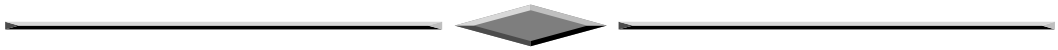
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
Государственное образовательное учреждение высшего профессионального образования
"ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ"



А.А. Михальчук, Е.Г. Язиков, В.В. Ершов

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЭКОЛОГО–ГЕОХИМИЧЕСКОЙ ИНФОРМАЦИИ

Учебно–методическое пособие

Издательство ТПУ
Томск 2006

УДК 681.3 : 519.2

Михальчук А.А.

Статистический анализ эколого–геохимической информации: учебное пособие / А.А. Михальчук, Е.Г. Язиков, В.В. Ершов – Томск: Изд.–во ТПУ, 2006. – 235 с.

Пособие содержит краткие теоретические сведения в области математической статистики; примеры и рекомендации по решению типовых задач с использованием современного компьютерного инструментария (систем STATISTICA и Excel) на уровне модульного анализа данных с помощью мастер-макросов; показывает особенности статистического анализа в случае малых выборок; снабжено наглядными графическими иллюстрациями, выполненными в системе STATISTICA 6.0; содержит список рекомендуемой литературы.

Настоящее учебное пособие предназначено для студентов и аспирантов по специальности 320300 (013600) – «Геоэкология» и может быть полезным при усвоении теоретического материала и овладении необходимыми практическими навыками при проведении сравнительного статистического анализа (ССА) эколого–геохимической информации.

УДК 681.3 : 519.2

Рекомендовано к печати Редакционно-издательским советом Томского политехнического университета

Рецензенты

Доктор геолого-минералогических наук, профессор ТГАСУ
А.В. Мананков

Кандидат геолого-минералогических наук, профессор ТГУ
А.И. Летувнинкас

© Томский политехнический университет, 2006
© А.А. Михальчук, Е.Г. Язиков, В.В. Ершов, 2006

Оглавление

Введение……..……..…….. 5
ЧАСТЬ 1. МАТЕМАТИЧЕСКИЕ ОСНОВЫ …..…….....….. 7
1.1. Введение в теорию вероятностей……..……... 7
1.1.1. Понятие вероятности случайного события……...…… 7
1.1.2. Случайная величина и законы ее распределения…….… 9
1.1.3. Основные характеристики случайной величины…… 11
1.1.4. Примеры законов распределения случайной величины ……. 16
1.1.5. Система двух случайных величин...……... 25
1.2. Элементы математической статистики…….. 27
1.2.1. Выборочный метод…….. 27
1.2.2. Корреляционно–регрессионный анализ……..…….. 33
1.2.3. Проверка статистических гипотез……. 37
1.2.4. Особенности применения статистического анализа эколого–геохимической информации в случае малых выборок…..... 46
ЧАСТЬ 2. КОМПЬЮТЕРНЫЙ ПРАКТИКУМ…….……..….. 58
2.1. Описательная статистика…….……..……. 59
2.1.1. Вычисление оценок числовых характеристик содержаний химических элементов……..…….…… 59
2.1.2. Построение диаграммы размаха……...…… 61
2.1.3. Построение гистограммы содержаний микроэлементов…..… 63
2.2. Проверка статистических гипотез……..…….....…. 69
2.2.1. Проверка гипотезы о законе распределении содержаний химических элементов……..……...…….…… 69
2.2.2. Оценка различия содержаний двух выборок…… 73
2.3. Корреляционно–регрессионный анализ…….....…. 81
2.3.1. Вычисление корреляционной матрицы ассоциации содержаний микроэлементов……..……..……. 81
2.3.2. Вычисление коэффициента корреляции Спирмена……. 86

2.3.3. Построение диаграммы рассеяния…… 88
2.4. Кластерный анализ ……. 101
2.5. Средства статистического анализа данных в системе Excel…….. 106
2.6. Модульный анализ данных в системе STATISTICA…. 127
ЧАСТЬ 3. ПРИМЕРЫ СРАВНИТЕЛЬНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА (ССА) ЭКОЛОГО–ГЕОХИМИЧЕСКОЙ ИНФОРМАЦИИ……..... 137
3.1. ССА эколого–геохимических оценок разных территорий ……. 137
3.2. ССА эколого–геохимической информации по данным различных съемок…..……..…….…..... 157
3.3. ССА содержаний микроэлементов в накипи разных территорий, полученных методами ИНАА и ISP…….. 173
ЧАСТЬ 4. ПРЕЗЕНТАЦИЯ ЛАБОРАТОРНОЙ РАБОТЫ…..... 186
Заключение…….…….…….……. 216
ПРИЛОЖЕНИЯ……..……... 217
$П_1$. Содержание микроэлементов в почве территории Томского региона, полученное методами ИНАА и ISP..... 217
$П_2$. Содержание микроэлементов в почве территорий гг. Междуреченска и Томска, полученное методом ЭСП..... 221
$П_3$. Содержание микроэлементов в почве, снеге и золе растений территории г. Стрежевого……..…..... 228
$П_4$. Категоризированные содержания микроэлементов в накипи территорий Томской и Челябинской областей, полученных методами ИНАА и ISP……..... 231
Библиографический список……..…… 233

