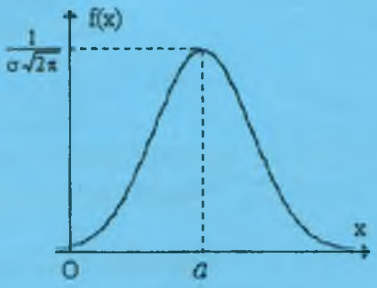
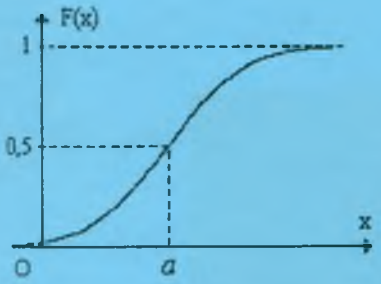
Министерство образования Республики Беларусь
БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ ПОЛИТЕХНИЧЕСКАЯ АКАДЕМИЯ

Кафедра «Высшая математика № 3»



В.В.Веременюк
В.В.Кожушко
О.А.Мороз



СТАТИСТИЧЕСКАЯ ОБРАБОТКА ВЫБОРКИ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Минск 2002

Министерство образования Республики Беларусь
БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ ПОЛИТЕХНИЧЕСКАЯ АКАДЕМИЯ

Кафедра «Высшая математика № 3»

В.В.Веременюк
В.В.Кожушко
О.А.Мороз

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ВЫБОРКИ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Учебно-методическое пособие
к лабораторной работе по высшей математике
для студентов строительных специальностей

Минск 2002

ББК 22.1 я 7
В31

Рецензент Г.Л.Бахмат

Веременюк В.В.
В31 Статистическая обработка выборки значений случайной величины: Учеб.-метод. пособие по высшей математике для студ. строит. спец. / В.В.Веременюк, В.В.Кожушко, О.А.Мороз. – Мн.: БГПА, 2002. - 102 с.

ISBN 985-6529-57-3.

Учебно-методическое пособие содержит теоретический материал, необходимый для выполнения лабораторной работы «Статистическая обработка выборки значений случайной величины». 3-й и 4-й разделы издания содержат дополнительный материал из курса теории вероятностей.

УДК 519.216 (075.8)
ББК 22.1 я 7

ISBN 985-6529-57-3

© Веременюк В.В., Кожушко В.В., Мороз О.А., 2002

Введение

В повседневной жизни, технике, научных исследованиях, бизнесе, иной профессиональной деятельности мы постоянно сталкиваемся с событиями и явлениями с неопределенным исходом. Например, торговец не знает, сколько посетителей придет к нему в магазин, бизнесмен - какой будет завтра или через месяц курс доллара, студент, проводя какой-то эксперимент, не может в силу самых различных причин точно предсказать показание прибора и т.д. При этом нам постоянно приходится в подобных неопределенных, связанных со многими случайностями ситуациях принимать решения, иногда очень важные.

В быту или в несложном бизнесе мы можем принимать такие решения на основе здравого смысла, интуиции, предыдущего опыта. Здесь мы можем создать некий "запас прочности" на действие случая: скажем, выходить из дома на десять минут раньше, чтобы уже почти наверняка не опаздывать на работу, и т.п.

Однако в важных научных исследованиях, серьезном бизнесе решения должны приниматься на основе тщательного анализа имеющейся информации, быть обоснованными и доказуемыми. Для решения задач, связанных с анализом данных при наличии случайных и непредсказуемых воздействий, математиками и другими исследователями (биологами, инженерами, экономистами и т.д.) за последние двести лет был выработан мощный и гибкий арсенал методов, называемых в совокупности *математической статистикой* (а также *прикладной статистикой*, или *анализом данных*).

Эти методы позволяют выявлять закономерности на фоне случайностей, делать обоснованные выводы и прогнозы, давать оценки вероятностей их выполнения или невыполнения.

Цель работы:

1. Изучить основные понятия математической статистики, применить их к анализу полученных данных. Как правило, результаты эксперимента получаются в виде ряда значений интересующего нас признака (обозначенного через X): $x_1, x_2, \ldots, x_n$.

2. Провести первичную обработку данных, по возможности, представив их в наглядном виде, используя при этом методы описатель-

ной статистики – группировки данных, их графического представления, вычисления различных показателей, описывающих положение данных на числовой оси, степень их разброса, симметрии и т.п.

3. Изучить основы теории оценивания и проверки статистических гипотез.

4. Научиться по результатам обработки данных делать закономерные выводы о поведении и характеристиках изучаемого признака X, т.е. по выборке делать выводы о генеральной совокупности.

Содержание работы

При изучении теоретического материала по данной работе необходимо:

1. Освоить основные понятия описательной статистики, применяемые для анализа экспериментальных данных:

1) генеральную совокупность (ГС) и выборку;

2) эмпирическую функцию распределения;

3) полигон и гистограмму;

4) выборочные характеристики ГС.

2. Получить представление о теории оценивания, в том числе:

1) статистической оценке, статистике;

2) основных свойствах статистических оценок;

3) методах построения статистических оценок;

4) оценках математического ожидания $M[X]$ и дисперсиях $D[X]$.

3. Научиться определять точность статистических оценок, уметь:

1) применять точечные и интервальные оценки статистических параметров, доверительные интервалы, доверительную вероятность, уровень значимости;

2) построить доверительный интервал для оценки $M[X]$, если известна дисперсия $D[X]$, а $D[X]$ неизвестна.

4. Изучить основы проверки статистических гипотез, иметь представление о:

1) статистической гипотезе;

2) критериях согласия для проверки статистических гипотез;

3) критерии согласия хи-квадрат (χ^2).

Порядок проведения работы

1. Изучить 1-й и 2-й разделы настоящего пособия.
2. Построить для своей задачи вариационный группированный статистический ряд.
3. Построить эмпирическую функцию распределения и гистограмму.
4. Найти выборочные значения средней, дисперсии, асимметрии и эксцесса заданной выборочной совокупности.
5. Провести оценку математического ожидания и дисперсии исследуемого признака и найти доверительные интервалы для полученных оценок (доверительную вероятность принять равной $\gamma = 0{,}95$).
6. На основании анализа выборочных характеристик, эмпирической функции распределения и гистограммы выдвинуть гипотезу о характере распределения исследуемого признака (случайной величины X).
7. Используя критерий согласия хи-квадрат (χ^2), проверить выдвинутую гипотезу (уровень значимости принять равным $\alpha = 0{,}05$).
8. Провести расчеты на ПЭВМ и сравнить результаты.
9. Составить отчет о работе.

Содержание отчета по работе

Отчет по работе должен состоять из следующих пунктов:

1. Постановка задачи.
2. Результаты построения вариационного группированного статистического ряда.
3. Графическая интерпретация данных.
4. Результаты вычисления и анализа выборочных характеристик.
5. Построение доверительных интервалов для оценок $M[X]$ и $D[X]$.
6. Анализ проверки гипотезы о законе распределения X.
7. Результаты расчета на ПЭВМ.
8. Основные выводы.

1. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ВЫБОРКИ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

1.1. Теория вероятностей и математическая статистика

Из основ теории вероятностей известно, что построение вероятностной модели того или иного случайного эксперимента (явления) начинается с введения понятия пространства элементарных исходов:

$$\Omega = \{\omega_1, \omega_2, ..., \omega_N\}.$$

Каждому элементарному исходу ω_i ставится в соответствие некоторое число $p(\omega_i) \in [0;1]$, называемое *вероятностью исхода*, причем для вероятностей элементарных исходов должна выполняться аксиома

$$\sum_{i=1}^{N} p(\omega_i) = 1.$$

Далее дается определение события A как подмножества множества Ω, и вероятность любого события A можно вычислить по формуле

$$P(A) = \sum_{\omega_i \in A} p(\omega_i).$$

Весьма существенно, что вероятности элементарных событий считаются заданными. В частности, во многих задачах, рассматриваемых теорией вероятностей, нахождение этих вероятностей основано на некоторых общих соображениях симметрии.

Однако в повседневной жизни, в технике, науке, естествознании, экономике такой симметрией, как при игре в карты или "Орлянку", элементарные события не обладают, и вычислить вероятность этих событий заранее (a priori) невозможно. Здесь остается, пожалуй, единственный путь - определить эти вероятности из опыта (a posteriori). Действительно, на основании теоремы Бернулли

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$$

для любого $\varepsilon>0$. Следовательно, при многократном повторении эксперимента (т.е. при $n\to\infty$) частота $W(A)=\frac{m}{n}$ наступления интересующего нас события A практически наверняка совпадает с вероятностью p наступления этого события.

Сама процедура проведения экспериментов и подсчета частот выходит за границы теории вероятностей и относится уже к другому разделу математики, называемому *математической статистикой*.

Задачи математической статистики далеко не ограничиваются подсчетом частот и оценкой на основании этого вероятности наступления интересующего нас события. Это - частная задача. Основным объектом изучения в математической статистике будет *случайная величина X*, над которой проводится *n* наблюдений (экспериментов, обследований, испытаний) с целью получения данных для анализа и принятия на основании этого анализа некоторого решения - "статистического вывода" о случайной величине *X* (например, о ее законе распределения, математическом ожидании, дисперсии и т.п.).

Ясно, что раз мы приняли вероятностную природу происхождения наших экспериментальных данных (т.е. считаем, что они подвержены влиянию случайных факторов), то все дальнейшие суждения, основанные на этих данных, будут иметь вероятностный характер. Это значит, что *всякое утверждение в рамках математической статистики будет верным лишь с некоторой вероятностью, а с некоторой вероятностью оно может оказаться неверным. Одна из центральных задач статистического анализа: важные выводы должны содержать оценку степени их неопределенности.*

Естественно, встает вопрос: будут ли полезными такие выводы, и можно ли вообще на таком пути получить достоверные результаты? Здесь следует руководствоваться *следующими правилами*:

1. Выводы математической статистики имеют значение *только для массовых случайных явлений*, а не для единичных.

2. Событие, вероятность которого близка к 1, считается *практически достоверным*, а событие, вероятность которого близка к 0, считается *практически невозможным*.

Конечно, такой подход не защищает нас полностью от ошибок, но эти ошибки будут проявляться редко.

Нам остается выяснить, какую же вероятность считать малой. На этот вопрос нельзя дать точного количественного ответа, пригодного во всех случаях. Ответ зависит от того, какой опасностью грозит нам ошибка. Довольно часто, - например, при проверке статистических гипотез, - полагают малыми вероятности, начиная с 0,01...0,05. Другое дело - надежность технических устройств, например, тормозов автомобиля. Здесь недопустимо большой будет вероятность отказа, скажем 0,001, т.к. выход из строя тормозов один раз на тысячу торможений повлечет большое число аварий. Поэтому при расчетах надежности нередко требуют, чтобы вероятность безотказной работы была бы порядка 10^{-6}.

Итак, под *математической статистикой* понимается раздел математики, посвященный методам систематизации, обработки и использования опытных данных для научных и практических выводов, а именно: статистических выводов о значениях числовых характеристик случайных величин (математического ожидания, дисперсии и т.д.) и об истинности тех или иных гипотез (гипотезы о законе распределения случайной величины X, о характере связи двух случайных величин и т.п.).

1.2. Генеральная совокупность и выборка

Значительная часть статистики связана с описанием больших совокупностей данных. Если интересующая нас совокупность слишком многочисленна (может быть, бесконечна), либо ее элементы малодоступны, либо имеются другие причины, не позволяющие изучать сразу все элементы (например, исследование качества большой партии консервов), прибегают к изучению какой-то части этой совокупности.

Определение. Множество всех изучаемых элементов называется *генеральной совокупностью* (ГС), а выбранная для исследования группа элементов называется *выборкой*, или *выборочной совокупностью.*

Статистикой называется та или иная числовая характеристика выборки, *параметрами* - числовые характеристики генеральной совокупности.

Эти понятия играют особо важную роль в теории статистических выводов. Из ГС случайным образом извлекается выборка и исходя из статистик, рассчитанных по этой выборке, делаются выводы о значении соответствующих параметров ГС.

1.3. Методы описательной статистики

Пусть из некоторой ГС извлечена выборка объема n со значениями исследуемого признака X: $x_1, x_2, ..., x_n$. Весьма полезную информацию о свойствах ГС можно получить уже на основе первичного анализа, используя методы описательной статистики - методы описания выборок $x_1, x_2, ..., x_n$ с помощью различных показателей и графиков. Полезность методов описательной статистики состоит в том, что несколько простых и довольно информативных статистических показателей способны избавить нас от просмотра сотен, а порой и тысяч значений выборки. Описывающие выборку показатели можно разбить на несколько групп.

1. *Показатели расположения* описывают положения данных на числовой оси. Это, например, минимальный и максимальный элементы выборки, выборочное среднее, медиана и др.

2. *Показатели разброса* описывают степень разброса данных относительно центра. По сути дела, они показывают, насколько кучно основная масса данных группируется около центра. В первую очередь, сюда относятся: дисперсия выборки, стандартные отклонения, размах выборки, коэффициент эксцесса и т.п.

3. *Показатели асимметрии* отвечают на вопрос о симметрии распределения данных около своего центра. К ним можно отнести: коэффициент асимметрии, положение выборочной медианы относительно выборочного среднего, гистограмму и т.д.

4. *Показатели, описывающие закон распределения*, дают представление собственно о законе распределения данных. Сюда относятся таблицы частот, гистограммы и эмпирические функции распределения.

Далее мы рассмотрим наиболее часто встречающиеся и наиболее информативные показатели описательной статистики. Начнем с показателей четвертой группы.

1.3.1. Вариационный ряд. Эмпирическая функция распределения

Для построения *выборочной (эмпирической) функции распределения* удобно от выборки $x_1, x_2, ..., x_n$ перейти к вариационному ряду $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

Определение. Вариационным рядом $x_{(1)}, x_{(2)}, ..., x_{(n)}$ называют выборку, перенумерованную в порядке неубывания.

Это следует понимать так: $x_{(1)}$ обозначает наименьшее из чисел $x_1, x_2, ..., x_n$; $x_{(2)}$ - наименьшее из оставшихся после удаления $x_{(1)}$ и т.д. В частности, $x_{(n)}$ есть наибольшее из чисел $x_1, x_2, ..., x_n$.

Вполне естественно, что среди чисел $x_{(1)}, x_{(2)}, ..., x_{(n)}$ могут встречаться одинаковые. Поэтому рассмотрим следующее определение.

Определение. Частотой элемента $x_{(i)}$ будем называть число m_i, которое показывает, сколько раз этот элемент встречается в данной выборке.

Теперь выделяем в выборке различные элементы и располагаем их в порядке возрастания:

$$x_{\min} = x_1^* < x_2^* < ... < x_k^* = x_{\max} \left(k \le n\right),$$

затем для каждого элемента x_i^* находим соответствующую частоту m_i. Распределение частот записывают в виде статистического ряда:

Элемент выборки	x_1^*	x_2^*	...	x_{k-1}^*	x_k^*
Частота	m_1	m_2	...	m_{k-1}	m_k

где $m_1 + m_2 + ... + m_k = n$ - объем выборки. Иногда и этот ряд также называют *вариационным рядом*, а значения x_i^* - *вариантами*.

Определение. Отношение $w_i = \frac{m_i}{n}$ частоты m_i к объему выборки n называется *относительной частотой* значения x_i^* $(i = 1, 2, ..., k)$.

Очевидно, что

$$\sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{m_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i = \frac{1}{n} \cdot n = 1.$$

Определение. Таблица, устанавливающая соответствие между вариантами x_i^* и их относительными частотами w_i, называется *статистическим распределением* выборки случайной величины X.

Определение. *Выборочной (эмпирической) функцией распределения* случайной величины X, построенной по статистическому распределению

Варианта	x_1^*	x_2^*	...	x_{k-1}^*	x_k^*
Относит. частота	w_1	w_2	...	w_{k-1}	w_k

называется функция

$$F_n(x) = \begin{cases} 0 \text{ при } x \le x_1^*; \\ \sum\limits_{i:x_i<x} w_i \text{ при } x > x_1^*. \end{cases}$$

Другими словами, значение выборочной функции распределения $F_n(x)$ есть сумма относительных частот вариант x_i^*, попадающих в интервал $(-\infty, x)$, т.е. доля в объеме выборки тех элементов выборки, которые попали в данный интервал.

Например, если

$$x_2 < x \le x_3,$$

то

$$F_n(x) = w_1 + w_2,$$

а при

$$x > x_k^* = x_{\max}$$

$$F_n(x) = 1.$$

Таким образом, $F_n(x)$ является кусочно-постоянной монотонно неубывающей функцией (ступенчатой функцией), имеющей в точках $x_1^*, x_2^*, ..., x_k^*$ разрывы 1-го рода (разрывы типа скачка), причем величина скачка в точке x_i^* равна относительной частоте $w_i = m_i/n$.

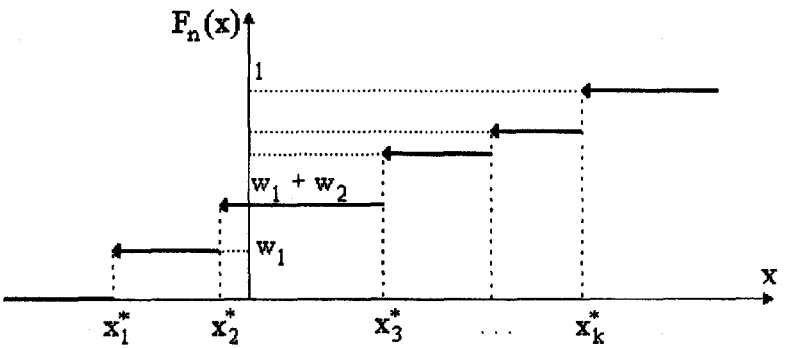


Рис. 1.1

Видно, что график эмпирической функции распределения (рис 1.1) напоминает график функции распределения дискретной случайной величины. Связь между эмпирической функцией распределения $F_n(x)$ и функцией распределения $F(x)$ исследуемой случайной величины X, которая определяется как $F(x) = P(X < x)$ (часто говорят - *теоретической функцией распределения*), основана на уже упомянутой теореме Бернулли. Она – такая же, как и связь между частотой события и его вероятностью. А именно, для любого числа x значение $F_n(x)$ представляет собой частоту появления события $\{X < x\}$, которое состоит в том, что случайная величина примет значение из интервала $(-\infty, x)$) в ряду из n независимых испытаний, следовательно, с вероятностью, равной 1, $F_n(x) \to F(x)$ при $n \to \infty$ для любого x. Или более точно: для

любого числа x и любого $\varepsilon > 0$ выполняется

$$\lim_{n\to\infty} P\left(\left| F(x) - F_n(x)\right| < \varepsilon\right) = 1.$$

1.3.2. Глазомерный метод обоснования гипотезы о законе распределения случайной величины

Эмпирическую функцию распределения можно использовать для обоснования гипотезы о законе распределения исследуемой случайной величины.

Пусть мы имеем основания считать, что выборка значений $x_1, x_2, ..., x_n$ сделана из ГС значений случайной величины X непрерывного типа. В этом случае можно использовать простой графический прием представления данных (так называемый *глазомерный метод*), который позволяет выдвинуть достаточно обоснованную гипотезу о виде закона распределения случайной величины X (нормальный, логнормальный и т.д.). В его основе лежат следующие рассуждения.

Пусть $y = F(x)$ - функция распределения случайной величины X. Ранее мы отмечали, что для эмпирической функции распределения с вероятностью, равной 1, должно выполняться условие: $F_n(x) \to F(x)$ при $n \to \infty$ для любого x (т.е. для любого x при больших объемах выборки событие $F_n(x) \approx F(x)$ является практически достоверным). Следовательно, для вариант x_i^* статистического ряда с большой вероятностью должно выполняться приближенное равенство

$$y_i^* = F\left(x_i^*\right) \approx \frac{F_n\left(x_i^*\right) + F_n\left(x_i^* + 0\right)}{2} = \bar{y}_i,$$

где $F_n\left(x_i^* + 0\right)$ - предел (справа) эмпирической функции распределения при $x \to x_i^* + 0$. Используя определение эмпирической функции распределения, легко получить:

$$\bar{y}_i = \sum_{k=1}^{i-1} w_k + \frac{w_i}{2},$$

в частности,

$$\bar{y}_1 = \frac{w_1}{2},\ \bar{y}_2 = w_1 + \frac{w_2}{2} \text{ и т.д.,}$$

где w_k - относительные частоты вариант x_k^* статистического ряда.

