**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**
**УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ**
**«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**Кафедра «Прикладная математика»**

**Е. Л. САЗОНОВА**

# ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

**Ч а с т ь 2**

## МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

**Учебно-методическое пособие**
**для студентов факультета безотрывного обучения**

**Гомель 2007**

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТРАНСПОРТА»

Кафедра «Прикладная математика»

Е. Л. САЗОНОВА

ТЕОРИЯ ВЕРОЯТНОСТЕЙ
И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Ч а с т ь 2

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие
для студентов факультета безотрывного обучения

Под редакцией канд. физ.-мат. наук, доцента *В. С. Серёгиной*

Одобрено методической комиссией
факультета безотрывного обучения

Гомель 2007

УДК 519.21/.22
ББК 22.171
С14

Рецензент – канд. физ.-мат. наук, доцент кафедры «Высшая математика» А. М. Щербо (УО «БелГУТ»).

**Сазонова, Е. Л.**

С14 Теория вероятностей и математическая статистика. В 2 ч. Ч. 2. Математическая статистика : учеб.-метод. пособие для студентов факультета безотрывного обучения / Е.Л. Сазонова ; под ред. В. С. Серёгиной; М-во образования Респ. Беларусь, Белорус. гос. ун-т трансп. – Гомель : БелГУТ, 2007. – 70 с.

ISBN 978-985-468-130-6

Содержит основные сведения курса математической статистики, примеры решения задач, задания для контрольной работы № 2 и пример выполнения контрольной работы. Часть 1 «Теория вероятностей» была издана в 2000 г. (переиздана в 2003 г).

Предназначено для студентов факультета безотрывного обучения всех специальностей.

**УДК 519.21/.22**
**ББК 22.171**

**ISBN 978-958-468-130-6**





## ПРЕДМЕТ И ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

В практической работе, при проведении научных и технических исследований, при изучении явлений окружающего мира часто приходится иметь дело с так называемыми **вероятностными (случайными) экспериментами**, то есть экспериментами, воспроизводимыми многократно при соблюдении одного и того же комплекса условий, результаты которых изменяются от опыта к опыту. Наблюдаемые различия в результатах экспериментов объясняются тем, что на исход испытания оказывает влияние очень большое число факторов, не заданных в числе его основных условий. Влияние этих факторов изменяется при переходе от одного испытания к другому и вносит практически непредсказуемые различия в их результаты.

Например, при изготовлении на станке серии однотипных деталей, несмотря на постулируемую неизменность условий их производства, размеры получаемых деталей всегда будут несколько отличаться друг от друга. Это является следствием влияния таких факторов как:

– неоднородность свойств материала, из которого изготавливается деталь;

– мельчайшие изменения положения детали относительно станка;

– мельчайшие изменения размеров рабочих частей станка (стачивание, изнашивание и т. д.);

– изменения параметров настройки станка;

– колебания температуры, влажности воздуха в помещении и т.п.

Влияние каждого из таких факторов на результат эксперимента ничтожно, но наложение влияний большого числа различных факторов в каждом конкретном случае формирует уже заметные вариации размера изготавливаемой детали.

Если при построении математической модели изучаемого вероятностного эксперимента пренебречь влиянием таких факторов, то полученная модель будет являться очень грубым приближением исследуемого явления и окажется мало пригодной для его описания. Попытка же учесть все эти факторы обычными детерминистскими методами приведет к непомерному усложнению и загромождению модели и, следовательно, к невозможности ее практического использования.

Поэтому для учета вклада большого числа факторов (эти факторы принято называть **случайными**, поскольку законы изменения их влияния в большинстве случаев не известны исследователю) необходимы другие, качественно отличающиеся от детерминистских методы. Такие методы и разрабатываются в теории вероятностей.

Замечательным экспериментальным фактом является то, что результаты большой серии испытаний, рассматриваемые «суммарно», мало отличаются друг от друга. Наличие подобных устойчивостей объясняется тем, что при рассмотрении результатов большого числа экспериментов в совокупности, влияние неконтролируемо меняющихся факторов в значительной мере взаимно погашается, и получаемые таким образом «суммарные» значения регистрируемых величин концентрируются около некоторых средних значений, причем отличие от этих средних значений тем меньше, чем больше число проведенных экспериментов.

Установленный факт является основой практического применения теории вероятностей: только такие случайные явления, обладающие свойством устойчивости частот, и являются объектом ее исследования. Напомним, что *основной задачей теории вероятностей* является изучение закономерностей, характеризующих проведение большого числа вероятностных экспериментов, и разработка математических моделей, предназначенных для описания случайных явлений. При этом исход каждого отдельного вероятностного эксперимента остается непредсказуемым, но результаты достаточно большой серии испытаний могут быть хорошо описаны с помощью построенной математической модели.

При проведении практических исследований, в большинстве случаев, математические модели изучаемых явлений неизвестны, в распоряжении исследователя имеются лишь результаты наблюдений. Для построения (подбора) на основании имеющихся опытных данных наилучшей математической модели используются методы математической статистики.

**Математическая статистика** – система основанных на теоретико-вероятностных моделях понятий, приемов и математических методов, предназначенных для сбора, систематизации, интерпретации и обработки статистических данных с целью получения научных и практических выводов.

Круг задач, решаемых методами математической статистики очень широк. В настоящее время практически не существует областей науки, техники и естествознания, где в той или иной мере не применялись бы вероятностно-статистические методы.

В пособии изложены некоторые методы математической статистики, позволяющие:

– компактно и наглядно представить имеющиеся опытные данные;

– оценить параметры теоретического распределения по экспериментальным данным;

– проверить статистические гипотезы о виде закона распределения исследуемой случайной величины;

– исследовать зависимость между двумя случайными величинами и осуществить прогноз значений одной из них по известным значениям другой переменной.



# 1 НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ КУРСА ТЕОРИИ ВЕРОЯТНОСТЕЙ

В этом разделе приведены основные сведения курса теории вероятностей, используемые при последующем изложении материала.

## 1.1 Закон распределения случайной величины

**Случайной величиной *X*** называется функция, определенная на заданном пространстве элементарных событий $\Omega$ и ставящая в соответствие каждому элементарному событию $\omega_i \in \Omega$ некоторое вещественное число $x$. Или иначе, случайную величину можно рассматривать как величину определенного физического смысла, которая в результате вероятностного эксперимента принимает одно из своих возможных значений, причем неизвестно заранее, какое именно.

Случайные величины, множество возможных значений которых является конечным или счетным множеством, называются **дискретными**; это означает, что дискретные случайные величины могут принимать только отдельные друг от друга значения с определенными вероятностями.

Случайные величины, множество значений которых непрерывно заполняет некоторый промежуток числовой оси, называются **непрерывными**. Очевидно, что множество значений непрерывной случайной величины является несчетным множеством.

**Законом распределения** случайной величины называется любое соотношение, устанавливающее связь между возможными значениями этой величины и соответствующими им вероятностями.

Универсальной формой задания закона распределения дискретных и непрерывных случайных величин является функция распределения.

**Функцией распределения случайной величины *X*** называется функция $F(x)$, определяющая для каждого значения $x$ вероятность того, что случайная величина $X$ примет значение меньшее, чем $x$:

$$F(x) = P(X < x)$$

(более строгое определение непрерывной случайной величины может быть дано следующим образом: непрерывной называется случайная величина, функция распределения которой непрерывна).

*Основные свойства функции распределения F(x):*

1 Все возможные значения функции распределения принадлежат отрезку [0; 1]:

$$0 \le F(x) \le 1.$$

2 $F(x)$ – неубывающая функция своего аргумента, то есть для любых значений $x_1$ и $x_2$, таких, что $x_1 < x_2$, справедливо соотношение: $F(x_1) \le F(x_2)$.

3 Вероятность того, что случайная величина примет значение, принадлежащее полуинтервалу $[a; b)$, равна приращению функции распределения на этом интервале:

$$P(a \le X < b) = F(b) - F(a).$$

4 Если все возможные значения случайной величины принадлежат отрезку $[a; b]$, то $F(x) = 0$, при $x \le a$; $F(x) = 1$, при $x > b$.

5 $\lim\limits_{n \to -\infty} F(x) = 0; \quad \lim\limits_{n \to \infty} F(x) = 1.$

Закон распределения *дискретной* случайной величины может быть задан также с помощью ряда распределения.

**Рядом распределения** дискретной случайной величины $X$ называется таблица, в первой строке которой указаны все возможные значения этой случайной величины, а во второй – соответствующие им вероятности:

| $x_i$ | $x_1$ | $x_2$ | $x_3$ | . . . |
|---|---|---|---|---|
| $p_i$ | $p_1$ | $p_2$ | $p_3$ | . . . |

Согласно определению, $\sum\limits_i p_i = \sum\limits_i P(X = x_i) = 1 \quad (p_i \ge 0).$

Другим возможным способом задания закона распределения *непрерывной* случайной величины является использование функции плотности распределения. Согласно определению, **функция плотности распределения *f*(*x*)** является производной функции распределения $F(x)$: $f(x) = F'(x)$. График функции плотности распределения называется **кривой распределения**.

*Основные свойства функции плотности распределения вероятностей:*

1 Плотность распределения вероятностей – неотрицательная функция:

$$f(x) \ge 0$$

(*геометрически:* кривая распределения лежит не ниже оси абсцисс).

2 Вероятность попадания значения случайной величины на участок от $\alpha$ до $\beta$ определяется по формуле

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx$$

(*геометрически:* эта вероятность равна площади криволинейной трапеции, ограниченной кривой $f(x)$, осью $Ox$ и прямыми $x = \alpha$ и $x = \beta$).

3 Функция распределения $F(x)$ может быть определена по известной функции плотности распределения следующим образом:



$$F(x) = P(X < x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(t)dt.$$

4 Площадь фигуры, ограниченной кривой распределения и осью абсцисс, равна единице:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

В частности, если все возможные значения случайной величины принадлежат отрезку $[a; b]$, то

$$\int_{a}^{b} f(x)dx = 1.$$

### 1.2 Числовые характеристики случайной величины

В теории вероятностей при описании случайных величин широко используются так называемые *числовые характеристики* – числа, характеризующие определенные свойства исследуемой случайной величины.

**Математическое ожидание (*M*[*X*])** характеризует среднее значение (центр распределения) случайной величины.

**Модой ($x_{\text{mod}}$)** *дискретной* случайной величины называется ее наиболее вероятное значение; а модой *непрерывной* величины – то ее значение $x$, при котором достигается максимум функции $f(x)$.

**Медиана ($x_{med}$)** – это то значение случайной величины, для которого выполняется соотношение $P(X < x_{med}) = P(X \ge x_{med}) = 0{,}5$, то есть одинаково вероятно, примет ли случайная величина $X$ значение большее или меньшее, чем $x_{med}$.

**Дисперсия (*D*[*X*]) и среднее квадратическое отклонение (σ[*X*])** являются мерами рассеивания значений случайной величины относительно математического ожидания.

**Коэффициент асимметрии (*A*[*X*])** является показателем «скошенности» распределения случайной величины относительно математического ожидания. Для симметричных распределений $A[X] = 0$.

**Коэффициент эксцесса (*Ex*[*X*])** может использоваться в качестве характеристики «островершинности» кривой распределения . Для нормально распределенной случайной величины $Ex[X] = 0$.

Расчетные формулы для вычисления числовых характеристик приведены в таблице 1.1.

*Таблица 1.1* – **Расчетные формулы для вычисления числовых характеристик**

| Числовая характеристика | Расчётная формула | |
|---|---|---|
| | для дискретных с. в. | для непрерывных с. в. |
| $M[X]$ | $\sum_i x_i p_i$ | $\int_{-\infty}^{\infty} x f(x)dx$ |
| $D[X]$ | $\sum_i (x_i - M[X])^2 p_i = \sum_i x_i^2 p_i - (M[X])^2$ | $\int_{-\infty}^{\infty} (x - M[X])^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - (M[X])^2$ |
| $\sigma[X]$ | $\sqrt{D[X]}$ | $\sqrt{D[X]}$ |
| $A[X]$ | $\frac{\sum_i (x_i - M[X])^3 p_i}{(\sigma[X])^3}$ | $\frac{\int_{-\infty}^{\infty} (x - M[X])^3 f(x)dx}{(\sigma[X])^3}$ |
| $Ex[X]$ | $\frac{\sum_i (x_i - M[X])^4 p_i}{(\sigma[X])^4} - 3$ | $\frac{\int_{-\infty}^{\infty} (x - M[X])^4 f(x)dx}{(\sigma[X])^4} - 3$ |

### 1.3 Основные законы распределения случайных величин, наиболее часто встречающиеся на практике

В таблицах 1.2 и 1.3 приведены основные сведения о наиболее часто встречающихся на практике законах распределения дискретных и непрерывных случайных величин.

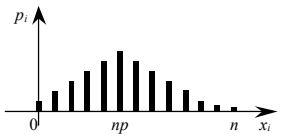
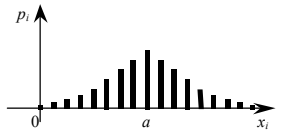
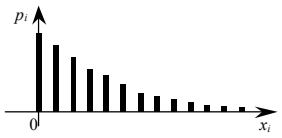
### 1.4 Законы распределения случайных величин, широко использующиеся в математической статистике

В этом подразделе кратко рассмотрены распределения случайных величин, играющих фундаментальную роль в математической статистике и составляющих основу проверки статистических гипотез и интервального оценивания параметров.

#### 1.4.1 Распределение $\chi^2$ (хи-квадрат)

Рассмотрим $n$ независимых стандартизованных нормально распределенных случайных величин $X_1, X_2, \ldots, X_n$ (случайная величина $X$ называется *стандартизованной*, если $M[X] = 0,\ \sigma[X] = 1$). Сумма квадратов этих переменных $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ называется **случайной величиной, распределенной по закону $\chi^2$** с $\nu = n$ степенями свободы.

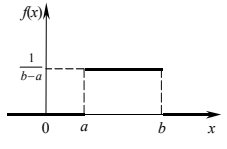
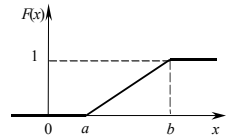
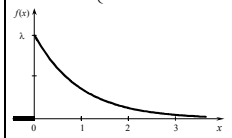
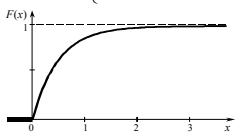
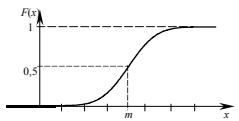
*Таблица 1.2* – **Основные сведения о наиболее часто встречающихся на практике законах распределения дискретных случайных величин**

| Название закона распределения | Возможные значения | Параметры | Статистическая оценка параметров | Числовые характеристики | | | Вероятности возможных значений, столбцовая диаграмма | Примечание |
|---|---|---|---|---|---|---|---|---|
| | | | | $M[X]$ | $D[X]$ | $\sigma[X]$ | | |
| Биномиальный | $X = 0, 1, 2, ..., n$ | $p$, $n$ | $\hat{p} = \hat{M}[X]/\hat{n}$ | $np$ | $npq$ | $\sqrt{npq}$ | $P(X=k) = C_n^k p^k q^{n-k}$   | Случайная величина $X$ характеризует число появлений события $A$ в серии из $n$ независимых испытаний, в каждом из которых это событие может осуществиться с вероятностью $p$ |
| Пуассона | $X = 0, 1, 2, ..., m, ...$ | $a$ | $\hat{a} = \hat{M}[X]$ | $a$ | $a$ | $\sqrt{a}$ | $P(X=k) = \frac{a^k}{k!} e^{-a}$   | **Пример:** число событий простейшего потока, характеризующегося интенсивностью $a$, где $a$ – число событий, произошедших в течение единицы времени |
| Геометрический | $X = 0, 1, 2, ..., m, ...$ | $p$ | $\hat{p} = 1/(\hat{M}[X]+1)$ | $\frac{1}{p} - 1$ | $\frac{q}{p^2}$ | $\sqrt{\frac{q}{p^2}}$ | $P(X=k) = q^k p$   | Случайная величина $X$ характеризует число независимых испытаний, произведённых до первого появления события $A$, которое в каждом из этих испытаний может произойти с вероятностью $p$ (при этом испытание, в котором появляется событие $A$, не учитывается) |


9

*Таблица 1.3* – **Наиболее часто встречающиеся на практике законы распределения непрерывных случайных величин**

| Закон распределения | Возможные значения | Параметры | Статистическая оценка параметров | Числовые характеристики | | | | | Вероятность попадания значений с. в. в отрезок [α; β] |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $M[X]$ | $D[X]$ | $\sigma[X]$ | $A[X]$ | $Ex[X]$ | |
| Равномерный | $X \in [a;b]$ | $a$<br>$b$ | $\hat{a} = x_{\min} - \frac{x_{\max} - x_{\min}}{n-1}$,<br>$\hat{b} = x_{\max} + \frac{x_{\max} - x_{\min}}{n-1}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ | $\frac{b-a}{2\sqrt{3}}$ | 0 | –1,2 | $P(\alpha \le X \le \beta) = \frac{\beta - \alpha}{b-a}$ |
| Экспоненциальный (показательный) | $X \in [0;\infty)$ | $\lambda$ | $\hat{\lambda} = 1/\hat{M}[X]$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ | $\frac{1}{\lambda}$ | 2 | 6 | $P(\alpha \le X \le \beta) = e^{-\lambda\alpha} - e^{-\lambda\beta}$ |
| Нормальный | $X \in R$ | $m$<br>$\sigma$ | $\hat{m} = \hat{M}[X]$<br>$\hat{\sigma} = \hat{\sigma}[X]$ | $m$ | $\sigma^2$ | $\sigma$ | 0 | 0 | $P(\alpha \le X \le \beta) = \Phi\left(\frac{\beta - m}{\sigma}\right) - \Phi\left(\frac{\alpha - m}{\sigma}\right)$;<br>$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ |

| Функция плотности распределения вероятностей, кривая распределения | Функция распределения | Примечание |
|---|---|---|
| $f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a;b]; \\ 0, & x \notin [a;b] \end{cases}$<br>  | $F(x) = \begin{cases} 0, & x \le a; \\ \frac{x-a}{b-a}, & a < x \le b; \\ 1, & x > b \end{cases}$<br>  | Если все возможные значения непрерывной случайной величины принадлежат отрезку [a; b], и все значения, попадающие на этот отрезок, равноправны, то данная случайная величина распределена по равномерному закону.<br>**Пример:** величина погрешности при округлении данных |
| $f(x) = \begin{cases} 0, & x < 0; \\ \lambda e^{-\lambda x}, & x \ge 0 \end{cases}$<br>  | $F(x) = \begin{cases} 0, & x \le 0; \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}$<br>  | **Примеры:**<br>– промежуток времени между моментами наступления двух последовательных событий простейшего потока;<br>– разнообразные временные характеристики функционирования технических устройств (время безотказной работы оборудования, время простоя в ожидании ремонта и т. д.) |
| $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-m)^2}{2\sigma^2}}$<br>  | $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{(t-m)^2}{2\sigma^2}} dt$<br>  | Если случайная величина $X$ представляет собой сумму большого числа независимых (или слабо зависимых) случайных величин, сопоставимых по уровню своего влияния на суммарный результат, то эта величина имеет распределение, близкое к нормальному.<br>**Пример:** реальные значения параметров изготовленного изделия |

Распределение $\chi^2$ зависит только от одного параметра $\nu$ – числа степеней свободы. На рисунке 1.1, *а* приведены графики функции плотности распределения $\chi^2$ в зависимости от числа степеней свободы $\nu$. Кривая распределения $\chi^2$ имеет положительную асимметрию. С увеличением числа степеней свободы она становится все более симметричной и при $\nu > 30$ это распределение практически совпадает с нормальным распределением. Известны значения числовых характеристик распределения $\chi^2$:

$$M[\chi^2] = \nu; \;\; D[\chi^2] = 2\nu^2; \;\; \sigma[\chi^2] = \sqrt{2}\nu.$$

На практике часто используются квантили $\chi^2_{\alpha,\nu}$ распределения $\chi^2$. **Квантилем** $\chi^2_{\alpha,\nu}$, отвечающим заданному уровню вероятности $\alpha$, называется такое значение $\chi^2 = \chi^2_{\alpha,\nu}$, при котором

$$P(\chi^2 > \chi^2_{\alpha,\nu}) = \int_{\chi^2_{\alpha,\nu}}^{\infty} f(\chi^2) d\chi^2 = \alpha .$$

С геометрической точки зрения, нахождение квантиля $\chi^2_{\alpha,\nu}$ заключается в выборе такого значения $\chi^2 = \chi^2_{\alpha,\nu}$, при котором площадь заштрихованной криволинейной трапеции (рисунок 1.1, *б*) равна $\alpha$.

В приложении А приведены значения квантилей $\chi^2_{\alpha,\nu}$ в зависимости от числа степеней свободы $\nu$ и заданного уровня вероятности $\alpha$.

а) б)

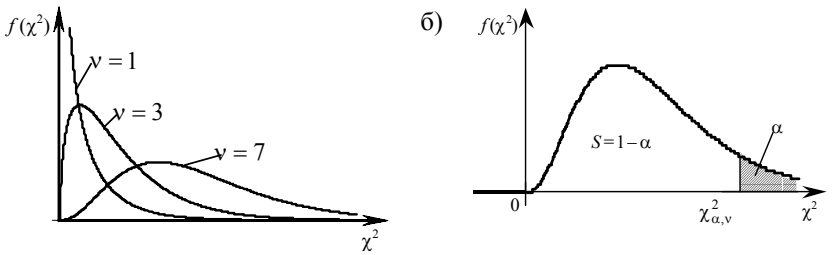



Рисунок 1.1 – Распределение $\chi^2$:
*а* – зависимость $f(\chi^2)$ от значения параметра ν; *б* – геометрическая интерпретация квантиля $\chi^2_{\alpha,\nu}$

### 1.4.2 *t*-распределение Стьюдента

Пусть $X$, $X_1$, $X_2$, …, $X_n$ – независимые случайные величины, имеющие стандартизованное нормальное распределение. Тогда случайная величина

$$t = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

имеет распределение, которое называется ***t*-распределением**, или **распределением Стьюдента** с $\nu = n$ степенями свободы. *t*-распределение симметрично относительно оси ординат при всех значениях ν, и при неограниченном увеличении числа степеней свободы (практически уже при $\nu > 30$) приближается к стандартизованному нормальному распределению (имеющему плотность распределения вероятностей $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$). На рисунке 1.2 изображены кривые распределения Стьюдента для нескольких значений параметра ν и кривая стандартизованного нормального распределения.

