# ЧАСТЬ 8

# МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

## Лекция 14

## ОСНОВНЫЕ ПОНЯТИЯ И ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

*ЦЕЛЬ ЛЕКЦИИ: определить понятие генеральной и выборочной совокупности и сформулировать три типичные задачи математической статистики; ввести понятия выборочной функции распределения, вариационного ряда и гистограммы; привести наиболее важные для математической статистики распределения.*

Математическая статистика – это математическая наука посвященная разработке методов описания и анализа статистических экспериментальных данных, полученных в результате наблюдений массовых случайных явлений.

### Генеральная и выборочная совокупности

Значительная часть математической статистики связана с описанием и анализом больших совокупностей объектов, объединенных по некоторому качественному или количественному признаку $X$. Такая группа объектов называется статистической совокупностью. Если исследуемая совокупность слишком многочисленна, либо ее элементы малодоступны, либо имеются другие причины, не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется выборочной совокупностью или выборкой, а все множество изучаемых элементов – генеральной совокупностью. Под выборкой понимается последовательность независимых, одинаково распределенных случайных величин, т. е. каждая выборка $(x_1, x_2, \ldots, x_n)$ значений случайной величины $X$ рассматривается как результат $n$ независимых повторных испытаний. Объемом совокупности называется число объектов, входящих в эту совокупность. Например, если из 10 000 микросхем для проверки качества отобрано 200 штук, то объем генеральной совокупности равен 10 000, а выборочной – 200.

Естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, т. е. была бы, как говорят, представительной (репрезентативной). Это обеспечивается как независимостью результатов наблюдений в выборке и случайностью выбора объектов из генеральной совокупности, так и правильным определением объема выборки с учетом всех конкретных условий. Чтобы этого добиться, применяются различные способы получения выборки или отбора.

- Отбор, не требующий разбиения генеральной совокупности на части, например простой случайный бесповторный отбор и простой случайный повторный отбор.
- Отбор, при котором генеральная совокупность разбивается на части, например типический, механический, серийный и комбинированный отборы.

На практике чаще всего используют бесповторный отбор, так как повторный отбор в некоторых случаях может оказаться нереализуемым из-за разрушения одного или нескольких элементов.

Статистическая совокупность, расположенная в порядке возрастания или убывания значений изучаемого признака $X$, называется вариационным рядом, а ее объекты – вариантами.

Вариационный ряд называется дискретным, если его члены принимают конкретные изолированные значения. Если элементы вариационного ряда заполняют некоторый интервал, то такой ряд называется непрерывным.

## Типичные задачи математической статистики

Методы математической статистики нашли широкое применение в различных областях науки (физике, биологии, медицине, экономике, социологии и др.) и могут применяться для решения различных задач. Однако можно сформулировать три основные (типичные) задачи математической статистики, наиболее часто встречающиеся на практике.

1. Определение закона распределения случайной величины. По результатам независимых наблюдений случайной величины $X$ требуется оценить неизвестную функцию распределения $F(x)$ или плотность вероятности $f(x)$ этой случайной величины.

2. Задача проверки правдоподобия гипотез. Из обширного круга задач, связанных с проверкой статистических гипотез, наиболее типичными являются две задачи. Первая: как согласуются результаты эксперимента с гипотезой о том, что исследуемая случайная величина имеет плотность распределения $f(x)$? Вторая: не противоречит ли полученная оценка неизвестного параметра выдвинутой гипотезе о значении данного параметра?

3. Задача оценки неизвестных параметров распределения. Предполагается, что закон распределения исследуемой случайной величины известен до опыта из физических или теоретических предположений (например, нормальный). Возникает более узкая задача – определить некоторые параметры (числовые характеристики) случайной величины, т. е. по экспериментальным данным необходимо оценить значения этих параметров. С этой задачей отыскания "подходящих значений" числовых характеристик тесно связана задача оценки их точности и надежности.

## Выборочная функция распределения

Пусть изучается некоторая случайная величина (признак) $X$ с неизвестным законом распределения. Нужно определить закон из опыта и проверить гипотезу о том, что распределение случайной величины $X$ подчиняется именно этому закону. Для этого над случайной величиной $X$ производится ряд независимых испытаний (наблюдений), в каждом из которых $X$ принимает то или иное значение $x_i$, $i = 1, 2, \ldots, n$; $n$ – количество проведенных опытов. Вот эта совокупность наблюдаемых значений случайной величины и есть выборочная совокупность или выборка, которая представляет собой первичный статистический материал, подлежащий обработке и анализу. Выборка оформляется в виде таблицы, в первом столбце которой записаны номера опытов $i$, а во втором – наблюдаемые значения случайной величины.