Теперь рассмотрим конкретные случаи, используя то, что для истинной функции распределения $F(x)$ с большой вероятностью должны выполняться приближенные равенства

$$y_i^* = F\left(x_i^*\right) \approx \bar{y}_i.$$

При этом для первых двух случаев приведем подробные рассуждения, а для остальных – только выводы (обоснования сделать самостоятельно). Прежде чем изучать дальнейший материал, следует ознакомиться с содержанием раздела 3.

Далее $\Phi(x)$ обозначает большую функцию Лапласа, а $\Phi^{-1}(y)$ - обратную к ней функцию, значения которых можно найти, используя табл. 3.2, 3.3. Кроме того, используем обозначения:

x_i^* - варианты статистического ряда;

$y_i^* = F\left(x_i^*\right)$;

$\bar{y}_i = \sum_{k=1}^{i-1} w_k + \frac{w_i}{2}$.

1. *Случайная величина X имеет нормальное распределение* $N(a,\sigma)$. Ее функция распределения равна

$$y = \Phi\left(\frac{x-a}{\sigma}\right),$$

откуда получаем

$$x = \sigma \cdot \Phi^{-1}(y) + a.$$

Следовательно, если на координатную плоскость Oxz нанести точки с координатами $\left(x_i^*, z_i^*\right)$, где $z_i^* = \Phi^{-1}\left(y_i^*\right)$, эти точки должны лежать на прямой, задаваемой уравнением

$$x = \sigma \cdot z + a .$$

Вывод: при условии нормальности распределения изучаемой случайной величины точки с координатами $\left(x_i^*, z_i^*\right)$, где $z_i^* = \Phi^{-1}(\bar{y}_i)$, на координатной плоскости Oxz должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

2. *Случайная величина* X *имеет логнормальное распределение.* Ее функция распределения равна

$$y = \Phi\left(\frac{\ln x - a}{\sigma}\right),$$

где $a \in R$ и $\sigma > 0$ - параметры распределения, откуда имеем:

$$\ln x = \sigma \cdot \Phi^{-1}(y) + a .$$

Значит, если на координатную плоскость Ouv нанести точки с координатами $\left(u_i^*, v_i^*\right)$, где $u_i^* = \ln x_i^*$, $v_i^* = \Phi^{-1}\left(y_i^*\right)$, эти точки должны лежать на прямой, задаваемой уравнением

$$u = \sigma \cdot v + a .$$

Вывод: при условии логнормальности распределения изучаемой случайной величины точки с координатами $\left(u_i^*, v_i^*\right)$, где $u_i^* = \ln x_i^*$, $v_i^* = \Phi^{-1}(\bar{y}_i)$, на координатной плоскости Ouv должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

3. *Случайная величина X имеет усеченное слева нормальное распределение* с заданной степенью усечения $\tau \in (0,1)$. Вывод: при условии, что изучаемая случайная величина имеет усеченное слева нормальное распределение со степенью усечения τ, точки с координатами (x_i^*, z_i^*), где $z_i^* = \Phi^{-1}((1-\tau)\bar{y}_i + \tau)$, на координатной плоскости *Oxz* должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

4. *Случайная величина X имеет усеченное справа нормальное распределение* с заданной степенью усечения $\tau \in (0,1)$. Вывод: при условии, что изучаемая случайная величина имеет усеченное справа нормальное распределение со степенью усечения τ, точки с координатами (x_i^*, z_i^*), где $z_i^* = \Phi^{-1}(\tau \cdot \bar{y}_i)$, на координатной плоскости *Oxz* должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

5. *Случайная величина X имеет равномерное распределение.* Вывод: при условии, что изучаемая случайная величина имеет равномерное распределение, точки с координатами $(x_i^*, \bar{y}_i)$ на координатной плоскости *Oxz* должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

6. *Случайная величина X имеет показательное распределение.* Вывод: при условии, что изучаемая случайная величина имеет показательное распределение, точки с координатами (x_i^*, z_i^*), где $z_i^* = -\ln(1-\bar{y}_i)$, на координатной плоскости *Oxz* должны лежать близко к некоторой прямой с положительным угловым коэффициентом, проходящей через начало координат.

7. *Случайная величина* X *имеет распределение Лапласа* . Вывод: при условии, что изучаемая случайная величина имеет распределение Лапласа, точки с координатами (x_i^*, z_i^*), где $z_i^* = \ln(2\bar{y}_i)$ при $\bar{y}_i \le 0{,}5$; $z_i^* = -\ln(2-2\bar{y}_i)$ при $\bar{y}_i > 0{,}5$, на координатной плоскости *Oxz* должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

8. *Случайная величина X имеет распределение Вейбулла* с заданной степенью $n \in N$. Вывод: при условии, что изучаемая случайная

величина имеет распределение Вейбулла с заданной степенью n, точки с координатами $\left(x_i^*, z_i^*\right)$, где $z_i^* = \left(-\ln\left(1-\bar{y}_i\right)\right)^{1/n}$, на координатной плоскости Oxz должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

9. *Случайная величина X имеет распределение Парето.* Вывод: при условии, что изучаемая случайная величина имеет распределение Парето, точки с координатами $\left(u_i^*, v_i^*\right)$, где $u_i^* = \ln x_i^*$; $v_i^* = -\ln\left(1-\bar{y}_i\right)$, на координатной плоскости Ouv должны лежать близко к некоторой прямой с положительным угловым коэффициентом.

1.3.3. Некоторые показатели расположения

Пусть после первичной обработки n элементов $x_1, x_2, \ldots, x_n$ выборки мы получили вариационный ряд $x_{\min} = x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} = x_{\max}$, а затем, выбрав различные варианты $x_{\min} = x_1^* < x_2^* < \ldots < x_k^* = x_{\max}$ и подсчитав их частоты, получили статистический ряд:

Элемент выборки	x_1^*	x_2^*	...	x_{k-1}^*	x_k^*
Частота	m_1	m_2	...	m_{k-1}	m_k

где $m_1 + m_2 + \ldots + m_k = n$ - объем выборки.

Показателем расположения выборки является *среднее значение выборки*, указывающее на то, где находится ее "центр". Но точно так же, как люди могут иметь различные мнения по поводу местонахождения центра города (в зависимости от того, что они собираются там делать), есть и различные способы оценки среднего значения выборки. Рассмотрим следующие определения:

1. *Полусумма крайних значений*

$$\frac{x_{\min} + x_{\max}}{2}.$$

2. *Выборочная медиана* есть число $h_в$, которое делит вариационный ряд на две части, содержащие равное число элементов. Если объем выборки $n = 2\ell + 1$ (т.е. n - нечетное число), то медиана равна $h_в = x_{(\ell+1)}$ - элементу вариационного ряда со средним номером. Если же $n = 2\ell$, то

$$h_в = \frac{x_{(\ell)} + x_{(\ell+1)}}{2}.$$

3. *Выборочная мода dв* есть варианта x_i^*, имеющая наибольшую частоту (и поэтому один и тот же статистический ряд может иметь более одной моды). Если выборка имеет одну моду, говорят, что статистическое распределение - *унимодальное*.

4. *Выборочным средним* (или выборочным аналогом математического ожидания) называется величина

$$\overline{X}_в = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \text{ или } \overline{X}_в = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i x_i^*,$$

если данные сведены в статистический ряд.

Последняя характеристика в дальнейшем будет использоваться наиболее часто.

Пример 1. В результате первичной обработки результатов измерений диаметров 50 подшипников получен следующий статистический ряд:

X, мм	10	12	14	16	17	19	20	21
m	2	3	4	5	6	8	10	12

Найти показатели положения.

Р е ш е н и е. Полусумма крайних значений равна

$$\frac{10+21}{2}=15{,}5.$$

Находим объем выборки:

$$n=2+3+4+5+6+8+10+12=50.$$

Т.к. n - четное число, то $\ell=n/2=25$, и выборочная медиана равна

$$h_в=\frac{x_{(25)}+x_{(26)}}{2}=\frac{19+19}{2}=19.$$

Выборочная мода, очевидно, равна 21 (унимодальное распределение).

Находим среднее выборочное:

$$\overline{X}_в=\frac{2\cdot10+3\cdot12+4\cdot14+5\cdot16+6\cdot17+8\cdot19+10\cdot20+12\cdot21}{50}=17{,}96\ .$$

1.3.4. Некоторые показатели разброса (рассеяния)

В ряде случаев единственной осмысленной статистикой является мера расположения, но в большинстве других необходимо, кроме этого, знать и меру рассеяния данных (называемую также *разбросом*, или *вариацией*). Если мы произвели замер 50 подшипников, при изготовлении которых требовалось, чтобы диаметр их равнялся 18 мм, и обнаружили, что средний диаметр составляет 17,96 мм, то нам не придется особо радоваться, если единичные замеры окажутся такими, как в приведенном примере 1. Мера рассеяния позволяет выяснить, как часто и насколько диаметр детали будет отклоняться от среднего значения.

Далее используем предыдущие обозначения.

Простейшей мерой рассеяния является *размах выборки*: $d = x_{\max} - x_{\min}$ (в примере 1 размах равен $d = 21 - 10 = 11$ мм). Однако размах выборки, сделанной из большой совокупности, окажется гораздо менее удовлетворительной оценкой рассеяния, чем оценка с помощью другой меры, учитывающей вместо двух экстремальных значений все без исключения наблюдения. Наилучшей такой характеристикой является *выборочная дисперсия*, которая представляет собой среднее значение квадратов отклонений элементов выборки от ее среднего выборочного:

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 \text{ или } D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \cdot \left(x_i^* - \overline{X}\right)^2,$$

если данные сведены в статистический ряд.

Для вычислений лучше использовать эквивалентные формулы, которые получаются из определения путем несложных преобразований (проверить это в качестве упражнения):

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{X}^2 \text{ или } D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \cdot x_i^{*2} - \overline{X}^2$$

для статистического ряда.

Для сравнения отметим, что дисперсия дискретной случайной величины X в теории вероятностей вычисляется по формулам

$$D = \sum_{i=1}^{n} \left(x_i - M[X]\right)^2 \cdot p_i = \sum_{i=1}^{n} x_i^2 \cdot p_i - M[X]^2,$$

где $M[X]$ - математическое ожидание случайной величины X.

Как и в теории вероятностей, *выборочным среднеквадратическим отклонением*, или *стандартным отклонением*, называется величина $\sigma_в = \sqrt{D_в}$.

В качестве примера вычислим выборочную дисперсию и среднеквадратичное отклонение по данным примера 1:

$$D_в = \frac{2\cdot 10^2 + 3\cdot 12^2 + 4\cdot 14^2 + 5\cdot 16^2 + 6\cdot 17^2 + 8\cdot 19^2 + 10\cdot 20^2 + 12\cdot 21^2}{50} -$$

$$-17{,}96^2 = 9{,}638,$$

тогда

$$\sigma_в = \sqrt{9{,}638} = 3{,}105.$$

1.3.5. Группированные данные

Рассмотрим следующий ряд данных, полученных в какой-либо серии наблюдений случайной величины X:

40, 43, 46, 59, 64, 67, 68, 69, 75, 76, 78, 80, 82, 82, 86, 90, 92, 127.

Эти 18 наблюдений принимают 17 различных значений, и поскольку варианта x = 82 есть единственное значение, встречающиеся более одного раза, оно и является модой. Представлять эти данные в виде статистического ряда было бы, по крайней мере, неразумно. С другой стороны, для удобства их можно сгруппировать, например, в шесть классов:

Класс	39...53	54...68	69...83	84...98	99...113	114...128
Частота	3	4	7	3	0	1

где под классами подразумеваются промежутки [39, 53], [54, 68] и т.д., а под частотами - количество элементов выборки, попавших в соответствующий промежуток. В результате мы получим так называемый *группированный статистический ряд*.

Группировать данные имеет смысл в том случае, если нам необходимо собирать и записывать информацию о большом количестве наблюдений (причем, когда велико не только число наблюдений, но и число различных значений среди них). Эта ситуация особенно часто встречается при наблюдении (измерении, регистрации и т.п.) непрерывных случайных величин. Действительно, вспомним, что в теории вероятностей мы встречаемся с двумя типами случайных величин: дискретными и непрерывными. Но в математической ста-

тистике одни и те же данные можно характеризовать как дискретные или непрерывные. Это зависит от природы этих данных. Так, например, если приведенные выше данные представляют собой число бракованных деталей в проверяемых 18-ти партиях, то это – дискретные данные. А если эти же данные представляют вес в килограммах 18-ти взрослых людей, то это будут непрерывные данные, хотя и здесь наблюдения довольно необычны. В каком смысле? Непрерывные данные состоят из наблюдений над непрерывной случайной величиной, т.е. над такой величиной, которая на интервале своего изменения может принимать любые значения - целые, дробные или иррациональные. Эти значения никогда не могут быть зафиксированы "точно", и мы обычно понимаем, что они округлены до ближайшего значения. Точность округления, очевидно, определяется возможностями измерительной аппаратуры либо задачами, которые ставятся в данном конкретном исследовании. Так, если вышеприведенные данные представляют собой вес в килограммах, то предполагается, что элемент выборки 40 означает вес между 39,5 и 40,5 кг. Да и вообще можно сказать, что результаты, полученные с помощью всякого рода измерений, обычно непрерывны, а результаты подсчетов - дискретны.

Методика построения группированного статистического ряда следующая. Обозначим через $x_{\min}$ и $x_{\max}$ минимальный и максимальный элементы выборки. Выберем числа $y_{\min} \le x_{\min}$ и $y_{\max} \ge x_{\max}$. Отрезок $[y_{\min}, y_{\max}]$ разбиваем на k частичных интервалов:

$$[y_0, y_1), [y_1, y_2), \dots, [y_{k-1}, y_k],$$

где $y_0 = y_{\min}$ и $y_k = y_{\max}$ (для упрощения вычислений длины этих интервалов часто берут одинаковыми).

Затем каждому интервалу ставят в соответствие частоту m_i^* - количество элементов выборки, попавших в этот интервал. Тогда *группированный статистический ряд* (*или интервальный ряд*) имеет вид

Интервал	$[y_0, y_1)$	$[y_1, y_2)$	$[y_2, y_3)$	...	$[y_{k-1}, y_k]$
Частота	m_1^*	m_2^*	m_3^*	...	m_k^*

В описанной выше методике есть неопределенность, которая заключается в выборе чисел $y_{\min}$, $y_{\max}$ и k. Число интервалов группировки k можно варьировать в разумных пределах. Эта "разумность" определяется 10...15 группировками, хотя бывают случаи, когда требуется больше 25 группировок или меньше 8 (но не меньше 4).

В литературе предлагается формула для оценки снизу числа интервалов группировки:

$$k \geq [\log_2 n] + 1,$$

где $[n]$ обозначает целую часть числа n.

Определив число k, находим длину интервалов группировки (если длины всех интервалов берутся одинаковыми):

$$r = \frac{x_{\max} - x_{\min}}{k}.$$

Число r можно округлить в большую сторону до нужного количества знаков после запятой. Затем выбираем $y_{\min}$ и $y_{\max}$ так, чтобы отрезок $[y_{\min}, y_{\max}]$ накрывал отрезок $[x_{\min}, x_{\max}]$.

Для данных, приведенных в начале этого раздела, описанная выше процедура выглядит так. Находим число интервалов k: т.к. $k \geq [\log_2 18] + 1 = 4 + 1 = 5$, то можно взять $k = 5$. Находим длину интервалов группировки:

$$r = \frac{127 - 40}{5} = \frac{87}{5} = 17{,}4.$$

Округлим это число до 17,5. Далее берем $y_{\min} = 40$ и $y_{\max} = 127{,}5$ (очевидно, что отрезок $[40, 127, 5]$ накрывает отрезок $[40, 127]$). Тогда группированный статистический ряд имеет вид:

Интервал	[39,5; 57)	[57;74,5)	[74,5; 92)	[92; 109,5)	[109,5;127]
Частота	3	5	8	1	1

Стоит заметить, что в ряде случаев (например, при проверке гипотезы о характере распределения исследуемой случайной величины) данные приходится распределять по классам с неравными интервалами.

В ряде случаев исходная статистическая информация поступает только в виде группированного (интервального) статистического ряда. Тогда для вычисления выборочных средней и дисперсии из группированного статистического ряда надо получить соответствующий вариационный статистический ряд. Делается это так: каждому интервалу группировки $[y_{i-1}, y_i)$ ставится в соответствие варианта $z_i = \frac{1}{2}(y_{i-1} + y_i)$, а затем этому числу z_i приписывается частота m_i^* - количество элементов выборки, попавших в данный интервал.

Например, если использовать данные полученного выше интервального ряда, то указанная процедура приведет к получению следующего вариационного ряда:

Z	48,25	65,75	83,25	100,75	118,25
Частота	3	5	8	1	1

По этим данным найдем выборочную среднюю и выборочную дисперсию:

$$\overline{X}_в = \frac{48{,}25 \cdot 3 + 65{,}75 \cdot 5 + 83{,}25 \cdot 8 + 100{,}75 \cdot 1 + 118{,}25 \cdot 1}{18} = 75{,}472;$$

$$D_в = \frac{48{,}25^2 \cdot 3 + 65{,}75^2 \cdot 5 + 83{,}25^2 \cdot 8 + 100{,}75^2 \cdot 1 + 118{,}25^2 \cdot 1}{18} -$$

$$- 75{,}472^2 = 313{,}8.$$

Отметим, что аналогичная методика используется для упрощения вычислений выборочных характеристик имеющейся выборки $x_1, x_2, \ldots, x_n$. Вначале по исходным данным строится интервальный статистический ряд, затем по этому интервальному ряду строится вариационный статистический ряд, где в качестве вариант берутся середины соответствующих интервалов, а затем по этим данным вычисляются выборочные характеристики. Естественно, мы получим значения, которые будут отличаться от таких же характеристик, вычисленных непосредственно по элементам заданной выборки $x_1, x_2, \ldots, x_n$. Но эти погрешности, как правило, бывают несущественными (особенно, если число интервалов берется достаточно большое). В качестве иллюстрации к сказанному приведем значения выборочных средней и дисперсии, найденной непосредственно по данным, указанным в начале этого пункта: $\overline{X}_в = 73{,}556$; $D_в = 391{,}69$. Мы видим, что точные значения выборочных характеристик не так уж существенно отличаются от соответствующих значений, которые мы получили чуть выше по вспомогательному вариационному ряду.

1.3.6. Графические представления выборки

Для того, чтобы получить наглядное представление о характере распределения генеральной совокупности по результатам выборки, используют такие графические объекты, как *гистограмма относительных частот* и *полигон относительных частот*. Для их построения имеющуюся выборку объема n надо представить в виде группированного статистического ряда с k частичными интервалами одинаковой длины (длину обозначим через h).

Гистограмма выборки (рис. 1.2) – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частич-

ные интервалы, а высоты равны отношениям $\frac{m_i}{nh}$, где $\frac{m_i}{n} = w_i$ - относительная частота попадания элементов выборки в i-й интервал (очевидно, что площадь i-го прямоугольника гистограммы равна относительной частоте w_i, а площадь всех прямоугольников - площадь гистограммы - равна единице).

Полигон – это ломаная, соединяющая середины верхних сторон прямоугольников гистограммы.

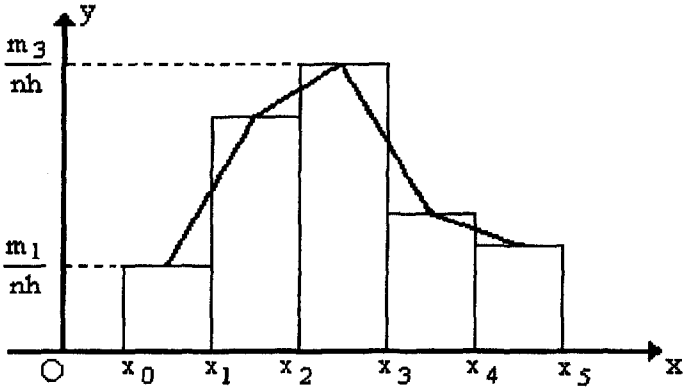


Рис. 1.2

Гистограмма и полигон относительных частот могут дать первое представление о характере закона распределения исследуемой случайной величины. Для непрерывных случайных величин гистограмма и полигон относительных частот являются, в определенном смысле, приближением для плотности $f(x)$ распределения случайной величины X. Сравнивая график плотности распределения известной случайной величины (оценки параметров распределения можно найти, используя метод моментов из раздела 3) и построенную гистограмму (полигон), мы можем сделать первое предположение о законе распределения изучаемой случайной величины.