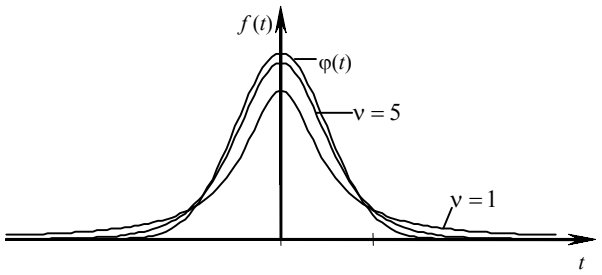



Рисунок 1.2 – Кривые распределения Стьюдента

Математическое ожидание, дисперсия и среднее квадратическое отклонение случайной величины *t*, соответственно, равны:



$$M[t] = 0;\ D[t] = \frac{\nu}{\nu - 2};\ \sigma[t] = \sqrt{\frac{\nu}{\nu - 2}}.$$

В приложении Б приведены значения квантилей распределения Стьюдента $t_{\alpha,\nu}$ в зависимости от числа степеней свободы ν и заданного уровня значимости α, определяемые условием:

$$P(t > t_{\alpha,\nu}) = \int_{t_{\alpha,\nu}}^{\infty} f(t)dt = \alpha\ .$$

На рисунке 1.3, *а* штриховкой отмечена область, площадь которой равна α.

При построении доверительных интервалов для неизвестных параметров распределения часто возникает задача определения таких значений $t_1$ и $t_2$, что $P(t_1 < t < t_2) = 1 - \alpha$. Обычно значения $t_1$ и $t_2$ выбираются симметричными относительно оси ординат, то есть $t_1 = -t_{\alpha/2,\nu}$, $t_2 = t_{\alpha/2,\nu}$. С геометрической точки зрения это означает, что суммарная площадь заштрихованных на рисунке 1.3, *б* криволинейных трапеций равна α.

а) б)

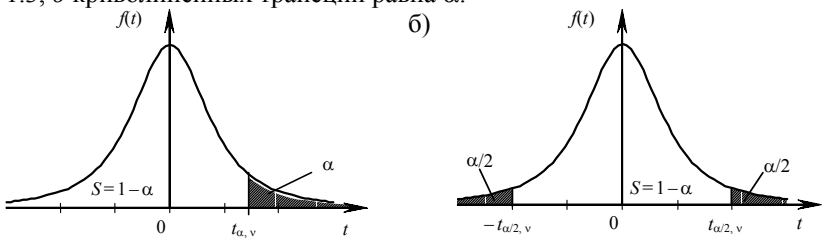



Рисунок 1.3 – Геометрическая интерпретация
квантилей распределения Стьюдента

### 1.4.3 *F*-распределение Фишера

Пусть случайные величины $X_1$, $X_2$, …, $X_m$; $Y_1$, $Y_2$, …, $Y_n$ независимы и имеют стандартизованное нормальное распределение. Тогда случайная величина

$$F = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}$$

имеет **распределение Фишера**, которое зависит от двух параметров $\nu_1 = m$ и $\nu_2 = n$, называемых числами степеней свободы.

Графики функции плотности *F*-распределения при различных значениях параметров $\nu_1$ и $\nu_2$ изображены на рисунке 1.4, *а*.

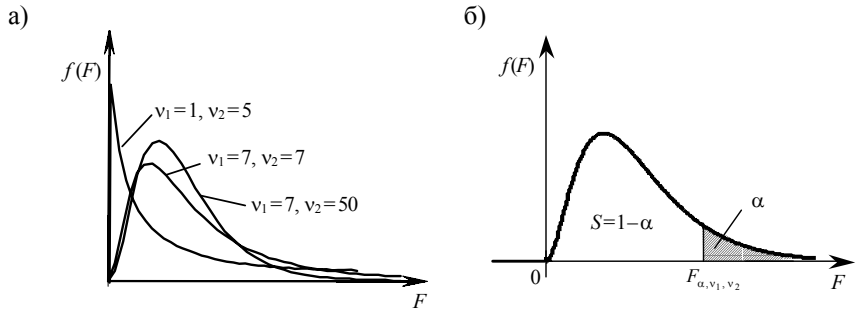



Рисунок 1.4 – Распределение Фишера:

*а* – зависимость *f*(*F*) от значений параметров $\nu_1$ и $\nu_2$; *б* – геометрическая интерпретация квантилей распределения

Кривая *F*-распределения имеет положительную асимметрию, при больших значениях $\nu_1$ и $\nu_2$ это распределение медленно приближается к нормальному.

В приложении В приведены значения квантилей *F*-распределения $F_{\alpha,\nu_1,\nu_2}$ в зависимости от числа степеней свободы и значения вероятности $\alpha$. Значения квантилей найдены путем решения уравнения

$$P(F > F_{\alpha,\nu_1,\nu_2}) = \int_{F_{\alpha,\nu_1,\nu_2}}^{\infty} f(F)dF = \alpha\ .$$

На рисунке 1.4, *б* штриховкой выделена фигура, площадь которой равна $\alpha$.

## 2 ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

### 2.1 Выборочный метод

При проведении статистического исследования множество всех объектов, подвергающихся обследованию называется **генеральной совокупностью**. Иначе говоря, генеральную совокупность можно рассматривать как множество всех мыслимых наблюдений, которые могли бы быть зафиксированы при воспроизведении изучаемого эксперимента.

Очевидно, что наиболее полное представление об изучаемом явлении мы получим в том случае, если сможем обследовать всю генеральную совокупность.

В качестве примеров сплошного обследования генеральной совокупности можно привести перепись населения, проведение референдумов и выборов (при этом в роли генеральной совокупности выступает множество жителей государства), контроль качества всех выпускаемых предприятием изделий и др.



Несмотря на привлекательность сплошного исследования с точки зрения точности получаемых результатов, на практике оно используется сравнительно редко в силу его дороговизны, ресурсоемкости, длительности проведения, а порой и невозможности его осуществления (например, если изучается продолжительность безотказной работы технических устройств определенного вида, то в принципе невозможно подвергнуть проверке такого рода все выпускаемые устройства).

Поэтому при проведении практических исследований широко используется *выборочный метод*, состоящий в том, что из генеральной совокупности определенным образом извлекается часть объектов, называемая **выборкой**, которая подвергается детальному изучению (число элементов выборки обозначается символом *n* и называется ее *объемом*). Используя известные теоретико-вероятностные соотношения, по результатам выборочного обследования формулируются выводы о свойствах всей генеральной совокупности. Можно сказать, что основное назначение математико-статистических методов именно в том и состоит, чтобы с их помощью на основании ограниченного набора выборочных данных получить как можно более полное представление о свойствах изучаемой случайной величины.

Приведем *математическую интерпретацию выборочного метода*. Для получения выборки значений исследуемой случайной величины вероятностный эксперимент воспроизводится в одних и тех же условиях независимым образом *n* раз. Результат каждого испытания может быть описан с помощью случайной величины $X_i$, $i = 1, 2, \ldots, n$, причем закон распределения каждой из этих величин совпадает с законом распределения вероятностей исследуемой случайной величины $X$. Таким образом, выборка представляет собой $n$-мерную случайную величину $(X_1, X_2, \ldots, X_n)$, компоненты которой независимы и одинаково распределены. Конкретные экспериментально полученные реализации выборки $(X_1, X_2, \ldots, X_n)$ обозначаются $(x_1, x_2, \ldots, x_n)$.

Для того, чтобы по имеющимся опытным данным можно было сделать обоснованный вывод о свойствах всей генеральной совокупности, исследуемая выборка должна быть репрезентативной (представительной), то есть хорошо отображать свойства изучаемой генеральной совокупности. Один из разделов математической статистики посвящен теории получения репрезентативных выборок. В частности, доказано, что для получения представительной выборки, все элементы должны извлекаться независимым образом и с *сохранением принципа случайности*, то есть каждый из элементов генеральной совокупности должен иметь равные с другими шансы быть представленным в выборке. Кроме того, большое значение имеет объем исследуемых данных. Увеличение объема выборки позволяет повысить точность получаемых результатов. Разработаны специальные

методы, позволяющие определить объем выборочных данных, необходимый для достижения заданной точности результатов исследования.

В дальнейшем, всегда будем предполагать, что все полученные для исследования выборки $(x_1, x_2, \ldots, x_n)$ являются репрезентативными и представляют собой экспериментальную реализацию многомерной случайной величины $(X_1, X_2, \ldots, X_n)$, компоненты которой независимы и закон распределения каждой из величин $X_i$ совпадает с законом распределения изучаемой случайной величины $X$.

### 2.2 Статистический закон распределения

Пусть для исследования свойств случайной величины $X$ получена выборка объема $n$: $(x_1, x_2, \ldots, x_n)$. Последовательность выборочных значений $x_1$, $x_2, \ldots, x_n$, записанных в порядке их появления, представляет собой исходный статистический материал и называется **простым статистическим рядом**.

Для компактного, удобного и наглядного представления имеющихся статистических данных необходимо произвести их первичную обработку. Естественным первым шагом такой обработки является упорядочение полученных значений. Последовательность элементов выборки, записанных в порядке возрастания (неубывания) их значений называется **вариационным рядом**, а каждый из элементов вариационного ряда называется **вариантой**.

Если изучается *дискретная случайная величина*, число различных наблюдаемых значений которой не велико, то для каждого из отличающихся друг от друга значений (обозначим их $\tilde{x}_i$) подсчитываются частоты $m_i$ и относительные частоты (частости) $m_i/n$ появления этих значений в выборке.*

Результаты вычислений заносятся в таблицу 2.1, которая называется **сгруппированным статистическим рядом**.

*Таблица 2.1* – **Сгруппированный статистический ряд**

| Наблюдаемые значения $\tilde{x}_i$ | $\tilde{x}_1$ | $\tilde{x}_2$ | … | $\tilde{x}_k$ | $k \le n$ |
|---|---|---|---|---|---|
| Частоты $m_i$ | $m_1$ | $m_2$ | … | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |
| Относительные частоты $m_i / n$ | $m_1 / n$ | $m_2 / n$ | … | $m_k / n$ | $\sum_{i=1}^{k} m_i / n = 1$ |

Для графического изображения сгруппированного статистического ряда обычно используется *столбцовая диаграмма*. Этот график представляет собой последовательность вертикальных отрезков длины $m_i / n$, отложенных от оси абсцисс в точках с координатами $\tilde{x}_i$ (рисунок 2.1).

* **Частотой** $m_i$ значения $\tilde{x}_i$ называется число повторений этого значения в выборке, а **относительной частотой** (**частостью**) – отношение частоты к объёму выборки $m_i/n$.



Если изучается *непрерывная случайная величина* (в этом случае все выборочные значения $x_i$ могут оказаться различными), либо дискретная случайная величина, число отличающихся друг от друга значений которой достаточно велико, то диапазон всех наблюдаемых значений $x_i$ разбивается на $k$ разрядов длины $h$ и подсчитывается число вариант, попавших в каждый из разрядов. Результаты расчетов заносятся в таблицу 2.2, которая называется **интервальным статистическим рядом**.

*Таблица 2.2* – **Интервальный статистический ряд**

| Интервалы выборочных значений $[C_i; C_{i+1})$ | $[C_1; C_2)$ | $[C_2; C_3)$ | … | $[C_k; C_{k+1})$ | |
|---|---|---|---|---|---|
| Среднее значение интервала $\tilde{x}_i$ | $\tilde{x}_1$ | $\tilde{x}_2$ | … | $\tilde{x}_k$ | |
| Частоты $m_i$ | $m_1$ | $m_2$ | … | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |
| Относительные частоты $m_i / n$ | $m_1 / n$ | $m_2 / n$ | … | $m_k / n$ | $\sum_{i=1}^{k} m_i / n = 1$ |

Для определения границ частичных интервалов $(C_i; C_{i+1})$ можно воспользоваться следующей методикой:

1 Вычислить размах варьирования выборочных значений: $R = x_{\max} - x_{\min}$, где $x_{\min}$ и $x_{\max}$, соответственно, минимальное и максимальное значения выборки.

2 Определить длину шага разбиения $h = R/k$, здесь $k$ – число разрядов разбиения (принято использовать значения $5 \le k \le 15$). Для примерной ориентации в выборе значения $k$ можно воспользоваться формулой Стерджесса $k \approx 1 + 3{,}322 \lg n$, где $n$ – объем выборки.

3 Определить границы интервалов разбиения: $C_1 = x_{\min} - h/2$, $C_2 = C_1 + h$, $C_3 = C_2 + h, \ldots$, $C_{k+1} = C_k + h$. Процесс разбиения продолжается до тех пор, пока очередное значение $C_{k+1}$ не превысит максимальный элемент выборки.

Среднее значение $\tilde{x}_i$ $i$-го частичного интервала можно определить как среднее арифметическое границ этого интервала: $\tilde{x}_i = (C_i + C_{i+1}) / 2$, $i = 1, 2, \ldots, k$.

Выборочные значения, попавшие на границы интервалов разбиения, либо могут быть приписаны к какому-то одному из этих интервалов (например, к правому, как это сделано в таблице 2.2), либо частоты этих значений могут быть разделены поровну между двумя соседними интервалами.

Нетрудно заметить, что сумма всех частот $m_i$ равна объему выборки $n$, а сумма частостей, согласно определению, равна единице: $\sum_{i=1}^{k} m_i = n$, $\sum_{i=1}^{k} \frac{m_i}{n} = 1$.

**Замечание** – На основании предельных теорем теории вероятностей доказано, что при неограниченном увеличении объема выборки $n$ относительные частоты со-

бытий $X = \widetilde{x}_i$ и $X \in [c_i; c_{i+1})$ сходятся к вероятностям соответствующих событий, то есть относительные частоты $m_i/n$, записанные в последних строках сгруппированного и интервального статистических рядов, являются некоторыми приближениями (оценками), соответственно, вероятностей событий $X = \widetilde{x}_i$ и $X \in [c_i; c_{i+1})$.

Для графического представления интервального статистического ряда используется *гистограмма относительных частот*. При построении гистограммы по оси абсцисс необходимо отложить границы интервалов выборочных значений $(C_i; C_{i+1})$, и на каждом из этих интервалов, как на основании, построить прямоугольники, площади которых равны $m_i/n$. Следовательно, высоты этих прямоугольников должны быть равны $m_i/(nh)$.

Легко заметить, что согласно определению, площадь всей гистограммы равна 1 (так как $\sum_{i=1}^{k} \frac{m_i}{n} = 1$). Это требование формулируется для того, чтобы построенная по опытным данным гистограмма относительных частот представляла собой экспериментальный аналог графика функции плотности распределения вероятностей $f(x)$ исследуемой случайной величины $X$ (напомним, что площадь фигуры, ограниченной графиком функции $f(x)$ и осью абсцисс, равна 1). С использованием предельных теорем теории вероятностей доказано, что при неограниченном увеличении объема выборки $n$ высоты полученных прямоугольников гистограммы, фиксируемые в средней точке каждого из частичных интервалов, приближаются к значению функции плотности распределения вероятностей $f(x)$ в этой точке.

**Пример 2.1.** Для исследования характеристик работы железнодорожной станции получена выборка объёма $n = 100$ значений случайной величины $X$, характеризующей число составов, прибывающих на эту станцию в течение часа.

Выборочные данные:

1, 1, 5, 1, 3, 3, 1, 5, 2, 6, 1, 6, 5, 2, 4, 3, 4, 2, 3, 2, 3, 3, 8, 4, 3, 2, 1, 1, 2, 5, 0, 3, 1, 7, 3, 2, 3, 3, 5, 1, 3, 1, 1, 4, 5, 1, 7, 1, 2, 1, 2, 3, 5, 2, 4, 2, 1, 1, 3, 2, 2, 3, 5, 1, 5, 4, 3, 1, 5, 3, 3, 3, 5, 3, 3, 1, 4, 4, 0, 2, 2, 2, 0, 1, 1, 6, 5, 4, 3, 3, 4, 2, 1, 1, 0, 0, 2, 5, 8, 2, 1.

Требуется на основании имеющихся опытных данных исследовать свойства случайной величины $X$.

Решение. Исследование будем проводить поэтапно, параллельно с изложением теоретического материала. Начнем с первичной обработки выборочных данных. Построим вариационный ряд, соответствующий данной выборке:

0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 8.

Сгруппированный статистический ряд имеет вид:

| $\widetilde{x}_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $m_i$ | 4 | 24 | 19 | 23 | 10 | 13 | 3 | 2 | 2 |
| $m_i / n$ | 0,04 | 0,24 | 0,19 | 0,23 | 0,1 | 0,13 | 0,03 | 0,02 | 0,02 |



Столбцовая диаграмма, соответствующая этому ряду, приведена на рисунке 2.1.

**Пример 2.2.** При исследовании параметров работы сортировочной станции произведены измерения случайной величины $X$, характеризующей промежуток времени между моментами прибытия грузовых поездов на данную станцию (в минутах):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 642,23 | 80,36 | 571,72 | 368,41 | 568,65 | 300,55 | 573,41 | 8,17 | 405,89 | 87,54 |
| 251,84 | 41,71 | 440,87 | 20,48 | 308,45 | 135,90 | 143,06 | 467,31 | 49,82 | 241,62 |
| 598,16 | 7,91 | 98,70 | 92,84 | 5,49 | 285,89 | 349,71 | 91,76 | 158,75 | 124,04 |
| 53,40 | 155,96 | 12,06 | 22,77 | 369,03 | 50,14 | 61,02 | 42,44 | 184,26 | 183,59 |
| 84,38 | 402,74 | 10,39 | 223,17 | 130,14 | 57,61 | 99,65 | 257,59 | 505,05 | 282,39. |

Требуется на основании полученных экспериментальных данных изучить свойства данной случайной величины .

Решение. Построим вариационный ряд:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5,49 | 7,91 | 8,17 | 10,39 | 12,06 | 20,48 | 22,77 | 41,71 | 42,44 | 49,82 |
| 50,14 | 53,40 | 57,61 | 61,02 | 80,36 | 84,38 | 87,54 | 91,76 | 92,84 | 98,70 |
| 99,65 | 124,04 | 130,14 | 135,90 | 143,06 | 155,96 | 158,75 | 183,59 | 184,26 | 223,17 |
| 241,62 | 251,84 | 257,59 | 282,39 | 285,89 | 300,55 | 308,45 | 349,71 | 368,41 | 369,03 |
| 402,74 | 405,89 | 440,87 | 467,31 | 505,05 | 568,65 | 571,72 | 573,41 | 598,16 | 642,23 |

Для записи интервального статистического ряда вычислим границы интервалов разбиения: $h = \frac{R}{n} = \frac{x_{\max} - x_{\min}}{1 + 3,322 \lg n} = \frac{642,23 - 5,49}{1 + 3,322 \lg 50} \approx 95,84$;

$$C_1 = x_{\min} - h/2 = 5,49 - 95,84/2 = -42,43;$$
$$C_2 = C_1 + h = -42,43 + 95,84 = 53,41;$$
$$C_3 = C_2 + h = 53,41 + 95,84 = 149,25;$$
$$C_4 = C_3 + h = 149,25 + 95,84 = 245,09;$$
$$C_5 = C_4 + h = 45,09 + 95,84 = 340,93;$$
$$C_6 = C_5 + h = 340,93 + 95,84 = 436,77;$$
$$C_7 = C_6 + h = 436,77 + 95,84 = 532,61;$$
$$C_8 = C_7 + h = 532,61 + 95,84 = 628,45;$$
$$C_9 = C_8 + h = 628,45 + 95,84 = 724,29.$$

Интервальный статистический ряд имеет вид:

| $[C_i, C_{i+1})$ | [–42,43; 53,41) | [53,41; 149,25) | [149,25; 245,09) | [245,09; 340,93) | [340,93; 436,77) | [436,77; 532,61) | [532,61; 628,45) | [628,45; 724,29) |
|---|---|---|---|---|---|---|---|---|
| $\widetilde{x}_i$ | 5,49 | 101,33 | 197,17 | 293,01 | 388,85 | 484,69 | 580,53 | 676,37 |
| $m_i$ | 12 | 13 | 6 | 6 | 5 | 3 | 4 | 1 |
| $m_i / n$ | 0,24 | 0,26 | 0,12 | 0,12 | 0,1 | 0,06 | 0,08 | 0,02 |

Для контроля убедимся, что $\sum_{i=1}^{8} m_i = n = 50$, $\sum_{i=1}^{8} m_i / n = 1$.

На рисунке 2.2 приведена гистограмма относительных частот, соответствующая этому интервальному статистическому ряду.

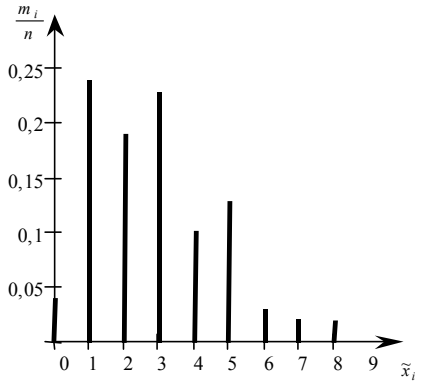



Рисунок 2.1 – Столбцовая диаграмма (пример 2.1)

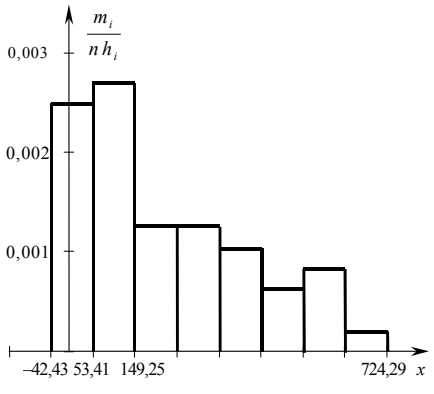



Рисунок 2.2 – Гистограмма относительных частот (пример 2.2)

**Замечание** – Графическое изображение статистического закона распределения является одним из основных аргументов при выдвижении гипотезы о виде закона распределения изучаемой случайной величины. Например, по виду столбцовой диаграммы, изображенной на рисунке 2.1, можно выдвинуть гипотезу о том, что исследуемая случайная величина $X$ – число составов, прибывающих на железнодорожную станцию в течение часа, подчиняется закону распределения Пуассона. Вид гистограммы, изображенной на рисунке 2.2, напоминает кривую экспоненциального распределения, что дает основания для выдвижения гипотезы о том, что случайная величина $X$, характеризующая промежуток времени между моментами прибытия грузовых поездов на сортировочную станцию, распределена по экспоненциальному закону. По виду гистограмм, изображенных на рисунке 2.3, *а* и *б*, можно предположить, что исследуемые случайные величины подчиняются, соответственно, равномерному и нормальному законам распределения вероятностей.