**Пример**. Случайная величина $X$ – значения напряжения на выходе генератора шума, взятые через 20 миллисекунд. Выборочная совокупность представлена в виде табл. 8.1.

*Таблица 8.1*

| $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ |
|---|---|---|---|---|---|
| 1 | -2 | 8 | 8 | 15 | 2 |
| 2 | -6 | 9 | -3 | 16 | -3 |
| 3 | 0 | 10 | 1 | 17 | 11 |
| 4 | -5 | 11 | 11 | 18 | 1 |
| 5 | -9 | 12 | 9 | 18 | 12 |
| 6 | 0 | 13 | -8 | 20 | -2 |
| 7 | -7 | 14 | -3 | | |

Упорядоченные в порядке возрастания значения признака $X$ дадут вариационный ряд, который может быть обработан различными методами.

Один из таких способов – построение выборочной функции распределения случайной величины.

Выборочной функцией распределения случайной величины $X$ называется частота события $\{X < x\}$

$$F^*(x) = P^*\{X < x\}.$$

Для получения значений $F^*(x)$ для заданного аргумента $x$ достаточно подсчитать число испытаний, в которых случайная величина $X$ приняла значение, меньшее чем $x$, и разделить на общее число $n$ проведенных экспериментов.

На рис. 8.1 представлен график выборочной функции распределения случайной величины $X$ – напряжения на выходе генератора шума.

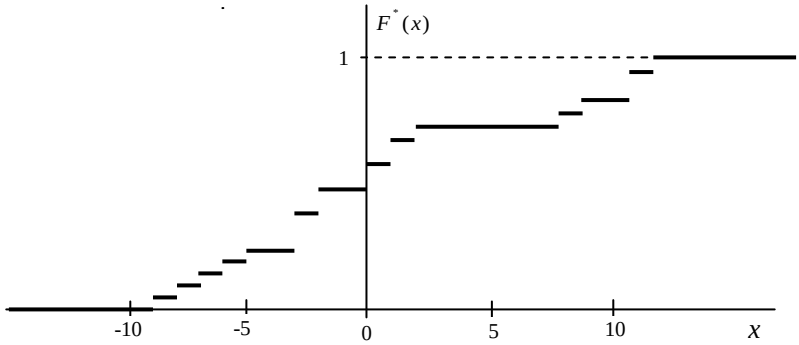


*Рис. 8.1.* Выборочная функция распределения

Выборочная функция распределения любой случайной величины, как непрерывной, так и дискретной, представляет собой неубывающую, прерывистую, ступенчатую функцию. При этом разрывы функции происходят при значениях аргумента, равных наблюдаемым значениям случайной величины, а величины разрывов равны частотам этих значений. Если каждое значение встречается по одному разу, то все скачки будут равны $1/n$.

При увеличении числа опытов $n$, согласно теореме Бернулли (следствие закона больших чисел), для любых $x$ частота события $\{X < x\}$ приближается (сходится по вероятности) к вероятности этого события. Таким образом, при увеличении $n$ выборочная функция распределения $F^*(x)$ сходится по вероятности к истинной функции распределения $F(x)$ случайной величины $X$.

Если $X$ – непрерывная случайная величина, то при увеличении числа наблюдений $n$ число скачков функции $F^*(x)$ увеличивается, а величина скачков уменьшается, и график функции $F^*(x)$ сходится к плавной кривой $F(x)$.

## Статистическое распределение выборки. Полигон и гистограмма

Практически построение $F^*(x)$ решает задачу описания экспериментального материала. Однако при больших $n$ построение $F^*(x)$ слишком трудоемко и не всегда наглядно по сравнению с другими видами закона распределения, например $f(x)$.

Для придания выборочной совокупности или вариационному ряду компактности и наглядности статистический материал подвергается дополнительной обработке, т. е. строится так называемое статистическое распределение выборки. Для дискретного вариационного ряда статистическое распределение представляется в виде табл. 8.2, в первой строке которой записываются в возрастающем порядке варианты (элементы выборки) $x_i$, а во второй – соответствующие им частоты $p_i^*$.

*Таблица 8.2*

| Варианты | $x_1$ | $x_2$ | … | $x_i$ | … | $x_n$ |
|---|---|---|---|---|---|---|
| Частоты | $p_1^*$ | $p_2^*$ | … | $p_i^*$ | … | $p_k^*$ |

Для непрерывного вариационного ряда весь диапазон наблюдаемых значений случайной величины $X$ разбивается на интервалы и подсчитывается количество значений $m_i$, приходящихся на каждый $i$-й интервал. После деления $m_i$ на общее число опытов $n$, получается частота, соответствующая каждому интервалу:

$$p_i^* = m_i / n .$$

Сумма этих частот должна быть равна единице.