Например, плотность распределения нормально распределенных случайных величин (такие случайные величины наиболее часто встречается в практических задачах) имеет вид

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}},$$

где a и $\sigma>0$ - некоторые параметры (см. подраздел 3.1).

Если при определенных значениях этих параметров кривая Гаусса (график данной функции) проходит достаточно близко от точек гистограммы и полигона (как это показано на рис. 1.3), вполне закономерно выдвинуть гипотезу о том, что изучаемая случайная величина имеет нормальное распределение. Было бы необоснованным предположить, что гистограмма и полигон, изображенные на рис. 1.3, соответствуют выборке из ГС значений случайной величины, имеющей, к примеру, показательное распределение или распределение Парето; с другой стороны, есть смысл рассмотреть также гипотезы о логнормальном распределении или усеченном слева нормальном распределении с малой степенью усечения (см. раздел 3).

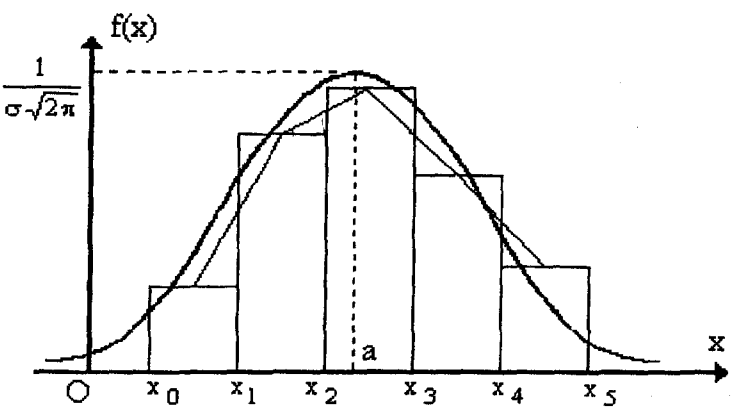


Рис. 1.3

Замечание. Сказанное выше можно пояснить следующим образом. Вероятность p_i того, что случайная величина X с плотностью распределения $f(x)$ примет значение из интервала $[x_{i-1}, x_i]$, равна площади криволинейной трапеции с основанием $[x_{i-1}, x_i]$, ограни-

ченной сверху графиком функции $f(x)$. В то же время площадь соответствующего прямоугольника гистограммы равна относительной частоте w_i попадания значений случайной величины в этот интервал.

1.3.7. Некоторые дополнительные характеристики выборки

Пусть мы имеем выборку $x_1, x_2, .., x_n$ объема n (которая может быть преобразована в статистический ряд с k вариантами x_i^* и соответствующими частотами m_i). Рассмотрим некоторые дополнительные числовые характеристики выборки.

1. *Выборочный начальный момент r - порядка* обозначается M_r' и определяется следующим образом:

$$M_r' = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^r ,$$

для статистического ряда

$$M_r' = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \left(x_i^*\right)^r .$$

Сравнивая эти выражения с формулами для выборочного среднего, видим, что M_1' есть выборочная средняя $\overline{X}$.

2. *Выборочный центральный момент r - порядка* обозначается M_r и определяется следующим образом:

$$M_r = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^r ;$$

для статистического ряда

$$M_r = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \left(x_i^* - \overline{X} \right)^r .$$

Очевидно, что значение M_2 равно выборочной дисперсии $D_в$.

3. *Выборочный коэффициент асимметрии* обозначается A_s и определяется по формуле

$$A_s = \frac{M_3}{\left(D_в\right)^{3/2}} = \frac{M_3}{\sigma_в^3},$$

где $\sigma_в = \sqrt{D_в}$ - выборочное среднеквадратическое отклонение.

Величина A_s является безразмерной, т.е. не зависит от выбора единицы измерения элементов выборки. Для упрощения вычислений A_s можно использовать следующую формулу:

$$M_3 = M_3' - 3M_2' \cdot \overline{X} + 2\overline{X}^2 .$$

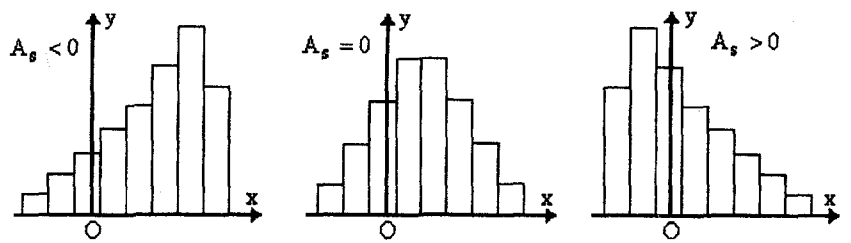


Рис. 1.4

Очевидно, что для симметричного распределения $A_s = 0$. Если $A_s < 0$, распределение имеет "скошенность влево", при $A_s > 0$ – "скошенность вправо".

4. *Выборочный коэффициент эксцесса* обозначается E_x и определяется по формуле

$$E_x = \frac{M_4}{\sigma_в^4} - 3,$$

где $\sigma_в = \sqrt{D_в}$ - выборочное среднеквадратическое отклонение.

Величина E_x так же, как и коэффициент асимметрии, является безразмерной, т.е. не зависит от выбора единицы измерения элементов выборки. Для упрощения вычислений E_x можно использовать следующую формулу:

$$M_4 = M_4' - 4M_3' \cdot \overline{X} + 6M_2' \cdot \overline{X}^2 - 3\overline{X}^4.$$

Этот показатель обладает теми же свойствами "формообразующей статистики", что и коэффициент асимметрии. Для "колоколообразного" нормального распределения $E_x = 0$. Для данных с идеально прямоугольной гистограммой $E_x < -1$. "Острый пик и положение окраины" распределения определяют значения эксцесса примерно 2...3.

Замечание. Для нормального распределения значения коэффициентов асимметрии и эксцесса равны 0. Поэтому, если по результатам выборки мы получили выборочные значения $A_s \approx 0$ и $E_x \approx 0$, то имеет смысл выдвинуть гипотезу о том, что генеральная совокупность, из которой сделана выборка, имеет нормальное распределение.

1.3.8. Некоторые замечания о числовых характеристиках выборки

Для получения более точных и достоверных выводов о генеральной совокупности особое внимание следует обратить на наличие в выборке так называемых *выбросов*, т.е. грубых (ошибочных), сильно отличающихся от основной массы наблюдений. Дело в том, что даже одно или несколько грубых наблюдений способны сильно исказить такие выборочные характеристики, как среднее, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцес-

са. Проще всего обнаружить такие наблюдения с помощью перехода от выборки к ее вариационному ряду или гистограмме с достаточно большим числом интервалов группировки. Подозрение о присутствии таких наблюдений может возникнуть, если выборочная медиана заметно отличается от выборочного среднего (хотя в целом совокупность симметрична), если положение медианы сильно несимметрично относительно минимального и максимального элементов выборки, и т.д.

Вообще следует иметь в виду, что для данных, имеющих хорошую форму распределения, медиана всегда лежит в промежутке между средним и модой. Примеры расположения для данных с хорошей формой распределения и отрицательной асимметрией (скошенность влево) выстраиваются по возрастанию следующим образом: среднее, медиана, мода, а для таких же данных с положительной асимметрией (скошенность вправо) они располагаются в обратном порядке.

1.4. Статистическое оценивание параметров

Главная цель, которую ставит перед собой исследователь, приступая к статистической обработке выборки, - это получение на основании имеющихся данных максимально достоверной информации о всей генеральной совокупности (т.е. о случайной величине X): о законе распределения этой величины, о параметрах этого распределения (математическом ожидании, дисперсии и др.). Конечно, по результатам конкретной выборки $x_1, x_2, \ldots, x_n$ можно вычислить различные ее характеристики, но они будут давать лишь приближенные значения каких-то параметров распределения случайной величины X. Так, мы уже встречались с выборочной средней и выборочной дисперсией выборки и можем предположить (пока только интуитивно), что эти величины будут неплохими оценками математического ожидания и дисперсии изучаемой случайной величины X. Наша задача теперь – познакомиться с понятием точечной оценки, выяснить, какие оценки чаще всего используются на практике, как они получаются и какими свойствами обладают, чтобы мы могли им доверять.

Итак, пусть нам дана выборка объема n из некоторой генеральной совокупности. Рассмотрим следующее определение (которое

дальше будет уточнено).

Определение. *Статистикой* (*точечной оценкой*) называется любая функция $\hat{\Theta}_n = U(x_1, x_2, \ldots, x_n)$ от элементов выборки $x_1, x_2, \ldots, x_n$.

Задача *статистического оценивания* неизвестного параметра Θ генеральной совокупности состоит в указании таких статистик $\hat{\Theta}_n = U(x_1, x_2, \ldots, x_n)$, что будет выполнено приближенное равенство $\Theta \approx \hat{\Theta}_n$.

Здесь же возникает вопрос, какие требования мы должны предъявить к статистике $\hat{\Theta}_n$, чтобы в понятие приближенного равенства $\Theta \approx \hat{\Theta}_n$ был вложен здравый смысл (ведь, в конце концов, можно сказать, что $1 \approx 1000$).

Нетрудно понять, что любая статистика в определенном смысле является случайной величиной: при переходе от одной выборки к другой (даже в рамках одной и той же генеральной совокупности) конкретные значения статистики (подсчитанные по одной и той же формуле) будут подвержены некоторому неконтролируемому разбросу - случайной изменчивости. Поэтому желательно, чтобы значения статистики, подсчитанные по разным выборкам из одной и той же генеральной совокупности, концентрировались около истинного значения оцениваемого параметра. Кроме того, вполне естественно требование, чтобы с увеличением объема выборки n погрешность в приближенном равенстве $\Theta \approx \hat{\Theta}_n$ уменьшалась. Эти требования заложены в определения следующих трех свойств точечных оценок: несмещенности , состоятельности и эффективности.

Но, прежде чем переходить к изучению этих свойств, мы должны уточнить общий принцип подхода к понятиям выборки и точечной оценки (статистики), принятый в математической статистике.

Пусть произведено n независимых измерений (наблюдений) случайной величины X и получен случайный набор ее значений $\{x_1, x_2, \ldots, x_n\}$. Логически мы можем представить этот набор как результат одновременного опыта над n независимыми случайными величинами $X_1, X_2, \ldots, X_n$, которые имеют тот же закон распреде-

ления, что и величина X. Для того, чтобы можно было применить для оценки степени неопределенности статистических оценок те или иные методы теории вероятностей, в математической статистике принято считать *выборкой* (в широком смысле) последовательность независимых одинаково распределенных случайных величин $\{X_1, X_2, ..., X_n\}$, а полученный в результате опыта набор чисел $\{x_1, x_2, ..., x_n\}$ - *реализацией этой выборки*. При таком подходе статистика (точечная оценка) – это функция $\hat{\Theta}_n = U(X_1, X_2, ..., X_n)$ от последовательности случайных величин $\{X_1, X_2, ..., X_n\}$, а величина $\hat{\Theta}_{n\text{выб}} = U(x_1, x_2, ..., x_n)$, полученная при подстановке в статистику вместо случайных величин X_i значений x_i из реализации выборки, есть *выборочное значение этой статистики.*

Функция от случайных величин сама является случайной величиной. Таким образом, во-первых, мы вложили точный смысл в интуитивные рассуждения о том, что точечные оценки являются случайными величинами, а во-вторых, теперь можем оперировать такими понятиями, как *математическое ожидание* $M[\hat{\Theta}_n]$ *и дисперсия* $D[\hat{\Theta}_n]$ *точечной оценки.*

При дальнейшем изложении, не оговаривая этого специально, будем предполагать, что у нас имеется выборка $\{X_1, X_2, ..., X_n\}$ объема n независимых случайных величин, одинаково распределенных с изучаемой случайной величиной X. Отсюда, в частности, следует, что если $m = M[X]$ - математическое ожидание; $\sigma^2 = D[X]$ – дисперсия величины X, то

$$M[X_1] = M[X_2] = ... = M[X_n] = m;$$
$$D[X_1] = D[X_2] = ... = D[X_n] = \sigma^2.$$

Замечание. В ряде учебных пособий по математической статистике зачастую не делают различия между понятиями выборки как последовательности независимых одинаково распределенных случайных величин $\{X_1, X_2, ..., X_n\}$ и ее конкретной реализации

как некой последовательности чисел $\{x_1, x_2, ..., x_n\}$, полученных в результате статистических испытаний. Обычно это отличие становится понятно из контекста, но при первом прочтении могут возникнуть определенные сложности для понимания.

1.4.1. Свойства точечных оценок

1. Оценка (статистика) $\hat{\Theta}_n$ неизвестного параметра Θ генеральной совокупности называется *несмещенной* (без систематической ошибки), если ее математическое ожидание равно оцениваемому параметру, т.е.

$$M\left[\hat{\Theta}_n\right] = \Theta.$$

В некоторых случаях для простоты вычислений или исходя из других соображений используется *асимптотически несмещенная* оценка, которая должна удовлетворять условию

$$\lim_{n\to\infty} M\left[\hat{\Theta}_n\right] = \Theta$$

(например, далее мы узнаем, что выборочная дисперсия не является несмещенной оценкой дисперсии, но является асимптотически несмещенной). Оценки такого типа содержат систематические ошибки, однако абсолютная величина этих ошибок с ростом объема выборки стремится к 0.

2. Оценку (статистику) $\hat{\Theta}_n$ неизвестного параметра Θ генеральной совокупности называют *состоятельной,* если для любого $\varepsilon > 0$ выполняется условие

$$\lim_{n\to\infty} P\left(\left|\hat{\Theta}_n - \Theta\right| < \varepsilon\right) = 1.$$

Определение состоятельности оценки $\hat{\Theta}_n$ говорит о том, что с вероятностью 1 (т.е. практически всегда) при увеличении объема

выборки n разница между значениями $\hat{\Theta}_n$ и Θ становится сколь угодно мала.

Таким образом, требование состоятельности и несмещенности (асимптотической несмещенности) представляется необходимым для того, чтобы данная оценка (статистика) имела практический смысл, т.к. в противном случае увеличение объема исходной информации не будет приближать нас к истине.

3. *Эффективность* оценок. Для оценки параметра Θ может быть предложено несколько несмещенных (и даже состоятельных) оценок. Мерой точности несмещенной оценки $\hat{\Theta}_n$ в математической статистике считают ее дисперсию $D\left[\hat{\Theta}_n\right]$. Наилучшей (эффективной) оценкой считают ту, для которой эта величина минимальна среди всех несмещенных оценок.

Вопрос об эффективности оценки является весьма сложным. В частности, одна и та же оценка может быть эффективной для выборок из генеральных совокупностей, подчиненных определенному закону распределения (например, нормальному), и неэффективной для других распределений (см. замечание 1 в 1.4.3).

Замечание. К сожалению, наилучших во всех отношениях оценок не бывает. Например, оценка, замечательно ведущая себя при некоторых предположениях об исходных данных, при отклонениях от этих предположений может приводить к сильно искаженным результатам. Например, выборочное среднее (как мы увидим ниже, это - оценка математического ожидания) обладает многими свойствами оптимальности, но очень плохо реагирует на наличие в выборке выбросов, т.е. резко выделяющихся значений (которые обычно порождены грубыми ошибками в измерениях и иными причинами). Поэтому в последнее время интенсивно развиваются методы устойчивого (робастного) оценивания, главная задача которых – получение надежных и эффективных оценок, пригодных для ситуаций, когда данные отклоняются от моделей выборок, содержат засорения или грубые ошибки наблюдения.

1.4.2. Метод моментов для нахождения оценок параметров распределения по выборке

В математической статистике есть много подходов, которые придают зависимости $\hat{\Theta}_n = U(X_1, X_2, ..., X_n)$ точную математическую форму. В настоящее время, как правило, используются три основных метода получения оценок: *метод моментов, метод наименьших квадратов, метод максимального правдоподобия.*

В дальнейшем мы будем применять для оценки неизвестных параметров распределения метод моментов, а для оценки неизвестных параметров модели – метод наименьших квадратов.

Суть этого метода состоит в том, что выборочные моменты (см. 1.3.6) принимаются за оценки соответствующих теоретических моментов. Так, за оценку математического ожидания случайной величины X принимается первый начальный момент, за оценку дисперсии – второй центральный момент и т.д. Вопрос о качестве некоторых из этих оценок (выборочной средней и выборочной дисперсии) будет рассмотрен далее.

В процессе рассмотрения гипотез о законе распределения ГС по результатам выборки нам придется иметь дело с оценками таких параметров этих распределений, которые не являются непосредственно начальными или центральными моментами. В этом случае поступают следующим образом. Начальные или центральные моменты распределения выражают через изучаемые параметры, затем заменяют соответствующими выборочными моментами. В результате получают систему уравнений, из которой находят оценки интересующих параметров, выраженные через значения выборочных моментов. Как это делается непосредственно для наиболее важных на практике распределений, описано в разделе 3.

1.4.3. Оценка математического ожидания случайной величины по результатам наблюдений

Согласно методу моментов, за оценку математического ожидания $m = M[X]$ случайной величины X берется первый начальный выборочный момент:

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

(выборочное значение этой статистики $\overline{X}_{nв}$ есть выборочная средняя).

Проверим, что $M[\overline{X}_n] = m$, т.е. $\overline{X}_n$ - оценка несмещенная (не дает систематической ошибки). Действительно, согласно свойствам математического ожидания имеем:

$$M[\overline{X}_n] = \frac{M[X_1] + M[X_2] + ... + M[X_n]}{n} = \frac{n \cdot m}{n} = m.$$

Теперь исследуем эту оценку на состоятельность. Согласно свойствам дисперсии (вспомнить их) имеем:

$$D[\overline{X}_n] = D\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2} \cdot \sum_{i=1}^{n} D[X_i] = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n},$$

откуда получаем:

$$\lim_{n \to \infty} D[\overline{X}_n] = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0.$$

Т.к. на основании неравенства Чебышева для любого $\varepsilon > 0$ выполняется условие

$$P\left(\left| \overline{X}_n - M[\overline{X}_n] \right| \geq \varepsilon\right) \leq \frac{D[\overline{X}_n]}{\varepsilon^2},$$

то, учитывая несмещенность оценки $\overline{X}_n$, имеем:

$$1 \geq \lim_{n\to\infty} P\left(\left|\overline{X}_n - m\right| < \varepsilon\right) \geq 1 - \frac{1}{\varepsilon^2} \lim_{n\to\infty} D\left[\overline{X}_n\right] = 1 \quad \Rightarrow$$

$$\Rightarrow \lim_{n\to\infty} P\left(\left|\overline{X}_n - m\right| < \varepsilon\right) = 1.$$

Следовательно, $\overline{X}_n$ - состоятельная оценка.

Замечание 1. Можно показать, что оценка $\overline{X}_n$ является эффективной для выборки из нормально распределенной генеральной совокупности. В то же время для равномерно распределенной генеральной совокупности несмещенная статистика

$$\hat{m}_n = \frac{\min\limits_i X_i + \max\limits_i X_i}{2}$$

(полусумма крайних значений) является более эффективной, чем статистика $\overline{X}_n$.

Вывод. Оценка $\overline{X}_n$ математического ожидания случайной величины X обладает необходимыми свойствами несмещенности и состоятельности (а в ряде случаев – и эффективности). Значит, этой оценкой можно смело пользоваться в практических расчетах.

Замечание 2. Попутно мы получили интересное для практики утверждение, что среднеквадратическая погрешность $\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ среднего арифметического n измерений меньше в $\sqrt{n}$ раз по отношению к среднеквадратической погрешности отдельного измерения $\sigma = \sqrt{D\left[X_i\right]}$ (закон возрастания точности при возрастании числа измерений).