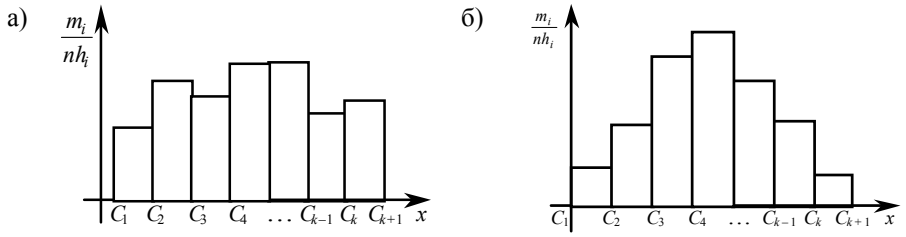



Рисунок 2.3 – Примеры гистограмм относительных частот

### 2.3 Эмпирическая функция распределения

Напомним, что универсальным способом задания закона распределения дискретных и непрерывных случайных величин является использование



*функции распределения F(x)*, определяющей для каждого значения $x \in R$ вероятность события $X < x$.

**Эмпирической функцией распределения $\hat{F}(x)$** * называется функция, которая каждому значению $x \in R$ ставит в соответствие относительную частоту события $X < x$,

$$\hat{F}(x) = \frac{n_x}{n},$$

где $n$ – объем исследуемой выборки;

$n_x$ – число элементов выборки, меньших данного фиксированного значения $x$.

Для вычисления эмпирической функции распределения по данным сгруппированного или интервального статистического рядов можно использовать соотношение

$$\hat{F}(x) = \sum_{\tilde{x}_i < x} \frac{m_i}{n}. \qquad (2.1)$$

Из определения эмпирической функции распределения следует, что она обладает всеми свойствами «теоретической» функции распределения $F(x)$:

1 Все возможные значения эмпирической функции распределения принадлежат отрезку [0; 1]: $0 \le \hat{F}(x) \le 1$.

2 $\hat{F}(x)$ – неубывающая функция своего аргумента, то есть $\hat{F}(x_1) \le \hat{F}(x_2)$ для любых значений $x_1$ и $x_2$, таких, что $x_1 < x_2$.

3 Если все выборочные значения исследуемой случайной величины принадлежат отрезку [$a$; $b$], то при $x \le a$ $\hat{F}(x) = 0$, при $x > b$ $\hat{F}(x) = 1$.

Важнейшее свойство эмпирической функции распределения состоит в том, что при увеличении объема выборки $n$, значение этой функции в каждой точке приближается к значению функции распределения $F(x)$ в той же точке. То есть эмпирическая функция распределения является экспериментальным аналогом (оценкой) неизвестной исследователю функции распределения $F(x)$.

**Пример 2.1.** (продолжение). Построим эмпирическую функцию распределения изучаемой дискретной случайной величины, используя соотношение (2.1):

При $x \le 0$ $\hat{F}(x) = 0$, поскольку нет выборочных значений случайной величины $X$, меньших рассматриваемых значений $x$, а значит, нет ни одного слагаемого в сумме (2.1).

При $0 < x \le 1$ меньше рассматриваемых значений $x$ только одно значение: $X = 0$.

* В математической статистике для обозначения выборочных аналогов рассматриваемых величин часто используются такие же обозначения, как и в теории вероятностей, но для указания их экспериментального «происхождения» они снабжаются значком «^» сверху.

По данным сгруппированного статистического ряда (см. пример 2.1), относительная частота значения $X = 0$ равна 0,04. Следовательно, $\hat{F}(x) = 0{,}04$ .

При $1 < x \le 2$ меньше рассматриваемых значений аргумента $x$ выборочные значения $X = 0$ и $X = 1$. Суммируя относительные частоты этих значений, получим: $\hat{F}(x) = 0{,}04 + 0{,}24 = 0{,}28$ и т.д.:

при $2 < x \le 3$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 = 0{,}47$ ;

$3 < x \le 4$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 = 0{,}7$ ;

$4 < x \le 5$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 + 0{,}1 = 0{,}8$ ;

$5 < x \le 6$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 + 0{,}1 + 0{,}13 = 0{,}93$ ;

$6 < x \le 7$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 + 0{,}1 + 0{,}13 + 0{,}03 = 0{,}96$ ;

$7 < x \le 8$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 + 0{,}1 + 0{,}13 + 0{,}03 + 0{,}02 = 0{,}98$ ;

$x > 8$: $\hat{F}(x) = 0{,}04 + 0{,}24 + 0{,}19 + 0{,}23 + 0{,}1 + 0{,}13 + 0{,}03 + 0{,}02 + 0{,}02 = 1$ .

Таким образом,

$$\hat{F}(x) = \begin{cases} 0, & x \le 0; \\ 0{,}04, & 0 < x \le 1; \\ 0{,}28, & 1 < x \le 2; \\ 0{,}47, & 2 < x \le 3; \\ 0{,}7, & 3 < x \le 4; \\ 0{,}8, & 4 < x \le 5; \\ 0{,}93, & 5 < x \le 6; \\ 0{,}96, & 6 < x \le 7; \\ 0{,}98, & 7 < x \le 8; \\ 1, & x > 8. \end{cases}$$

График функции $\hat{F}(x)$ приведён на рисунке 2.4.

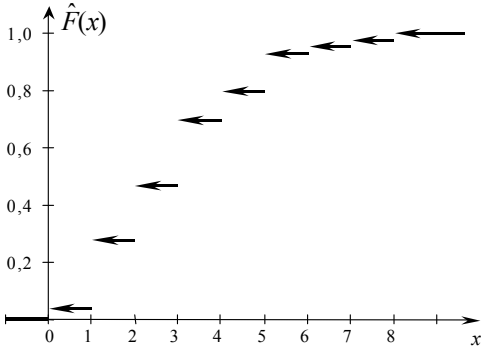



Рисунок 2.4 – График эмпирической функции распределения (пример 2.1)



**Пример 2.2.** (продолжение). Для приближённого построения эмпирической функции распределения воспользуемся соотношением: $\hat{F}(x) = \sum_{\tilde{x}_i < x} \frac{m_i}{n}$ *,

$$\hat{F}(x) = \begin{cases} 0, & x \le 5{,}49; \\ 0{,}24, & 5{,}49 < x \le 101{,}33; \\ 0{,}5, & 101{,}33 < x \le 197{,}17; \\ 0{,}62, & 197{,}17 < x \le 293{,}01; \\ 0{,}74, & 293{,}01 < x \le 388{,}85; \\ 0{,}84, & 388{,}85 < x \le 484{,}69; \\ 0{,}9, & 484{,}69 < x \le 580{,}53; \\ 0{,}98, & 31{,}257 < x \le 676{,}33; \\ 1 & x > 676{,}33. \end{cases}$$

График вычисленной эмпирической функции распределения приведен на рисунке 2.5.

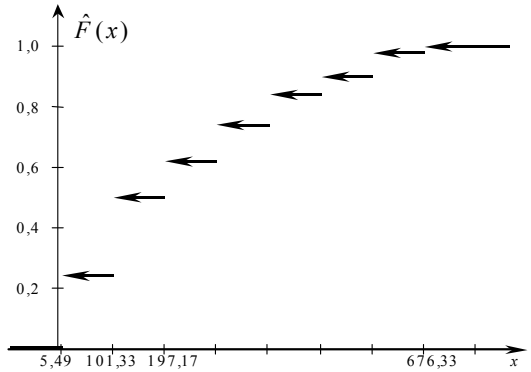



Рисунок 2.5 – Эмпирическая функция распределения (пример 2.2)

## 3 СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ

### 3.1 Основные понятия. Свойства точечных оценок

Любая величина $\hat{\Theta}$, представляющая собой функцию элементов выборки $\hat{\Theta} = f(X_1, X_2, \ldots, X_n)$, называется ***выборочной статистикой*** или просто ***статистикой***. Статистика $\hat{\Theta}$, используемая в качестве приближённого

* При использовании этой формулы принимаем допущение о том, что исследуемая случайная величина принимает только значения, соответствующие серединам интервалов $[C_i; C_{i+1})$ с частотами, равными $m_i$ ($i = 1, 2, \ldots, k$).

значения неизвестного параметра Θ, называется ***статистической оценкой параметра*** Θ.

Различают два вида статистических оценок: точечные и интервальные. ***Точечные оценки*** позволяют определить точку $\hat{\Theta}$, являющуюся некоторым приближением оцениваемого параметра Θ. ***Интервальная оценка*** представляет собой интервал $(\hat{\Theta}_1, \hat{\Theta}_2)$, который с заданной (как правило, близкой к единице) вероятностью накрывает неизвестное исследователю значение параметра Θ.

Познакомимся сначала с точечными оценками. Поскольку любая выборка является конечной и случайной, все выборочные функции $\hat{\Theta} = f(X_1, X_2, ..., X_n)$ являются случайными величинами, то есть при переходе от одной выборки к другой вычисленные значения оценки $\hat{\Theta}$ будут несколько отличаться друг от друга. При этом желательно, чтобы получаемые значения $\hat{\Theta}$ располагались как можно ближе к оцениваемому значению Θ. Это достигается в тех случаях, когда статистическая оценка $\hat{\Theta} = f(X_1, X_2, ..., X_n)$ обладает такими свойствами точечных оценок как состоятельность, несмещённость и эффективность.

Статистическая оценка называется **состоятельной**, если ее вычисляемое по опытным данным значение при увеличении объема выборки сходится по вероятности к истинному значению оцениваемого параметра, то есть, для любого, сколь угодно малого $\varepsilon > 0$ $\lim\limits_{n\to\infty} P(|\Theta - \hat{\Theta}| < \varepsilon) = 1$ (то есть при неограниченном увеличении объема выборки с вероятностью, равной 1, оценка $\hat{\Theta}$ отклонится от оцениваемого значения $\Theta$ не более, чем на малую величину $\varepsilon$).

Оценка называется **несмещенной** (или **оценкой без систематической ошибки**), если ее математическое ожидание совпадает со значением оцениваемого параметра: $M[\hat{\Theta}] = \Theta$.

Несмещенная оценка называется **эффективной**, если по сравнению с другими оценками этого параметра, вычисляемыми на основании выборок одинакового объема *n*, данная оценка обладает наименьшей дисперсией.

### 3.2 Точечные оценки числовых характеристик

Вычисление по выборочным данным оценок числовых характеристик изучаемой случайной величины является очень важным и, в большинстве случаев, необходимым этапом статистического исследования. Полученные оценки позволяют в количественной форме описать характерные черты статистического распределения и являются базой для построения математической модели исследуемого случайного явления.



В частности, для большинства используемых на практике теоретических распределений оценки параметров могут быть однозначно определены по известным оценкам числовых характеристик (для рассматриваемых в пособии законов распределения такие соотношения приведены в графе «Статистическое оценивание параметров» таблиц 1.2 и 1.3). Таким образом, задача нахождения оценок параметров распределения может быть сведена к задаче вычисления по выборочным данным оценок числовых характеристик исследуемой случайной величины.

В этом подразделе приведены расчетные формулы для вычисления точечных оценок числовых характеристик и некоторые свойства этих оценок.

В качестве оценки математического ожидания используется среднее арифметическое $\bar{x}$ опытных значений. Эта статистика называется **выборочным средним**

$$\hat{M}[X] = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Доказано, что $\bar{x}$ является несмещенной и состоятельной оценкой математического ожидания исследуемой случайной величины. Если случайная величина *X* распределена по нормальному закону, то $\bar{x}$ является и эффективной оценкой математического ожидания.

Для оценивания по выборочным данным моды распределения, используется то значение *сгруппированного статистического ряда* $\hat{x}_{\text{mod}}$, которому соответствует наибольшее значение частоты. По *интервальному статистическому ряду* определяется модальный интервал, в который попало наибольшее число выборочных значений, и в качестве точечной оценки моды может использоваться среднее значение этого интервала.

Для определения выборочного значения медианы используется вариационный ряд. В качестве оценки медианы $\hat{x}_{\text{med}}$ принимают средний (то есть $\frac{1}{2}(n+1)$-й) член этого ряда, если значение *n* – нечётно и среднее арифметическое между двумя средними (то есть между $\frac{1}{2}n$-м и $\left(\frac{1}{2}n+1\right)$-м) членами этого ряда, если *n* – чётно.

В качестве оценки дисперсии можно использовать статистику $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ и соответствующую оценку среднего квадратического отклонения

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

Исследования показали, что эта выборочная статистика является состоятельной, но смещенной оценкой дисперсии. Для устранения выявленного смещения, ее домножают на «исправляющий коэффициент», равный $\frac{n}{n-1}$, получая таким образом «исправленную» оценку дисперсии:

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

Очевидно, что при больших значениях $n$ различие между оценками $s^2$ и $\bar{s}^2$ не значительно, но при исследовании выборок небольшого объема в качестве оценки дисперсии принято использовать ее *несмещенную оценку*:

$$\hat{D}[X] = \bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ,$$

и соответствующую оценку для среднего квадратического отклонения:

$$\hat{\sigma}[X] = \bar{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \ .$$

В качестве оценок коэффициентов асимметрии и эксцесса используются следующие выборочные статистики:

$$A[X] = \beta_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\bar{s}^3};$$

$$Ex[X] = \beta_2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\bar{s}^4} - 3 \ .$$

**Замечание** – Если выборочные данные представлены в виде сгруппированного статистического ряда, то для вычисления оценок числовых характеристик удобно использовать приведенные ниже формулы:

$$\hat{M}[X] = \bar{x} = \frac{1}{n}\sum_{i=1}^{k}\tilde{x}_i m_i \ ;$$

$$\hat{D}[X] = \bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(\tilde{x}_i - \bar{x})^2 m_i \ ;$$

$$\hat{\sigma}[X] = \bar{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2 m_i} \ ;$$



$$\hat{A}[X] = \hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^3 m_i}{\bar{s}^3} \ .$$

$$\hat{E}x[X] = \hat{\beta}_2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^4 m_i}{\bar{s}^4} - 3 \ .$$

**Пример 3.1.** Вычислим оценки числовых характеристик случайной величины $X$, характеризующей число составов, прибывающих на сортировочную станцию в течение часа (см. пример 2.1).

Оценка математического ожидания

$$\hat{M}[X] = \sum_{i=1}^{100} x_i = \bar{x} = \frac{1}{n}\sum_{i=1}^{7}\tilde{x}_i m_i =$$

$$= \frac{1}{100}(0\cdot 4 + 1\cdot 24 + 2\cdot 19 + 3\cdot 23 + 4\cdot 10 + 5\cdot 13 + 6\cdot 3 + 7\cdot 2 + 8\cdot 2) = 2{,}84 \ \text{(состава)}.$$

В качестве оценки моды данной случайной величины можно принять значение 1: $\hat{x}_{\text{mod}} = 1$ (состав), так как этому значению соответствует наибольшее значение частоты.

Оценка дисперсии

$$\hat{D}[X] = \frac{1}{n-1}\sum_{i=1}^{7}(x_i - \bar{x})^2 m_i = \frac{1}{99}((0-2{,}84)^2\cdot 4 + (1-2{,}84)^2\cdot 24 + (2-2{,}84)^2\cdot 19 +$$

$$+ (3-2{,}84)^2\cdot 23 + (4-2{,}84)^2\cdot 10 + (5-2{,}84)^2\cdot 13 + (6-2{,}84)^2\cdot 3 + (7-2{,}84)^2\cdot 2 +$$

$$+ (8-2{,}84)^2\cdot 2) = \frac{319{,}44}{99} = 3{,}226667 \ \text{(составов}^2\text{)}.$$

Оценка среднего квадратического отклонения

$$\sigma[X] = \sqrt{D[X]} = \sqrt{3{,}226667} = 1{,}796292 \ \text{(состава)}.$$

**Пример 3.2.** Вычислим оценки числовых характеристик случайной величины $X$, описывающей промежуток времени между моментами прибытия товарных составов на сортировочную станцию (см. пример 2.2).

Оценка математического ожидания

$$\hat{M}[X] = \bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{1}{50}(642{,}23 + 80{,}36 + \cdots + 283{,}39) = 214{,}17 \ \text{(мин)}.$$

В качестве оценки моды примем среднее значение модального интервала (53,41; 149,25): $\hat{x}_{\text{mod}} = 101{,}33$ (мин).

Оценка медианы $\hat{x}_{\text{med}} = \frac{x_{(25)} + x_{(26)}}{2} = \frac{143{,}06 + 155{,}96}{2} = 149{,}51$ (мин).

Оценка дисперсии $\hat{D}[X] = \frac{1}{49}\sum_{i=1}^{50}(x_i - \bar{x})^2 = 34606{,}7$ (мин$^2$).

Оценка среднего квадратического отклонения $\hat{\sigma}[X] = \sqrt{\hat{D}[X]} = 186{,}029$ (мин).

### 3.3 Понятие об интервальном оценивании

Заменяя при проведении статистического исследования неизвестное значение параметра Θ его точечной оценкой $\hat{\Theta}$, мы отдаем себе отчет в том, что при этом совершаем некоторую ошибку. Большое практическое значение имеет информация о величине этой ошибки. Другими словами, возникает вопрос об определении **точности** оценки $\hat{\Theta}$, то есть о таком значении ε, что $|\Theta - \hat{\Theta}| < \varepsilon$.

Поскольку в нашем распоряжении имеются лишь выборочные данные, то можно определить только вероятность ***P*** осуществления этого неравенства, которая называется **доверительной вероятностью**:

$$P\left(\left|\Theta - \hat{\Theta}\right| < \varepsilon\right) = \boldsymbol{P}. \qquad (3.1)$$

Соотношение (3.1) может быть записано следующим образом:

$$P(\hat{\Theta}_1 < \Theta < \hat{\Theta}_2) = \boldsymbol{P}.$$

Это означает, что неизвестное значение оцениваемого параметра Θ с доверительной вероятностью ***P*** будет накрыто интервалом $(\hat{\Theta}_1; \hat{\Theta}_2)$, который называется **доверительным интервалом** (или **интервальной оценкой параметра Θ**).

Значение доверительной вероятности выбирается исходя из целей исследования и ответственности при принятии решения в конкретной задаче. Обычно доверительная вероятность ***P*** принимается равной 0,95, 0,99, реже – 0,999.

В литературе часто используется еще одно обозначение доверительной вероятности $\boldsymbol{P} = 1 - \alpha$, где α – некоторое малое число (например, $\alpha = 0{,}05$, или $\alpha = 0{,}01$), задающее вероятность того, что оцениваемый параметр окажется за пределами доверительного интервала. Это означает, что при извлечении большого числа выборок объема *n* из одной и той же генеральной совокупности, только в $\alpha \cdot 100$ % случаев построенный по выборочным данным доверительный интервал не будет накрывать значение оцениваемого параметра Θ.

Подчеркнем еще раз, что границы доверительного интервала являются случайными величинами (так как они определяются на основании выборочных данных). Именно поэтому мы можем говорить только о *вероятности* накрыть доверительным интервалом некоторую (неслучайную!) точку Θ. Ширина доверительного интервала существенно зависит от объема выборки *n* (уменьшается с ростом *n*) и от величины



доверительной вероятности (увеличивается с приближением ***P*** к единице).

В общем случае, задача построения доверительных интервалов является сложной математической задачей, допускающей сравнительно простое аналитическое решение лишь для некоторых частных случаев, подобных рассмотренным ниже.

### 3.4 Построение доверительных интервалов для математического ожидания и среднего квадратического отклонения нормально распределенной случайной величины

Пусть на основании выборочных данных ($x_1$, $x_2$, …, $x_n$), полученных при исследовании нормально распределенной случайной величины *X*, вычислены точечные оценки математического ожидания и среднего квадратического отклонения этой случайной величины:

$$\hat{M}[X] = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \qquad \hat{\sigma}[X] = \bar{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \; .$$

Ставится задача определения на основании имеющихся опытных данных интервальных оценок параметров $m = M[X]$ и $\sigma = \sigma[X]$ изучаемой величины.

**Построение доверительного интервала для математического ожидания случайной величины, распределенной по нормальному закону**, основано на том факте, что распределение выборочной статистики $t = \frac{\bar{x} - m}{\bar{s}}\sqrt{n}$ не зависит от значений $\bar{x}$, *m*, $\bar{s}$ и подчиняется закону распределения Стьюдента (см. подразд. 1.4.2) с $\nu = n - 1$ степенями свободы.

Это дает возможность для заданного значения доверительной вероятности $\boldsymbol{P} = 1 - \alpha$ и числа степеней свободы ν определить такие значения $t_1$ и $t_2$, что

$$P(t_1 < t < t_2) = 1 - \alpha \, . \qquad (3.2)$$

В геометрической интерпретации эта вероятность численно равна площади фигуры, ограниченной кривой *t*-распределения и осью абсцисс, заключенной между значениями $t_1$ и $t_2$ (рисунок 3.1).

Как указывалось в подразд. 1.4.2, значения $t_1$ и $t_2$ определяются по таблице квантилей распределения Стьюдента: $t_1 = -t_{\alpha/2;\nu}$, $t_2 = t_{\alpha/2;\nu}$.

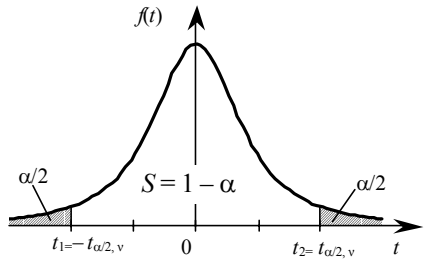



Рисунок 3.1 – К определению доверительного интервала для параметра $m$

Преобразуем соотношение (3.2):

$$P\left(-t_{\alpha/2;\nu} < \frac{\bar{x}-m}{\bar{s}}\sqrt{n} < t_{\alpha/2;\nu}\right) = 1-\alpha,$$

отсюда

$$P\left(\bar{x} - \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}} < m < \bar{x} + \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}\right) = 1-\alpha.$$

То есть интервал $I_m = \left(\bar{x} - \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}; \bar{x} + \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}\right)$ является $(1-\alpha)\cdot 100\%$-ным доверительным интервалом для неизвестного математического ожидания $m$ нормально распределенной случайной величины. Точность этой оценки $\varepsilon = \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}$. Легко заметить, что при увеличении объема выборки $n$ ширина доверительного интервала, равная $2\varepsilon$, будет уменьшаться.

**Замечание 1** – Как известно, при больших значениях $\nu = n-1$ (уже при $\nu > 30$) распределение Стьюдента приближается к нормальному распределению. В этом случае, при построении доверительного интервала для неизвестного математического ожидания вместо распределения Стьюдента можно приближенно использовать стандартизованное нормальное распределение. Соответствующий доверительный интервал будет иметь вид

$$I_m = \left(\bar{x} - \frac{u_{\alpha/2}}{\sqrt{n}}\bar{s}; \bar{x} + \frac{u_{\alpha/2}}{\sqrt{n}}\bar{s}\right), \qquad (3.3)$$

где $u_{\alpha/2}$ – квантиль стандартизованного нормального распределения. Для наиболее часто используемых значений доверительной вероятности $\boldsymbol{P} = 1-\alpha$ значения $u_{\alpha/2}$ приведены в таблице 3.1.