Затем строится таблица, в первой строке которой приводятся в порядке возрастания интервалы, а во второй – соответствующие частоты. Табл. 8.3 и есть статистическое распределение непрерывной выборки.

*Таблица 8.3*

| Интервалы | $x_1 : x_2$ | $x_2 : x_3$ | … | $x_i : x_{i+1}$ | … | $x_k : x_{k+1}$ |
|---|---|---|---|---|---|---|
| Частоты | $p_1^*$ | $p_2^*$ | … | $p_i^*$ | … | $p_k^*$ |

Если наблюдаемое значение случайной величины попадает точно на границу двух интервалов, то такая величина в равной степени принадлежит к обоим интервалам, и поэтому к значениям $m_i$ того и другого разряда прибавляется по $1/2$.

Число интервалов, на которые необходимо группировать статистические данные, не должно быть слишком большим, так как в этом случае статистический ряд становится невыразительным, а частоты в нем могут иметь незакономерные колебания. Но, с другой стороны, количество интервалов не должно быть и слишком малым, потому что в этом случае особенности распределения описываются статистическим распределением лишком грубо. Из практических соображений число интервалов выбирается порядка $10 \div 20$.

Графически статистическое распределение дискретного статистического ряда представляют в виде полигона (см. рис. 8.2), который строится следующим образом. На оси абсцисс откладываются значения варианта $x_i$, а на оси ординат соответствующие им частоты $p_i^*$. Полученные точки соединяются ломаной линией

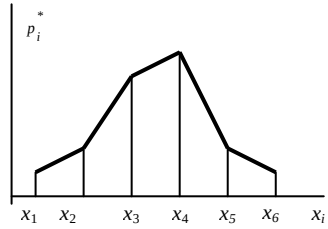


*Рис. 8.2. Полигон*

Графическое представление статистического распределения непрерывного вариационного ряда называется гистограммой (см. рис. 8.3). На оси абсцисс откладываются интервалы, и на каждом из них, как на основании, строится прямоугольник, площадь которого равна частоте соответствующего разряда. Для одинаковых по ширине интервалов высоты прямоугольников пропорциональны соответствующим частотам. Полная площадь гистограммы равна единице. При увеличении числа испытаний

можно выбирать все меньшую и меньшую ширину интервалов и гистограмма будет приближаться к кривой распределения $f(x)$.

Статистическое распределение выборки можно использовать для приближенного построения выборочной функции распределения случайной величины, так как построение точной $F^*(x)$ с несколькими сотнями скачков для всех наблюдаемых значений случайной величины $X$ очень трудоемко. На практике достаточно построить $F^*(x)$ по нескольким точкам, в качестве которых выбираются границы интервалов $x_1, x_2; \ldots$, находящиеся в первой строке статистического распределения. Таким образом, имеем:

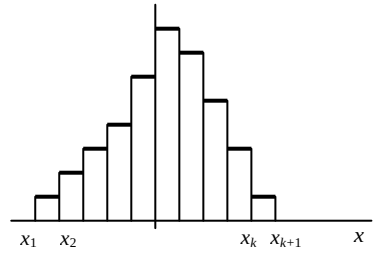

*Рис. 8.3.* Гистограмма

$$F^*(x_1) = 0;\ F^*(x_2) = p_1^*;\ \ldots;\ F^*(x_k) = \sum_{i=1}^{k-1} p_i^*;\ F^*(x_{k+1}) = \sum_{i=1}^{k} p_i^* = 1.$$

Соединив полученные точки ломаной линией или плавной кривой, получим приближенный график выборочной функции распределения (см. рис. 8.4).

В зависимости от конкретного содержания задачи в схему построения гистограммы могут быть внесены некоторые изменения. Например, в некоторых задачах целесообразно отказаться от требований равной длины интервалов.

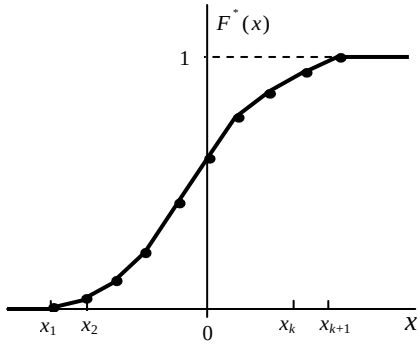

*Рис. 8.4.* Приближенная кривая выборочной функции распределения

## Наиболее важные распределения

Несколько примеров распределений дискретных и непрерывных случайных величин было приведено в лекциях 7 и 8. Важнейшим с точки зрения приложений математической статистики является нормальное (гауссово) распределение. В статистике широко используются еще три распределения, связанные с нормально распределенными случайными величинами. К ним относятся распределение $\chi^2$ (Пирсона), $t$-распределение (Стьюдента) и $F$-распределение (Снедекора – Фишера).