1.4.4. Оценка дисперсии и среднеквадратического отклонения случайной величины по результатам наблюдений

Следуя методу моментов, за оценку дисперсии $\sigma^2 = D[X]$ случайной величины X мы берем второй центральный выборочный момент

$$D_n = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 ,$$

а за оценку среднеквадратического отклонения σ - величину $\sigma_n = \sqrt{D_n}$ (ясно, что выборочные значения этих статистик $D_{nв}$ и $\sigma_{nв}$ есть соответственно выборочные дисперсия и среднеквадратическое отклонение). Рассмотрим свойства оценки D_n.

Выясним вопрос о несмещенности оценки D_n.

Вначале выполним следующие преобразования (напомним, что $m = M[X]$):

$$D_n = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left[(X_i - m) - (\overline{X}_n - m) \right]^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[(X_i - m)^2 - 2(X_i - m)(\overline{X}_n - m) + (\overline{X}_n - m)^2 \right] =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - m)^2 - 2(\overline{X}_n - m) \cdot \frac{1}{n} \sum_{i=1}^{n} (X_i - m) + \frac{1}{n} \cdot (\overline{X}_n - m)^2 \cdot n =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - m)^2 - 2(\overline{X}_n - m)^2 + (\overline{X}_n - m)^2 =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - m)^2 - (\overline{X}_n - m)^2 .$$

Т.к. математическое ожидание (по определению дисперсии)

$$M(X_i - m)^2 = D[X_i] = \sigma^2,$$

а математическое ожидание

$$M(\overline{X}_n - m)^2 = D[\overline{X}_n] = \frac{\sigma^2}{n}$$

(это равенство получено в предыдущем пункте), получим:

$$M[D_n] = \frac{1}{n} \cdot \sum_{i=1}^{n} M(X_i - m)^2 - M(\overline{X}_n - m)^2 = \frac{\sigma^2 \cdot n}{n} - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Таким образом,

$$M[D_n] \neq \sigma^2 = D[X],$$

но

$$\lim_{n\to\infty} M[D_n] = \sigma^2 = D[X].$$

Следовательно, оценка D_n не является несмещенной, но является асимптотически несмещенной.

Причина этого кроется в том, что одна и та же выборка используется дважды: во-первых, для нахождения оценки математического ожидания $\overline{X}_n$, во-вторых, для нахождения оценки самой дисперсии. Мы знаем, что несмещенность оценки указывает на отсутствие систематической ошибки, поэтому весьма желательно устранить возникшую неприятность.

Из расчетов, приведенных выше, видно, что это исправляется довольно легко. Действительно, положим

$$s_n^2 = \frac{n}{n-1} D_n = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Проверим, что оценка s_n^2 является несмещенной оценкой дисперсии $\sigma^2 = D[X]$ случайной величины X.

Мы имеем:

$$M\left[s_n^2\right]=\frac{n}{n-1}\cdot M\left[D_n\right]=\frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2=\sigma^2.$$

Оценка s_n^2 называется *исправленной выборочной дисперсией*, а оценка $s_n=\sqrt{s_n^2}$ - *исправленной оценкой среднеквадратического отклонения*.

Оценки D_n и s_n^2 являются состоятельными. Доказательство этого факта осуществить самостоятельно, используя полученное выше представление для D_n и теорему Чебышева.

Вывод. Оценка s_n^2 дисперсии случайной величины X обладает необходимыми свойствами несмещенности и состоятельности. Значит, этой оценкой можно пользоваться в практических расчетах. Оценка D_n является состоятельной и асимптотически несмещенной. Поэтому ее также можно использовать (на практике ее можно считать несмещенной для достаточно больших n, например при $n>30$).

Замечание. Следует подчеркнуть, что мы рассматривали тот случай, когда математическое ожидание случайной величины X до опыта (a priori) было неизвестно и само находилось по результатам выборки. Если же математическое ожидание a priori известно, то за оценку дисперсии следует взять обычную выборочную дисперсию

$$D_n=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-a\right)^2,$$

где

$$a=M\left[X\right].$$

Легко убедиться, что в данном случае такая оценка будет несмещенной.

1.5. Точность статистических оценок

Итак, мы показали, что выбранные нами точечные оценки $\overline{X}_n$ и s_n^2 (или D_n) математического ожидания и дисперсии случайной величины X практически всегда (т.е. с вероятностью 1) должны давать хорошие результаты для очень больших объемов выборки n (т.е. при $n \to \infty$). Но, к сожалению, они не позволяют судить о степени близости их выборочных значений к истинному значению оцениваемого параметра при конкретном значении объема выборки. Естественно, возникает вопрос о мере доверия к полученным оценкам. Погрешности $\left|\overline{X}_n - m\right|$, $\left|s_n^2 - \sigma^2\right|$ (или в общем случае $\left|\hat{\Theta}_n - \Theta\right|$) неизбежны, но не окажутся ли они недопустимо высокими?

1.5.1. Доверительное оценивание

Пусть $\hat{\Theta}_n$ есть точечная оценка неизвестного параметра Θ генеральной совокупности. *Задача доверительного оценивания* состоит в следующем: требуется по оценке $\hat{\Theta}_n$ определить такое значение $\delta > 0$, что вероятность

$$P\left(\left|\hat{\Theta}_n - \Theta\right| < \delta\right) = p_o,$$

где $p_o \in (0,1)$ - наперед заданное число (и, следовательно, при использовании этой точечной оценки $\hat{\Theta}_n$ для нахождения приближенного значения неизвестного параметра Θ с вероятностью p_o погрешность $\left|\hat{\Theta}_n - \Theta\right|$ не превысит величины $\delta > 0$).

Число p_o называют *доверительной вероятностью*. Величина $\alpha = 1 - p_o$ называется *уровнем значимости*.

Условие $\left|\hat{\Theta}_n - \Theta\right| < \delta$, очевидно, означает, что

$$\Theta \in \left(\hat{\Theta}_n - \delta, \hat{\Theta}_n + \delta\right).$$

Этот интервал называется *доверительным интервалом* для параметра Θ при доверительной вероятности p_o. Таким образом, доверительная вероятность есть вероятность того, что доверительный интервал $\left(\hat{\Theta}_n - \delta, \hat{\Theta}_n + \delta\right)$ содержит (накрывает) истинное значение параметра Θ. Соответственно, уровень значимости есть вероятность того, что произошла ошибка и истинное значение параметра Θ не попадает в данный интервал. Доверительную вероятность $p_o = 1 - \alpha$ иногда называют *надежностью*.

Доверие, разумеется, не следует обесценивать. Поэтому значения доверительной вероятности $p_o = 1 - \alpha$ следует выбирать близкими к 1 (а значения уровня значимости, соответственно, близкими к 0): $p_o = 1 - \alpha$ = 0,9; 0,95; 0,99; 0,995. В этом случае событие, состоящее в том, что истинное значение оцениваемого параметра лежит в найденном доверительном интервале, является практически достоверным.

Замечание 1. При извлечении выборок объема n из одной и той же генеральной совокупности в $p_o \cdot 100\%$ случаях параметр Θ будет накрываться доверительным интервалом, найденным по доверительной вероятности p_o и выборочному значению оценки $\hat{\Theta}_{n\text{выб}}$.

Замечание 2. Длина доверительного интервала (например, для математического ожидания), найденная по конкретной реализации выборки, является, в определенной мере, показателем качества проведенного статистического исследования. Если она получилась слишком большой, следует проанализировать имеющиеся выборочные значения на предмет наличия грубых погрешностей измерения или провести дополнительные опыты с целью увеличения объема выборки.

Замечание 3. К сожалению, методика нахождения доверительных интервалов в полной мере разработана для нормальных выборок (т.е. выборок из нормально распределенных ГС), которые наиболее часто встречаются на практике. Далее мы приведем формулы для нахождения границ доверительных интервалов математического ожидания и дисперсии по результатам нормальных выборок. Для других типов распределений эти формулы следует рассматривать как определенное приближение истинных значений.

1.5.2. Доверительный интервал для математического ожидания нормально распределенной случайной величины с известным среднеквадратическим отклонением

Пусть случайная величина X распределена нормально (см. раздел 3), причем известно ее среднеквадратическое отклонение (стандартная ошибка измерений) σ. Требуется при доверительной вероятности p_o по выборке $X_1, X_2, ..., X_n$ (представляющей собой n независимых случайных величин, имеющих тот же закон распределения, что и величина X) найти доверительный интервал для математического ожидания $a = M[X]$.

В качестве оценки математического ожидания берем, как и ранее, среднее арифметическое

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

В курсе теории вероятностей доказывается, что если независимые случайные величины имеют одно и то же нормальное распределение $N(a,\sigma)$, то их среднее арифметическое имеет нормальное распределение $N\left(a, \frac{\sigma}{\sqrt{n}}\right)$. Тогда случайная величина $\hat{X}_n = \frac{\overline{X}_n - a}{\sigma/\sqrt{n}}$ имеет стандартизированное нормальное распределение $N(0,1)$.

Определим величину погрешности $\delta > 0$, исходя из уравнения $P\left(\left|\hat{X}_n\right| < \delta\right) = p_o$:

$$p_o = P\left(\left|\hat{X}_n\right| < \delta\right) = P\left(\hat{X}_n < \delta\right) - P\left(\hat{X}_n < -\delta\right) =$$

$$= P\left(\hat{X}_n < \delta\right) - \left(1 - P\left(\hat{X}_n < \delta\right)\right) = 2P\left(\hat{X}_n < \delta\right) - 1,$$

откуда $P\left(\hat{X}_n < \delta\right) = \dfrac{1+p_O}{2}$. Следовательно, $\delta = u_{\frac{1+p_O}{2}}$ - квантиль распределения $N(0,1)$ порядка $\dfrac{1+p_O}{2}$ (квантили стандартизированного нормального распределения см.в табл. 3.3). При выполнении преобразований мы использовали свойство симметричности распределения $N(0,1)$.

Очевидно, что неравенство $\left|\hat{X}_n\right| < \delta$ эквивалентно неравенству

$$\left|\overline{X}_n - a\right| < \delta \cdot \frac{\sigma}{\sqrt{n}}.$$

Следовательно,

$$P\left(\left|\overline{X}_n - a\right| < u_{\frac{1+p_O}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = p_o,$$

т.е. интервал

$$\left(\overline{X}_n - u_{\frac{1+p_O}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X}_n + u_{\frac{1+p_O}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

является доверительным интервалом для математического ожидания $a = M[X]$ при доверительной вероятности p_o.

Замечание 1. Если мы имеем конкретную реализацию $x_1, x_2, \ldots, x_n$ выборки $X_1, X_2, \ldots, X_n$, то в формулу доверительного интервала надо подставить выборочное значение $\overline{X}_{n\,выб}$.

Замечание 2. Очевидно, что

$$u_{\frac{1+p_O}{2}} = u_{1-\frac{\alpha}{2}},$$

где $\alpha = 1 - p_o$ - уровень значимости.

Замечание 3. Конечно, желательно получить доверительный интервал возможно более узким, т.е. уменьшить величину $u_{\frac{1+p_o}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Но мы видим, что если заданы доверительная вероятность p_o и стандартная ошибка σ каждого измерения, то единственное, чем мы можем манипулировать, - это число измерений. За увеличение доверительной вероятности и за уменьшение доверительного интервала приходится платить увеличением числа измерений или повышением их точности.

Замечание 4. При достаточно большом объеме выборки n полученные формулы можно применять и при приближенной оценке границ доверительного интервала для $M[X]$ случайной величины X, не имеющей нормального распределения. Практически во многих случаях это приближение уже при $n > 10$ является хорошим, а при $n > 30$ - очень хорошим.

Основанием для этого служит ***центральная предельная теорема*** – наиболее важная теорема статистики. В ее разработке принимали участие крупнейшие математики – Муавр, Лаплас, Гаусс, Чебышев, Ляпунов и др. Ее краткая формулировка: для любой случайной величины X с конечными математическим ожиданием a и дисперсией σ^2 при стремлении объема выборки n к бесконечности распределение среднего арифметического $\overline{X}_n$ стремится к нормальному закону $N\left(a, \frac{\sigma}{\sqrt{n}}\right)$.

1.5.3. Доверительный интервал для математического ожидания нормально распределенной случайной величины с неизвестным среднеквадратическим отклонением

В практических задачах чаще всего среднеквадратическое отклонение исследуемой случайной величины X неизвестно, и его также нужно оценивать по результатам выборки. Как приблизиться к истине в такой ситуации?

Будем использовать предыдущие обозначения. Т.к. среднеквад-

ратическое отклонение σ неизвестно, в качестве его оценки возьмем несмещенную и состоятельную оценку (см. 1.4.4)

$$s_n = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2} \ .$$

Аналогично 1.5.2 рассмотрим величину

$$\hat{X}_n = \frac{\overline{X}_n - a}{s_n / \sqrt{n}} .$$

В данном случае, особенно при относительно малых значениях n, нет никакой гарантии, что случайная величина $\hat{X}_n$ распределена по нормальному закону (даже приближенно), поэтому использование для нахождения доверительного интервала большой функции Лапласа было бы некорректно и могло привести к большим ошибкам. Как же поступить? В 1908 году английский химик и математик В.Госсет, публиковавший свои труды под псевдонимом "Стьюдент", установил, что эта случайная величина распределена по закону Стьюдента с n -1 степенью свободы (см. раздел 5).

Теперь, используя свойство симметричности распределения Стьюдента (см. 1.5.2), находим доверительный интервал для математического ожидания $a = M[X]$ при доверительной вероятности p_o:

$$\left(\overline{X}_n - t_{n-1, \frac{1+p_o}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X}_n + t_{n-1, \frac{1+p_o}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right),$$

где $t_{n-1, \frac{1+p_o}{2}}$ - квантиль порядка $\frac{1+p_o}{2}$ распределения Стьюдента с n - 1 степенью свободы.

Замечание 1. Если мы имеем конкретную реализацию $x_1, x_2, \ldots, x_n$ выборки $X_1, X_2, \ldots, X_n$, в формулу доверительного интервала надо подставить выборочное значение $\overline{X}_{n\,выб}$.

Замечание 2. Очевидно, что

$$t_{n-1,\frac{1+p_o}{2}} = t_{n-1,1-\frac{\alpha}{2}},$$

где $\alpha = 1 - p_o$ - уровень значимости.

Замечание 3. Чем выше надежность (коэффициент доверия) p_o, тем больше квантиль $t_{n-1,\frac{1+p_o}{2}}$, т.е. тем ниже точность оценки. При этом значение квантиля $t_{n-1,\frac{1+p_o}{2}}$ сильно увеличивается с уменьшением значения n. Поэтому при малых объемах выборки мы можем гарантировать лишь относительно невысокую точность.

1.5.4. Доверительный интервал для оценки дисперсии нормально распределенной случайной величины

Будем использовать обозначения, принятые в 1.5.2. Требуется при доверительной вероятности p_o по выборке $X_1, X_2, ..., X_n$ найти доверительный интервал для дисперсии $\sigma^2 = D[X]$ случайной величины X, имеющей распределение $N(a,\sigma)$.

Рассмотрим несмещенную и состоятельную оценку дисперсии

$$s_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

(см. 1.4.4). Напомним, что элементы выборки X_i так же, как и величина X, имеют распределение $N(a,\sigma)$. Представим их в виде

$$X_i = a + \sigma \cdot \xi_i,$$

где $\xi_1, \xi_2, ..., \xi_n$ - независимые величины, имеющие стандартизированное нормальное распределение $N(0,1)$. Тогда

$$\overline{X}_n = a + \overline{\xi}_n,$$

где

$$\overline{\xi}_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n},$$

и мы получаем

$$s_n^2 = \frac{\sigma^2}{n-1} \cdot \sum_{i=1}^{n} \left(\xi_i - \overline{\xi}_n\right)^2 .$$

Рассмотрим случайную величину

$$\hat{D}_n = \frac{s_n^2 \cdot (n-1)}{\sigma^2} = \sum_{i=1}^{n} \left(\xi_i - \overline{\xi}_n\right)^2 .$$

Доказано, что сумму $\sum_{i=1}^{n} \left(\xi_i - \overline{\xi}_n\right)^2$ можно представить в виде

$$\sum_{i=1}^{n} \left(\xi_i - \overline{\xi}_n\right)^2 = \sum_{i=1}^{n-1} \eta_i^2 ,$$

где $\eta_1, \eta_2, \ldots, \eta_{n-1}$ - независимые случайные величины, имеющие стандартизированное нормальное распределение $N(0,1)$.

Следовательно (см. раздел 4), случайная величина $\hat{D}_n$ имеет χ^2 - распределение с $n-1$ степенями свободы.

Положим $\alpha = \dfrac{1-p_o}{2}$ и определим (по табл. 4.1) квантили $\chi^2_{n-1,\,\alpha}$ и $\chi^2_{n-1,\,1-\alpha}$ этого распределения. Находим вероятность:

$$P\left(\chi^2_{n-1,\,\alpha} < \hat{D}_n < \chi^2_{n-1,\,1-\alpha}\right) = P\left(\hat{D}_n < \chi^2_{n-1,\,1-\alpha}\right) - P\left(\hat{D}_n < \chi^2_{n-1,\,\alpha}\right) =$$

$$= 1 - 2\alpha = p_o .$$

Выполняя простые преобразования, видим, что неравенство

$$\chi^2_{n-1,\alpha} < \hat{D}_n < \chi^2_{n-1,1-\alpha}$$

эквивалентно неравенству

$$\frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,1-\alpha}} < \sigma^2 < \frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,\alpha}}.$$

Следовательно,

$$P\left(\frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,1-\alpha}} < \sigma^2 < \frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,\alpha}} \right) = p_o,$$

т.е. интервал

$$\left(\frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,1-\alpha}}; \frac{s_n^2 \cdot (n-1)}{\chi^2_{n-1,\alpha}} \right)$$

является доверительным интервалом для дисперсии $\sigma^2 = D[X]$ при доверительной вероятности p_o.

Замечание. Если мы имеем конкретную реализацию $x_1, x_2, ..., x_n$ выборки $X_1, X_2, ..., X_n$, то в формулу доверительного интервала надо подставить выборочное значение $s^2_{n\,выб}$.

2. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Во многих случаях нам требуется на основе тех или иных данных решить вопрос об истинности некоторого суждения. Например, верно ли, что два набора данных исходят из одного и того же источника, что *A* - лучший стрелок, чем *B*, что от дома до работы быстрее доехать на метро, а не на автобусе, и т.д. Если мы считаем, что исходные данные для таких суждений в той или иной мере но-

сят случайный характер, то ответы можно дать лишь с определенной степенью уверенности, и существует некоторая вероятность ошибиться. Например, предложив двум персонам A и B выстрелить по три раза в мишень и осмотрев результаты стрельбы, мы лишь предположительно можем сказать, кто из них лучший стрелок: ведь возможно, что победителю просто повезло, и он по чистой случайности стрелял намного точнее, чем обычно. Поэтому при ответе на подобные вопросы хотелось бы не только уметь принимать наиболее обоснованные решения, но и оценивать вероятность ошибочности принятого решения.

Рассмотрение таких задач в строго математической постановке приводит к понятию статистической гипотезы. Далее мы узнаем, что такое статистические гипотезы, какие существуют способы их проверки, каковы наилучшие методы действий.

2.1. Статистические гипотезы

Определение. *Статистической гипотезой* называется любое предположение относительно закона распределения или параметров некоторой генеральной совокупности (ГС), которое мы хотим проверить по результатам выборки.

Проверяемая (основная) гипотеза называется *нулевой гипотезой* и обычно обозначается H_0. Отрицание нулевой гипотезы называется *альтернативной гипотезой* H_a.

Естественно, в некотором смысле гипотезы H_0 и H_a совершенно равносильны: одна является отрицанием другой, и поэтому справедливой может оказаться одна и только одна из них. Какую же принять в качестве основной? Обычно основной гипотезой считают ту, которую наиболее важно не отвергнуть в случае, если она на самом деле верна (т.е. не совершить большой ошибки и не выплеснуть из ванны, как говорил Гегель, вместе с грязной водой и ребенка).

К выбору основной гипотезы H_0 следует подходить следующим образом. Во-первых, принимать во внимание косвенные факторы (например, графические представления выборки, ее выборочные характеристики). Например, если в выборке есть отрицательные элементы, то уже нельзя выдвинуть гипотезу, что это выборка из

ГС, распределенной по логнормальному или экспоненциальному распределению; если элементами выборки являются только натуральные числа, то маловероятно, что мы имеем дело со случайной величиной непрерывного типа. Во-вторых, при выдвижении основной гипотезы часто используются различные соображения, подтвержденные практикой в данной отрасли науки (см. раздел 3).