Введение

Резкое увеличение количественной информации, получаемой в процессе эколого-геохимических исследований, вызвало необходимость использования современных способов ее обработки и анализа с помощью ЭВМ. В последнее время наблюдается глубокое проникновение математических методов исследования во все отрасли геологических наук (геохимии в частности). Для успешного развития эколого-геохимических исследований необходимо также использовать полный арсенал существующих прогрессивных научных и технических средств, включая методы статистического анализа и ЭВМ.

Современная геохимия уже не может ограничиться изучением лишь качественных сторон явлений и процессов, а должна активно и всесторонне выявлять их количественные характеристики, обеспечивая тем самым более высокий научный уровень исследований экологии окружающей среды.

Прежде чем приступить к статистической обработке, следует особое внимание обратить на составление выборок. Именно грамотная формулировка задачи и формирование массива данных в соответствии с ней будет определять эффективность применения статистического анализа.

Подборка выборки проб определяется в первую очередь задачей исследований. Например, при изучении различных почвенных разрезов, следует объединять пробы по горизонтам. При изучении воздействия предприятия можно объединять в выборки пробы, отобранные на разных расстояниях и направлениях с учётом розы ветров. Кроме того, необходимо учитывать способ отбора проб, метод аналитических исследований и специфику лаборатории, в которой проводился анализ.

Количество проб, объединённых в выборку, может быть различным. Однако для получения достоверных статистических данных желательно иметь не менее 30 проб. Количество элементов, принимаемых к расчёту, определяется задачей исследования.

Важным моментом в настоящее время является использование эколого-геохимической информации в небольшом объёме выборки. Основным фактором в данном случае являются дорогостоящие методы анализа. В задачу данной работы входит также возможность применения методов статистической обработки при небольшом объеме выборок для сопоставления результатов исследований.

Содержания химических элементов желательно выражать в единицах одной размерности. Следует помнить, что 1 г/т = 1 мг/кг = 1 ppm= $1x10^{-4}$ % (0,0001 %); 1 мг/т = 0,001 г/т = 1 ppb = $1x10^{-7}$ % (0,000001 %).

В том случае, когда содержание элемента в пробе не превышает порога чувствительности анализа, можно заменить содержание на половину порога чувствительности.

Учебно–методическое пособие состоит из двух частей. В первой части рассматриваются математические основы с введением в теорию вероятностей и подробная характеристика элементов математической статистики. Вторая часть включает компьютерный практикум с рассмотрением описательной статистики, проверки статистических гипотез и корреляционно-регрессионного анализа. В третьей части подробно рассматриваются примеры сравнительного статистического анализа эколого-геохимической информации на реальных материалах.

Авторы преследовали цель не только создать учебно-методическое пособие, но и на конкретных примерах показать возможность применения современного компьютерного инструментария (системы STATISTIKA 6.0) для овладения необходимыми практическими навыками при проведении сравнительного статистического анализа эколого-геохимической информации. Конечно, всё изложить не представляется возможным, поэтому предполагается изучение литературы, приведённой в библиографическом списке.

Авторы будут признательны читателям за отзывы, критические замечания и полезные советы, которые помогут устранить имеющиеся в пособии недостатки и улучшить в будущем его содержание.

ЧАСТЬ 1.
МАТЕМАТИЧЕСКИЕ ОСНОВЫ

1.1. Введение в теорию вероятностей

Теория вероятностей – математическая наука, позволяющая по вероятностям одних случайных событий находить вероятности других случайных событий, связанных каким-либо образом между собой.

В этом определении есть целый ряд понятий: случайное событие, вероятность случайного события, связь между случайными событиями. Все эти понятия нуждаются в определении и разъяснении. В усвоении этого круга вопросов и состоит первое знакомство с теорией вероятностей.

Теория вероятностей изучает свойства массовых случайных событий, способных многократно повторяться при воспроизведении определенного комплекса условий. Основное свойство любого случайного события независимо от его природы – вероятность его осуществления.

Предметом теории вероятностей является изучение вероятностных закономерностей массовых однородных случайных событий.