*Таблица 3.1* – **Значения квантилей стандартизованного нормального распределения**

| Доверительная вероятность $P = 1-\alpha$ | Значение $\alpha$ | Значение квантиля $u_{\alpha/2}$ |
|---|---|---|
| 0,9 | 0,1 | 1,645 |
| 0,95 | 0,05 | 1,96 |
| 0,99 | 0,01 | 2,58 |
| 0,999 | 0,001 | 3,28 |

**Замечание 2** – Если случайная величина $X$ имеет произвольный закон распределения вероятностей, то при достаточно большом объеме выборки соотношение (3.3) можно использовать для построения приближенного доверительного интервала для математического ожидания этой случайной величины.

**Построение доверительного интервала для среднего квадратического отклонения нормально распределенной случайной величины** базируется на использовании статистики $\chi^2 = \frac{n-1}{\sigma^2}\bar{s}^2$. Доказано, что распределение этой величины не зависит от значений $\sigma^2$ и $\bar{s}^2$ и подчиняется закону распределения $\chi^2$ с $\nu = n-1$ степенями свободы (см. подразд. 1.4.1).

Это позволяет определить такие значения $\chi_1^2$ и $\chi_2^2$, для которых выполняется соотношение

$$P(\chi_1^2 < \chi^2 < \chi_2^2) = 1-\alpha, \qquad (3.4)$$

где $\boldsymbol{P} = 1-\alpha$ – заданный уровень доверительной вероятности.

Поскольку распределение $\chi^2$ не симметрично, для определения значений $\chi_1^2$ и $\chi_2^2$ обычно используются дополнительные условия: $P(\chi^2 < \chi_1^2) = \alpha/2$, $P(\chi^2 > \chi_2^2) = \alpha/2$ (на рисунке 3.2 штриховкой выделены фигуры, площади которых равны указанным вероятностям).

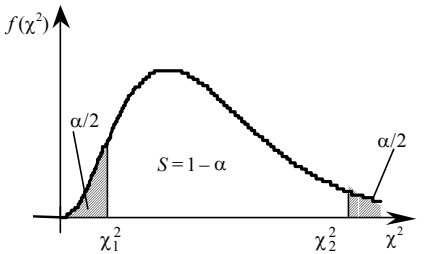



Рисунок 3.2 – К определению доверительного интервала для параметра σ

Полагая значения $\chi_1^2 = \chi_{1-\alpha/2;\nu}^2$ и $\chi_2^2 = \chi_{\alpha/2;\nu}^2$ известными, преобразуем формулу (3.4): $P\left(\chi_1^2 < \frac{n-1}{\sigma^2}\bar{s}^2 < \chi_2^2\right) = 1-\alpha$.

Отсюда $P\left(\bar{s}\sqrt{\frac{n-1}{\chi_2^2}} < \sigma < \bar{s}\sqrt{\frac{n-1}{\chi_1^2}}\right) = 1-\alpha$,

или иначе: $P(\bar{s}\gamma_1 < \sigma < \bar{s}\gamma_2) = 1-\alpha$, где $\gamma_1 = \sqrt{\frac{n-1}{\chi_2^2}}$, $\gamma_2 = \sqrt{\frac{n-1}{\chi_1^2}}$. То есть интервал $I_\sigma = (\bar{s}\gamma_1; \bar{s}\gamma_2)$ является $(1-\alpha)\cdot 100\,\%$-ным доверительным интервалом для среднего квадратического отклонения нормально распределенной случайной величины *X*.

Значения коэффициентов $\gamma_1$ и $\gamma_2$ в зависимости от доверительной вероятности $\boldsymbol{P} = 1-\alpha$ и числа степеней свободы $\nu = n-1$ можно определить по таблице, приведенной в приложении Г.

**Пример 3.3.** В результате обследования партии, состоящей из 10 деталей, изготовленных на определённом станке, получены оценки математического ожидания и среднего квадратического отклонения случайной величины *X*, характеризующей внутренний диаметр исследуемых деталей: $\bar{x} = 10{,}1$ мм, $\bar{s} = 0{,}4$ мм. Предполагая, что данная случайная величина подчиняется нормальному закону распределения, построить 95%-ный и 99%-ный доверительные интервалы для неизвестных значений математического ожидания и среднего квадратического отклонения изучаемой случайной величины *X*.

Решение. Имеем $n = 10$; $\bar{x} = 10{,}1$; $\bar{s} = 0{,}4$.

При построении доверительных интервалов для математического ожидания воспользуемся соотношением

$$I_m = (\bar{x} - \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}; \bar{x} + \frac{t_{\alpha/2;\nu}\bar{s}}{\sqrt{n}}).$$

По таблице квантилей распределения Стьюдента (см. приложение Б) для $\nu = n-1 = 9$ определяем критические значения $t_{\alpha/2;\nu}$:

при $\alpha = 0{,}05$ $t_{0{,}025;9} = 2{,}262$;
при $\alpha = 0{,}01$ $t_{0{,}005;9} = 3{,}25$.

Таким образом, 95%-ный доверительный интервал для математического ожидания имеет вид:

$$I_m = \left(10{,}1 - \frac{2{,}262\cdot 0{,}4}{\sqrt{10}};\ 10{,}1 + \frac{2{,}262\cdot 0{,}4}{\sqrt{10}}\right) = (9{,}8139; 10{,}3861);\ (\varepsilon = 0{,}2861).$$

99%-ный доверительный интервал для математического ожидания

$$I_m = \left(10{,}1 - \frac{3{,}25\cdot 0{,}4}{\sqrt{10}};\ 10{,}1 + \frac{3{,}25\cdot 0{,}4}{\sqrt{10}}\right) = (9{,}689; 10{,}511);\ (\varepsilon = 0{,}411).$$

При построении доверительных интервалов для среднего квадратического отклонения используем соотношение $I_\sigma = (\bar{s}\gamma_1; \bar{s}\gamma_2)$.



Значения $\gamma_1$ и $\gamma_2$, соответствующие числу степеней свободы $\nu = n-1 = 9$, определяем с помощью приложения Г.

В случае $p = 0{,}95$ $\gamma_1 = 0{,}688$, $\gamma_2 = 1{,}826$;
при $p = 0{,}99$ $\gamma_1 = 0{,}618$, $\gamma_2 = 2{,}277$.

Получаем 95%-ный доверительный интервал для среднего квадратического отклонения:

$$I_\sigma = (0{,}4\cdot 0{,}688;\ 0{,}4\cdot 1{,}826) = (0{,}2752;\ 0{,}7304);\ (\varepsilon = 0{,}2276).$$

99%-ный доверительный интервал

$$I_\sigma = (0{,}4\cdot 0{,}618;\ 0{,}4\cdot 2{,}277) = (0{,}2472;\ 0{,}9108);\ (\varepsilon = 0{,}3318).$$

Сопоставляя полученные результаты, легко заметить, что, как и следовало ожидать, увеличение значения доверительной вероятности приводит к расширению доверительного интервала.

**Пример 3.4.** Используя данные примера 3.3, определить границы 95%-ных доверительных интервалов для математического ожидания и среднего квадратического отклонения исследуемой случайной величины, предполагая, что точечные оценки $\bar{x} = 10{,}1$ и $\bar{s} = 0{,}4$ получены в результате исследования выборки деталей объёма 100.

Решение. В данном случае имеем: $n = 100$; $\bar{x} = 10{,}1$; $\bar{s} = 0{,}4$. Поскольку используемое значение объёма выборки *n* достаточно велико ($n > 30$), при построении интервальной оценки для математического ожидания можно воспользоваться соотношением (3.3):

$$I_m = \left(\bar{x} - \frac{u_{\alpha/2}\bar{s}}{\sqrt{n}}; \bar{x} + \frac{u_{\alpha/2}\bar{s}}{\sqrt{n}}\right).$$

При $p = 0{,}95$ значение квантиля стандартизованного нормального распределения $u_{\alpha/2}$ определим с помощью таблицы 3.1: $u_{\alpha/2} = 1{,}96$.

Таким образом, 95%-ный доверительный интервал для математического ожидания

$$I_m = \left(10{,}1 - \frac{1{,}96\cdot 0{,}4}{\sqrt{100}}; 10{,}1 + \frac{1{,}96\cdot 0{,}4}{\sqrt{100}}\right) = (10{,}0216; 10{,}1784);\ (\varepsilon = 0{,}0784).$$

При построении интервальной оценки для среднего квадратического отклонения так же, как и при решении примера 3.3, используем соотношение $I_\sigma = (\bar{s}\gamma_1; \bar{s}\gamma_2)$.

По приложению Г определяем для $p = 0{,}95$ и $\nu = n-1 = 100-1 = 99$: $\gamma_1 = 0{,}878$; $\gamma_2 = 1{,}161$. Отсюда

$$I_\sigma = (0{,}4\cdot 0{,}878;\ 0{,}4\cdot 1{,}161) = (0{,}3512;\ 0{,}4644);\ (\varepsilon = 0{,}0566).$$

Сопоставляя доверительные интервалы, построенные в примерах 3.3 и 3.4, необходимо отметить, что увеличение объёма выборочных данных позволяет значительно уменьшить ширину доверительного интервала, т.е. повысить точность интервальной оценки.

# 4 СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

## 4.1 Основные понятия теории статистической проверки гипотез

При проведении статистических исследований часто возникает необходимость проверки согласования с экспериментальными данными некоторых предположений о свойствах изучаемой случайной величины. Формулируемые предположения называются **гипотезами**, а методы, позволяющие решить поставленную задачу, составляют раздел математической статистики, который называется **теорией статистической проверки гипотез**.

Проверяемая гипотеза называется **основной** (или **нулевой**) и обозначается $H_0$. В случае отклонения гипотезы $H_0$ полагают, что справедлива некоторая **альтернативная** гипотеза, обозначаемая символом $H_a$ или $H_1$.

Принято различать параметрические и непараметрические гипотезы. *Непараметрические гипотезы* представляют собой утверждения о виде закона распределения исследуемой случайной величины. *В параметрических гипотезах* сформулированы предположения о значениях параметров функции распределения заданного вида.

Результат сопоставления гипотезы $H_0$ с опытными данными может быть либо отрицательным, либо неотрицательным. При получении отрицательного результата полагают, что наблюдаемое несоответствие гипотезы $H_0$ опытным данным вызвано ошибочностью этой гипотезы, и в этом случае проверяемая гипотеза отклоняется как не согласующаяся с результатами эксперимента. Получение неотрицательного результата, в общем случае, не является свидетельством истинности проверяемой гипотезы, а указывает лишь на приемлемое согласование сформулированного предположения с имеющимся экспериментальным материалом. Подобное согласование может наблюдаться при проверке двух и более взаимоисключающих гипотез на основании конкретного набора выборочных данных. Для получения более надёжных выводов, необходимо произвести исследование с использованием нескольких различных выборок, извлеченных из одной и той же генеральной совокупности.

Методика проверки согласования гипотез с опытными данными основана на использовании специальных выборочных статистик, называемых **статистическими критериями** (или просто **критериями**). Критерий конструируется таким образом, что позволяет оценить меру отклонения эмпирического распределения (то есть выборочных данных) от предполагаемого теоретического (то есть от проверяемой гипотезы). При этом множество возможных значений критерия разбивается на 2 непересекающиеся части: *критическую область* и *область допустимых значений*.

Критическая область определяется таким образом, что попадание значения критерия в эту область при справедливости проверяемой гипотезы является маловероятным событием (вероятность которого обозначается $\alpha$), и полагают, что осуществление этого события вызвано ошибочностью выдвинутой гипотезы. Поэтому при попадании расчетного значения критерия в критическую область, проверяемая гипотеза отвергается как не согласующаяся с опытными данными. Если значение критерия принадлежит области допустимых значений, то делают вывод об удовлетворительном согласовании выдвинутого предположения с выборочными данными и об отсутствии оснований для отклонения проверяемой гипотезы. Другими словами, в этом случае полагают, что наблюдаемое различие между эмпирическим и теоретическим распределениями может быть объяснено только случайным характером имеющихся выборочных данных.

## 4.2 Ошибки, допускаемые при проверке гипотез

Поскольку решение об отклонении или неотклонении проверяемой гипотезы принимается на основании выборочных данных, при этом всегда существует риск совершения ошибки. Допускаемые ошибки могут быть двух видов:

1) отклонение верной гипотезы $H_0$ (вероятность совершения этой ошибки обозначается $\alpha$ и называется *уровнем значимости критерия*);

2) принятие ложной гипотезы $H_0$ (вероятность этой ошибки обозначается $\beta$).

Схематически возможные ошибки и их вероятности удобно представить в виде следующей таблицы:

| Справедливость гипотезы $H_0$ | Принимаемое решение | |
|---|---|---|
| | Принять $H_0$ | Отвергнуть $H_0$ |
| $H_0$ верна | Верное решение с вероятностью $1 - \alpha$ | Ошибочное решение с вероятностью $\alpha$ |
| $H_0$ не верна | Ошибочное решение с вероятностью $\beta$ | Верное решение с вероятностью $1 - \beta$ |

Вопрос о том, какие значения вероятностей $\alpha$ и $\beta$ являются приемлемыми при решении поставленной задачи, решается исследователем в каждой конкретной ситуации исходя из практических целей.

В общем случае, одновременно, при фиксированном объеме выборки обеспечить достижение сколь угодно малых значений $\alpha$ и $\beta$ не представляется возможным. Действительно, для уменьшения значения $\alpha$ необходимо сужать размеры критической области, но при этом уменьшается вероятность попадания значения критерия в эту область и при справедливости какой-либо другой гипотезы $H_a$. Это означает, что, сужая критическую область, мы увеличиваем вероятность совершения ошибки второго рода $\beta$. Таким обра-

зом, ошибки первого и второго рода являются конкурирующими между собой и, чаще всего, единственным способом уменьшения вероятностей этих обеих ошибок является увеличение объема выборки, то есть использование дополнительной информации об исследуемом явлении.

В следующем подразделе приведена методика применения одного из так называемых *критериев согласия* – критериев, использующихся для проверки гипотезы о виде закона распределения изучаемой случайной величины. При применении критериев согласия заранее фиксируется значение $\alpha$ и проверяется согласование теоретического и эмпирического распределений при заданном уровне значимости $\alpha$.

### 4.3 Применение критерия Пирсона $\chi^2$ для проверки гипотезы о виде закона распределения случайной величины

Одним из наиболее широко используемых на практике критериев согласия является критерий $\chi^2$ Пирсона. Он может применяться для проверки гипотез о распределении исследуемой случайной величины по любому из известных законов распределения (как дискретных, так и непрерывных случайных величин).

Для применения этого критерия необходимо представление эмпирического (т.е. экспериментального) распределения в виде интервального или сгруппированного статистического ряда. Обозначим, как и ранее, число используемых разрядов разбиения $k$, а частоту $i$-го разряда – $m_i$ ($i = 1, 2, \ldots, k$).

При применении этого критерия на основании выдвинутой гипотезы вычисляются теоретические частоты $np_i$ попадания значений случайной величины $X$, распределенной по предполагаемому закону в частичные разряды разбиения. Критерий $\chi^2$ основан на сопоставлении эмпирических ($m_i$) и теоретических ($np_i$) частот попадания значений исследуемой величины в рассматриваемые интервалы. В качестве меры расхождения экспериментального и теоретического распределений используется статистика

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

которая при $n \to \infty$ независимо от вида предполагаемого распределения стремится к распределению $\chi^2$ с $\nu = k - r - 1$ степенями свободы. Здесь $r$ – число параметров гипотетического (т.е. предполагаемого) распределения, оцениваемых по выборочным данным.

Легко заметить, что при незначительных отклонениях значений $m_i$ от $np_i$ значение критерия $\chi^2$ будет близким к нулю. И наоборот, большое значение критерия $\chi^2$ свидетельствует о существенном отклонении экспериментально полученного распределения от предполагаемого. Поэтому критическая область в данном случае определяется условием: $\chi^2 > \chi^2_{\alpha,\nu}$ ( рисунок 4.1).



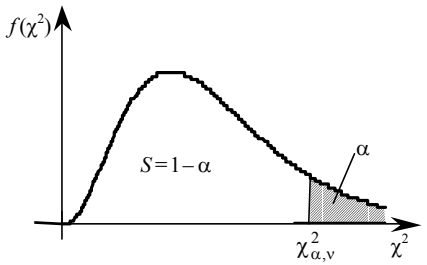



Рисунок 4.1 – Критическая область критерия $\chi^2$

Критическое значение $\chi^2_{\alpha,\nu}$ определяется по таблице квантилей распределения $\chi^2$ в зависимости от уровня значимости $\alpha$ и числа степеней свободы $\nu$.

**Алгоритм применения критерия $\chi^2$ для проверки гипотезы о виде закона распределения исследуемой случайной величины**

1 Выборочные данные представляют в виде сгруппированного или интервального статистического ряда.

2 Выбирают уровень значимости $\alpha$.

3 Формулируют гипотезу о виде закона распределения исследуемой случайной величины (при выдвижении гипотезы могут использоваться сведения о физической природе и механизме формирования значений этой случайной величины, вид графического изображения статистического распределения, значения оценок числовых характеристик).

4 На основании выдвинутой гипотезы вычисляют вероятности $p_i$ попадания значений случайной величины, распределенной по предполагаемому закону в рассматриваемые разряды разбиения (необходимые расчетные формулы приведены в таблицах 1.2 и 1.3).

**Замечание** – Если изучается непрерывная случайная величина, то при вычислении значений $p_i = P[C_i \le X < C_{i+1})$ необходимо изменить границы первого и последнего частичных разрядов разбиения таким образом, чтобы учесть все возможные значения, которые может принять данная случайная величина. В зависимости от вида проверяемой гипотезы границы частичных интервалов определяются следующим образом:

| Вид закона распределения | Первый интервал разбиения | Последний интервал разбиения |
|---|---|---|
| Равномерный | $[\hat{a}; C_2)$ | $\lfloor C_1; \hat{b} \rfloor$ |
| Экспоненциальный | $[0; C_2)$ | $[C_k; \infty)$ |
| Нормальный | $(-\infty; C_2)$ | $[C_k; \infty)$ |

5 Определяют значения теоретических частот $np_i$. Разряды разбиения,

характеризующиеся малыми значениями теоретических частот, объединяются с соседними, но с соблюдением условия $k \geq 5$.

6 Вычисляют наблюдаемое значение критерия $\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}$.

7 По таблицам квантилей распределения $\chi^2$ определяют критическое значение $\chi^2_{\alpha,\nu}$, соответствующее заданному уровню значимости $\alpha$ и числу степеней свободы $\nu = k - r - 1$. Сравнивая расчетное значение критерия $\chi^2$ с критическим значением $\chi^2_{\alpha,\nu}$, делают вывод об отклонении или принятии проверяемой гипотезы.

Если $\chi^2 > \chi^2_{\alpha,\nu}$, то выдвинутая гипотеза отклоняется как не согласующаяся с опытными данными. В случае $\chi^2 \leq \chi^2_{\alpha,\nu}$ делают вывод о том, что выборочные данные не дают оснований для отклонения этой гипотезы. Подчеркнем еще раз, что полученный результат свидетельствует лишь о приемлемом согласовании выдвинутой гипотезы с имеющимися выборочными данными и, в общем случае, не является доказательством ее истинности.

**Пример 4.1** (см. примеры 2.1 и 3.1 ).

Проверим с помощью критерия $\chi^2$ Пирсона гипотезу о том, что случайная величина *X*, характеризующая число составов, прибывающих на сортировочную станцию в течение часа, подчиняется закону распределения Пуассона.

Вычислим оценку параметра *a* распределения Пуассона:

$$\hat{a} = \hat{M}(X) = \bar{x} = 2{,}84 .$$

По формуле Пуассона вычислим вероятности возможных значений данной случайной величины:

$$P(X = m) = \frac{a^m}{m!} e^{-a} ;$$

$$P(X = 0) = \frac{2{,}84^0}{0!} e^{-2{,}84} = 0{,}0584 ; \; P(X = 1) = \frac{2{,}84^1}{1!} e^{-2{,}84} = 0{,}1659 ;$$

$$P(X = 2) = \frac{2{,}84^2}{2!} e^{-2{,}84} = 0{,}2356 \text{ и т. д.}$$

$$\sum_{i=0}^{8} P(X = i) = 0{,}9974 .$$

Следовательно, учитывая диапазон возможных значений случайной величины, распределённой по закону Пуассона, вероятность события $X > 8$ можно определить следующим образом:

$$P(X > 8) = 1 - P(X \leq 8) = 1 - \sum_{i=0}^{8} P(X = i) = 1 - 0{,}9974 = 0{,}0026 .$$



Занесём это значение в последний столбец расчётной таблицы:

| $\widetilde{x}_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | [9; ∞) |
|---|---|---|---|---|---|---|---|---|---|---|
| $m_i$ | 4 | 24 | 19 | 23 | 10 | 13 | 3 | 2 | 2 | 0 |
| $p_i$ | 0,0584 | 0,1659 | 0,2356 | 0,2231 | 0,1584 | 0,090 | 0,0426 | 0,0173 | 0,0061 | 0,0026 |
| $np_i$ | 5,84 | 16,59 | 23,56 | 22,31 | 15,84 | 9,0 | 4,26 | 1,73 | 0,61 | 0,26 |

Поскольку значения $np_i$ в четырёх последних разрядах разбиения не превышают пяти единиц, объединим эти разряды в один:

| $\widetilde{x}_i$ | 0 | 1 | 2 | 3 | 4 | 5 | [6; ∞) |
|---|---|---|---|---|---|---|---|
| $m_i$ | 4 | 24 | 19 | 23 | 10 | 13 | 7 |
| $p_i$ | 0,0584 | 0,1659 | 0,2356 | 0,2231 | 0,1584 | 0,090 | 0,0686 |
| $np_i$ | 5,84 | 16,59 | 23,56 | 22,61 | 15,84 | 9,0 | 6,86 |

Вычислим значение критерия:

$$\chi^2 = \sum_{i=1}^{7} \frac{(m_i - np_i)^2}{np_i} = \frac{(4 - 5{,}84)^2}{5{,}84} + \frac{(24 - 16{,}59)^2}{16{,}59} + \frac{(19 - 23{,}56)^2}{23{,}56} + \frac{(23 - 22{,}31)^2}{22{,}31} +$$
$$+ \frac{(10 - 15{,}84)^2}{15{,}84} + \frac{(13 - 9)^2}{9} + \frac{(7 - 6{,}86)^2}{6{,}86} \approx 9{,}23.$$

По таблицам квантилей распределение $\chi^2$ (см. приложение А) определим критическое значение $\chi^2_{\alpha,\nu}$, соответствующее уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = k - r - 1 = 7 - 1 - 1 = 5$: $\chi^2_{0{,}05;5} = 11{,}070$.