Принимая решение об истинности той или иной статистической гипотезы, мы можем совершить ошибку, связанную с тем, что вывод делается на основании случайно полученной выборки. Можно выделить два вида ошибок. Во-первых, основная гипотеза H_0 может быть отклонена, хотя в действительности она верна, и принята гипотеза H_a. Эту ошибку называют *ошибкой первого рода.* Во-вторых, гипотеза H_0 может быть принята, хотя в действительности она неверна (т.е. фактически верна гипотеза H_a). Такая ошибка называется *ошибкой второго рода.* В математической статистике, в первую очередь, стремятся, чтобы вероятность возникновения ошибки первого рода была мала. Конечно, было бы хорошо, если бы при этом и вероятность ошибки второго рода была невелика. Но, как правило, оценить эту вероятность не удается.

2.2. Проверка статистических гипотез

При рассмотрении статистических гипотез в математической статистике используются косвенные проверки: проверяются следствия, логически вытекающие из содержания гипотезы, и применяется правило: если по результатам выборки мы получили соотношения, практически невероятные при условии истинности гипотезы, то гипотезу следует отвергнуть. В противном случае гипотеза принимается.

Ясно, что подтверждение следствия не означает однозначно справедливости гипотезы, поскольку правильное следствие может вытекать и из неверной предпосылки. Поэтому правила принятия статистических гипотез носят название критериев согласия, – когда мы согласны с тем, что гипотеза не противоречит реальности, и не отвергаем ее.

2.2.1. Критерии согласия

Суждения о справедливости основной гипотезы H_o или альтернативной гипотезы H_a делаются на основании реализации $\{x_1, x_2, ..., x_n\}$ выборки $\{X_1, X_2, ..., X_n\}$ объема n независимых случайных величин, одинаково распределенных с изучаемой случайной величиной X (см. 1.4). При этом правило, с помощью которого принимается решение о справедливости одной из этих гипотез, называется *статистическим критерием*, или *критерием согласия*.

Каковы же основные принципы построения статистических критериев?

Выбирается малое число $\alpha \in (0,1)$. Условимся считать событие практически невозможным, если вероятность его появления меньше, чем α. Число α называют *уровнем значимости*. Естественно, этот уровень надо выбирать достаточно маленьким. По традиции его берут равным одному из чисел: 0,005; 0,01; 0,025; 0,05; 0,10 (хотя это не означает, что нельзя взять какое-то другое значение, например, 0,03).

Далее в зависимости от конкретной задачи выбирается функция $T_n = T_n(X_1, X_2, ..., X_n)$ от элементов выборки, которая называется *статистикой критерия*. Используя эту функцию, мы можем определить множество V_a, исходя из равенства $P(T_n \in V_a | H_o) = \alpha$ (которое означает, что вероятность попадания значения статистики T_n во множество V_a при условии истинности основной гипотезы H_o равна уровню значимости α). Множество V_a называют *критической областью критерия*. Поскольку попадание значения статистики T_n в критическую область в предположении, что верна гипотеза H_o, есть событие практически невозможное, то в случае наступления этого события гипотеза H_o должна быть отклонена. Это означает, что следует отвергнуть основную гипотезу, если выборочное значение статистики критерия $T_{n\,выб} = T_n(x_1, x_2, ..., x_n)$, найденное

по реализации выборки $\{x_1, x_2, ..., x_n\}$, удовлетворяет условию: $T_{n\,выб} \in V_a$.

Замечание 1. При такой конструкции критерия согласия мы с вероятностью α можем отклонить основную (нулевую) гипотезу H_o при условии, что она является истинной. Иными словами, уровень значимости α есть вероятность совершения ошибки первого рода.

Замечание 2. Если выборочное значение статистики критерия не попадает в критическую область (т.е. $T_{n\,выб} \notin V_a$), то нет оснований для того, чтобы отвергнуть основную гипотезу. Другими словами, в данном случае мы принимаем гипотезу H_o. При этом существует вероятность совершить ошибку второго рода, но оценить эту вероятность практически невозможно. Уменьшить вероятность ошибки второго рода можно, используя для проверки гипотезы несколько различных критериев или же увеличивая объем выборки.

Замечание 3. Если уровень значимости увеличивать, то, очевидно, и критическая область будет увеличиваться. Следовательно, при прочих равных условиях гипотеза будет чаще отвергаться, – даже в том случае, когда она верна (т.е. вероятна ошибка первого рода), что чревато большими потерями: выпуском бракованной продукции, пропуском самолета противника и т.п. Если же уровень значимости уменьшать, область принятия гипотезы увеличивается, а критическая область суживается, и гипотеза H_o будет все реже отвергаться, – даже в тех случаях, когда она не является справедливой. Критерий в этом случае становится малочувствительным.

Таким образом, увеличение уровня значимости ведет к увеличению вероятности ошибки первого рода, называемой "пропуском", уменьшение - к увеличению вероятности ошибки второго рода – принятия гипотезы в случаях, когда она не является справедливой, – так называемой "ложной тревоги", т.е. к уменьшению *мощности критерия*.

Можно еще раз отметить: единственный способ уменьшить вероятность обеих ошибок состоит в увеличении размера выборки n.

2.2.2. Некоторые замечания к практическому использованию критериев согласия

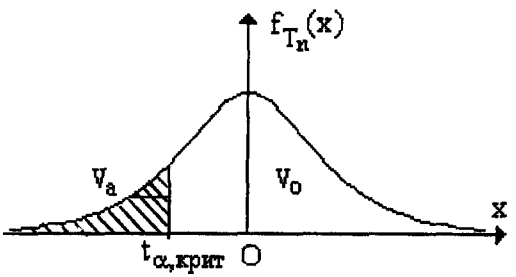
При практическом использовании критерия согласия надо знать, как определяется критическая область V_a. Что для этого необходимо?

Прежде всего, надо учитывать закон распределения статистики

$$T_n = T_n(X_1, X_2, ..., X_n)$$

критерия при условии, что верна гипотеза H_o. Другими словами, надо знать плотность $f_{T_n}(x)$ распределения статистики при условии истинности нулевой гипотезы. В зависимости от того, какие значения может принимать статистика T_n, критическая область V_a может быть правосторонней, левосторонней, двусторонней. Рассмотрим возможные случаи.

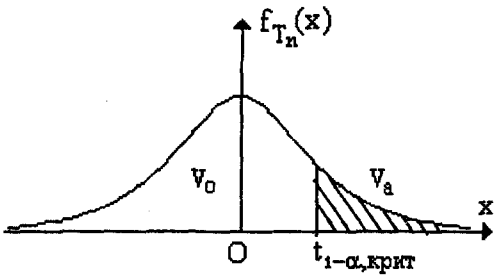
1. *В случае левосторонней критической области* критическая область V_a представляет собой интервал $(-\infty, t_{\alpha,крит})$, где $t_{\alpha,крит}$ - квантиль порядка α распределения случайной величины T_n. Напомним определение из теории вероятностей:



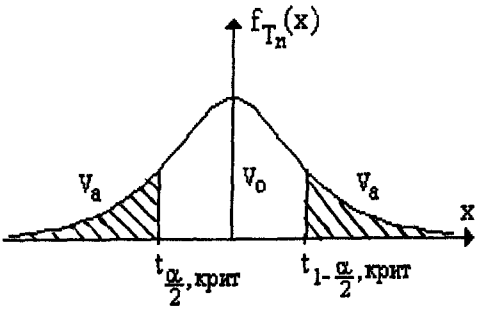
квантилем порядка α называется такое число $t_{\alpha,крит}$, при котором вероятность того, что случайная величина T_n примет значение, не превосходящее $t_{\alpha,крит}$, равна α (т.е. квантиль $t_{\alpha,крит}$ фактически является решением уравнения $P(T_n \le t_{\alpha,крит}) = \alpha$ с заданным значением α). Напомним также факт из теории вероятностей: $P(T_n \le t_{\alpha,крит}) = \int_{-\infty}^{t_{\alpha,крит}} f_{T_n}(x)\,dx$, т.е. площадь заштрихованной фигуры на рисунке равна α.

Как на практике принимается решение по выдвинутой гипотезе в данном случае? Вначале по реализации выборки $\{x_1, x_2, ..., x_n\}$ находится выборочное значение статистики критерия $T_{n\,выб} = T_n(x_1, x_2, ..., x_n)$ (*наблюдаемое значение*). Затем в случае выполнения неравенства $T_{nвыб} < t_{\alpha,крит}$ (когда статистика критерия T_n приняла значение из критической области V_a) принимают решение, что основную гипотезу следует отвергнуть. В случае же выполнения неравенства $T_{nвыб} \geq t_{\alpha,крит}$ основную гипотезу отвергать нет основания.

2. *В случае правосторонней критической области* критическая область V_a представляет собой интервал $(t_{1-\alpha,крит}, +\infty)$, где $t_{1-\alpha,крит}$ - квантиль порядка 1 - α распределения случайной величины T_n (площадь заштрихованной фигуры на рисунке равна α). В этом случае при выполнении неравенства $T_{nвыб} > t_{1-\alpha,крит}$ основную гипотезу следует отвергнуть. В случае же выполнения неравенства $T_{nвыб} \leq t_{1-\alpha,крит}$ основную гипотезу отвергать нет основания.



3. *В случае двусторонней критической области* критическая область V_a представляет собой объединение интервалов $(-\infty, t_{\frac{\alpha}{2},крит})$ и $(t_{1-\frac{\alpha}{2},крит}, +\infty)$, где $t_{\frac{\alpha}{2},крит}$ и $t_{1-\frac{\alpha}{2},крит}$ - соответственно квантили порядка $\frac{\alpha}{2}$ и



$1-\frac{\alpha}{2}$ распределения случайной величины T_n (суммарная площадь заштрихованных фигур на рисунке равна α). В этом случае при выполнении одного из неравенств $T_{nвыб} > t_{1-\frac{\alpha}{2},крит}$ или $T_{nвыб} < t_{\frac{\alpha}{2},крит}$ основную гипотезу следует отвергнуть. В случае же выполнения двойного неравенства $t_{\frac{\alpha}{2},крит} \le T_{nвыб} \le t_{1-\frac{\alpha}{2},крит}$ основную гипотезу отвергать нет основания.

В заключение отметим, что если график плотности $f_{T_n}(x)$ статистики критерия симметричен относительно оси ординат (как, например, это имеет место для стандартизированного нормального распределения $N(0,1)$ и t-распределения Стьюдента), то для квантилей распределения T_n справедливо соотношение: $t_{\frac{\alpha}{2},крит} = -t_{1-\frac{\alpha}{2},крит}$ (это хорошо иллюстрирует третий из рисунков, приведенных выше). В этом случае условие, при котором гипотезу H_o следует отвергнуть, записывается следующим образом:

$$\left|T_{nвыб}\right| > t_{1-\frac{\alpha}{2},крит}.$$

2.2.3. Проверка гипотезы о значении математического ожидания нормального распределения

Рассмотрим в этом пункте в качестве примера следующую задачу. Предположим, нам известно, что случайная величина X имеет нормальное распределение. Требуется по выборке значений этой величины проверить гипотезу о том, что ее математическое ожидание (среднее значение) равно заданной величине a_o, т.е.

$$M[X] = a_o.$$

Этой задаче можно придать следующий конкретный смысл. Предположим, мы купили некий автомат, штампующий детали, и в технической документации говорится, что номинальный диаметр

этих деталей должен равняться a_o. Нам требуется по n изготовленным деталям проверить, действительно ли номинальный диаметр равен a_o (т.е. хорошо ли проведена настройка этого автомата). В такой постановке задачи (см. подраздел 3.1) мы можем считать, что диаметр изготавливаемых деталей есть нормально распределенная случайная величина.

Рассмотрим два случая.

1. Среднеквадратическое отклонение σ (стандартная ошибка) известно. Другими словами, в задаче об автомате, описанной выше, величина стандартной ошибки указана в документации на этот автомат.

Итак, мы имеем выборку $\{X_1, X_2, ..., X_n\}$ объема n независимых случайных величин, имеющих, как и исследуемая случайная величина X, распределение $N(a, \sigma)$. Требуется, используя эту выборку, при уровне значимости α проверить нулевую гипотезу $H_o: a = a_o$ при альтернативной гипотезе $H_a: a \neq a_o$.

Для проверки гипотезы H_0 выбираем статистику критерия в следующем виде:

$$T_n = \sqrt{n} \cdot \frac{\overline{X}_n - a_o}{\sigma},$$

где $\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$.

Выбор именно такой статистики можно обосновать следующим образом. Во-первых, статистика $\overline{X}_n$ является, как мы знаем, оценкой математического ожидания $M[X] = a$, и, значит, разность $\overline{X}_n - a_o$ естественно характеризует степень близости величин $M[X]$ и a_o. Во-вторых, из курса теории вероятностей известно, что случайная величина T_n (при условии правильности нулевой гипотезы!) имеет стандартизированное нормальное распределение $N(0, 1)$.

Далее, статистика T_n может принимать как большие положительные значения, так и большие по модулю отрицательные значе-

ния. Следовательно, мы должны рассматривать случай двусторонней критической области.

Учитывая вышесказанное, нулевую гипотезу можно принять, если выборочное (наблюдаемое) значение статистики T_n критерия удовлетворяет неравенству

$$\left|T_{n\text{выб}}\right| \le t_{1-\frac{\alpha}{2},\text{крит}},$$

где $t_{1-\frac{\alpha}{2},\text{крит}} = u_{1-\frac{\alpha}{2}}$ - квантиль порядка $1-\frac{\alpha}{2}$ стандартизированного нормального распределения (см. табл. 3.2). В противном случае принимается альтернативная гипотеза (и, соответственно, отвергается основная гипотеза).

2. Среднеквадратическое отклонение σ (стандартная ошибка) неизвестно. Другими словами, в задаче об автомате, описанной выше, величина стандартной ошибки должна также определяться по результатам выборки.

Однако в статистике T_n вместо среднеквадратического отклонения σ мы должны взять его оценку $s_n = \sqrt{s_n^2}$ - исправленное среднеквадратическое (см. 1.4.4). В этом случае статистика T_n при условии правильности нулевой гипотезы имеет распределение Стьюдента с $n-1$ степенями свободы, и неравенство, при котором нулевая гипотеза принимается, выглядит так:

$$\left|T_{n\text{выб}}\right| \le t_{1-\frac{\alpha}{2},\text{крит}},$$

где $t_{1-\frac{\alpha}{2},\text{крит}} = t_{n-1,1-\frac{\alpha}{2}}$ - квантиль порядка $1-\frac{\alpha}{2}$ распределения Стьюдента с $n-1$ степенями свободы (см. раздел 5).

2.3. Проверка гипотезы об общем виде закона распределения случайной величины

Важной задачей статистической обработки выборок случайной величины является определение ее закона распределения, т.к. это позволяет делать прогнозы относительно попаданий значений этой величины в те или иные множества, моделировать выборки ее значений на компьютере и т.д.

Конкретно задача ставится так. Обозначим через $F^*(x)$ истинную функцию распределения случайной величины X, а через $F_o(x)$ - гипотетическую функцию распределения. Основная гипотеза H_o состоит в следующем:

$$F^*(x) = F_o(x).$$

Требуется на основании реализации выборки $\{x_1, x_2, ..., x_n\}$ при некотором уровне значимости α сделать заключение о соответствии этой гипотезы реальной действительности.

В вопросе о выборе гипотетической функции $F_o(x)$ надо задаться ее формульным выражением. Общие принципы решения данной проблемы уже обсуждались.

Гипотеза о законе распределения ГС является одной из обычных статистических гипотез. В этом разделе мы приведем наиболее часто применяющиеся на практике критерии проверки гипотез о законе распределения.

Гипотезы о законе распределения разделяют на два типа: простые и сложные. Гипотезу называют *простой*, если параметры, от которых должна зависеть гипотетическая функция распределения, являются заданными. Если же мы знаем только вид этой функции, но параметры, от которых она зависит, должны определять по результатам выборки, гипотеза называется *сложной*.

2.3.1. Критерий согласия Пирсона (критерий хи-квадрат)

В критерии согласия Пирсона в качестве меры расхождения между теоретическим и эмпирическим распределениями выбирается специальная статистика, закон распределения которой приближенно равен известному χ^2- распределению (см. раздел 4).

Применение данного критерия для проверки простой гипотезы (случай, когда гипотетическая функция распределения $F_o(x)$ полностью задана) обосновано следующей теоремой.

Теорема К. Пирсона. Пусть n - число независимых повторений некоторого опыта, который может закончиться одним из r элементарных исходов $A_1, A_2, ..., A_r$; положительные числа $p_1, p_2, ..., p_r$ - вероятности появления этих исходов; $m_1, m_2, ..., m_r$ - количества опытов, которые заканчиваются исходами $A_1, A_2, ..., A_r$ соответственно, при этом должны быть выполнены условия

$$p_1 + p_2 + ... + p_r = 1; \qquad m_1 + m_2 + ... + m_r = n.$$

Введем случайную величину

$$\chi^2_{n,r} = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}.$$

Справедливо следующее утверждение: при $n \to \infty$ случайная величина $\chi^2_{n,r}$ асимптотически подчиняется распределению χ^2_{r-1} (хи-квадрат) с $r-1$ степенями свободы.

Случайная величина $\chi^2_{n,r}$ как раз и является статистикой критерия согласия Пирсона. При проверке гипотезы о законе распределения по этому критерию числовую прямую разбивают на r непересекающихся интервалов (полуинтервалов). Под событием A_i понимают событие, состоящее в том, что значение случайной величины попадает в i–й интервал, а вероятности p_i определяются с исполь-

зованием гипотетической функции распределения $F_o(x)$. Более подробно этот вопрос будет рассмотрен ниже.

Обсудим поведение статистики $\chi^2_{n,r}$ критерия Пирсона. Сделаем преобразование:

$$\chi^2_{n,r} = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} = n \cdot \sum_{i=1}^{r} \frac{1}{p_i} \cdot \left(\frac{m_i}{n} - p_i \right)^2 . \qquad (2.1)$$

Если верна гипотеза H_o, то по закону больших чисел с вероятностью 1 величины $\frac{m_i}{n} \to p_i$ при $n \to \infty$, причем с учетом теоремы Муавра-Лапласа

$$\left| \frac{m_i}{n} - p_i \right| \le 3 \cdot \sqrt{\frac{p_i(1-p_i)}{n}} . \qquad (2.2)$$

Следовательно, при больших n статистика $\chi^2_{n,r}$ с вероятностью, близкой к 1, не должна принимать большие положительные значения.

Это позволяет сделать вывод, что критическая область этой статистики – правосторонняя (см. 2.2.2). Таким образом, гипотеза H_o должна быть отвергнута, если полученное в результате опыта выборочное (наблюдаемое) значение $\chi^2_{n,r\,выб}$ слишком велико. Здесь, как всегда, слова "слишком велико" означают, что данное значение превосходит критическое значение $\chi^2_{n,r\,крит}$ статистики для заданного уровня значимости. Согласно теореме Пирсона, для простой гипотезы можно брать $\chi^2_{n,r\,крит} = \chi^2_{r-1,\,1-\alpha}$ - квантиль порядка $1 - \alpha$ распределения хи-квадрат с $r - 1$ степенями свободы.

Итак, простая гипотеза о законе распределения отвергается, если выполняется неравенство $\chi^2_{n,r\,выб} > \chi^2_{r-1,\,1-\alpha}$, и принимается в противном случае.

В заключение отметим, что числа p_i в критерии Пирсона называют *теоретическими вероятностями* попадания случайной величины в соответствующие интервалы разбиения числовой прямой, произведения np_i - *теоретическими частотами* наступления этих событий, а величины m_i - *эмпирическими частотами*. С учетом этого можно сказать, что статистика $\chi^2_{n,r}$ есть мера отклонения эмпирических частот и теоретических частот.

Замечание 1. Асимптотический характер теоремы К.Пирсона требует осторожности при его практическом использовании. На нее можно полагаться только при достаточно больших значениях n. Судить же о том, достаточно ли n велико, надо с учетом вероятностей $p_1, p_2, ..., p_r$. Совокупность теоретических и экспериментальных доводов приводит к убеждению, что критерий применим, если все теоретические частоты $np_i \geq 5$. Чтобы соблюсти это требование, на практике приходится объединять некоторые интервалы разбиения.

