крови).

Пусть объем выборки — $n$. Каждой паре значений $(x_i, y_i)$ на плоскости $x0y$ соответствует одна точка. Всего будет $n$ точек.

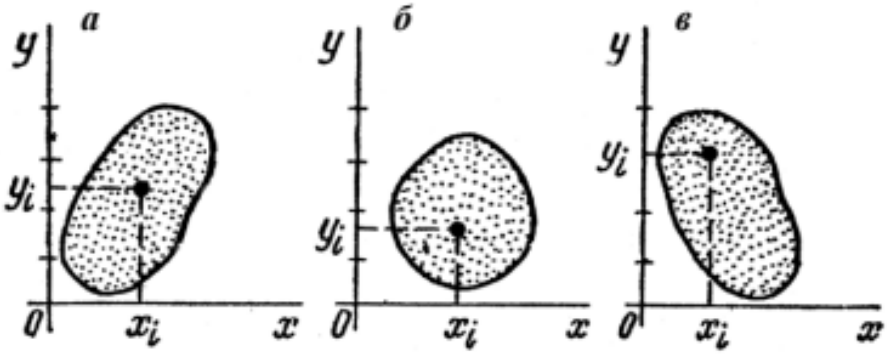



*Рис. 17*

Область на графике $y(x)$, занятая этими точками, образует *диаграмму расеяния* (*корреляционное поле)*. Разные виды таких диаграмм (полей) показаны на рис. 17. Если форма корреляционного поля близка к кругу (рис. 17, *б*), то связи между признаками $X$ и $Y$ нет. Если же корреляционное поле вытянуто (рис. 17, *а*, 17, *в*), то корреляционная связь между признаками $X$ и $Y$ есть, и она тем сильнее, чем более вытянуто корреляционное поле.

[1] Корреляция (от англ. correlation) — согласование, связь, взаимозависимость.

По экспериментальным данным, для каждого значения признака $X$ можно найти $\overline{Y}$. Зависимость $\overline{Y}_x = f(x)$ называется *эмпирическим уравнением регрессии Y на X*. Аналогично можно получить зависимость $\overline{X}_y = \varphi(y)$ — *уравнение регрессии X на Y*. Графики этих функций называются *линиями регрессии*. Экспериментальные данные более или менее тесно группируются вдоль этих линий. Если они представляют собой прямые, то корреляционная связь между признаками $X$ и $Y$ называется *линейной* и оценивается с помощью *выборочного коэффициента корреляции r*.

Значения $r$ по модулю не превышают 1, но могут быть как положительными, так и отрицательными:

$$-1 \le r \le 1 \text{ или } |r| \le 1.$$

При $r = 0$ линейная связь между $X$ и $Y$ отсутствует; при значениях $|r|$ до 0,3 — связь слабая; от 0,3 до 0,7 — умеренная; от 0,7 до 1 — сильная; если $|r| \approx 1$ — связь полная или, иначе, функциональная — в этом случае существует функция $Y = f(X)$, связывающая значения $Y$ и $X$.

Вернемся к линиям регрессии. Рассмотрим, например, зависимость $Y = f(X)$, Если ее график — прямая, то ее уравнение ($\overline{Y}_x = f(x)$) можно записать в виде:

$$\overline{Y}_x = kx + b, \qquad (31)$$

где постоянные $k$ и $b$ определяются по выборке.

По линиям регрессии можно оценить тенденцию (тренд) изменения одной величины при изменении значений другой. Коэффициент $k$ в (31) называется коэффициентом регрессии. Если $k > 0$, то при увеличении (уменьшении) значений одной величины, например $X$, увеличиваются (уменьшаются) значения другой $Y$. При $k < 0$, с увеличением (уменьшением) значений $X$, значения $Y$ уменьшаются (увеличиваются).

Коэффициент корреляции $r$ имеет тот же знак, что и $k$. При $r > 0$ связь между признаками $X$ и $Y$ называется *прямой*, при $r < 0$ — *обратной*.

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте зависимости между признаками, полученное по данным выборки, можно распространить и на генеральную совокупность. Например, для оценки коэффициента корреляции $r_г$ нормально распределенной генеральной совокупности (при $n > 50$) можно воспользоваться формулой

$$r - 3\frac{1-r^2}{\sqrt{n}} < r_г < r + 3\frac{1-r^2}{\sqrt{n}}.$$