**Стандартное нормальное распределение**. Плотность распределения вероятности и функция распределения нормальной случайной величины $X$ определяются выражениями соответственно (4.29) и (4.38). Удобнее пользоваться стандартной случайной величиной

$$Z = \frac{X - m}{\sigma}, \tag{8.1}$$

где $m$ – математическое ожидание, $\sigma$ – среднее квадратичное отклонение нормальной случайной величины $X$.

После подстановки выражения (8.1) в формулы (4.29) и (4.38) получим плотность распределения и функцию распределения стандартной гауссовой случайной величины $Z$ с нулевым математическим ожиданием и единичной дисперсией ( $m_Z = 0, \sigma_Z^2 = 1$):

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \varphi(z), \tag{8.2}$$

$$F(z) = \frac{1}{2} + \Phi(z). \tag{8.3}$$

Таблица функции $\varphi(z)$ приведены в прил. 1, а $\Phi(z)$ – в прил. 2.

Значение $z_\alpha$, удовлетворяющее уравнениям:

$$F(z_\alpha) = P\{z < z_\alpha\} = p = 1 - \alpha, \tag{8.4а}$$

$$\int_{z_\alpha}^{\infty} f(z)dz = P\{z \geq z_\alpha\} = \alpha, \tag{8.4б}$$

где вероятность $0 < \alpha < 1$, называется квантилем порядка $p = 1 - \alpha$ или $100\alpha$-процентной точкой стандартного нормального распределения.

**Распределение** $\chi^2$. Есть $Z_1, Z_2, \ldots, Z_n$ – $n$ независимых случайных величин, каждая из которых имеет нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Определяют новую случайную величину

$$\chi_n^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2 .$$

Величина $\chi_n^2$ называется хи-квадрат случайной величиной с $n$ степенями свободы. Число степеней свободы $n$ определяет число независимых, или "свободных", квадратов входящих в сумму. Плотность распределения $\chi_n^2$ имеет следующий вид:

$$f(\chi_n^2) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} (\chi_n^2)^{\frac{n}{2}-1} e^{-\frac{\chi_n^2}{2}}, \chi_n^2 > 0 \\ 0, \chi_n^2 \le 0 \end{cases}, \qquad (8.5)$$

где $\Gamma(\beta) = \int_0^\infty t^{\beta-1} e^{-t} dt$ – гамма-функция.

Математическое ожидание случайной величины имеющей распределение $\chi_n^2$, равно $n$, а дисперсия – $2n$. Кривые распределения $\chi^2$ для трех значений $r$ представлены на рис. 8.5.

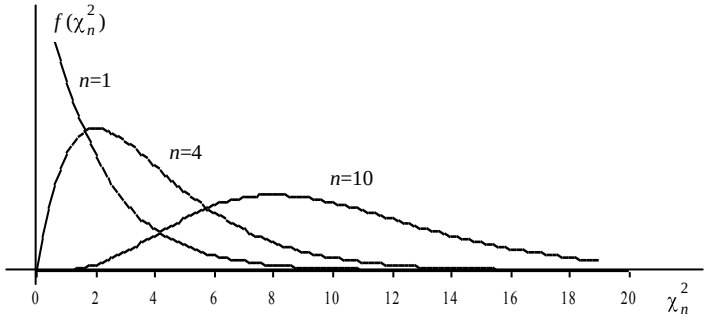


*Рис. 8.5.* Кривые распределения $\chi_n^2$

При увеличении числа степеней свободы $\chi^2$-распределение приближается к нормальному. Для $n > 30$ случайная величина $\sqrt{2\chi_n^2}$ почти нормальна с математическим ожиданием $M[\chi_n^2] = \sqrt{2n-1}$ и дисперсией $D[\chi_n^2] = 1$.

Процентные точки $\chi^2$-распределения обозначают через $\chi^2_{n;\alpha}$ – это решения уравнения

$$F(\chi^2_{n,\alpha}) = \int_{\chi^2_{n;\alpha}}^{\infty} f(\chi_n^2) d\chi_n^2 = P\{\chi_n^2 \geq \chi^2_{n;\alpha}\} = \alpha .$$

Таблица процентных точек распределения $\chi_n^2$ приведена в прил. 3.