Поскольку $\chi^2 \leq \chi^2_{\alpha,\nu}$, можно сделать вывод о том, что гипотеза о распределении изучаемой случайной величины по закону Пуассона не противоречит выборочным данным.

**Пример 4.2** (см. примеры 2.2 и 3.2).

Проверим с помощью критерия $\chi^2$ Пирсона гипотезу о том, что случайная величина *X*, характеризующая промежуток времени между моментами прибытия товарных составов на сортировочную станцию подчиняется экспоненциальному закону распределения. Эта гипотеза была выдвинута на основании вида полученной гистограммы с учётом диапазона возможных значений этой величины ( $x \in [0; \infty)$ ) и сведений о физическом смысле полученных значений (промежуток времени между моментами наступления событий простейшего потока обычно описывается показательно распределённой случайной величиной).

Вычислим оценку параметра экспоненциального закона распределения:

$$\hat{\lambda} = \frac{1}{\hat{M}(x)} = \frac{1}{214{,}17} = 0{,}0047 .$$

Изменим границы первого и последнего интервалов разбиения в соответствии с диапазоном возможных значений показательно распределённой случайной величины и вычислим вероятности попадания значений изучаемой величины в каждый из частичных интервалов $[C_i; C_{i+1})$ :

$$P(C_i \le X < C_{i+1}) = e^{-\hat{\lambda}C_i} - e^{-\hat{\lambda}C_{i+1}};$$

$P(0 \le X < 53{,}41) = e^{-0{,}0047\cdot 0} - e^{-0{,}0047\cdot 53{,}41} = 1 - 0{,}778 = 0{,}222$;

$P(53{,}41 \le X < 149{,}25) = e^{-0{,}0047\cdot 53{,}41} - e^{-0{,}0047\cdot 149{,}25} = 0{,}778 - 0{,}496 = 0{,}282$;

$P(149{,}25 \le X < 245{,}09) = e^{-0{,}0047\cdot 149{,}25} - e^{-0{,}0047\cdot 245{,}09} = 0{,}496 - 0{,}316 = 0{,}18$;

$P(245{,}09 \le X < 340{,}93) = e^{-0{,}0047\cdot 245{,}09} - e^{-0{,}0047\cdot 340{,}93} = 0{,}316 - 0{,}201 = 0{,}115$;

$P(340{,}93 \le X < 436{,}77) = e^{-0{,}0047\cdot 340{,}93} - e^{-0{,}0047\cdot 436{,}77} = 0{,}201 - 0{,}128 = 0{,}073$;

$P(436{,}77 \le X < 532{,}61) = e^{-0{,}0047\cdot 436{,}77} - e^{-0{,}0047\cdot 532{,}61} = 0{,}128 - 0{,}082 = 0{,}046$;

$P(532{,}61 \le X < 628{,}45) = e^{-0{,}0047\cdot 532{,}61} - e^{-0{,}0047\cdot 628{,}45} = 0{,}082 - 0{,}052 = 0{,}03$;

$P(628{,}45 \le X < \infty) = e^{-0{,}0047\cdot 628{,}45} - e^{-0{,}0047\cdot \infty} = 0{,}052 - 0 = 0{,}052$.

Занесём полученные значения в расчётную таблицу:

| $[C_i; C_{i+1})$ | [0; 53,41) | [53,41; 149,25) | [149,25; 245,09) | [245,09; 340,93) | [340,93; 436,77) | [436,77; 532,61) | [532,61; 628,45) | [628,45; ∞) |
|---|---|---|---|---|---|---|---|---|
| $m_i$ | 12 | 13 | 6 | 6 | 5 | 3 | 4 | 1 |
| $p_i$ | 0,222 | 0,282 | 0,180 | 0,115 | 0,073 | 0,046 | 0,030 | 0,052 |
| $np_i$ | 11,1 | 14,1 | 9 | 5,75 | 3,65 | 2,3 | 1,5 | 2,6 |

Три последних интервала, характеризующиеся малыми значениями $np_i$, объединим в один:

| $[C_i; C_{i+1})$ | [0; 53,41) | [53,41; 149,25) | [149,25; 245,09) | [245,09; 340,93) | [340,93; 436,77) | [436,77; ∞) |
|---|---|---|---|---|---|---|
| $m_i$ | 12 | 13 | 6 | 6 | 5 | 8 |
| $p_i$ | 0,222 | 0,282 | 0,180 | 0,115 | 0,073 | 0,128 |
| $np_i$ | 11,1 | 14,1 | 9 | 5,75 | 3,65 | 6,4 |

Вычислим значение критерия:

$$\chi^2 = \sum_{i=1}^{5} \frac{(m_i - np_i)^2}{np_i} = \frac{(12-11{,}1)^2}{11{,}1} + \frac{(13-14{,}1)^2}{14{,}1} + \frac{(6-9)^2}{9} + \frac{(6-5{,}75)^2}{5{,}75} + \frac{(5-3{,}65)^2}{3{,}65} + \frac{(8-6{,}4)^2}{6{,}4} \approx 2{,}069.$$

Критическое значение $\chi^2_{\alpha,\nu}$ определим по таблице квантилей распределения $\chi^2$ (см. приложение А) для значений $\alpha = 0{,}05$, $\nu = k - r - 1 = 6 - 1 - 1 = 4$: $\chi^2_{0{,}05;4} = 9{,}488$.

Поскольку $\chi^2 \le \chi^2_{\alpha,\nu}$, экспериментальные данные не дают оснований для отклонения гипотезы об экспоненциальном распределении случайной величины $X$, характеризующей промежуток времени между моментами прибытия товарных поездов на сортировочную станцию.



# 5 ЭЛЕМЕНТЫ РЕГРЕССИОННОГО АНАЛИЗА

## 5.1 Основные понятия регрессионного и корреляционного анализа

Предыдущие разделы пособия были посвящены рассмотрению такой ситуации, когда при проведении вероятностного эксперимента фиксировались значения только одной случайной величины. Однако, как правило, большинство представляющих практический интерес вероятностных явлений имеют более сложный, комплексный характер и описываются некоторым множеством величин, которые, вообще говоря, могут быть взаимосвязаны. В дальнейшем ограничимся рассмотрением только того частного случая, когда при проведении вероятностного эксперимента фиксируются значения двух случайных переменных: $X$ и $Y$.

Две случайные величины могут быть связаны функционально, статистически, либо быть независимыми. Говорят, что две переменные связаны **функционально**, если каждому значению одной величины соответствует единственное значение другой.

Например, функциональная зависимость существует между радиусом и площадью круга, силой тока и напряжением в электрической цепи при известном сопротивлении, давлением и объемом газа в сосуде при постоянной температуре и т.п.

Однако чаще на практике встречаются зависимости другого рода, когда при фиксированном значении одной переменной другая имеет некоторую свободу и принимает не заранее определенное, а одно из своих возможных значений. Такая зависимость называется **статистической** (или **вероятностной**). Она состоит в том, что при изменении значений одной величины происходит изменение закона распределения второй.

Для иллюстрации понятия статистической зависимости приведем несколько примеров:

- зависимость расхода топлива от скорости движения поезда на заданном перегоне;
- связь между производительностью труда и себестоимостью продукции;
- зависимость между временем протекания химической реакции и массой выделившегося вещества;
- связь между временем обработки детали на станке и отклонением размеров детали от номинала и т.п.

**Статистической зависимостью** между случайными величинами $X$ и $Y$ называется правило $F$, позволяющее каждому возможному значению одной величины поставить в соответствие условное распределение другой. Например, статистическая зависимость $Y$ от $X$ может быть представлена в виде соотношения $x \to F(Y|X=x)$.

Наиболее важные особенности статистической зависимости находят отражение в тех изменениях, которые претерпевает центр условного распределения одной величины при изменении значения другой. Поэтому на практи-

ке обычно ограничиваются рассмотрением частного случая статистической зависимости, а именно зависимостью математического ожидания одной величины от значения другой. Такая зависимость называется **регрессионной**.

Например, регрессионная зависимость $Y$ от $X$ может быть представлена в виде

$$M(Y|X = x) = f_1(x). \qquad (5.1)$$

Уравнение (5.1) называется **уравнением регрессии *Y* на *X***. Функция $f_1(x)$ называется **функцией регрессии *Y* на *X***, а ее график – **линией регрессии**.

Аналогичным образом можно определить регрессию $X$ на $Y$:

$$M(X|Y = y) = f_2(y).$$

Если уравнение регрессии известно, то с его помощью можно осуществить прогноз математических ожиданий зависимой переменной, соответствующих заданным значениям независимой переменной.

Для построения уравнения регрессии необходимо знание совместного закона распределения вероятностей случайных величин $X$ и $Y$. На практике, при обработке экспериментальных данных это распределение, как правило, неизвестно. В распоряжении исследователя имеется только двумерная выборка объема $n$ значений $(x_i, y_i)$, $i = 1, 2, \ldots, n$ изучаемых случайных величин. Поэтому возникает задача построения на основании имеющихся статистических данных приближенного уравнения регрессии (то есть его оценки) вида $\bar{y}(x) = \varphi(x)$, которое называется **эмпирическим уравнением регрессии**. В качестве экспериментального аналога условного математического ожидания величины $Y$ используется условное среднее $\bar{y}(x)$.

При исследовании статистической зависимости между переменными $X$ и $Y$ на основании опытных данных различают задачи ***регрессионного и корреляционного анализа***.

Методы регрессионного анализа предназначены для построения по имеющимся выборочным данным эмпирического уравнения регрессии, проверки адекватности полученного уравнения опытным данным и использования этого уравнения для осуществления обоснованного прогноза значений зависимой случайной величины.

Методы корреляционного анализа позволяют установить, существует ли какая-либо зависимость между исследуемыми величинами и оценить тесноту существующей связи (то есть близость этой связи к функциональной).

### 5.2 Эмпирическое уравнение регрессии

Важнейшим этапом регрессионного анализа является выбор подходящей регрессионной модели, то есть общего вида функции регрессии. Обычно



этот выбор осуществляется на основании знаний о физической сущности задачи и опыта предыдущих исследований. Если имеющейся информации недостаточно, то, как правило, помогает графическое представление экспериментальных данных в виде **диаграммы рассеивания** (этот график еще называют **корреляционным полем**). Для построения диаграммы рассеивания необходимо в декартовой системе координат точками изобразить выборочные значения $(x_i, y_i)$, $i = 1, 2, \ldots, n$. Вид функции регрессии выбирается таким образом, чтобы она отражала наиболее существенные и характерные особенности расположения этих точек.

Определившись с видом функции регрессии, мы фактически получаем класс сравниваемых функций определенного типа, зависящих от параметров. В общем случае можно записать $\bar{y}(x) = \varphi(x, \beta_0, \beta_1, ..., \beta_k)$, где $\beta_i$ $(i = 0, 1, ..., k)$ – параметры функции $\varphi$.

Для выбора из этого класса функции, наилучшим образом описывающей наблюдаемую закономерность, используют метод наименьших квадратов. Согласно этому методу значения параметров $\beta_0, \beta_1, ..., \beta_k$ выбираются таким образом, чтобы сумма квадратов отклонений выборочных значений $y_i$ от соответствующих им значений на регрессионной кривой $\bar{y}(x_i)$ была бы минимальной. В геометрической интерпретации это означает, что сумма квадратов длин вертикальных отрезков $e_i = |y_i - \bar{y}(x_i)|$ $(i = 1, 2, ..., n)$, изображенных на рисунке 5.1, для «оптимальной» линии регрессии является минимальной.

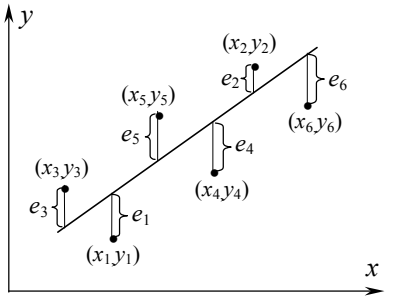



Рисунок 5.1 – Корреляционное поле и линия регрессии

Оценки параметров $\beta_i$ по методу наименьших квадратов определяются из условия:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \varphi(x, \beta_0, \beta_1, ..., \beta_k))^2 \to \min. \qquad (5.2)$$

Вычисляя частные производные функции $S(\beta_0, \beta_1, ..., \beta_k)$ по переменным $\beta_i$ $(i = 0, 1, ..., k)$ и приравнивая их к нулю, получим систему так называемых

нормальных уравнений. Решив эту систему, будут найдены «оптимальные», то есть соответствующие условию (5.2) значения коэффициентов $\hat{\beta}_i$.

После построения эмпирического уравнения регрессии $\bar{y}(x) = \varphi(x)$ можно осуществить прогноз ожидаемых средних значений величины $Y$, соответствующих заданным значениям величины $X$. Понятно, что наблюдаемые значения $y_i$ из-за влияния случайных факторов будут отличаться от прогнозных значений $\bar{y}(x_i)$. В общем случае, зависимость между наблюдаемыми значениями переменных $X$ и $Y$ может быть представлена в виде: $y_i = \varphi(x_i) + \varepsilon_i$, где случайная компонента $\varepsilon_i$ характеризует влияние неучтенных в данной модели факторов и неконтролируемых изменений условий проведения вероятностного эксперимента (условия, накладываемые на значения случайных величин $\varepsilon_i$ будут подробно рассмотрены в подразд. 5.3 на примере простой линейной регрессии).

### 5.3 Простая линейная регрессия

Простейшим видом регрессионной зависимости между переменными $X$ и $Y$ является простая линейная регрессия, которая может быть определена следующим образом $M(Y|X = x) = \beta_1 x + \beta_0$. Регрессионная модель такого вида достаточно часто встречается на практике. Исследования показывают, что в большой части случаев наблюдаемая зависимость между изучаемыми величинами по крайней мере приближенно может быть описана с помощью линейной регрессионной модели. В частности, в курсе теории вероятностей доказано, что если случайные величины $X$ и $Y$ подчиняются закону двумерного нормального распределения вероятностей*, то линии регрессии $Y$ на $X$ и $X$ на $Y$ являются линейными функциями.

Зависимость значений величины $Y$ от $X$ в этом случае может быть представлена в виде $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$. Доказано, что при справедливости предположения о двумерном нормальном распределении величин $X$ и $Y$, отклонения $\varepsilon_i$ являются независимыми случайными величинами, имеющими нормальное распределение, причем $M(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma_\varepsilon{}^2 = \text{const}$

---

* Функция плотности двумерного нормального распределения имеет вид:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp\left(-\frac{1}{2(1-r^2)}\left(\frac{(x-m_x)^2}{\sigma_x{}^2} - 2r\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y{}^2}\right)\right), \ (x,y) \in R^2,$$

где $m_x$, $m_y$ – математические ожидания составляющих случайных величин $X$ и $Y$; $\sigma_x$, $\sigma_y$ – их средние квадратические отклонения; $r$ – коэффициент корреляции между переменными $X$ и $Y$.



(эти свойства $\varepsilon_i$ являются необходимыми условиями для проведения последующего анализа эмпирического уравнения линейной регрессии).

Итак, пусть в результате $n$ независимых испытаний получена двумерная выборка значений изучаемых величин $(x_i, y_i)$, $i = 1, 2, \ldots, n$, на основании которой необходимо построить уравнение линейной регрессии $Y$ от $X$. Соотношение (5.2) в этом случае примет вид:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2 = \min .$$

Вычислим частные производные функции $S(\beta_0, \beta_1)$ по переменным $\beta_0$ и $\beta_1$ и приравняем их к нулю:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = (-2)\left[\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_0 n\right] = 0;$$

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ;$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\Big|_{\beta_1=\hat{\beta}_1} = (-2)\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = (-2)\left[\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} x_i\right] = 0;$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i .$$

Таким образом, система нормальных уравнений имеет вид:

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ; \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i . \end{cases}$$

Решив эту систему уравнений, получим значения коэффициентов эмпирического уравнения линейной регрессии $\bar{y}(x) = \hat{\beta}_1 x + \hat{\beta}_0$.

**Пример 5.1.** В книге «Основы химии» Д. И. Менделеева приведены следующие данные о количестве азотнонатриевой соли (переменная $Y$), которое можно растворить в 100 г воды в зависимости от температуры раствора (переменная $X$):

| $X$ | 0 | 4 | 10 | 15 | 21 | 29 | 36 | 51 | 68 |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 66,7 | 71,0 | 76,3 | 80,6 | 85,7 | 92,9 | 99,4 | 113,6 | 125,1 |

Здесь $X$ – температура раствора, °C;

$Y$ – количество растворившегося вещества.

Требуется на основании имеющихся экспериментальных данных исследовать зависимость значений переменной $Y$ от $X$.

Р е ш е н и е . Для наглядного представления выборочных данных построим корреляционное поле (рисунок 5.2, *а*). По виду полученного графика можно судить о том, что между исследуемыми переменными существует достаточно тесная линейная зависимость. Для построения эмпирического уравнения линейной регрессии $\bar{y}(x) = \hat{\beta}_1 x + \hat{\beta}_0$ вычислим необходимые значения сумм:

$$\sum_{i=1}^{9} x_i = 0 + 4 + 10 + 15 + 21 + 29 + 36 + 51 + 68 = 234;$$

$$\sum_{i=1}^{9} x_i^2 = 0 + 4^2 + 10^2 + 15^2 + 21^2 + 29^2 + 36^2 + 51^2 + 68^2 = 10144;$$

$$\sum_{i=1}^{9} y_i = 66{,}7 + 71{,}0 + 76{,}3 + 80{,}6 + 85{,}7 + 92{,}9 + 99{,}4 + 113{,}6 + 125{,}1 = 811{,}3;$$

$$\sum_{i=1}^{9} x_i y_i = 0 \cdot 66{,}7 + 4 \cdot 71{,}0 + 10 \cdot 76{,}3 + 15 \cdot 80{,}6 + 21 \cdot 85{,}7 + 29 \cdot 92{,}9 +$$

$$+ 36 \cdot 99{,}4 + 51 \cdot 113{,}6 + 68 \cdot 125{,}1 = 24628{,}6.$$

Таким образом, система нормальных уравнений для определения значений $\hat{\beta}_i$ имеет вид:

$$\begin{cases} 9\hat{\beta}_0 + 234\hat{\beta}_1 = 811{,}3; \\ 234\hat{\beta}_0 + 10144\hat{\beta}_1 = 24628{,}6. \end{cases}$$

Решая ее, получим: $\hat{\beta}_0 \approx 67{,}5$; $\hat{\beta}_1 \approx 0{,}87$. Итак, уравнение линейной регрессии, описывающее зависимость количества растворенного вещества от температуры раствора, имеет вид $\bar{y}(x) = 0{,}87x + 67{,}5$. График соответствующей линии регрессии изображен на рисунке 5.2, *б*.

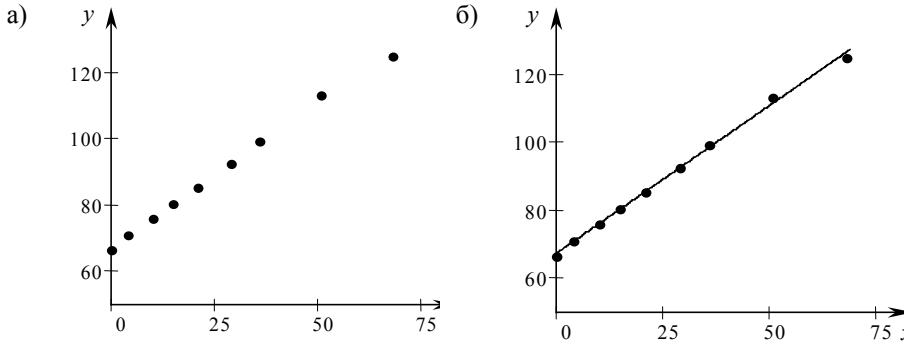



Рисунок 5.2 – Диаграмма рассеивания и график линейной регрессии (пример 5.1)



### 5.4 Использование эмпирического уравнения линейной регрессии для прогноза

Поскольку коэффициенты эмпирического уравнения регрессии $\hat{\beta}_0$ и $\hat{\beta}_1$, вычисленные на основании опытных данных, являются случайными величинами, уравнение регрессии также содержит в себе некоторый элемент случайности. Причем изменения значения $\hat{\beta}_0$ приводят к смещению линии регрессии относительно оси ординат, а варьирование значения $\hat{\beta}_1$ влечет за собой «покачивание» линии регрессии относительно центра распределения $(\bar{x}, \bar{y})$. Поэтому точечные оценки условных средних значений переменной *Y*, соответствующие заданным значениям *x*, могут значительно отличаться от неизвестных исследователю точных значений $M(Y|X = x)$.

Доказано, что в случае двумерного нормального распределения изучаемых случайных величин (или, иначе говоря, при справедливости сделанных в подразд. 5.3 предположений относительно значений $\varepsilon_i$), статистика $\dfrac{M(Y|X = x) - \bar{y}(x)}{S_{\bar{y}(x)}}$ для каждого фиксированного значения *x* имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы.

Здесь

$$S_{\bar{y}(x_0)} = S_{\text{ост}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (5.3)$$

представляет собой оценку среднего квадратического отклонения значения регрессии в точке $x_0$. Величина $S_{\text{ост}} = \sqrt{\dfrac{1}{n-2} \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2}$ характеризует рассеяние выборочных значений $y_i$ относительно линии регрессии.

Для построения доверительного интервала для неизвестного условного математического ожидания величины *Y* в точке $x_0$ необходимо по таблице распределения Стьюдента в зависимости от уровня значимости $\alpha$ найти критическое значение $t_{\alpha/2,\, n-2}$, соответствующее условию $P(|t| < t_{\alpha/2,\, n-2}) = 1 - \alpha$.