Все это предопределяет необходимость овладения методами теории вероятностей и математической статистики как инструментом статистического анализа и прогнозирования явлений и процессов.

1.1.1. Понятие вероятности случайного события

Осуществление каждого отдельного наблюдения, опыта или измерения при изучении эксперимента называют испытанием. Результат испытания назовем событием. Различают события: достоверные, невозможные и случайные.

Достоверное событие – это такое событие, которое всегда происходит в рассматриваемом эксперименте.

Невозможное событие – это такое событие, которое никогда не происходит в рассматриваемом эксперименте.

Случайное событие – событие, которое при воспроизведении опыта может наступить, а может и не наступить**.**

События обозначаются большими латинскими буквами $A,B,C,\ldots$, невозможное – $\varnothing$, достоверное – Ω.

Сравнивать случайные события естественно по степени возможности их наступления. С этой целью вводится числовая характеристика этой степени возможности (случайности), называемая вероятностью со-

бытия. Для события A, вероятность принято обозначать $P(A)$. Существует несколько подходов, поясняющих понятие вероятности. В каждом из этих подходов указываются правила, по которым случайному событию ставится в соответствие положительное число, объективно характеризующее степень возможности появления этого события.

С практической точки зрения представляет интерес статистическое определение математической вероятности.

Многочисленными наблюдениями над самыми разнообразными случайными событиями установлен следующий достоверный факт: если над одним и тем же случайным событием в одних и тех же условиях проводить много серий из большого числа испытаний, то каждая наблюдаемая в такой серии частота появления события будет колебаться от серии к серии в сравнительно узких пределах, будет, как говорят в теории вероятностей "устойчивой". При этом пределы, в которых будет колебаться устойчивая частота случайного события, будут тем теснее, чем большее число испытаний в каждой серии. Это свидетельствует о наличии статистической закономерности в изучаемом явлении.

Пусть в одних и тех же условиях проведена серия из n^* испытаний, в каждом из которых могло появиться или не появиться интересующее нас событие A. Пусть событие A появилось при этом в m^* испытаниях. ***Относительной частотой*** $P^*(A)$ события A в данной серии испытаний называется отношение m* (числа испытаний, в которых появилось событие A) к n^* (общему числу проведенных испытаний), то есть

$$P^*(A) = \frac{m^*}{n^*}. \qquad (1.1)^*$$

Из данного определения следует, что относительная частота случайного события всегда заключена между нулем и единицей:

$$0 \le P^*(A) \le 1.$$

Статистической вероятностью $P(A)$ события A называется предел, к которому стремится относительная частота $P^*(A)$ при неограниченном увеличении числа испытаний, то есть

$$P(A) = \lim_{n\to\infty} P^*(A) = \lim_{n\to\infty} \frac{m^*}{n^*}. \qquad (1.1)$$

При больших n статистическое определение позволяет в приблизительных расчетах относительную частоту $P^*(A)$ принимать за вероятность события A. Недостатком этого определения вероятности является необходимость проведения большого числа опытов в одинаковых условиях.

1.1.2. Случайная величина и законы ее распределения

Случайной величиной X называется величина, которая в результате опыта может принять то или иное значение x_i. Выпадение некоторого значения случайной величины X – есть случайное событие: $X = x_i$.

Функцией распределения случайной величины X называется функция $F(x)$, значение которой в точке x равно вероятности того, что случайная величина X будет меньше этого значения x, то есть

$$F(x) = P(X < x). \qquad (1.2)$$

Среди случайных величин выделяют прерывные (дискретные) и непрерывные случайные величины.

Дискретной называют случайную величину, которая может принимать отдельные, изолированные значения с определенными вероятностями.

Дискретная случайная величина X может быть задана рядом распределения или функцией распределения (интегральным законом распределения).

Рядом распределения называется совокупность всех возможных значений x_i и соответствующих им вероятностей $p_i = P(X = x_i)$, он может быть задан в виде таблицы.

Таблица 1.1

Ряд распределения дискретной случайной величины X

x_i	x_1	x_2	. . .	x_k
p_i	p_1	p_2	. . .	p_k

При этом вероятности p_i удовлетворяют условию

$$\sum_{i=1}^{k} p_i = 1,$$

где число возможных значений k может быть конечным или бесконечным.

Графическое изображение ряда распределения называется ***многоугольником распределения***. Для его построения возможные значения случайной величины (x_i) откладываются по оси абсцисс, а вероятности p_i – по оси ординат; точки A_i с координатами (x_i , p_i) соединяются ломаными линиями.