Замечание 2. Мы изложили применение критерия Пирсона для проверки простых гипотез. Но на практике простые гипотезы встречаются реже, чем сложные, ведь в большинстве случаев теоретические соображения или традиции не идут далее указания типа распределения (нормальный, показательный, пуассоновский), а параметры его остаются неопределенными. Оказывается, критерий Пирсона по сравнению с другими критериями имеет то преимущество, что его статистика вычисляется для сложной гипотезы так же, как и для простой. Отличие в том, что в данной ситуации статистика $\chi^2_{n,r}$ асимптотически подчиняется распределению χ^2_{r-k-1} с $r-k$ -1 степенями свободы, где k – число неизвестных параметров распределения. Эта корректировка связана с необходимостью оценивать неизвестные параметры (и, соответственно, определять гипотетическую функцию распределения $F_o(x)$) по результатам выборки. Например, в случае нормального распределения при двух неизвестных параметрах a и σ число степеней свободы будет равно r - 3.

Вывод: сложная гипотеза о законе распределения с k неизвестными параметрами отвергается, если выполняется неравенство $\chi^2_{n,r\ выб} > \chi^2_{r-k-1,\,1-\alpha}$, и принимается в противном случае.

Замечание 3. Из двух предыдущих замечаний можно сделать вывод, что при проверке гипотезы о законе распределения по критерию Пирсона для сложной гипотезы с двумя неизвестными параметрами объем выборки не может быть меньше 20.

2.3.2. Схема применения критерия Пирсона для проверки сложной гипотезы о законе распределения

Пусть дана выборка (реализация выборки) $\{x_1, x_2, ..., x_n\}$ из n независимых наблюдений случайной величины X.

Выдвинута *гипотеза* H_o: функция распределения случайной величины X имеет вид $F_o(x)$.

Требуется: при уровне значимости α проверить эту гипотезу, используя критерий Пирсона.

Для определенности рассмотрим случай непрерывной случайной величины X, распределение которой зависит от двух параметров Θ_1 и Θ_2 (и, следовательно, ее гипотетическая функция распределения $F_o(x)$ должна быть непрерывной и зависеть от этих параметров $F_o(x) = F_o(x, \Theta_1, \Theta_2)$). Для проверки гипотезы по критерию Пирсона надо произвести следующие действия:

1. Вычислить выборочные среднее $\overline{X}_n$ и исправленную выборочную дисперсию s_n^2. Далее, используя метод моментов (см. 1.4.2 и раздел 3), вычислить оценки параметров гипотетического распределения $\Theta_{1выб}$ и $\Theta_{2выб}$.

Замечание 1. Более строгий подход требует вычисления оценок параметров распределения методом максимального правдоподобия, а не методом моментов. Но для наиболее часто используемых распределений (нормального, показательного, Пуассона) эти оценки совпадают, а достаточная простота метода моментов компенсирует его недостатки.

2. Выборку представить в виде группированного (интервального) статистического ряда (методику этого процесса см. в 1.3.4, только следует положить $y_o = -\infty$, $y_r = +\infty$).

Интервал	$(-\infty, y_1)$	$[y_1, y_2)$	...	$[y_{r-1}, +\infty)$
Частота	m_1	m_2	...	m_r

3. Подсчитать теоретические вероятности попадания значений случайной величины в интервалы группировки

$$p_i = F_o(y_i) - F_o(y_{i-1}), i = 1,2,...,r,$$

где в формульное определение гипотетической функции распределения $F_o(x)$ должны быть подставлены найденные ранее выборочные оценки неизвестных параметров. Напомним, что $F_o(y_o) = F_o(-\infty) = 0$ и $F_o(y_r) = F_o(+\infty) = 1$. Кстати, мы должны были положить $y_o = -\infty$, $y_r = +\infty$ для того, чтобы соблюсти требование $p_1 + p_2 + ... + p_r = 1$ теоремы Пирсона.

Замечание 2. Для дискретной случайной величины X теоретические вероятности лучше находить по формуле

$$p_i = \sum_{x_j \in [y_{i-1}, y_i)} P(X = x_j),$$

где $P(X = x_j)$ - вероятность того, что случайная величина X примет значение x_j, а суммирование ведется по тем значениям x_j, которые попадают в полуинтервал $[y_{i-1}, y_i)$.

4. Для всех интервалов должно выполняться условие: $np_i \geq 5$. Если для какого-либо интервала это условие нарушается, его надо объединить с соседним интервалом, при этом следует просуммировать их частоты, а также теоретические вероятности (число r интервалов группировки естественно уменьшится).

5. Вычислить выборочное (наблюдаемое) значение $\chi^2_{n,r\text{выб}}$ статистики критерия (см. формулу (2.1)). По табл. 4.1 квантилей χ^2 - распределения найти критические значения при заданном уровне зна-

чимости α и числе степеней свободы $r-3$ (напомним, что мы рассматриваем случай $k=2$): $\chi^2_{n,r\ крит} = \chi^2_{r-3,1-\alpha}$.

6. Если $\chi^2_{n,r\ крит} > \chi^2_{r-3,1-\alpha}$, то нулевую гипотезу отвергают. В противном случае оснований отвергать нулевую гипотезу нет.

Замечание 3. Если проверяется простая гипотеза, то пункт 1 выполнять не надо, а в пунктах 5 и 6 число степеней свободы будет равно $r-1$.

2.3.3. Примеры применения критерия Пирсона для проверки гипотезы о законе распределения

Пример 1. На компьютере проведено моделирование выборки из 500 значений случайной величины, равномерно распределенной на отрезке [0; 12]. По результатам выборки составлен группированный статистический ряд.

Интервал	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)	[8, 9)	[9,10)	[10, 11)	[11, 12]
Частота	41	34	54	39	49	45	41	33	37	41	47	39

Мы должны выяснить, согласуются ли эти данные с гипотезой H_o о том, что мы действительно имеем дело с выборкой значений случайной величины, равномерно распределенной на отрезке [0; 12] (т.е. проверить качество моделирования). Уровень значимости примем $\alpha = 0{,}05$.

Решение. В данном примере мы рассматриваем простую гипотезу, т.к. гипотетическая функция распределения однозначно определяется из условия (см. 3.5):

$$F_o(x) = \begin{cases} 0, x < 0; \\ \dfrac{x}{12}, x \in [0,12]; \\ 1, x > 12. \end{cases}$$

Нетрудно видеть, что все теоретические вероятности равны $p_i = \frac{1}{12}$, объем выборки равен $n = 500$ (следовательно, условие $np_i \geq 5$ выполняется для всех интервалов).

Находим наблюдаемое (выборочное) значение статистики критерия Пирсона (формула (2.1)):

$$\chi^2_{500,12\ выб} = 500 \cdot \sum_{i=1}^{12} 12 \cdot \left(\frac{m_i}{500} - \frac{1}{12} \right)^2 = 9{,}04 .$$

Число степеней свободы равно $r - 1 = 12 - 1 = 11$. По табл. 4.1 квантилей χ^2-распределения находим критическое значение статистики $\chi^2_{n,r\ крит} = \chi^2_{11,0,95} = 19{,}7$. Мы видим, что $\chi^2_{500,12\ выб} < 19{,}7$.

Вывод. Гипотетическое распределение согласуется с экспериментальными данными, т.е. нет оснований отвергать гипотезу H_o.

Пример 2. Через равные промежутки времени в тонком слое раствора золота регистрировалось число частиц золота, попадавших в поле зрения микроскопа. Результаты наблюдений приведены в следующей таблице.

x_i	0	1	2	3	4	5	6	7
m_i	112	168	130	68	32	5	1	1

В первой строке приведены регистрировавшиеся значения x_i частиц золота, во второй – соответствующие частоты m_i (число интервалов времени, в течение которых в поле зрения попало ровно x_i частиц).

Требуется: используя критерий Пирсона и приняв за уровень значимости $\alpha = 0{,}05$, проверить согласие полученных экспериментальных данных с законом распределения Пуассона.

Решение. Итак, нам надо проверить сложную гипотезу H_o о том, что исследуемая величина X распределена по закону Пуассона с некоторым параметром λ (см. 3.4):

$$H_o : P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!},\ k = 0,1,2,3,\ldots .$$

Поскольку параметр λ распределения Пуассона неизвестен, то, согласно методу моментов, в качестве оценки этого параметра возьмем выборочное среднее

$$\lambda = \overline{X}_{выб} = 1{,}544.$$

Составим интервальный ряд.

Интервал	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7)	$[7, +\infty)$
Частота m_i	112	168	130	68	32	5	1	1
Вероятности p_i	0,2135	0,3297	0,2545	0,1310	0,0506	0,0156	0,0040	0,0011

Теоретические вероятности p_i находим, используя формулу Пуассона при $\lambda = 1, 544$:

$$p_o = P(X = 0) = \frac{1{,}544^o \cdot e^{-1{,}544}}{0!} = 0{,}2135;$$

$$p_1 = P(X = 1) = \frac{1{,}544^1 \cdot e^{-1{,}544}}{1!} = 0{,}3297;$$

$$p_2 = P(X = 2) = 0{,}2545; \qquad p_3 = P(X = 3) = 0{,}1310;$$

$$p_4 = P(X = 4) = 0{,}0506; \qquad p_5 = P(X = 5) = 0{,}0156;$$

$$p_6 = P(X = 6) = 0{,}0040;$$

$$p_7 = P(X \geq 7) = \sum_{k=1}^{\infty} \frac{1{,}544^k \cdot e^{-1{,}544}}{k!} = 1 - \sum_{k=o}^{6} \frac{1{,}544^k \cdot e^{-1{,}544}}{k!} = 0{,}0011.$$

Объем выборки равен $n = 517$. Т.к. требование $np_i \geq 5$ не выполняется для последних трех интервалов, их следует объединить (при этом просуммировав их частоты, а также теоретические вероятности). В результате объединения получим интервальный ряд.

Интервал	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)	$[5, +\infty)$
Частота m_i	112	168	130	68	32	7
Вероятности p_i	0,2135	0,3297	0,2545	0,1310	0,0506	0,0207

По этим данным находим наблюдаемое (выборочное) значение статистики Пирсона (формула (2.1) в 2.3.1):

$$\chi^2_{517,6\ выб} = \sum_{i=1}^{6} \frac{(m_i - 517p_i)^2}{517p_i} = 2{,}663.$$

По табл. 4.1 при уровне значимости $\alpha = 0{,}05$ и числу степеней свободы, равном $r - k - 1 = 6 - 2 = 4$ ($k = 1$ – число неизвестных параметров), находим критическое значение статистики

$$\chi^2_{n,r\ крит} = \chi^2_{4,0.95} = 9{,}49.$$

Мы видим, что $\chi^2_{517,6\ выб} < 9{,}49$.

Вывод. Гипотетическое распределение согласуется с экспериментальными данными, т.е. нет оснований отвергать гипотезу H_o.

2.3.4. Некоторые другие критерии согласия

Оценка вероятности ошибки второго рода (т.е. принятия нулевой гипотезы при условии истинности гипотезы альтернативной) при проверке гипотезы о законе распределения является очень сложной задачей. Поэтому, чтобы после подтверждения нулевой гипотезы по критерию Пирсона иметь большую уверенность в правильности

выбора, имеет смысл проверить эту гипотезу, используя другие критерии согласия.

Из других критериев согласия, наиболее часто применяющихся на практике, можно выделить *критерии согласия Колмогорова и омега-квадрат*. Сразу отметим, что эти критерии применимы только для непрерывных распределений, что несколько сужает область их применения.

Кроме того, распределение статистик этих критериев устроено достаточно просто только для простых гипотез. В случае сложных гипотез их распределение в большей степени зависит от вида гипотетического распределения. Напомним для сравнения, что для статистики критерия Пирсона появление неизвестных параметров влечет за собой только уменьшение числа степеней свободы в предельном распределении хи-квадрат.

Другими словами, статистики критериев Колмогорова и омега-квадрат в случае сложных гипотез не обладают столь привлекательным свойством "свободы от распределения выборки", как их прототипы для простой гипотезы (поэтому для каждого параметрического семейства распределений используются свои таблицы, т.е. надо отдельно определять критическое значение статистики критерия). Тем не менее, рассмотрим кратко суть этих критериев, предварительно сделав следующее замечание.

Если сложная гипотеза подтверждается по критерию Пирсона, то имеет смысл проверить ее с использованием критериев согласия Колмогорова и омега-квадрат, но при этом рассматривая гипотезу как простую (т.е. с уже заданными параметрами).

2.3.5. Критерий согласия Колмогорова для простой гипотезы

Итак, проверяется гипотеза H_o о том, что генеральная совокупность, из которой произведена выборка значений $\{x_1, x_2, ..., x_n\}$, подчиняется закону с непрерывной функцией распределения $F_o(x)$.

Пусть $F_n(x)$ - эмпирическая функция распределения. Для оценки степени отличия функций $F_o(x)$ и $F_n(x)$ вводится величина:

$$D_n = \sup_{x \in R} \left| F_o(x) - F_n(x) \right|.$$

Очевидно, что D_n - случайная величина (статистика), поскольку ее значение зависит от случайного объекта $F_n(x)$. Статистику D_n называют *статистикой Колмогорова.*

Надо отметить, что эмпирическая функция $F_n(x)$ должна определяться только по статистическому ряду (см. 1.3.1); нельзя использовать интервальный ряд.

Если гипотеза H_o справедлива, то в силу теоремы Бернулли для любого числа $x \in R$ и любого $\varepsilon > 0$ выполняется условие

$$\lim_{n \to \infty} P\left(\left| F_o(x) - F_n(x) \right| < \varepsilon \right) = 1.$$

Поэтому с вероятностью, близкой к 1, при больших объемах выборки n значение статистики должно быть мало.

Отсюда следует вывод: гипотеза H_o должна быть отвергнута, если полученное в результате эксперимента выборочное (наблюдаемое) значение статистики $D_{n выб}$ окажется неправдоподобно большим (т.е. больше некоторого критического значения статистики D_n, определенного с учетом уровня значимости). Другими словами, критическая область статистики критерия – правосторонняя (см. 2.2.2).

Естественно, для того, чтобы иметь возможность находить критические значения статистики D_n, надо знать ее распределение. Свойство статистики Колмогорова состоит в том, что ее закон распределения (если гипотеза H_o верна) зависит только от объема выборки и не зависит от функции $F_o(x)$. Асимптотические свойства статистики D_n (при условии истинности нулевой гипотезы) описывает найденная в 1933 г. А.Н. Колмогоровым предельная теорема.

Теорема Колмогорова утверждает, что при условии справедливости гипотезы H_o для любого $\lambda > 0$ существует предел

$$\lim_{n\to\infty} P\left(\sqrt{n}\cdot D_n < \lambda\right) = K(\lambda),$$

где

$$K(\lambda) = 1 + 2\sum_{j=1}^{\infty}(-1)^j e^{-2j^2\lambda^2}.$$

Квантили λ_{np} распределения случайной величины $\sqrt{n}\cdot D_n$ (напомним, что эти числа определяются из уравнения $P\left(\sqrt{n}\cdot D_n < \lambda_{np}\right) = p$) имеются в таблицах.

Таким образом, алгоритм проверки гипотезы следующий. По имеющимся результатам выборки находим выборочное значение статистики критерия $D_{nвыб}$ (что является весьма громоздкой задачей), затем находим величину $\lambda_{выб} = \sqrt{n}\cdot D_{nвыб}$ и сравниваем ее с критическим значением $\lambda_{крит} = \lambda_{n\,1-\alpha}$, где $\lambda_{n\,1-\alpha}$ - квантиль распределения $\sqrt{n}\cdot D_n$, найденный из таблиц по заданному уровню значимости α и объему выборки n. Гипотезу H_0 приходится отвергать при выполнении неравенства $\lambda_{выб} > \lambda_{крит}$.

2.3.6. Критерий согласия омега-квадрат для простой гипотезы

Не вдаваясь в подробности, отметим, что этот критерий основан на так называемой статистике омега-квадрат

$$\omega_n^2 = \int_{-\infty}^{+\infty}\left[F_n(x) - F_o(x)\right]^2 dF_o(x).$$

Известно, что если гипотеза H_o верна, то закон распределения статистики ω_n^2 зависит только от объема выборки и не зависит от функции $F_o(x)$. Н.В.Смирновым в 1939 г. найдено предельное рас-

пределение статистики $n \cdot \omega_n^2$ при условии истинности нулевой гипотезы, которое и используется при практическом применении критерия омега-квадрат. Имеются подробные таблицы квантилей ω_{np} этого распределения.

Для нахождения выборочного значения $\omega^2_{n\,выб}$ статистики $n \cdot \omega_n^2$ по элементам выборки, представленной в виде вариационного ряда $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ (см. 1.3.1), можно использовать формулу

$$\omega^2_{n\,выб} = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_n\left(x_{(i)}\right) - \frac{2i-1}{2n}\right]^2.$$

Гипотезу H_o приходится отвергать при выполнении неравенства $\omega^2_{n\,выб} > \omega^2_{n\,крит}$, где $\omega^2_{n\,крит} = \omega^2_{n\,1-\alpha}$ - квантиль распределения $n \cdot \omega_n^2$, найденный из таблиц по заданному уровню значимости α и объему выборки n.

3. НЕКОТОРЫЕ ЧАСТО ИСПОЛЬЗУЮЩИЕСЯ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

3.1. Нормальное распределение

Случайная величина X непрерывного типа распределена по *нормальному закону с параметрами* $a \in R$; $\sigma > 0$ (сокращенная запись $\mathrm{N}(a,\sigma)$), если ее плотность вероятности (рис. 3.1) задается формулой

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Функция распределения этой случайной величины имеет вид

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}}dt.$$

На рис. 3.1, 3.2 приведены графики плотности и функции распределения нормального распределения.

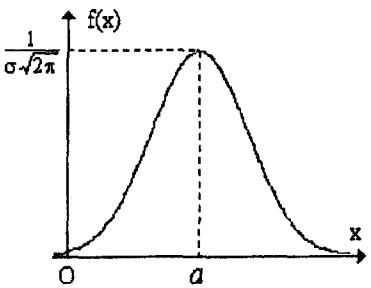


Рис. 3.1

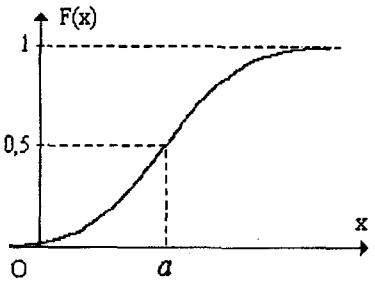


Рис. 3.2

График плотности $f(x)$ называется *кривой Гаусса* (прямая $x=a$ является осью симметрии графика). В связи с этим нормальное распределение еще называют *гауссовским*.

Известно, что параметр $a=M[X]$ есть математическое ожидание, а $\sigma^2=D[X]$ - дисперсия случайной величины X (соответственно σ - ее среднеквадратическое отклонение). Поэтому согласно методу моментов (см. 1.4.2 – 1.4.4) в качестве оценок этих параметров следует брать: $a=\overline{X}_n$, $\sigma^2=s_n^2$ (естественно, если мы имеем конкретную реализацию выборки, то должны положить параметры a и σ^2 равными выборочным значениям статистик $\overline{X}_{выб}$ и $s^2_{выб}$).

Нормальное распределение играет особую роль в теории вероятностей и математической статистике. Как показывает практика, самые разнообразные статистические данные с хорошей степенью точности можно считать выборками из нормально распределенной генеральной совокупности. Примерами этого могут служить помехи в электроаппаратуре, ошибки измерений, разброс попадания снарядов при стрельбе по заданной цели, рост наудачу взятого человека, скорость реакции на раздражитель и т.д. На практике считают (что,

в принципе, обосновано центральной предельной теоремой теории вероятностей): *если случайная величина формируется под воздействием большого числа независимых малых влияний, из которых ни одно не доминирует над остальными, то она подчинена нормальному распределению*. Например, большое число не зависящих друг от друга причин влияют на размер изготавливаемой керамической плитки, диаметр проволоки, разрушающую нагрузку для образца бетона и т.п. Поэтому неудивительно, что все эти виды технических измерений очень хорошо описываются нормальным распределением (со своими характерными значениями).