Тогда

$$P(-t_{\alpha/2,\, n-2} < \frac{M(Y|X = x_0) - \bar{y}(x_0)}{S_{\bar{y}(x_0)}} < t_{\alpha/2,\, n-2}) = 1 - \alpha;$$

$$P(\bar{y}(x_0) - t_{\alpha/2,\, n-2} S_{\bar{y}(x_0)} < M(Y|X = x) < \bar{y}(x_0) + t_{\alpha/2,\, n-2} S_{\bar{y}(x_0)}) = 1 - \alpha.$$

Таким образом, $(1-\alpha)$-процентный доверительный интервал для значения $M(Y|X=x_0)$ имеет вид:

$$(\bar{y}(x_0) - t_{\alpha/2,\,n-2} S_{\bar{y}(x_0)},\ \bar{y}(x_0) + t_{\alpha/2,\,n-2} S_{\bar{y}(x_0)}). \qquad (5.4)$$

На основании соотношений (5.3) и (5.4) нетрудно заметить, что ширина доверительного интервала существенно зависит от значения $x_0$. Наименьшее значение величины $S_{\bar{y}(x_0)}$, а следовательно, и ширины доверительного интервала соответствует значению $x_0 = \bar{x}$, при удалении значения $x_0$ от $\bar{x}$ доверительный интервал расширяется. Также легко заметить, что при увеличении объема выборки $n$ ширина доверительного интервала уменьшается, приближаясь к нулю при $n \to \infty$.

Если изобразить на графике границы доверительных интервалов для всех возможных значений $x$, то получим две гиперболы, между ветвями которых с вероятностью $(1-\alpha)$ находится «теоретическая» линия регрессии $Y$ на $X$ (рисунок 5.3).

**Замечание** – Из методики построения эмпирического уравнения регрессии следует, что прогнозы по этому уравнению правомерны только в том случае, если значения $x_0$ не выходят за пределы выборочных значений ($x_{\min}$, $x_{\max}$), на основании которых построено это уравнение. То есть экстраполяция по уравнению регрессии может привести к значительным погрешностям.

**Пример 5.1 (продолжение)**. На основании опытных данных было построено уравнение $\bar{y}(x) = 0{,}87x + 67{,}5$, описывающее зависимость количества растворенного $NaNO_3$ от температуры раствора.

Требуется оценить ожидаемые значения количества растворенного вещества при $x_0' = 20\ ^\circ C$ и $x_0'' = 60\ ^\circ C$. Построить 95%-ные доверительные интервалы для ожидаемых средних значений количества растворенного вещества при этих значениях температуры.

Р е ш е н и е. На основании построенного эмпирического уравнения регрессии вычислим:

$$\bar{y}(x_o') = \bar{y}(20) = 0{,}87 \cdot 20 + 67{,}5 = 84{,}9;$$
$$\bar{y}(x_o'') = \bar{y}(60) = 0{,}87 \cdot 60 + 67{,}5 = 119{,}7,$$

а также $\bar{y}(x_i)$:

| $x_i$ | 0 | 4 | 10 | 15 | 21 | 29 | 36 | 51 | 68 |
|---|---|---|---|---|---|---|---|---|---|
| $\bar{y}(x_i)$ | 67,5 | 70,98 | 76,2 | 80,55 | 85,77 | 92,73 | 98,82 | 111,87 | 126,66 |

Вычислим вспомогательные величины:

$$S_{\text{ост}} = \sqrt{\frac{1}{9-2}\sum_{i=1}^{9}(y_i - \bar{y}(x_i))^2} = \sqrt{\frac{1}{7}6{,}5907} = \sqrt{0{,}9415} = 0{,}97032 \approx 0{,}97;$$



$$S_{\bar{y}(20)} = S_{\text{ост}}\sqrt{\frac{1}{9} + \frac{(20-23{,}4)^2}{\sum_{i=1}^{9}(x_i - 23{,}4)^2}} = 0{,}97\sqrt{\frac{1}{9} + \frac{(20-23{,}4)^2}{4120{,}84}} = 0{,}3274;$$

$$S_{\bar{y}(60)} = S_{\text{ост}}\sqrt{\frac{1}{9} + \frac{(60-23{,}4)^2}{\sum_{i=1}^{9}(x_i - 23{,}4)^2}} = 0{,}97\sqrt{\frac{1}{9} + \frac{(60-23{,}4)^2}{4120{,}84}} = 0{,}6406.$$

По таблице распределения Стьюдента найдем критическое значение $t_{\alpha/2,\,n-2} = t_{0{,}025;\,7} = 2{,}365$.

При $x = 20$ 95%-ный доверительный интервал для $\bar{y}(x)$ имеет вид:

$(84{,}9 - 2{,}365 \cdot 0{,}3274;\ 84{,}9 + 2{,}365 \cdot 0{,}3274) = (84{,}1257;\ 85{,}6743);\ (\varepsilon = 0{,}7743).$

При $x = 60$

$(119{,}7 - 2{,}365 \cdot 0{,}6406;\ 119{,}7 + 2{,}365 \cdot 0{,}6406) = (118{,}185;\ 121{,}6215);\ (\varepsilon = 1{,}515).$

Можно отметить, что ширина доверительного интервала в точке $x_0'' = 60\ ^\circ C$ $(\varepsilon = 1{,}515)$ заметно больше, чем при $x_0' = 20\ ^\circ C$ $(\varepsilon = 0{,}7743)$, так как значение $x_0'' = 60$ отстоит гораздо дальше от среднего значения $\bar{x} = 23{,}4$, чем $x_0' = 20$.

На рисунке 5.3 пунктиром изображены границы доверительных интервалов для $M(Y|X=x)$, построенные для всего диапазона наблюдаемых значений $X$.

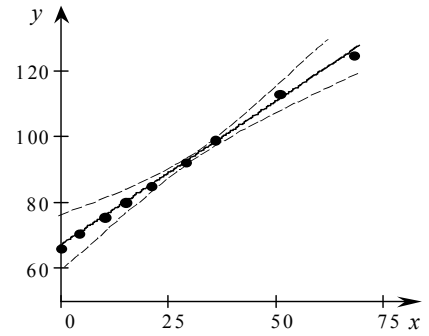



Рисунок 5.3 – Доверительные интервалы для $M(Y|X=x)$ (пример 5.1)

### 5.5 Построение эмпирического уравнения регрессии нелинейного вида

Линейные регрессионные модели, хотя и очень широко применяются на практике, но не всегда способны достаточно хорошо описать наблюдаемую зависимость между изучаемыми величинами.

В соответствии с методикой, приведенной в подразд. 5.2, теоретически можно построить эмпирические уравнения регрессии произвольного вида,

использование которых сдерживается лишь трудностями, возникающими при решении системы нормальных уравнений. Рассмотрим несколько примеров построения уравнений регрессии нелинейного вида, часто используемых на практике.

**1 Гиперболическая регрессионная модель.**

Предположим, что исследователь для описания наблюдаемой зависимости между переменными $X$ и $Y$ считает нужным использовать функцию гиперболического вида $\bar{y}(x) = \dfrac{\hat{\beta}_1}{x} + \hat{\beta}_0$.

Для нахождения оптимальных значений параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ вычислим частные производные функции

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2 = \sum_{i=1}^{n} (y_i - \frac{\beta_1}{x_i} - \beta_0)^2$$

и приравняем их к нулю:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0} = (-2)\sum_{i=1}^{n}(y_i - \frac{\hat{\beta}_1}{x_i} - \hat{\beta}_0) = (-2)\left[\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n}\frac{1}{x_i} - \hat{\beta}_0 n\right] = 0;$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\Big|_{\beta_1=\hat{\beta}_1} = (-2)\sum_{i=1}^{n}\frac{1}{x_i}(y_i - \frac{\hat{\beta}_1}{x_i} - \hat{\beta}_0) = (-2)\left[\sum_{i=1}^{n}\frac{y_i}{x_i} - \hat{\beta}_1 \sum_{i=1}^{n}\frac{1}{x_i^2} - \hat{\beta}_0 \sum_{i=1}^{n}\frac{1}{x_i}\right] = 0.$$

Таким образом, система нормальных уравнений для определения коэффициентов эмпирического уравнения регрессии гиперболического типа имеет вид:

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{-1} = \sum_{i=1}^{n} y_i; \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i^{-1} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{-2} = \sum_{i=1}^{n} x_i^{-1} y_i. \end{cases}$$

**2 Полиномиальная регрессионная модель.**

Пусть для описания зависимости между исследуемыми переменными предполагается использовать полиномиальную модель вида $\bar{y}(x) = \beta_k x^k + \beta_{k-1} x^{k-1} + ... + \beta_1 x + \beta_0$. Обычно на практике используются полиномы не выше третьей-четвертой степени.

В данном случае функция $S(\beta_0, \beta_1, ..., \beta_k)$ будет иметь вид:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_i^k)^2.$$

Вычислим частные производные функции $S$ по переменным $\beta_i$ и приравняем их к нулю:



$$\frac{\partial S(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - ... - \hat{\beta}_k x^k) =$$

$$= (-2)\left[\sum_{i=1}^{n} y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^{n} x_i - ... - \hat{\beta}_k \sum_{i=1}^{n} x_i^k\right] = 0;$$

$$\frac{\partial S(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_1}\Big|_{\beta_1=\hat{\beta}_1} = (-2)\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - ... - \hat{\beta}_k x^k) =$$

$$= (-2)\left[\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - ... - \beta_k \sum_{i=1}^{n} x_i^{k+1}\right] = 0;$$

$$\frac{\partial S(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_k}\Big|_{\beta_k=\hat{\beta}_k} = (-2)\sum_{i=1}^{n} x_i^k (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - ... - \hat{\beta}_k x^k) =$$

$$= (-2)\left[\sum_{i=1}^{n} x_i^k y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i^k - \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} - ... - \beta_k \sum_{i=1}^{n} x_i^{2k}\right] = 0.$$

Итак, система нормальных уравнений для определения параметров $\hat{\beta}_i$, $i = 1, 2, ..., k$ эмпирического уравнения полиномиальной регрессии имеет вид:

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i; \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i y_i; \\ \quad ... \\ \beta_0 \sum_{i=1}^{n} x_i^k + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} x_i^k y_i. \end{cases}$$

**Пример 5.2.** При исследовании работы предприятия получены следующие данные, характеризующие зависимость себестоимости выпускаемой продукции от объема производства:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 0,085 | 0,146 | 0,237 | 0,414 | 0,551 | 0,653 | 0,82 | 0,915 | 1,043 | 1,150 |
| $Y$ | 18,4 | 14,3 | 12,1 | 11,5 | 10,3 | 9,4 | 9,6 | 9,0 | 8,8 | 8,1 |

Здесь $X$ – объем выпускаемой продукции в течение года, млн т;

$Y$ – фактическая средняя себестоимость одной тонны продукции, у. ден. ед.

Требуется на основании приведенных опытных данных исследовать зависимость переменной $Y$ от $X$.

Решение. Для наглядного изображения имеющихся выборочных данных построим корреляционное поле (рисунок 5.4, *а*).

По виду расположения точек на корреляционном поле можно выдвинуть предположение о том, что между переменными $X$ и $Y$ существует зависимость гиперболиче-

ского типа. То есть для описания наблюдаемой зависимости будем использовать уравнение $\bar{y}(x) = \dfrac{\hat{\beta}_1}{x} + \hat{\beta}_0$.

Вычислим коэффициенты системы нормальных уравнений:

$\sum_{i=1}^{10} x_i^{-1} = 32{,}726$; $\sum_{i=1}^{10} y_i = 111{,}5$; $\sum_{i=1}^{10} x_i^{-2} = 218{,}9551$; $\sum_{i=1}^{10} x_i^{-1} y_i = 463{,}3617$.

Система нормальных уравнений имеет вид:

$$\begin{cases} 10\hat{\beta}_0 + 32{,}726\hat{\beta}_1 = 111{,}5; \\ 32{,}726\hat{\beta}_0 + 218{,}9551\hat{\beta}_1 = 463{,}9551. \end{cases}$$

Решая ее, получим: $\hat{\beta}_0 \approx 8{,}27$, $\hat{\beta}_1 \approx 0{,}88$, то есть искомое эмпирическое уравнение регрессии имеет вид: $\bar{y}(x) = \dfrac{0{,}88}{x} + 8{,}27$.

График соответствующей линии регрессии изображен на рисунке 5.4, *б*.

Для сравнения на основании того же набора экспериментальных данных можно построить регрессионную модель линейного типа. Уравнение линейной регрессии будет иметь вид: $\bar{y}(x) = -7{,}26x + 15{,}52$. График соответствующей линии регрессии также приведен на рисунке 5.4, *б*.

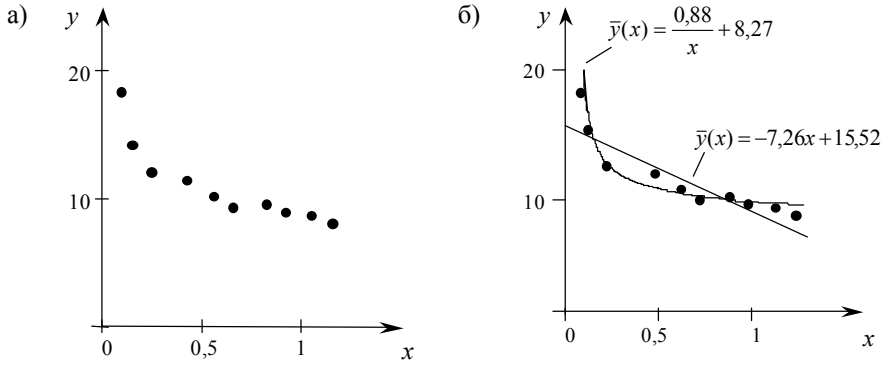



Рисунок 5.4 – Корреляционное поле и линии регрессии (пример 5.2)

### 5.6 Проверка адекватности эмпирического уравнения регрессии выборочным данным

Пусть на основании двумерной выборки $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ построено эмпирическое уравнение регрессии $\bar{y}(x) = \varphi(x, \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ (зависящее от $k+1$ параметров $\beta_0$, $\beta_1, \ldots, \beta_k$). Прежде чем использовать полученное



уравнение для описания зависимости между изучаемыми величинами, необходимо проверить адекватность этого уравнения выборочным данным, то есть проверить согласование с экспериментальными данными гипотезы $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, утверждающей, что между изучаемыми величинами отсутствует зависимость предполагаемого вида и отличие полученных оценок коэффициентов регрессии $\hat{\beta}_i$ от нуля объясняется только влиянием случайных факторов. Альтернативная гипотеза $H_1$ состоит в том, что хотя бы один из коэффициентов $\beta_i$ не равен нулю.

Для осуществления проверки указанной гипотезы используется дисперсионный анализ. Доказано, что общая сумма квадратов отклонений $S^2_{\text{общ}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ наблюдаемых значений зависимой переменной $y_i$ от среднего значения $\bar{y}$ может быть разделена на две составляющие: $S^2_{\text{общ}} = S^2_{\text{регр}} + S^2_{\text{ост}}$, где

$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2$ характеризует влияние переменной $X$ (описываемое соотношением $\bar{y}(x) = \varphi(x)$);

$S^2_{\text{ост}} = \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2$ характеризует влияние неучтенных в данной регрессионной модели случайных факторов.

Для проверки гипотезы об отсутствии между переменными $X$ и $Y$ регрессионной зависимости предполагаемого вида используется статистика $F = \dfrac{S^2_{\text{регр}} / \nu_1}{S^2_{\text{ост}} / \nu_2}$, которая при справедливости этой гипотезы имеет распределение Фишера с $\nu_1 = k$ и $\nu_2 = n - k - 1$ степенями свободы.

Очевидно, что чем больше доля $S^2_{\text{регр}}$ в значении суммы $S^2_{\text{общ}}$, и, соответственно, чем меньше доля $S^2_{\text{ост}}$, тем бóльшая часть рассеяния значений $y_i$ вокруг линии регрессии объясняется зависимостью от значений $X$, (задаваемой соотношением $\bar{y}(x) = \varphi(x)$) и, соответственно, тем интенсивнее связь данного вида между $X$ и $Y$. Таким образом, чем больше значение статистики $F$, тем менее вероятна справедливость проверяемой гипотезы. И наоборот, получение близких к нулю значений статистики $F$ свидетельствует о слабой зависимости значений переменной $Y$ от $X$.

Критическая область в данном случае определяется условием $P(F \geq F_{\alpha,\,\nu_1,\,\nu_2}) = \alpha$. Если расчетное значение критерия $F$ окажется больше критического значения $F_{\alpha,\,\nu_1,\,\nu_2}$ (определяемого по таблицам критических точек распределения Фишера), то проверяемую гипотезу об отсутствии между изучаемыми переменными зависимости предполагаемого вида следует отвергнуть. Это означает, что наблюдаемая зависимость между величинами $X$ и $Y$ не может быть объяснена влиянием только случайных факторов, и построенное эмпирическое уравнение регрессии можно считать адекватным опытным данным.

В противном случае, если $F < F_{\alpha,\,\nu_1,\,\nu_2}$, то нет оснований для отклонения выдвинутой гипотезы и полученное уравнение регрессии не является адекватным имеющимся выборочным данным. Следовательно, оно не может использоваться для описания изучаемого явления.

**Пример 5.1 (продолжение).** Пользуясь критерием Фишера, проверим адекватность построенной регрессионной модели $\bar{y}(x) = 0{,}87x + 67{,}5$ выборочным данным.

Определив на основании эмпирического уравнения регрессии значения $\bar{y}(x_i)$:

| $x_i$ | 0 | 4 | 10 | 15 | 21 | 29 | 36 | 51 | 68 |
|---|---|---|---|---|---|---|---|---|---|
| $\bar{y}(x_i)$ | 67,5 | 70,98 | 76,2 | 80,55 | 85,77 | 92,73 | 98,82 | 111,87 | 126,66 |

вычислим значения сумм:

$$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2 = 3800{,}395; \qquad S^2_{\text{ост}} = \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2 = 6{,}59.$$

Выборочное значение критерия Фишера: $F = \dfrac{S^2_{\text{регр}} / \nu_1}{S^2_{\text{ост}} / \nu_2} = \dfrac{3800{,}395/1}{6{,}59/7} = 4036{,}4.$

По приложению В определим критическое значение критерия Фишера, соответствующее значениям $\alpha = 0{,}05;\ \nu_1 = 1;\ \nu_2 = n - 2 = 9 - 2 = 7$: $F_{0{,}05;\,1;\,7} = 5{,}59$.

Поскольку $F >> F_{\alpha;\,\nu_1;\,\nu_2}$, гипотезу об отсутствии между изучаемыми величинами линейной зависимости следует отвергнуть. Построенное эмпирическое уравнение линейной регрессии является адекватным опытным данным.

**Пример 5.2 (продолжение).** Проверим соответствие опытным данным эмпирических уравнений регрессии гиперболического и линейного видов, описывающих зависимость себестоимости продукции от объемов производства.



Сначала подвергнем анализу гиперболическую модель $\bar{y}(x) = \dfrac{0{,}88}{x} + 8{,}27$:

| $x_i$ | 0,085 | 0,146 | 0,237 | 0,414 | 0,551 | 0,653 | 0,82 | 0,915 | 1,043 | 1,150 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\bar{y}(x_i)$ | 18,62 | 14,30 | 11,98 | 10,39 | 9,87 | 9,62 | 9,34 | 9,23 | 9,11 | 9,03 |

$$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2 = 86{,}5722; \quad S^2_{\text{ост}} = \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2 = 2{,}5728.$$

Отсюда получаем $F_{\text{гиперб}} = \dfrac{S^2_{\text{регр}} / \nu_1}{S^2_{\text{ост}} / \nu_2} = \dfrac{86{,}5722/1}{2{,}5728/8} = 269{,}192$.

Аналогичные вычисления выполним для линейной модели $\bar{y}(x) = -7{,}26x + 15{,}51$:

| $x_i$ | 0,085 | 0,146 | 0,237 | 0,414 | 0,551 | 0,653 | 0,82 | 0,915 | 1,043 | 1,150 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\bar{y}(x_i)$ | 14,89 | 14,45 | 13,79 | 12,5 | 11,51 | 10,76 | 9,55 | 8,87 | 7,94 | 7,16 |

Значения сумм:

$$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2 = 61{,}982; \quad S^2_{\text{ост}} = \sum_{i=1}^{n} (y_i - \bar{y}(x_i))^2 = 21{,}163.$$

$$F_{\text{лин}} = \frac{S^2_{\text{регр}} / \nu_1}{S^2_{\text{ост}} / \nu_2} = \frac{67{,}982/1}{21{,}163/8} = 25{,}698.$$

С помощью таблицы квантилей распределения Фишера (см. приложение В) определим критическое значение $F_{\alpha;\,\nu_1;\,\nu_2} = F_{0{,}05;\,1;\,8} = 5{,}32$.

Поскольку $F_{\text{гиперб}} > F_{\alpha;\,\nu_1;\,\nu_2}$ и $F_{\text{лин}} > F_{\alpha;\,\nu_1;\,\nu_2}$, обе построенные регрессионные модели являются адекватными опытным данным и могут быть использованы для описания изучаемого явления. Тот факт, что $F_{\text{гиперб}} >> F_{\text{лин}}$ свидетельствует о том, что построенная гиперболическая модель более полно описывает наблюдаемую зависимость между изучаемыми величинами.

## 6 ЭЛЕМЕНТЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Термин «корреляция» происходит от латинского correlatio, что означает «соотношение, взаимосвязь». Содержание корреляционного анализа составляют методы, позволяющие установить, существует ли зависимость между исследуемыми переменными и оценить тесноту этой зависимости.

### 6.1 Коэффициент корреляции

В курсе теории вероятностей для оценки тесноты связи между случайными величинами $X$ и $Y$, подчиняющимися закону двумерного нормального

распределения, используется коэффициент корреляции $r$, значение которого вычисляется по формуле

$$r = \frac{M\left[(X - M[X])(Y - M[Y])\right]}{\sigma[X]\sigma[Y]}. \qquad (6.1)$$

Напомним, что в этом случае зависимость между переменными $X$ и $Y$ имеет линейный вид. Таким образом, коэффициент корреляции является мерой тесноты именно линейной зависимости. Этот коэффициент может использоваться в качестве меры коррелированности (то есть линейной зависимости) любых изучаемых случайных величин.

Известны следующие свойства коэффициента корреляции $r$:

1 Возможные значения коэффициента корреляции принадлежат отрезку $[-1, 1]$: $-1 \le r \le 1$.

2 Необходимым и достаточным условием отсутствия линейной зависимости между исследуемыми величинами является равенство нулю соответствующего коэффициента корреляции.

3 Если корреляция между переменными $X$ и $Y$ положительна (то есть, если при увеличении значений одной переменной, значения другой также имеют тенденцию к возрастанию), то $r > 0$; если имеет место отрицательная корреляция (при увеличении значений одной переменной значения другой, в среднем, убывают), то $r < 0$.

4 Чем ближе по модулю значение коэффициента корреляции к единице, тем теснее линейная зависимость между изучаемыми величинами.

5 $|r| = 1$ тогда и только тогда, когда между переменными $X$ и $Y$ существует линейная функциональная зависимость.

6 Значение коэффициента корреляции не зависит от выбора начала отсчета и единиц измерения исследуемых величин.