***t*-распределение Стьюдента**. Есть $Y$ и $Z$ –независимые случайные величины, при этом $Y$ имеет $\chi_n^2$-распределение, а $Z$ – нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Определяется новая случайная величина

$$t_n = \frac{Y}{\sqrt{Z/n}} .$$

Случайная величина $t_n$ подчиняется закону распределения Стьюдента с $n$ степенями свободы, плотность распределения которого имеет вид

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty . \tag{8.6}$$

Кривые распределения Стьюдента для трех значений $r$ приведены на рис. 8.6.

Математическое ожидание и дисперсия случайной величины $t_n$ равны

$$M[t_n] = 0, n > 1,$$

$$D[t_n] = \frac{n}{n-2}, n > 2 .$$

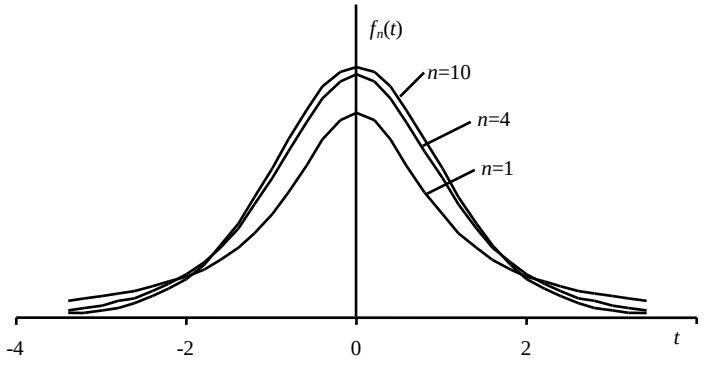


*Рис. 8.6.* Кривые *t*-распределения Стьюдента

Процентные точки *t*-распределения обозначают через $t_{n;\alpha}$ – это шения уравнения

$$F(t_{n,\alpha}) = \int_{t_{n;\alpha}}^{\infty} f_n(t)dt = P\{t \geq t_{n;\alpha}\} = \alpha .$$

Таблица процентных точек *t*-распределения приведена в прил. 4.

При увеличении числа степеней свободы $n$ *t*-распределение приближается к стандартному гауссовому распределению.

***F*-распределение Снедекора – Фишера**. Есть $Y_1$ и $Y_2$ – независимые случайные величины, подчиняющиеся распределению $\chi^2$ с $n_1$ и $n_2$ степенями свободы соответственно. Определяется новая случайная величина

$$F_{n_1,n_2} = \frac{Y_1/n_1}{Y_2/n_2} = \frac{Y_1 n_2}{Y_2 n_1}.$$

Случайная величина $F_{n_1,n_2}$ называется величиной $F$ с $n_1$ и $n_2$ степенями свободы, ее плотность распределения имеет вид

$$p(f) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} f^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}f\right)^{-\frac{n_1+n_2}{2}}, \qquad (8.7)$$

где $0 < f < \infty$.

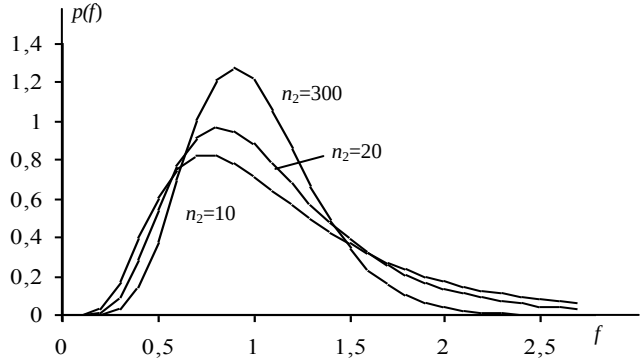


*Рис. 8.7.* Кривые *F*-распределения для $n_1$=20

Кривые *F*-распределения при $n_1 = 20$ и трех значениях $n_2$ приведены на рис. 8.7.

Математическое ожидание и дисперсия случайной величины $F_{n_1,n_2}$ определяются формулами

$$M[F_{n_1;n_2}] = \frac{n_2}{n_2 - 2},\ n_2 > 2$$

$$D[F_{n_1;n_2}] = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)},\ n_2 > 4.$$

Процентные точки *F*-распределения обозначают через $f_{n_1,n_2;\alpha}$. Эти точки являются решениями уравнения

$$F(f_{n_1,n_2;\alpha}) = \int_{f_{n_1,n_2;\alpha}}^{\infty} p(f)df = P\{F_{n_1,n_2} \geq f_{n_1,n_2;\alpha}\} = \alpha .$$

Таблица процентных точек *F*-распределения приведена в прил. 5.