Замечание 1. Случайная величина X, распределенная по закону $N(0,1)$, называется *стандартизованной нормальной величиной*. Ее плотность вероятности равна

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

и называется *малой функцией Лапласа*.

Функция распределения равна

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

и называется *функцией нормального распределения*, или *большой функцией Лапласа*.

Значения этой функции приведены в табл. 3.2. При использовании данной таблицы следует помнить следующие правила: $\Phi(-x)=1-\Phi(x)$; $\Phi(x)\approx 1$ при $x>3{,}5$; $\Phi(x)\approx 0$ при $x<-3{,}5$ (причем погрешность в этих приближенных равенствах – менее чем 10^{-4}).

Замечание 2. Используя замену переменной в интеграле, легко получить для любой случайной величины, распределенной по закону $N(a,\sigma)$, ее функцию распределения, равную

$$F(x)=\Phi\left(\frac{x-a}{\sigma}\right).$$

Замечание 3. Иногда в литературе используются таблицы значений функции

$$\Phi_o(x)=\frac{1}{\sqrt{2\pi}}\int_o^x e^{-\frac{t^2}{2}}dt.$$

В этом случае надо помнить следующие правила: $\Phi_o(-x)=-\Phi_o(x)$; $\Phi_o(x)=\Phi(x)-0{,}5$; $\Phi_o(0)=0$; $\Phi_o(x)\approx 0{,}5$ при $x\geq 4$.

3.2. Логнормальное распределение

Положительная случайная величина X непрерывного типа распределена по *логнормальному* (*логарифмически нормальному*) *закону с параметрами* $a\in R$ и $\sigma>0$, если ее плотность вероятности задается формулой

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}.$$

Функция распределения этой логнормальной случайной величины имеет вид

$$F(x)=\Phi\left(\frac{\ln x-a}{\sigma}\right),$$

где $\Phi(x)$ - большая функция Лапласа. На рис. 3.3, 3.4 приведены графики этих функций.

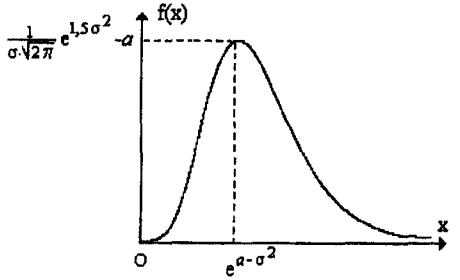


Рис. 3.3

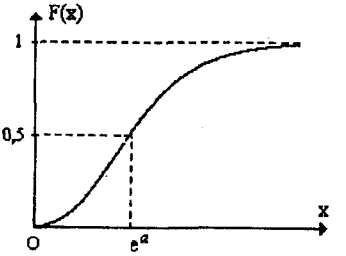


Рис. 3.4

Можно сказать, что случайная величина X подчиняется логнормальному распределению, если ее логарифм, т.е. случайная величина $Y = \ln X$, имеет нормальное распределение.

Известно, что математическое ожидание и дисперсия логнормальной случайной величины X вычисляются по формулам

$$M[X] = e^{a+0,5\sigma^2}; \quad D[X] = e^{2a+\sigma^2} \cdot \left(e^{\sigma^2} - 1\right).$$

Далее согласно методу моментов (см. 1.4.2 – 1.4.4) в качестве оценок этих параметров берутся

$$M[X] = \overline{X}_n, \quad D[X] = s_n^2.$$

В результате получим систему двух уравнений с двумя неизвестными для оценки параметров a и σ по результатам выборки, откуда находим

$$a = \ln \frac{\overline{X}_n^2}{\sqrt{\overline{X}_n^2 + s_n^2}}, \qquad \sigma^2 = \ln\left(\frac{s_n^2}{\overline{X}_n^2} + 1\right)$$

(естественно, если мы имеем конкретную реализацию выборки, то получим конкретную оценку этих параметров $a_{выб}$ и $\sigma_{выб}$).

Логнормальное распределение возникает при изучении моделей дробления частиц, моделей роста и т.д. А.Н. Колмогоров показал, что логарифмически нормальному закону подчинены размеры частиц при дроблении.

3.3. Усеченные нормальные распределения

Случайная величина X непрерывного типа имеет *усеченное слева нормальное распределение с параметрами* $a \in R$, $\sigma > 0$ и $\tau \in (0,1)$ (далее в этом пункте мы используем обозначения: $\varphi(x)$ - малая функция Лапласа, $\Phi(x)$ - большая функция Лапласа), если ее плотность вероятностей имеет вид

$$f(x) = \begin{cases} 0, x < x_o; \\ \dfrac{1}{\sigma(1-\tau)} \cdot \varphi\left(\dfrac{x-a}{\sigma}\right), x \geq x_o \end{cases},$$

где значение x_o определяется из соотношения $\tau = \Phi\left(\dfrac{x_o - a}{\sigma}\right)$ (в принципе, можно задавать значение x_o, а параметр τ находить из указанного соотношения).

Параметр τ называется *степенью усечения*. Функция распределения имеет вид

$$F(x) = \begin{cases} 0, x < x_o; \\ \dfrac{\Phi\left(\dfrac{x-a}{\sigma}\right) - \tau}{1-\tau}, x \geq x_o. \end{cases}$$

На рис. 3.5, 3.6 приведены графики этих функций.

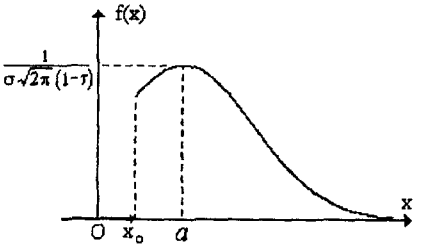


Рис. 3.5

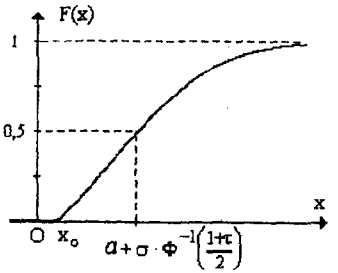


Рис. 3.6

Известно, что математическое ожидание и дисперсия усеченного слева нормального распределения вычисляются по формулам

$$M[X]=a+\sigma^2 f(x_o);\qquad D[X]=\sigma^2 f(x_o)(x_o-M[X])+\sigma^2,$$

где $f(x)$ – плотность распределения. Согласно методу моментов (см. п. 1.4.2 – 1.4.4), в качестве оценок этих параметров следует брать $M[X]=\overline{X}_n$; $D[X]=s_n^2$.

Считая заданной степень усечения τ, получим систему трех уравнений с тремя неизвестными для оценки параметров a, σ и x_o по результатам выборки (ниже обозначено $\gamma=\Phi^{-1}(\tau)$)

$$\begin{cases}\overline{X}_n=a+\sigma\cdot\dfrac{\varphi(\gamma)}{1-\tau};\\ s_n^2=\sigma^2\cdot\left(1+\dfrac{\varphi(\gamma)}{1-\tau}\cdot\left(\gamma-\dfrac{\varphi(\gamma)}{1-\tau}\right)\right);\\ x_o=a+\sigma\cdot\gamma.\end{cases}$$

Ясно, что из второго уравнения легко можно найти оценку для параметра σ, затем из первого – оценку для параметра a, после этого вычислить x_o.

Случайная величина X непрерывного типа имеет *усеченное справа нормальное распределение с параметрами* $a \in R$, $\sigma > 0$ и $\tau \in (0,1)$, если ее плотность вероятностей имеет вид

$$f(x) = \begin{cases} 0, x > x_o; \\ \dfrac{1}{\sigma\tau} \cdot \varphi\left(\dfrac{x-a}{\sigma}\right), x \le x_o, \end{cases}$$

где значение x_o определяется из соотношения $\tau = \Phi\left(\dfrac{x_o - a}{\sigma}\right)$ (можно задавать значение x_o, а степень усечения τ находить из указанного соотношения).

Функция распределения имеет вид

$$F(x) = \begin{cases} 1, x > x_o; \\ \dfrac{1}{\tau}\Phi\left(\dfrac{x-a}{\sigma}\right), x \le x_o. \end{cases}$$

На рис. 3.7, 3.8 приведены графики плотности и функции распределения усеченного справа нормального распределения.

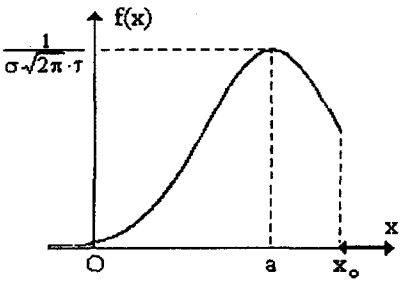


Рис. 3.7

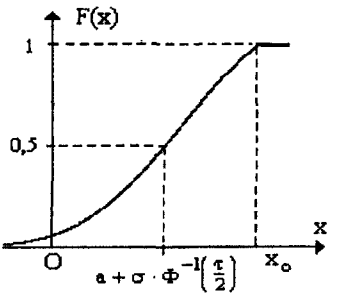


Рис. 3.8

Известно, что математическое ожидание и дисперсия усеченного справа нормального распределения вычисляются по формулам

$$M[X]=a-\sigma^2 f(x_o);\ D[X]=\sigma^2 f(x_o)(M[X]-x_o)+\sigma^2,$$

где $f(x)$ – плотность распределения.

Аналогично усеченному слева нормальному распределению (считая заданной степень усечения τ) получим систему трех уравнений с тремя неизвестными для оценки параметров a, σ и x_o по результатам выборки

$$\begin{cases} \overline{X}_n = a-\sigma\cdot\dfrac{\varphi(\gamma)}{\tau}; \\ s_n^2=\sigma^2\cdot\left(1-\dfrac{\varphi(\gamma)}{\tau}\cdot\left(\gamma+\dfrac{\varphi(\gamma)}{\tau}\right)\right); \\ x_o=a+\sigma\cdot\gamma. \end{cases}$$

Здесь $\gamma=\Phi^{-1}(\tau)$.

3.4. Распределение Пуассона

Случайная величина X дискретного типа имеет *распределение Пуассона с параметром* $\lambda>0$, если она принимает целые значения 0, 1, 2, 3, ... с вероятностями

$$P(X=k)=\frac{\lambda^k\cdot e^{-\lambda}}{k!},\ k=0,1,2,3,...$$

(напомним, что по определению 0! = 1).

Известно, что ее математическое ожидание и дисперсия равны параметру распределения:

$$\lambda=M[X]=D[X].$$

Поэтому согласно методу моментов (см. 1.4.2) в качестве оценки этого параметра следует брать $\lambda = \overline{X}_n$ (т.е. для конкретной реализации выборки $\lambda \approx \overline{X}_{выб}$).

На рис. 3.9 показаны значения вероятностей $P(X = k)$ для различных значений λ.

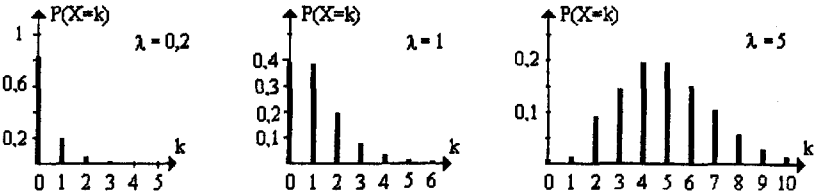


Рис. 3.9

При $\lambda > 9$ распределение Пуассона может быть аппроксимировано нормальным распределением со средним λ и дисперсией $\sqrt{\lambda}$. Известно, что при $\lambda \to \infty$ случайная величина $\dfrac{X-\lambda}{\sqrt{\lambda}}$, где X - пуассоновская случайная величина с параметром λ, имеет в пределе стандартное нормальное распределение $N(0,1)$. При достаточно больших λ можно использовать приближенную формулу

$$P(X=k) \approx \frac{1}{\sqrt{\lambda}} \varphi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right),$$

где $\varphi(x)$ - малая функция Лапласа.

Распределение Пуассона играет важную роль в ряде вопросов физики, теории связи, теории надежности, теории массового обслуживания и т.д., словом, всюду, где идет речь о распределении числа $X(t)$ некоторых случайных событий (радиоактивных распадов, телефонных вызовов, отказов оборудования, несчастных случаев и т.п.), происходящих в течение фиксированного интервала времени t:

$$P\big(X(t)=k\big)=\frac{(\lambda\cdot t)^k\cdot e^{-\lambda t}}{k!}.$$

Здесь параметр λ играет роль среднего числа (плотности) событий в единицу времени.

3.5. Равномерное распределение

Случайная величина X непрерывного типа имеет *равномерное распределение на отрезке* $[a, b]$, $a < b$, если ее плотность вероятности задается формулой

$$f(x)=\begin{cases}\dfrac{1}{b-a}, x\in[a,b];\\ 0, x\notin[a,b].\end{cases}$$

Функция распределения этой случайной величины равна

$$F(x)=\begin{cases}0, x<a;\\ \dfrac{x-a}{b-a}, x\in[a,b];\\ 1, x>b.\end{cases}$$

На рис. 3.10, 3.11 приведены графики плотности и функции распределения равномерного распределения.

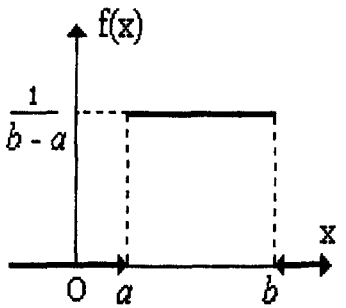


Рис. 3.10

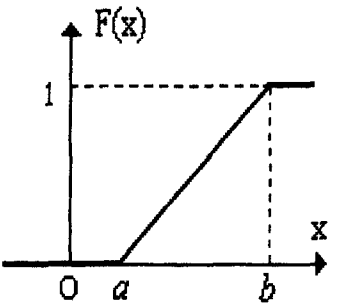


Рис. 3.11

Математическое ожидание и дисперсия равномерно распределенной на отрезке $[a, b]$ случайной величины X вычисляются по формулам

$$M[X]=\frac{a+b}{2}; \qquad D[X]=\frac{(b-a)^2}{12}.$$

Далее согласно методу моментов (см. 1.4.2–1.4.4) в качестве оценок этих параметров следует брать

$$M[X]=\overline{X}_n; \qquad D[X]=s_n^2.$$

В итоге мы имеем систему двух уравнений с двумя неизвестными для оценки границ отрезка по результатам выборки, откуда находим:

$$b=\overline{X}_n+\sqrt{3}\cdot s_n; \qquad a=\overline{X}_n-\sqrt{3}\cdot s_n$$

(естественно, если мы имеем конкретную реализацию выборки, то получим конкретную оценку границ $a_{выб}$ и $b_{выб}$).

Замечание. Метод наибольшего правдоподобия дает следующие оценки границ отрезка:

$$a=\min\{X_1, X_2, ..., X_n\}; \quad b=\max\{X_1, X_2, ..., X_n\},$$

т.е. a и b - соответственно минимальный и максимальный элементы выборки $\{X_1, X_2, ..., X_n\}$.

Равномерное распределение возникает при распространении идеи "равномерности" на непрерывный случай. Равномерное распределение имеют случайные величины, характеризующие ошибки измерений при помощи инструмента с круглыми делениями, когда значение округляется до ближайшего целого. Например, равномерное распределение имеют ошибки указания времени часами со скачущей стрелкой.

3.6. Показательное распределение

Случайная величина X непрерывного типа, принимающая только положительные значения, имеет *показательное* (или *экспоненциальное*) *распределение с параметром* $\lambda > 0$, если ее плотность задается формулой

$$f(x) = \begin{cases} 0, x < 0; \\ \lambda \cdot e^{-\lambda x}, x \geq 0. \end{cases}$$

Функция распределения этой случайной величины равна

$$F(x) = \begin{cases} 0, x < 0; \\ 1 - e^{-\lambda x}, x \geq 0. \end{cases}$$

На рис. 3.12, 3.13 приведены графики этих функций.

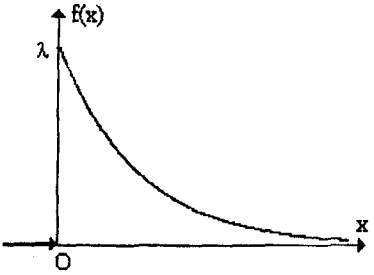


Рис. 3.12

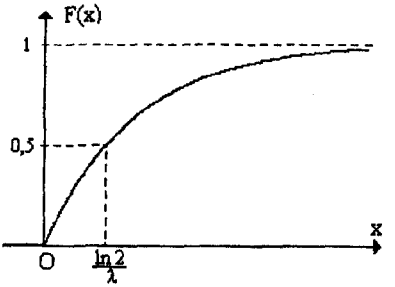


Рис. 3.13

Математическое ожидание и дисперсия этой случайной величины соответственно равны

$$M[X] = \frac{1}{\lambda}; \; D[X] = \frac{1}{\lambda^2}.$$

Поэтому согласно методу моментов (см. 1.4.2) в качестве оценки

параметра λ следует брать $\lambda = \frac{1}{\overline{X}_n}$ (для конкретной реализации выборки $\lambda \approx \frac{1}{\overline{X}_{выб}}$).

Укажем две области применения статистических методов, в которых показательное распределение играет базовую роль.

К первой из них относятся задачи типа "времени жизни". Понимать этот термин следует достаточно широко. В медико-биологических исследованиях под ним может подразумеваться продолжительность жизни больных при клинических исследованиях, в технике - продолжительность безотказной работы устройств, в психологии - время, затраченное испытуемым на выполнение тестовых задач, и т.д.

Второй областью активного использования показательного распределения являются задачи массового обслуживания. Здесь речь может идти об интервалах времени между вызовами "скорой помощи", телефонными звонками или обращениями клиентов и т.д. Длина интервала времени между появлениями последовательных событий имеет показательное распределение.

Показательное распределение среди всех других выделяется, как иногда говорят, отсутствием "памяти", т.е. отсутствием последействия. Это означает, что для изделия, прослужившего время t, вероятность прослужить дополнительное время s совпадает с вероятностью прослужить то же время s для нового (только начавшего работу) изделия, т.е. как бы исключается износ и старение. Поэтому в статистических моделях срока службы, если мы хотим учесть старение, приходится привлекать различного рода обобщения показательного распределения.

3.7. Распределение Лапласа

Случайная величина X непрерывного типа имеет *распределение Лапласа* с параметрами $a \in R$ и $\sigma > 0$, если ее плотность задается формулой

$$f(x)=\frac{1}{\sigma\sqrt{2}}e^{-\frac{|x-a|\cdot\sqrt{2}}{\sigma}}.$$

Ее функция распределения равна

$$F(x)=\begin{cases}\frac{1}{2}e^{\frac{(x-a)\cdot\sqrt{2}}{\sigma}}, x\le a;\\ 1-\frac{1}{2}e^{-\frac{(x-a)\cdot\sqrt{2}}{\sigma}}, x\ge a.\end{cases}$$

На рис. 3.14, 3.15 приведены графики этих функций.

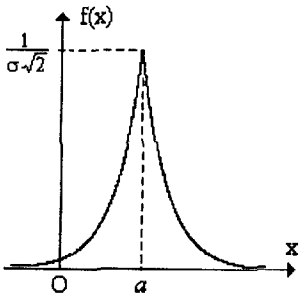


Рис. 3.14

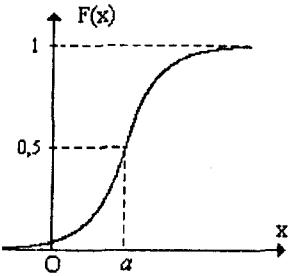


Рис. 3.15

Математическое ожидание случайной величины X, имеющей распределение Лапласа, и ее дисперсия вычисляются по формулам

$$M[X]=a;\ D[X]=\sigma^2.$$

Поэтому согласно методу моментов (см. 1.4.2 – 1.4.4) в качестве оценок этих параметров следует брать

$$a=\overline{X}_n;\qquad \sigma^2=s_n^2$$

(естественно, если мы имеем конкретную реализацию выборки, то должны положить параметры a и σ^2 равными выборочным значениям статистик $\overline{X}_{выб}$ и $s^2_{выб}$).

Распределение Лапласа было впервые введено П.Лапласом и часто называется *первым законом распределения* в отличие от *второго закона распределения* Лапласа, как иногда называют нормальное распределение. Распределение Лапласа называют также *двусторонним показательным распределением*.