### 6.2 Эмпирический коэффициент корреляции

Для вычисления значения коэффициента корреляции по формуле (6.1) необходимо знание совместного закона распределения вероятностей изучаемых случайных величин. Обычно при исследовании экспериментальных данных в нашем распоряжении имеется только двумерная выборка объема $n$, на основании которой необходимо оценить неизвестное значение $r$.

В качестве точечной оценки коэффициента корреляции используется статистика $\hat{r}$ которая называется **эмпирическим коэффициентом корреляции**:



$$\hat{r} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Эмпирический коэффициент корреляции обладает всеми свойствами «теоретического» коэффициента корреляции. Его важнейшее свойство состоит в том, что при увеличении объема выборки значение $\hat{r}$ приближается к оцениваемому значению коэффициента корреляции $r$ (то есть $\hat{r}$ является состоятельной оценкой $r$).

**Пример 6.1.** На основании данных, приведенных в примере 5.1, оценим с помощью эмпирического коэффициента корреляции тесноту связи между переменными $X$ и $Y$, характеризующими, соответственно, температуру раствора и количество растворенной азотнонатриевой соли:

$$\hat{r} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0{,}935.$$

Поскольку вычисленное значение эмпирического коэффициента корреляции очень близко к единице, можно сделать вывод о том, что между исследуемыми переменными существует тесная положительная линейная зависимость.

**Пример 6.2.** На сталеплавильном заводе при обследовании 14 плавок определенного сорта стали получен следующий набор экспериментальных значений:

| Номер плавки | $X$ – угар кремния, % | $Y$ – выход стали, % | Номер плавки | $X$ – угар кремния, % | $Y$ – выход стали, % |
|---|---|---|---|---|---|
| 1 | 7,9 | 70,3 | 8 | 7,2 | 86,8 |
| 2 | 0,9 | 85,0 | 9 | 8,8 | 70,1 |
| 3 | 3,7 | 100,0 | 10 | 11,2 | 81,9 |
| 4 | 8,1 | 78,1 | 11 | 0,5 | 97,1 |
| 5 | 6,9 | 77,9 | 12 | 4,6 | 68,2 |
| 6 | 0,8 | 98,4 | 13 | 9,7 | 92,1 |
| 7 | 6,0 | 59,2 | 14 | 1,0 | 91,2 |

Требуется на основании имеющихся опытных данных исследовать зависимость между случайными величинами $X$ и $Y$.

Р е ш е н и е. По виду корреляционного поля (рисунок 6.1) трудно сделать вывод о существовании какой-либо ярко выраженной зависимости между изучаемыми переменными. Однако можно заметить, что бóльшим значениям одной величины, в среднем, соответствуют несколько меньшие значения другой,

то есть между переменными $X$ и $Y$ имеет место некоторая отрицательная корреляция.

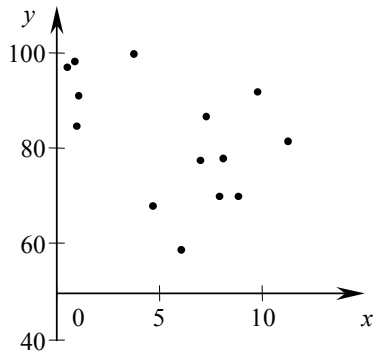



Рисунок 6.1 – Корреляционное поле (пример 6.2)

Вычислим значение эмпирического коэффициента корреляции:

$$\hat{r} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \approx -0{,}46.$$

Полученное значение $\hat{r}$ указывает на существование не тесной отрицательной линейной зависимости между изучаемыми величинами (угаром кремния и выходом стали).

### 6.3 Проверка значимости эмпирического коэффициента корреляции

Вычисляя на основании выборочных данных значение эмпирического коэффициента корреляции $\hat{r}$, мы понимаем, что получаем таким образом лишь некоторое приближение неизвестного значения $r$, которое, особенно при исследовании выборок небольшого объема, может содержать значительную погрешность. В частности, может оказаться, что при $r = 0$, то есть при изучении некоррелированных случайных величин вычисленное значение $\hat{r}$ в силу влияния случайных факторов будет значительно отличаться от нуля. В этом случае, опираясь только на полученное значение оценки коэффициента корреляции $\hat{r}$, мы можем сделать ошибочный вывод о коррелированности исследуемых случайных величин.

Для получения более надежных выводов о существовании линейной зависимости между изучаемыми переменными необходимо проверить значимость полученного значения $\hat{r}$, то есть проверить с помощью вероятностно-статистических методов согласование с опытными данными гипотезы $H_0$ :



$r = 0$ (альтернативная гипотеза: $r \neq 0$). Проверка этой гипотезы осуществляется в предположении о двумерном нормальном распределении исследуемых случайных величин.

Известно, что статистика

$$t = \hat{r}\sqrt{\frac{n-2}{1-\hat{r}^2}} \qquad (6.2)$$

при справедливости сформулированной гипотезы ($r = 0$) независимо от значений $\hat{r}$ и $n$ подчиняется закону распределения Стьюдента с $\nu = n - 2$ степенями свободы. Для заданного уровня значимости $\alpha$ по таблице квантилей распределения Стьюдента (см. приложение Б) можно определить критическое значение $t_{\alpha/2,\,n-2}$, такое, что $P(|t| > t_{\alpha/2;\,n-2}) = \alpha$ .

Если окажется, что $|t| > t_{\alpha/2;\,n-2}$, то проверяемая гипотеза $H_0$ : $r = 0$ должна быть отвергнута, то есть наблюдаемое отличие эмпирического коэффициента корреляции от нуля не может быть объяснено только случайностью выборки и является значимым при уровне значимости $\alpha$. Если $|t| \le t_{\alpha/2;\,n-2}$, то делают вывод от том, что нет оснований для отклонения гипотезы об отсутствии линейной зависимости между переменными $X$ и $Y$.

**Замечание** – Преобразовав выражение (6.2), можно получить соотношение, позволяющее определить минимальное «пороговое» значение $\hat{r}$, которое будет являться значимым при используемых значениях уровня значимости $\alpha$ и объема выборки $n$:

$$\hat{r} \ge \sqrt{\frac{t_{\alpha/2;\,n-2}^{\,2}}{t_{\alpha/2;\,n-2}^{\,2} + n - 2}}\,. \qquad (6.3)$$

**Пример 6.1 (продолжение)**. Приняв значение $\alpha$ равным 0,05, проверим значимость эмпирического коэффициента корреляции $\hat{r} = 0{,}935$ , характеризующего тесноту линейной связи между температурой раствора и массой растворенного $NaNO_3$.

Вычислим значение статистики: $t = \hat{r}\sqrt{\frac{n-2}{1-\hat{r}^2}} = 0{,}935\sqrt{\frac{9-2}{1-0{,}935^2}} = 6{,}975$ .

Критическое значение критерия Стьюдента $t_{\alpha/2;\,n-2} = t_{0{,}025;7} = 2{,}365.$

Поскольку $|t| > t_{\alpha/2,n-2}$ , то гипотезу об отсутствии между изучаемыми переменными линейной зависимости нужно отвергнуть.

Используя соотношение (6.3), можно установить, что в данном случае (то есть при $n = 9$ и $\alpha = 0{,}05$ и, соответственно, при $t_{\alpha/2;\,n-2} = 2{,}365$ ) минимальное значимо отличающее от нуля значение эмпирического коэффициента корреляции, составляет 0,666:

$$\hat{r} = \sqrt{\frac{t_{\alpha/2;n-2}{}^2}{t_{\alpha/2;n-2}{}^2 + n - 2}} = \sqrt{\frac{2{,}365^2}{2{,}365^2 + 9 - 2}} = 0{,}666 .$$

**Пример 6.2 (продолжение)**. Проверим при $\alpha = 0{,}05$ значимость выборочного коэффициента корреляции $\hat{r} = -0{,}46$, оценивающего тесноту линейной связи между угаром кремния и выходом стали в процессе плавки. В данном случае: $n = 14$, $\alpha = 0{,}05$, $t_{\alpha/2;\,n-2} = t_{0{,}025;12} = 2{,}18$.

Значение статистики $t = \hat{r}\sqrt{\frac{n-2}{1-\hat{r}^2}} = -0{,}46\sqrt{\frac{14-2}{1-(-0{,}46)^2}} = -1{,}415.$

Поскольку $|t| < t_{\alpha/2;n-2}$, то нет оснований для отклонения гипотезы о некоррелированности случайных величин *X* и *Y*. Это означает, что в данном случае отличие эмпирического коэффициента корреляции от нуля может быть объяснено только влиянием случайных факторов, и это значение не является достаточным для признания факта существования линейной зависимости между угаром кремния и выходом стали.

Заметим, что в этом случае «порогом значимости» является значение

$$\hat{r} = \sqrt{\frac{t_{\alpha/2;n-2}{}^2}{t_{\alpha/2;n-2}{}^2 + n - 2}} = \sqrt{\frac{2{,}18^2}{2{,}18^2 + 14 - 2}} = 0{,}533 .$$

### 6.4 Оценка тесноты связи при использовании нелинейных регрессионных моделей

Пусть для описания зависимости между исследуемыми случайными величинами на основании имеющихся выборочных данных построено эмпирическое уравнение регрессии $\bar{y}(x) = \varphi(x)$ (произвольного вида). В общем случае, для оценки тесноты связи между переменными *X* и *Y*, описываемой уравнением $\bar{y}(x) = \varphi(x)$, можно использовать так называемый коэффициент детерминации. Вычисление этого коэффициента основано на упоминавшемся в подразд. 5.6 разложении общей суммы квадратов отклонений $S^2_{\text{общ}}$ выборочных значений зависимой переменной $y_i$ от среднего значения $\bar{y}$ на две составляющие: $S^2_{\text{регр}}$ и $S^2_{\text{ост}}$. Коэффициент детерминации вычисляется по формуле $\hat{R}^2 = \frac{S^2_{\text{регр}}}{S^2_{\text{общ}}}$ и указывает, какая часть общего рассеяния переменной *Y* может быть объяснена зависимостью от переменной *X* на основании построенного уравнения регрессии.



Свойства коэффициента детерминации $\hat{R}^2$:

1 Возможные значения этого коэффициента принадлежат отрезку [0; 1]: $0 \le \hat{R}^2 \le 1$.

2 Если $\hat{R}^2 = 0$, то между исследуемыми величинами отсутствует зависимость, описываемая эмпирическим уравнением регрессии $\bar{y}(x) = \varphi(x)$.

3 Если $\hat{R}^2 = 1$, то между переменными *X* и *Y* существует функциональная зависимость предполагаемого вида.

4 Чем ближе значение коэффициента детерминации к единице, тем интенсивнее зависимость между величинами *X* и *Y,* описываемая соотношением $\bar{y}(x) = \varphi(x)$.

5 Значение коэффициента детерминации $\hat{R}^2$ не зависит от выбора начала отсчета и единиц измерения исследуемых величин *X* и *Y*.

**Замечание –** Доказано, что при использовании линейных регрессионных моделей $\hat{R}^2 = \hat{r}^2$.

Для проверки значимости $\hat{R}^2$, т.е. для проверки гипотезы об отсутствии между изучаемыми переменными зависимости предполагаемого вида ($H_0 : \hat{R}^2 = 0$) против альтернативной гипотезы $H_1 : \hat{R}^2 >> 0$ используется статистика

$$F = \hat{R}^2 \frac{n-m}{(m-1)\left(1-\hat{R}^2\right)},$$

имеющая распределение Фишера с $\nu_1 = m - 1$ и $\nu_2 = n - m$ степенями свободы. Здесь *m* – число неизвестных параметров построенного уравнения регрессии $\beta_i$. Вычисленное значение статистики *F* сравнивается с критическим значением $F_{\alpha,\,\nu_1,\,\nu_2}$, полученным по таблицам квантилей распределения Фишера ( см. приложение В) в зависимости от значений $\alpha$, $\nu_1$ и $\nu_2$.

Если $F \ge F_{\alpha,\,\nu_1,\,\nu_2}$, то проверяемая гипотеза об отсутствии между переменными *X* и *Y* зависимости предполагаемого вида отклоняется. Значение коэффициента детерминации признается значимым при заданном уровне значимости $\alpha$ и эмпирическое уравнение регрессии может использоваться для описания изучаемого явления.

**Пример 6.3** (см. пример 5.2). Оценим с помощью коэффициента детерминации тесноту зависимости себестоимости выпускаемой продукции от объемов производства, описываемую уравнением гиперболической регрессии $\bar{y}(x) = \frac{0{,}88}{x} + 8{,}27$.

Вычисляя значения сумм

$$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2 = 86{,}5722; \quad S^2_{\text{общ}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 89{,}145,$$

получим: $\hat{R}^2{}_{\text{гиперб}} = \dfrac{S^2_{\text{регр}}}{S^2_{\text{общ}}} = \dfrac{86{,}5722}{89{,}145} \approx 0{,}97$.

То есть данная гиперболическая регрессионная модель описывает 97 % общего рассеяния значений переменной $Y$ от среднего значения $\bar{y}$.

Для сравнения вычислим значение коэффициента детерминации для построенной на основании этого же набора опытных данных линейной модели $\bar{y}(x) = -7{,}26x + 15{,}51$.

В этом случае

$$S^2_{\text{регр}} = \sum_{i=1}^{n} (\bar{y}(x_i) - \bar{y})^2 = 67{,}982; \; \hat{R}^2_{\text{лин}} = \frac{S^2_{\text{регр}}}{S^2_{\text{общ}}} = \frac{67{,}982}{89{,}145} \approx 0{,}76.$$

Это означает, что на основании регрессионной модели линейного типа можно объяснить только 76 % наблюдаемого рассеяния значений переменной $Y$.

Проверим значимость $R^2{}_{\text{гиперб}}$ и $R^2{}_{\text{лин}}$. В данном случае $m = 2$, поскольку на основании выборочных данных мы оценивали значения двух параметров ($\hat{\beta}_0$ и $\hat{\beta}_1$).

$$F_{\text{гиперб}} = \hat{R}^2{}_{\text{гиперб}} \frac{n-m}{(m-1)(1-\hat{R}^2{}_{\text{гиперб}})} = 0{,}97 \frac{10-2}{(2-1)(1-0{,}97)} \approx 258{,}67;$$

$$F_{\text{лин}} = \hat{R}^2{}_{\text{лин}} \frac{n-m}{(m-1)(1-\hat{R}^2{}_{\text{лин}})} = 0{,}76 \frac{10-2}{(2-1)(1-0{,}76)} \approx 25{,}33.$$

По таблицам квантилей распределения Фишера (приложение В) определим для $\alpha = 0{,}05$, $\nu_1 = 2-1 = 1$, $\nu_2 = 10-2 = 8$ $\quad F_{0{,}05,\,1,\,8} = 5{,}32$.

Поскольку $F_{\text{гиперб}} > F_{0{,}05,\,1,\,8}$ и $F_{\text{лин}} > F_{0{,}05,\,1,\,8}$, значения коэффициентов детерминации, оценивающих тесноту зависимости гиперболического и линейного видов, значимо отличаются от нуля и построенные уравнения могут быть использованы для описания изучаемой зависимости.



*ПРИЛОЖЕНИЕ А*

(*справочное*)

**Критические точки распределения $\chi^2$**

| Степени свободы ν | Уровень значимости α | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0,001 | 0,01 | 0,025 | 0,05 | 0,1 | 0,9 | 0,95 | 0,975 | 0,99 |
| 1 | 10,827 | 6,635 | 5,024 | 3,841 | 2,706 | 0,016 | 0,0039 | 0,00098 | 0,00016 |
| 2 | 13,815 | 9,210 | 7,378 | 5,991 | 4,605 | 0,211 | 0,103 | 0,051 | 0,020 |
| 3 | 16,266 | 11,345 | 9,348 | 7,815 | 6,251 | 0,584 | 0,352 | 0,216 | 0,115 |
| 4 | 18,466 | 13,277 | 11,143 | 9,488 | 7,779 | 1,064 | 0,711 | 0,484 | 0,297 |
| 5 | 20,515 | 15,086 | 12,832 | 11,070 | 9,236 | 1,610 | 1,145 | 0,831 | 0,554 |
| 6 | 22,457 | 16,812 | 14,449 | 12,592 | 10,645 | 2,204 | 1,635 | 1,237 | 0,872 |
| 7 | 24,321 | 18,475 | 16,013 | 14,067 | 12,017 | 2,833 | 2,167 | 1,690 | 1,239 |
| 8 | 26,124 | 20,090 | 17,535 | 15,507 | 13,362 | 3,490 | 2,733 | 2,180 | 1,647 |
| 9 | 27,877 | 21,666 | 19,023 | 16,919 | 14,684 | 4,168 | 3,325 | 2,700 | 2,088 |
| 10 | 29,588 | 23,209 | 20,483 | 18,307 | 15,987 | 4,865 | 3,940 | 3,247 | 2,558 |
| 11 | 31,264 | 24,725 | 21,920 | 19,675 | 17,275 | 5,5748 | 4,575 | 3,816 | 3,053 |
| 12 | 32,909 | 26,217 | 23,337 | 21,026 | 18,549 | 6,304 | 5,226 | 4,404 | 3,571 |
| 13 | 34,527 | 27,688 | 24,736 | 22,362 | 19,812 | 7,041 | 5,892 | 5,009 | 4,107 |
| 14 | 36,124 | 29,141 | 26,119 | 23,685 | 21,064 | 7,790 | 6,571 | 5,629 | 4,660 |
| 15 | 37,698 | 30,578 | 27,488 | 24,996 | 22,307 | 8,547 | 7,261 | 6,262 | 5,229 |
| 16 | 39,252 | 32,000 | 28,845 | 26,296 | 23,542 | 9,312 | 7,962 | 6,908 | 5,812 |
| 17 | 40,791 | 33,409 | 30,191 | 27,587 | 24,769 | 10,085 | 8,672 | 7,564 | 6,408 |
| 18 | 42,312 | 34,805 | 31,526 | 28,869 | 25,989 | 10,865 | 9,390 | 8,231 | 7,015 |
| 19 | 43,819 | 36,191 | 32,852 | 30,144 | 27,204 | 11,651 | 10,117 | 8,907 | 7,633 |
| 20 | 45,314 | 37,566 | 34,170 | 31,410 | 28,412 | 12,443 | 10,851 | 9,591 | 8,260 |
| 21 | 46,796 | 38,932 | 35,479 | 32,671 | 29,615 | 13,240 | 11,591 | 10,283 | 8,897 |
| 22 | 48,268 | 40,289 | 36,781 | 33,924 | 30,813 | 14,041 | 12,338 | 10,982 | 9,542 |
| 23 | 49,728 | 41,638 | 38,076 | 35,172 | 32,007 | 14,848 | 13,091 | 11,689 | 10,196 |
| 24 | 51,179 | 42,980 | 39,364 | 36,415 | 33,196 | 15,659 | 13,848 | 12,401 | 10,856 |
| 25 | 52,619 | 44,314 | 40,646 | 37,652 | 34,382 | 16,473 | 14,611 | 13,120 | 11,524 |
| 26 | 54,051 | 45,642 | 41,923 | 38,885 | 35,563 | 17,292 | 15,379 | 13,844 | 12,198 |
| 27 | 55,475 | 46,963 | 43,195 | 40,113 | 36,741 | 18,114 | 16,151 | 14,573 | 12,878 |
| 28 | 56,892 | 48,278 | 44,461 | 41,337 | 37,916 | 18,939 | 16,928 | 15,308 | 13,565 |
| 29 | 58,301 | 49,588 | 45,722 | 42,557 | 39,087 | 19,768 | 17,708 | 16,047 | 14,256 |
| 30 | 59,702 | 50,892 | 46,979 | 43,773 | 40,256 | 20,599 | 18,493 | 16,791 | 14,953 |
| 31 | 61,098 | 52,191 | 48,232 | 44,985 | 41,422 | 21,434 | 19,281 | 17,539 | 15,655 |
| 32 | 62,487 | 53,486 | 49,480 | 46,194 | 42,585 | 22,271 | 20,072 | 18,291 | 16,362 |
| 33 | 63,869 | 54,775 | 50,725 | 47,400 | 43,745 | 23,110 | 20,867 | 19,047 | 17,073 |
| 34 | 65,247 | 56,061 | 51,966 | 48,602 | 44,903 | 23,952 | 21,664 | 19,806 | 17,789 |
| 35 | 66,619 | 57,342 | 53,203 | 49,802 | 46,059 | 24,797 | 22,465 | 20,569 | 18,509 |
| 36 | 67,985 | 58,619 | 54,437 | 50,998 | 47,212 | 25,643 | 23,269 | 21,336 | 19,233 |
| 37 | 69,348 | 59,893 | 55,668 | 52,192 | 48,363 | 26,492 | 24,075 | 22,106 | 19,960 |
| 38 | 70,704 | 61,162 | 56,895 | 53,384 | 49,513 | 27,343 | 24,884 | 22,878 | 20,691 |
| 39 | 72,055 | 62,428 | 58,120 | 54,572 | 50,660 | 28,196 | 25,695 | 23,654 | 21,426 |
| 40 | 73,403 | 63,691 | 59,342 | 55,758 | 51,805 | 29,051 | 26,509 | 24,433 | 22,164 |

*ПРИЛОЖЕНИЕ Б*

(*справочное*)

**Критические точки распределения Стьюдента**

| Степени свободы ν | Уровень значимости α (односторонняя критическая область) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0,2 | 0,1 | 0,05 | 0,025 | 0,01 | 0,005 | 0,001 | 0,0005 |
| 1 | 1,376 | 3,078 | 6,314 | 12,706 | 31,821 | 63,656 | 318,29 | 636,58 |
| 2 | 1,061 | 1,886 | 2,920 | 4,303 | 6,965 | 9,925 | 22,328 | 31,600 |
| 3 | 0,978 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 | 10,214 | 12,924 |
| 4 | 0,941 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 | 7,173 | 8,610 |
| 5 | 0,920 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 | 5,894 | 6,869 |
| 6 | 0,906 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 | 5,208 | 5,959 |
| 7 | 0,896 | 1,415 | 1,895 | 2,365 | 2,998 | 3,499 | 4,785 | 5,408 |
| 8 | 0,889 | 1,397 | 1,860 | 2,306 | 2,896 | 3,355 | 4,501 | 5,041 |
| 9 | 0,883 | 1,383 | 1,833 | 2,262 | 2,821 | 3,250 | 4,297 | 4,781 |
| 10 | 0,879 | 1,372 | 1,812 | 2,228 | 2,764 | 3,169 | 4,144 | 4,587 |
| 11 | 0,876 | 1,363 | 1,796 | 2,201 | 2,718 | 3,106 | 4,025 | 4,437 |
| 12 | 0,873 | 1,356 | 1,782 | 2,179 | 2,681 | 3,055 | 3,930 | 4,318 |
| 13 | 0,870 | 1,350 | 1,771 | 2,160 | 2,650 | 3,012 | 3,852 | 4,221 |
| 14 | 0,868 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 | 3,787 | 4,140 |
| 15 | 0,866 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 | 3,733 | 4,073 |
| 16 | 0,865 | 1,337 | 1,746 | 2,120 | 2,583 | 2,921 | 3,686 | 4,015 |
| 17 | 0,863 | 1,333 | 1,740 | 2,110 | 2,567 | 2,898 | 3,646 | 3,965 |
| 18 | 0,862 | 1,330 | 1,734 | 2,101 | 2,552 | 2,878 | 3,610 | 3,922 |
| 19 | 0,861 | 1,328 | 1,729 | 2,093 | 2,539 | 2,861 | 3,579 | 3,883 |
| 20 | 0,860 | 1,325 | 1,725 | 2,086 | 2,528 | 2,845 | 3,552 | 3,850 |
| 21 | 0,859 | 1,323 | 1,721 | 2,080 | 2,518 | 2,831 | 3,527 | 3,819 |
| 22 | 0,858 | 1,321 | 1,717 | 2,074 | 2,508 | 2,819 | 3,505 | 3,792 |
| 23 | 0,858 | 1,319 | 1,714 | 2,069 | 2,500 | 2,807 | 3,485 | 3,768 |
| 24 | 0,857 | 1,318 | 1,711 | 2,064 | 2,492 | 2,797 | 3,467 | 3,745 |
| 25 | 0,856 | 1,316 | 1,708 | 2,060 | 2,485 | 2,787 | 3,450 | 3,725 |
| 26 | 0,856 | 1,315 | 1,706 | 2,056 | 2,479 | 2,779 | 3,435 | 3,707 |
| 27 | 0,855 | 1,314 | 1,703 | 2,052 | 2,473 | 2,771 | 3,421 | 3,689 |
| 28 | 0,855 | 1,313 | 1,701 | 2,048 | 2,467 | 2,763 | 3,408 | 3,674 |
| 29 | 0,854 | 1,311 | 1,699 | 2,045 | 2,462 | 2,756 | 3,396 | 3,660 |
| 30 | 0,854 | 1,310 | 1,697 | 2,042 | 2,457 | 2,750 | 3,385 | 3,646 |
| 40 | 0,851 | 1,303 | 1,684 | 2,021 | 2,423 | 2,704 | 3,307 | 3,551 |
| 50 | 0,849 | 1,299 | 1,676 | 2,009 | 2,403 | 2,678 | 3,261 | 3,496 |
| 60 | 0,848 | 1,296 | 1,671 | 2,000 | 2,390 | 2,660 | 3,232 | 3,460 |
| 70 | 0,847 | 1,294 | 1,667 | 1,994 | 2,381 | 2,648 | 3,211 | 3,435 |
| 80 | 0,846 | 1,292 | 1,664 | 1,990 | 2,374 | 2,639 | 3,195 | 3,416 |
| 90 | 0,846 | 1,291 | 1,662 | 1,987 | 2,368 | 2,632 | 3,183 | 3,402 |
| 100 | 0,845 | 1,290 | 1,660 | 1,984 | 2,364 | 2,626 | 3,174 | 3,390 |
| 150 | 0,844 | 1,287 | 1,655 | 1,976 | 2,351 | 2,609 | 3,145 | 3,357 |
| 200 | 0,843 | 1,286 | 1,653 | 1,972 | 2,345 | 2,601 | 3,131 | 3,340 |
| 500 | 0,842 | 1,283 | 1,648 | 1,965 | 2,334 | 2,586 | 3,107 | 3,310 |
| ∞ | 0,842 | 1,282 | 1,645 | 1,960 | 2,326 | 2,576 | 3,090 | 3,290 |
| Степени свободы ν | 0,4 | 0,2 | 0,1 | 0,05 | 0,02 | 0,01 | 0,002 | 0,001 |
| | Уровень значимости α (двусторонняя критическая область) | | | | | | | |



*ПРИЛОЖЕНИЕ В*

(*справочное*)

**Критические точки распределения Фишера**

(число степеней свободы бóльшей дисперсии – $\nu_1$, меньшей – $\nu_2$)

| Уровень значимости α = 0,05 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\nu_2$ | $\nu_1$ | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 12 | 16 | 24 | 50 | ∞ |
| 1 | 161,45 | 199,50 | 215,71 | 224,58 | 230,16 | 233,99 | 238,88 | 243,90 | 246,47 | 249,05 | 251,77 | 254,31 |
| 2 | 18,513 | 19,000 | 19,164 | 19,247 | 19,296 | 19,329 | 19,371 | 19,412 | 19,433 | 19,454 | 19,476 | 19,496 |
| 3 | 10,128 | 9,552 | 9,277 | 9,117 | 9,013 | 8,941 | 8,845 | 8,745 | 8,692 | 8,638 | 8,581 | 8,526 |
| 4 | 7,709 | 6,944 | 6,591 | 6,388 | 6,256 | 6,163 | 6,041 | 5,912 | 5,844 | 5,774 | 5,699 | 5,628 |
| 5 | 6,608 | 5,786 | 5,409 | 5,192 | 5,050 | 4,950 | 4,818 | 4,678 | 4,604 | 4,527 | 4,444 | 4,365 |
| 6 | 5,987 | 5,143 | 4,757 | 4,534 | 4,387 | 4,284 | 4,147 | 4,000 | 3,922 | 3,841 | 3,754 | 3,669 |
| 7 | 5,591 | 4,737 | 4,347 | 4,120 | 3,972 | 3,866 | 3,726 | 3,575 | 3,494 | 3,410 | 3,319 | 3,230 |
| 8 | 5,318 | 4,459 | 4,066 | 3,838 | 3,688 | 3,581 | 3,438 | 3,284 | 3,202 | 3,115 | 3,020 | 2,928 |
| 9 | 5,117 | 4,256 | 3,863 | 3,633 | 3,482 | 3,374 | 3,230 | 3,073 | 2,989 | 2,900 | 2,803 | 2,707 |
| 10 | 4,965 | 4,103 | 3,708 | 3,478 | 3,326 | 3,217 | 3,072 | 2,913 | 2,828 | 2,737 | 2,637 | 2,538 |
| 11 | 4,844 | 3,982 | 3,587 | 3,357 | 3,204 | 3,095 | 2,948 | 2,788 | 2,701 | 2,609 | 2,507 | 2,404 |
| 12 | 4,747 | 3,885 | 3,490 | 3,259 | 3,106 | 2,996 | 2,849 | 2,687 | 2,599 | 2,505 | 2,401 | 2,296 |
| 13 | 4,667 | 3,806 | 3,411 | 3,179 | 3,025 | 2,915 | 2,767 | 2,604 | 2,515 | 2,420 | 2,314 | 2,206 |
| 14 | 4,600 | 3,739 | 3,344 | 3,112 | 2,958 | 2,848 | 2,699 | 2,534 | 2,445 | 2,349 | 2,241 | 2,131 |
| 15 | 4,543 | 3,682 | 3,287 | 3,056 | 2,901 | 2,790 | 2,641 | 2,475 | 2,385 | 2,288 | 2,178 | 2,066 |
| 16 | 4,494 | 3,634 | 3,239 | 3,007 | 2,852 | 2,741 | 2,591 | 2,425 | 2,333 | 2,235 | 2,124 | 2,010 |
| 17 | 4,451 | 3,592 | 3,197 | 2,965 | 2,810 | 2,699 | 2,548 | 2,381 | 2,289 | 2,190 | 2,077 | 1,960 |
| 18 | 4,414 | 3,555 | 3,160 | 2,928 | 2,773 | 2,661 | 2,510 | 2,342 | 2,250 | 2,150 | 2,035 | 1,917 |
| 19 | 4,381 | 3,522 | 3,127 | 2,895 | 2,740 | 2,628 | 2,477 | 2,308 | 2,215 | 2,114 | 1,999 | 1,878 |
| 20 | 4,351 | 3,493 | 3,098 | 2,866 | 2,711 | 2,599 | 2,447 | 2,278 | 2,184 | 2,082 | 1,966 | 1,843 |
| 21 | 4,325 | 3,467 | 3,072 | 2,840 | 2,685 | 2,573 | 2,420 | 2,250 | 2,156 | 2,054 | 1,936 | 1,812 |
| 22 | 4,301 | 3,443 | 3,049 | 2,817 | 2,661 | 2,549 | 2,397 | 2,226 | 2,131 | 2,028 | 1,909 | 1,783 |
| 23 | 4,279 | 3,422 | 3,028 | 2,796 | 2,640 | 2,528 | 2,375 | 2,204 | 2,109 | 2,005 | 1,885 | 1,757 |
| 24 | 4,260 | 3,403 | 3,009 | 2,776 | 2,621 | 2,508 | 2,355 | 2,183 | 2,088 | 1,984 | 1,863 | 1,733 |
| 25 | 4,242 | 3,385 | 2,991 | 2,759 | 2,603 | 2,490 | 2,337 | 2,165 | 2,069 | 1,964 | 1,842 | 1,711 |
| 26 | 4,225 | 3,369 | 2,975 | 2,743 | 2,587 | 2,474 | 2,321 | 2,148 | 2,052 | 1,946 | 1,823 | 1,691 |
| 27 | 4,210 | 3,354 | 2,960 | 2,728 | 2,572 | 2,459 | 2,305 | 2,132 | 2,036 | 1,930 | 1,806 | 1,672 |
| 28 | 4,196 | 3,340 | 2,947 | 2,714 | 2,558 | 2,445 | 2,291 | 2,118 | 2,021 | 1,915 | 1,790 | 1,654 |
| 29 | 4,183 | 3,328 | 2,934 | 2,701 | 2,545 | 2,432 | 2,278 | 2,104 | 2,007 | 1,901 | 1,775 | 1,638 |
| 30 | 4,171 | 3,316 | 2,922 | 2,690 | 2,534 | 2,421 | 2,266 | 2,092 | 1,995 | 1,887 | 1,761 | 1,622 |
| 40 | 4,085 | 3,232 | 2,839 | 2,606 | 2,449 | 2,336 | 2,180 | 2,003 | 1,904 | 1,793 | 1,660 | 1,509 |
| 50 | 4,034 | 3,183 | 2,790 | 2,557 | 2,400 | 2,286 | 2,130 | 1,952 | 1,850 | 1,737 | 1,599 | 1,438 |
| 60 | 4,001 | 3,150 | 2,758 | 2,525 | 2,368 | 2,254 | 2,097 | 1,917 | 1,815 | 1,700 | 1,559 | 1,389 |
| 70 | 3,978 | 3,128 | 2,736 | 2,503 | 2,346 | 2,231 | 2,074 | 1,893 | 1,790 | 1,674 | 1,530 | 1,353 |
| 80 | 3,960 | 3,111 | 2,719 | 2,486 | 2,329 | 2,214 | 2,056 | 1,875 | 1,772 | 1,654 | 1,508 | 1,325 |
| 90 | 3,947 | 3,098 | 2,706 | 2,473 | 2,316 | 2,201 | 2,043 | 1,861 | 1,757 | 1,639 | 1,491 | 1,302 |
| 100 | 3,936 | 3,087 | 2,696 | 2,463 | 2,305 | 2,191 | 2,032 | 1,850 | 1,746 | 1,627 | 1,477 | 1,283 |
| 150 | 3,904 | 3,056 | 2,665 | 2,432 | 2,274 | 2,160 | 2,001 | 1,817 | 1,711 | 1,590 | 1,436 | 1,223 |
| 200 | 3,888 | 3,041 | 2,650 | 2,417 | 2,259 | 2,144 | 1,985 | 1,801 | 1,694 | 1,572 | 1,415 | 1,189 |
| 500 | 3,860 | 3,014 | 2,623 | 2,390 | 2,232 | 2,117 | 1,957 | 1,772 | 1,664 | 1,539 | 1,376 | 1,113 |
| ∞ | 3,841 | 2,996 | 2,605 | 2,372 | 2,214 | 2,099 | 1,938 | 1,752 | 1,644 | 1,517 | 1,350 | 1,000 |

*ПРИЛОЖЕНИЕ Г*

(*справочное*)

**Нижние $\gamma_1$ и верхние $\gamma_2$ границы доверительного интервала $I_\sigma=(\gamma_1 s, \gamma_1 s)$**

| $\nu = n-1$ | Доверительная вероятность $P = 1 - \alpha$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0,99 | | 0,98 | | 0,95 | | 0,9 | |
| | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_1$ | $\gamma_2$ |
| 1 | 0,356 | 159 | 0,388 | 79,8 | 0,446 | 31,9 | 0,510 | 15,9 |
| 2 | 0,434 | 14,1 | 0,466 | 9,97 | 0,521 | 6,28 | 0,578 | 4,40 |
| 3 | 0,483 | 6,47 | 0,514 | 5,11 | 0,566 | 3,73 | 0,620 | 2,92 |
| 4 | 0,519 | 4,39 | 0,549 | 3,67 | 0,599 | 2,87 | 0,649 | 2,37 |
| 5 | 0,546 | 3,48 | 0,576 | 3,00 | 0,624 | 2,45 | 0,672 | 2,090 |
| 6 | 0,569 | 2,98 | 0,597 | 2,62 | 0,644 | 2,202 | 0,690 | 1,916 |
| 7 | 0,588 | 2,66 | 0,616 | 2,377 | 0,661 | 2,035 | 0,705 | 1,797 |
| 8 | 0,604 | 2,440 | 0,631 | 2,205 | 0,675 | 1,916 | 0,718 | 1,711 |
| 9 | 0,618 | 2,277 | 0,644 | 2,076 | 0,688 | 1,826 | 0,729 | 1,645 |
| 10 | 0,630 | 2,154 | 0,656 | 1,977 | 0,699 | 1,755 | 0,739 | 1,593 |
| 11 | 0,641 | 2,056 | 0,667 | 1,898 | 0,708 | 1,698 | 0,748 | 1,550 |
| 12 | 0,651 | 1,976 | 0,677 | 1,833 | 0,717 | 1,651 | 0,755 | 1,515 |
| 13 | 0,660 | 1,910 | 0,685 | 1,779 | 0,725 | 1,611 | 0,762 | 1,485 |
| 14 | 0,669 | 1,854 | 0,693 | 1,733 | 0,732 | 1,577 | 0,769 | 1,460 |
| 15 | 0,676 | 1,806 | 0,700 | 1,694 | 0,7398 | 1,548 | 0,775 | 1,437 |
| 16 | 0,683 | 1,764 | 0,707 | 1,659 | 0,745 | 1,522 | 0,780 | 1,418 |
| 17 | 0,690 | 1,727 | 0,713 | 1,629 | 0,750 | 1,499 | 0,785 | 1,400 |
| 18 | 0,696 | 1,695 | 0,719 | 1,602 | 0,756 | 1,479 | 0,790 | 1,385 |
| 19 | 0,702 | 1,666 | 0,725 | 1,578 | 0,760 | 1,460 | 0,794 | 1,370 |
| 20 | 0,707 | 1,64 | 0,730 | 1,556 | 0,765 | 1,444 | 0,798 | 1,358 |
| 21 | 0,712 | 1,617 | 0,734 | 1,536 | 0,769 | 1,429 | 0,802 | 1,346 |
| 22 | 0,717 | 1,595 | 0,739 | 1,519 | 0,773 | 1,416 | 0,805 | 1,335 |
| 23 | 0,722 | 1,576 | 0,743 | 1,502 | 0,777 | 1,402 | 0,809 | 1,3,26 |
| 24 | 0,726 | 1,558 | 0,747 | 1,487 | 0,781 | 1,391 | 0,812 | 1,316 |
| 25 | 0,730 | 1,541 | 0,751 | 1,473 | 0,784 | 1,380 | 0,815 | 1,308 |
| 26 | 0,734 | 1,526 | 0,755 | 1,460 | 0,788 | 1,371 | 0,818 | 1,300 |
| 27 | 0,737 | 1,512 | 0,758 | 1,448 | 0,791 | 1,361 | 0,820 | 1,293 |
| 28 | 0,741 | 1,499 | 0,762 | 1,436 | 0,794 | 1,352 | 0,823 | 1,2866 |
| 29 | 0,744 | 1,487 | 0,765 | 1,426 | 0,796 | 1,344 | 0,825 | 1,279 |
| 30 | 0,748 | 1,475 | 0,768 | 1,417 | 0,799 | 1,337 | 0,828 | 1,274 |
| 40 | 0,774 | 1,390 | 0,792 | 1,344 | 0,821 | 1,279 | 0,847 | 1,228 |
| 50 | 0,793 | 1,336 | 0,810 | 1,297 | 0,837 | 1,243 | 0,8661 | 1,199 |
| 60 | 0,808 | 1,299 | 0,824 | 1,265 | 0,849 | 1,217 | 0,871 | 1,179 |
| 70 | 0,820 | 1,272 | 0,835 | 1,241 | 0,858 | 1,198 | 0,879 | 1,163 |
| 80 | 0,829 | 1,250 | 0,844 | 1,222 | 0,866 | 1,83 | 0,886 | 1,151 |
| 90 | 0,838 | 1,233 | 0,852 | 1,207 | 0,873 | 1,171 | 0,892 | 1,141 |
| 100 | 0,845 | 1,219 | 0,858 | 1,195 | 0,878 | 1,161 | 0,897 | 1,133 |
| 200 | 0,887 | 1,15 | 0,897 | 1,13 | 0,912 | 1,11 | 0,925 | 1,09 |



*ПРИЛОЖЕНИЕ Д*

(*обязательное*)

**Рабочая программа по дисциплине**
**«Теория вероятностей и математическая статистика»**

**1 Случайные события и их вероятности**

1.1 Предмет и задачи ТВ и МС. Вероятностный эксперимент. Пространство элементарных событий. Классификация событий. Операции над событиями. Противоположные события. Несовместные события.

1.2 Вероятность. Аксиомы теории вероятностей. Свойства вероятностей. Методы определения вероятностей (классический, статистический). Элементы комбинаторики.

1.3 Теоремы сложения вероятностей. Условная вероятность. Теоремы умножения вероятностей. Независимость событий.

1.4 Формула полной вероятности. Формула Байеса.

1.5 Последовательности независимых испытаний. Формулы Бернулли, Пуассона, Муавра–Лапласа.

**2 Случайные величины**

2.1 Понятие случайной величины. Дискретные и непрерывные случайные величины.

2.2 Закон распределения случайной величины. Формы задания закона распределения дискретных и непрерывных случайных величин (ряд распределения, функция распределения, функция плотности распределения вероятностей).

2.3 Числовые характеристики дискретных и непрерывных случайных величин (математическое ожидание, мода, медиана, дисперсия, среднее квадратическое отклонение, коэффициент асимметрии, коэффициент эксцесса).

2.4 Законы распределения случайных величин, наиболее часто встречающиеся на практике (биномиальное, Пуассона, геометрическое, равномерное, экспоненциальное, нормальное распределения).

2.5 Основные законы распределения случайных величин, используемые в математической статистике («хи-квадрат», *t*–распределение Стьюдента, *F*–распределение Фишера).

2.6 Системы случайных величин. Совместные распределения системы случайных величин. Условные математические ожидания. Независимые случайные величины. Ковариация. Коэффициент корреляции.

**3 Математическая статистика**

3.1 Предмет и задачи математической статистики. Выборочный метод (генеральная и выборочная совокупности, репрезентативность выборки).

3.2 Первичная обработка статистических данных (построение вариационного ряда, статистического закона распределения, эмпирической функции распределения).

3.3 Статистическое оценивание параметров распределения. Постановка задачи. Точечные оценки числовых характеристик. Понятие об интервальном оценивании. Построение доверительных интервалов для параметров нормально распределенной случайной величины.

3.4 Проверка статистических гипотез. Статистический критерий. Ошибки, допускаемые при проверке гипотез. Непараметрические гипотезы. Применение критерия «хи-квадрат» Пирсона для проверки гипотезы о виде закона распределения одномерной случайной величины.

3.5 Элементы регрессионного анализа. Построение эмпирического уравнения регрессии. Проверка адекватности построенного уравнения регрессии выборочным данным.

3.6 Элементы корреляционного анализа. Использование эмпирических коэффициентов корреляции и детерминации для оценки тесноты зависимости между исследуемыми переменными. Проверка значимости этих коэффициентов.

## СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ


1 **Гмурман, В. Е**. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высш. шк., 1998. – 478 с.

2 **Гмурман, В. Е.** Руководство к решению задач по теории вероятностей и математической статистике / В. Е. Гмурман. – М. : Высш. шк., 1998. – 399 с.

3 **Вентцель, Е. С.** Теория вероятностей и ее инженерные приложения / Е. С. Вентцель, Л. А. Овчаров. – М. : Наука, 1988. – 480 с.

4 **Герасимович, А. И.** Математическая статистика / А. И. Герасимович. – Минск : Выш. шк., 1983. – 279 с.

5 **Мацкевич, И. П.** Высшая математика: теория вероятностей и математическая статистика / И. П. Мацкевич, Г. П. Свирид. – Минск : Выш. шк., 1993. – 268 с.

6 **Смирнов, Н. В.** Курс теории вероятностей и математической статистики для технических приложений / Н. В. Смирнов, И. В. Дунин-Барковский. – М.: Наука, 1969. – 511 с.

7 **Четыркин, Е. М.** Вероятность и статистика / Е. М. Четыркин, И. Л. Калихман. – М. : Финансы и статистика, 1982. – 319 с.

8 **Айвазян, С. А.** Прикладная статистика: основы моделирования и первичная обработка данных : справочное издание / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1983. – 471 с.

9 **Сазонова, Е. Л.** Теория вероятностей и математическая статистика. Ч. 1. Теория вероятностей : пособие для студентов факультета безотрывного обучения / Е. Л. Сазонова ; под ред. В. С. Серегиной. – Гомель : БелГУТ, 2000. – 95 с.




# ОГЛАВЛЕНИЕ

Учебное издание

С а з о н о в а Елена Леонидовна

**ТЕОРИЯ ВЕРОЯТНОСТЕЙ**
**И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА**
**Часть 2. Математическая статистика**

Учебно-методическое пособие для студентов факультета безотрывного обучения

Редактор Т. М. Р и з е в с к а я
Технический редактор В. Н. К у ч е р о в а
Компьютерный набор и вёрстка кафедры «Прикладная математика»

Подписано в печать
Формат 60 × 84 $^1/_{16}$. Бумага . Гарнитура Таймс.
Усл. печ. л. Уч.-изд. л. . Тираж экз.
Зак. № . Изд. №