3.8. Распределение Вейбулла

Случайная величина X непрерывного типа имеет *распределение Вейбулла* с параметрами $a \in R$, $b > 0$ и $n \in N$, если ее плотность имеет вид

$$f(x) = \begin{cases} 0, x < a; \\ \dfrac{n}{b}\left(\dfrac{x-a}{b}\right)^{n-1} e^{-\left(\frac{x-a}{b}\right)^n}, x \geq a. \end{cases}$$

Функция распределения этой случайной величины равна

$$F(x) = \begin{cases} 0, x < a; \\ 1 - e^{-\left(\frac{x-a}{b}\right)^n}, x \geq a. \end{cases}$$

На рис. 3.16, 3.17 приведены графики этих функций (случай $n = 2$).

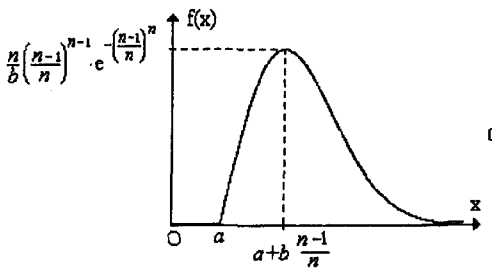


Рис. 3.16

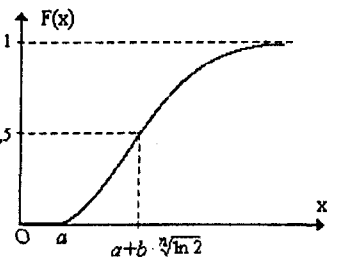


Рис. 3.17

Заметим, что в частном случае при $n = 1$, $a = 0$ распределение Вейбулла совпадает с уже известным нам показательным распределением.

Математическое ожидание и дисперсия данной случайной величины X вычисляются по формулам

$$M[X] = a + b \cdot \Gamma\left(1+\frac{1}{n}\right); \qquad D[X] = b^2 \cdot \left(\Gamma\left(1+\frac{2}{n}\right) - \Gamma^2\left(1+\frac{1}{n}\right)\right),$$

где $\Gamma(x)$ - гамма-функция, которая для $x > 0$ определяется равенством

$$\Gamma(x) = \int_0^{\infty} t^{x-1} \cdot e^{-t}\, dt.$$

В табл. 3.1 приведены ее значения, необходимые для работы.

Таблица 3.1

Значения гамма-функции

n	1	2	3	4	5	6	7	8	9	10
$\Gamma\left(1+\frac{1}{n}\right)$	1	0,8862	0,8930	0,9064	0,9182	0,9277	0,9395	0,9417	0,9470	0,9514
$\Gamma\left(1+\frac{2}{n}\right)$	2	1	0,9028	0,8862	0,8873	0,8930	0,8997	0,9064	0,°126	0,9182

Согласно методу моментов (см. 1.4.2-1.4.4) для оценки параметров распределения (при заданном n) мы имеем систему уравнений

$$\begin{cases} \overline{X}_k = a + b \cdot \Gamma\left(1+\frac{1}{n}\right); \\ s_k^2 = b^2 \cdot \left(\Gamma\left(1+\frac{2}{n}\right) - \Gamma^2\left(1+\frac{1}{n}\right)\right), \end{cases}$$

откуда

$$b=\sqrt{\frac{s_k^2}{\Gamma\left(1+\frac{2}{n}\right)-\Gamma^2\left(1+\frac{1}{n}\right)}};\qquad a=\overline{X}_k-b\cdot\Gamma\left(1+\frac{1}{n}\right).$$

Заметим, что при одних и тех же значениях выборочного среднего $\overline{X}_k$ и выборочной исправленной дисперсии s_k^2 с ростом значения параметра n увеличивается скошенность влево графиков плотности (если параметры a и b найдены по формулам, указанным выше).

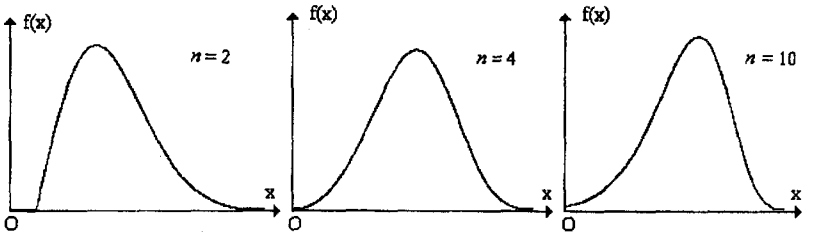


Рис. 3.18

Впервые данное распределение было использовано В.Вейбуллом для аппроксимации экспериментальных данных о прочности стали на разрыв при усталостных испытаниях. Оно широко используется для описания закономерностей отказов шарикоподшипников, вакуумных приборов, элементов электроники, при исследовании на прочность различных строительных и дорожных материалов.

3.9. Распределение Парето

Случайная величина X непрерывного типа имеет *распределение Парето* с параметрами $\alpha > 2$ и $x_o > 0$, если ее плотность задается формулой

$$f(x)=\begin{cases}0, x<x_o;\\ \dfrac{\alpha\cdot x_o^{\alpha}}{x^{\alpha+1}}, x\geq x_o.\end{cases}$$

Функция распределения этой случайной величины равна

$$F(x)=\begin{cases}0, x<x_o;\\ 1-\left(\dfrac{x_o}{x}\right)^{\alpha}, x\geq x_o.\end{cases}$$

На рис. 3.19, 3.20 приведены графики этих функций.

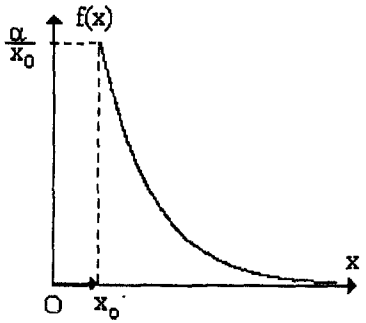


Рис. 3.19

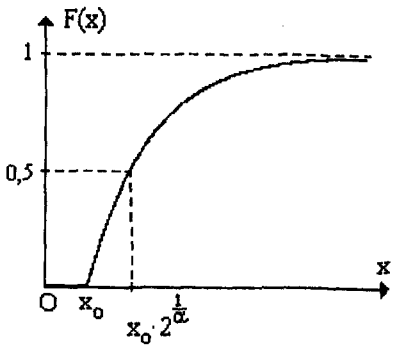


Рис. 3.20

Математическое ожидание и дисперсия случайной величины X, имеющей распределение Парето, вычисляются по формулам

$$M[X]=\frac{\alpha x_o}{\alpha-1};\qquad D[X]=\frac{\alpha\cdot x_o^2}{(\alpha-2)(\alpha-1)^2}$$

(условие $\alpha>2$ необходимо для существования дисперсии).

Согласно методу моментов (см. 1.4.2-1.4.4), в качестве оценок этих параметров следует брать

$$M[X]=\overline{X}_n;\qquad D[X]=s_n^2.$$

В результате получим систему двух уравнений с двумя неизвестными для оценки параметров α и x_o по результатам выборки, откуда находим:

$$\alpha = 1 + \sqrt{1 + \frac{\overline{X}_n^2}{s_n^2}}; \qquad x_o = \overline{X}_n \frac{\sqrt{1 + \frac{\overline{X}_n^2}{s_n^2}}}{1 + \sqrt{1 + \frac{\overline{X}_n^2}{s_n^2}}}.$$

Распределение Парето получило широкое распространение в различных задачах экономической статистики, начиная с работ В.Парето (1897 г.) о распределении доходов. Считалось, что распределение Парето достаточно хорошо описывает распределение доходов, превышающих некоторый уровень.

Таблица 3.2

Значения большой функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

(функции распределения стандартизованной нормальной случайной величины $N(0,1)$)

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1	2	3	4	5	6	7	8	9	10	11
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9438	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545

1	2	3	4	5	6	7	8	9	10	11
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Примечание.

При использовании табл. 3.2 следует помнить следующие правила: $\Phi(-x) = 1 - \Phi(x)$; $\Phi(x) \approx 1$ при $x > 3{,}5$; $\Phi(x) \approx 0$ при $x < -3{,}5$ (причем погрешность в этих приближенных равенствах – менее 10^{-4}).

Таблица 3.3

Квантили u_p стандартизованного нормального распределения $N(0,1)$

p	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
u_p	1,282	1,645	1,96	2,326	2,576	3,09	3,291

Примечания:

1. Квантилем порядка p называется такое число u_p, что $P(X < u_p) = p$).

2. При использовании табл. 3.3 следует помнить следующее правило: $u_{1-p} = -u_p$.

4. НЕКОТОРЫЕ СВЕДЕНИЯ О ХИ-КВАДРАТ РАСПРЕДЕЛЕНИИ

Важную роль в математической статистике наряду с нормальным распределением играет так называемое хи-квадрат распределение (χ^2- распределение). Оно определяется следующим образом. Пусть имеется n независимых случайных величин $Z_1, Z_2, \ldots, Z_n$, каждая из которых имеет стандартизованное нормальное распределение, т.е. нормальное распределение $N(0; 1)$ с нулевым средним и единичной дисперсией (см. подраздел 1.1). Определим новую случайную величину вида $\chi^2_n = Z_1^2 + Z_2^2 + \ldots + Z_n^2$. Эта величина называется χ^2 - *случайной величиной с n степенями свободы*.

Число степеней свободы n определяет число независимых квадратов, входящих в сумму. Ясно, что величина χ^2_n для любого $n \geq 1$ принимает только положительные значения.

Функция плотности χ^2 -распределения равна

$$f_n(x) = \begin{cases} 0, x \leq 0; \\ \dfrac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x > 0, \end{cases}$$

где $\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t}\, dt$ - гамма-функция. Не надо пугаться столь громоздкого вида определения плотности $f_n(x)$, т.к. на практике она редко используется непосредственно. Наглядное представление о схематическом поведении этой функции дает рис. 4.1.

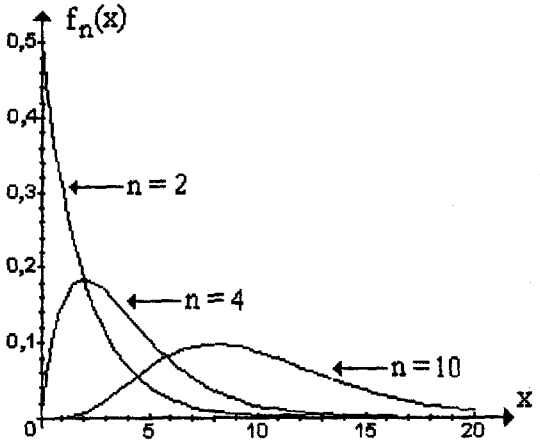


Рис. 4.1

Известно, что математическое ожидание и дисперсия случайной величины χ_n^2 равны:

$$M[\chi_n^2] = n,\ D[\chi_n^2] = 2n.$$

Таблица 4.1

Таблица квантилей $\chi^2_{n,p}$ хи-квадрат распределения

$\downarrow n \quad p \rightarrow$	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
1	2	3	4	5	6	7	8	9	10	11
1	0,000	0,000	0,001	0,004	0,016	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,051	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,071	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3

1	2	3	4	5	6	7	8	9	10	11
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	7,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8
16	5,14	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0	34,3
17	5,70	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4	35,7
18	6,26	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8	37,2
19	6,84	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2	38,6
20	7,43	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6	40,0
21	8,03	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9	41,4
22	8,64	9,54	11,011,7	12,3	14,0	30,8	33,9	36,8	40,3	42,8
23	9,26	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6	44,2
24	9,89	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0	45,6
25	10,5	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3	46,9
26	11,2	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6	48,3
27	11,8	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0	49,6
28	12,5	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3	51,0
29	13,1	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6	52,3
30	13,8	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9	53,7
35	17,2	18,5	20,6	22,5	24,8	46,1	49,8	53,2	57,3	60,3
40	20,7	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7	66,8
45	24,3	25,9	28,4	30,6	33,4	57,5	61,7	65,4	70,0	73,2
50	28,0	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2	79,5
75	47,2	49,5	52,9	56,1	59,8	91,1	96,2	100,8	106,4	110,3
100	67,3	70,1	74,2	77,9	82,4	118,5	124,3	129,6	135,6	140,2

Примечание.

Квантилем порядка p называется такое число $\chi^2_{n,p}$, что $P\left(\chi^2_n < \chi^2_{n,p}\right) = p$.

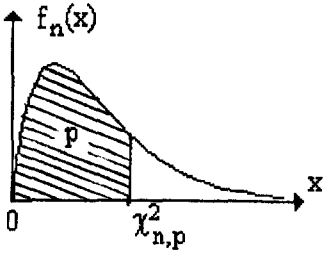


Рис. 4.2

5. НЕКОТОРЫЕ СВЕДЕНИЯ О РАСПРЕДЕЛЕНИИ СТЬЮДЕНТА

Распределение Стьюдента играет важную роль в математической статистике. Оно определяется следующим образом. Пусть Y и Z - независимые случайные величины, причем величина Y имеет χ^2 - распределение с n степенями свободы, а величина Z - стандартизованное нормальное распределение N (0; 1). Определим новую случайную величину:

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

Распределение этой величины носит название *распределение Стьюдента* (t - *распределение*) *с n* степенями свободы. Ее плотность вероятности имеет вид

$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

где $\Gamma(x) = \int_o^\infty t^{x-1}\cdot e^{-t}\,dt$ - гамма-функция.

Очевидно, что график плотности симметричен относительно оси ординат, и поэтому ее математическое ожидание

$$M[t_n] = 0.$$

Известно, что дисперсия

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

График плотности случайной величины t_n похож на график малой функции Лапласа, а при больших значениях n практически совпадает с ним.

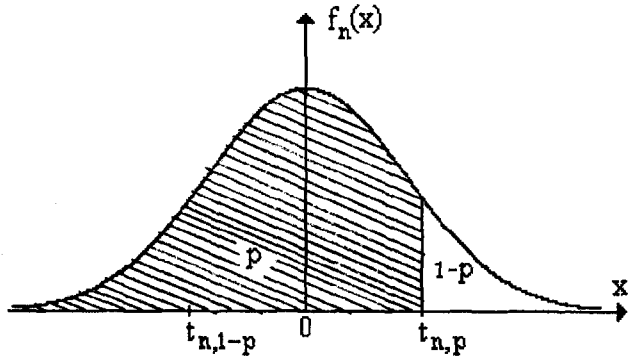


Рис. 5.1

Закон распределения случайной величины t_n установил в 1908 г. английский химик и математик У.Госсет, публиковавший свои труды под псевдонимом "Стьюдент".

Ниже приведена таблица квантилей $t_{n,p}$ распределения Стьюдента (напомним, что квантилем порядка p называется такое число $t_{n,p}$, что $P(t_n < t_{n,p}) = p$). При использовании этой таблицы следует иметь в виду, что $t_{n,1-p} = -t_{n,p}$ (это хорошо видно из рисунка).

Таблица 5.1

Таблица квантилей $t_{n,p}$ распределения Стьюдента

$\downarrow n \quad p \rightarrow$	0,9	0,95	0,975	0,99	0,995
1	2	3	4	5	6
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841

1	2	3	4	5	6
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,226	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1.721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1.318	1,711	2,064	2,492	2,797
25	1.316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

Литература

1. Айвазян С.А. Статистическое исследование зависимостей. – М.: Металлургия, 1968.

2. Булдык Г.М. Теория вероятностей и математическая статистика. – М.: Выш. школа, 1989.

3. Герасимович А.И. Математическая статистика. - Мн.: Выш. школа, 1983.

4. Колде Я.К. Практикум по теории вероятностей и математической статистике. – М.: Высш. школа, 1991.

5. Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика. – М.: Индгра, 1997.

6. Лозинский С.Н. Сборник задач по теории вероятностей и математической статистике. – М.: Статистика, 1975.

7. Львовский Е.Н. Статистические методы построения эмпирических формул. – Мн.: Выш. школа, 1988.

8. Микулик Н.А., Рейзина Г.Н. Решение технических задач по теории вероятностей и математической статистике. – Мн.: Выш. школа, 1991.

9. Хольд А. Математическая статистика с техническими приложениями. – М.: Иностранная литература, 1956.

10. Худсон Д. Статистика для физиков. – М.: Мир, 1970.

Содержание

Введение 3
1. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ВЫБОРКИ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ 6
1.1. Теория вероятностей и математическая статистика 6
1.2. Генеральная совокупность и выборка 8
1.3. Методы описательной статистики 9
1.3.1. Вариационный ряд. Эмпирическая функция распределения 10
1.3.2. Глазомерный метод обоснования гипотезы о законе распределения случайной величины 13
1.3.3. Некоторые показатели расположения 17
1.3.4. Некоторые показатели разброса (рассеяния) 19
1.3.5. Группированные данные 21
1.3.6. Графические представления выборки 25
1.3.7. Некоторые дополнительные характеристики выборки . . . 28
1.3.8. Некоторые замечания о числовых характеристиках выборки 30
1.4. Статистическое оценивание параметров 31
1.4.1. Свойства точечных оценок 34
1.4.2. Метод моментов для нахождения оценок параметров распределения по выборке 36
1.4.3. Оценка математического ожидания случайной величины по результатам наблюдений 36
1.4.4. Оценка дисперсии и среднеквадратического отклонения случайной величины по результатам наблюдений 39
1.5. Точность статистических оценок 42
1.5.1. Доверительное оценивание 42
1.5.2. Доверительный интервал для математического ожидания нормально распределенной случайной величины с известным среднеквадратическим отклонением 44
1.5.3. Доверительный интервал для математического ожидания нормально распределенной случайной величины с неизвестным среднеквадратическим отклонением 46

1.5.4. Доверительный интервал для оценки дисперсии нормально распределенной случайной величины. 48
2. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ. 50
2.1. Статистические гипотезы. 51
2.2. Проверка статистических гипотез. 52
2.2.1. Критерии согласия. 53
2.2.2. Некоторые замечания к практическому использованию критериев согласия. 55
2.2.3. Проверка гипотезы о значении математического ожидания нормального распределения. 57
2.3. Проверка гипотезы об общем виде закона распределения случайной величины. 60
2.3.1. Критерий согласия Пирсона (критерий хи-квадрат). 61
2.3.2. Схема применения критерия Пирсона для проверки сложной гипотезы о законе распределения. 64
2.3.3. Примеры применения критерия Пирсона для проверки гипотезы о законе распределения. 66
2.3.4. Некоторые другие критерии согласия. 69
2.3.5. Критерий согласия Колмогорова для простой гипотезы. . 70
2.3.6. Критерий согласия омега-квадрат для простой гипотезы. 72
3. НЕКОТОРЫЕ ЧАСТО ИСПОЛЬЗУЮЩИЕСЯ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ. 73
3.1. Нормальное распределение. 73
3.2. Логнормальное распределение. 76
3.3. Усеченные нормальные распределения. 78
3.4. Распределение Пуассона. 81
3.5. Равномерное распределение. 83
3.6. Показательное распределение. 85
3.7. Распределение Лапласа. 86
3.8. Распределение Вейбулла. 88
3.9. Распределение Парето. 90
4. НЕКОТОРЫЕ СВЕДЕНИЯ О ХИ-КВАДРАТ РАСПРЕДЕЛЕНИИ. 94
5. НЕКОТОРЫЕ СВЕДЕНИЯ О РАСПРЕДЕЛЕНИИ СТЬЮДЕНТА. 97
Литература. 100

Учебное издание

ВЕРЕМЕНЮК Валентин Валентинович
КОЖУШКО Валерий Васильевич
МОРОЗ Ольга Александровна

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ВЫБОРКИ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Учебно-методическое пособие
к лабораторной работе по высшей математике
для студентов строительных специальностей

Редактор Т.А.Палилова. Корректор М.П.Антонова
Компьютерная верстка Л.М.Чернышевич

Подписано в печать 26.01.2002.
Формат 60x84 1/16. Бумага типографская № 2.
Печать офсетная. Гарнитура Таймс.
Усл. печ. л. 6,1. Уч.-изд. л. 4,7. Тираж 100. Заказ 174.

Издатель и полиграфическое исполнение:
Белорусская государственная политехническая академия.
Лицензия ЛВ №155 от 30.01.98. 220027, Минск, проспект Ф.Скорины, 65.