Функция $F(x)$ для **дискретной случайной величины** вычисляется по формуле

$$F(x)= \sum_{x_i<x} p_i \ , \qquad (1.2)^*$$

где суммирование ведется по всем i, для которых $x_i < x$.

Непрерывной называют случайную величину, возможные значения которой непрерывно заполняют некоторые промежутки.

Непрерывная случайная величина характеризуется прежде всего заданием неотрицательной функции $f(x)$, называемой ***плотностью вероятности*** и определяемой соотношением:

$$f(x) = \lim_{\Delta x \to \infty} \frac{P(x < X < x + \Delta x)}{\Delta x}. \qquad (1.3)$$

При любых $\boldsymbol{x}$ плотность вероятности $f(x)$ удовлетворят равенству

$$F(x)= \int_{-\infty}^{x} f(\tilde{x})d\tilde{x} \ , \qquad (1.2)^{**}$$

связывающему ее с функцией распределения $F(x)$.

Непрерывная случайная величина задается, таким образом, либо функцией распределения $F(x)$ (интегральным законом), либо плотностью вероятности $f(x)$ (дифференциальным законом).

Функция распределения $F(x)$ имеет следующие свойства:

1) $P(a \le X < в) = F(в) - F(a)$;

2) $F(x_1) \le F(x_2)$, *если* $x_1 < x_2$;

3) $\lim\limits_{x \to +\infty} F(x) = 1;$ (1.4)

4) $\lim\limits_{x \to +\infty} F(x) = 0.$

Плотность вероятности $f(x)$ (дифференциальный закон распределения) обладает следующими основными свойствами:

1) $f(x) \ge 0;$

2) $f(x) = \dfrac{\boldsymbol{dF(x)}}{\boldsymbol{dx}} = F'(x);$

3) $\int\limits_{-\infty}^{\boldsymbol{x}} \boldsymbol{f(t)dt} = F(x);$ (1.5)

4) $\int\limits_{-\infty}^{\infty} \boldsymbol{f(x)dx} = 1;$

5) $P(a \le X < в) = \int\limits_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{f(x)dx}$.

Геометрически вероятность попадания величины X на участок *(a,b)* равна площади криволинейной трапеции, соответствующей определенному интегралу $\int_a^b f(x)dx$ (см. пример на рис. 1.1).

Рис.1.1. Графики плотности f(x) и функции F(x) распределения Стьюдента с числом степеней свободы df= 10.

Площадь затемненной области равна $0{,}75 = p = F(0{,}7) = \int_{-\infty}^{0{,}7} f(t)\,dt$.

1.1.3. Основные характеристики случайной величины

Свойства случайной величины могут характеризоваться различными параметрами. Важнейшие из них – ***математическое ожидание*** случайной величины, которое обозначается через M*(X),* и ***дисперсия*** D*(X)* = $\sigma^2(X)$, корень квадратный из которой $\sigma(X)$ называют ***среднеквадратическим отклонением*** или стандартом .

Математическим ожиданием M*(X)* (*средним по распределению*) **дискретной** (прерывной) случайной величины *X* называют сумму произведений всех возможных значений случайной величины на соответствующие им вероятности.

$$M(X) = x_1p_1 + x_2p_2 + \ldots + x_k p_k = \sum_{i=1}^{k} x_i p_i \ , \qquad (1.6)$$

Учитывая предыдущие записи, и что $\sum_{i=1}^{k} p_i = 1$, иногда пишут:

$$\mathrm{M}(X) = \sum_{i=1}^{k} x_i p_i \;/\; \sum_{i=1}^{k} p_i \,.$$

Эта запись позволяет дать ***механическую интерпретацию математического ожидания***: $\mathrm{M}(X)$ – абсцисса центра тяжести системы точек, абсциссы которых равны возможным значениям случайной величины, а массы, помещенные в эти точки, равны соответствующим вероятностям.

Математическим ожиданием непрерывной случайной величины X называется интеграл

$$\mathrm{M}(X) = \int_{-\infty}^{\infty} xf(x)dx \;, \qquad (1.6)^*$$

причем предполагается, что интеграл сходится абсолютно; здесь $f(x)$ – плотность вероятности распределения случайной величины X.

Математическое ожидание $M(X)$ можно понимать как «теоретическое среднее значение случайной величины».

Рассмотрим ***свойства математического ожидания***:

1. Математическое ожидание имеет ту же размерность, что и сама случайная величина.

2. Математическое ожидание может быть как положительным, так и отрицательным числом.

3. Математическое ожидание постоянной величины C равно этой постоянной, т.е. $\mathrm{M}(C) = C$.

4. Математическое ожидание суммы нескольких случайных величин равно сумме математических ожиданий этих величин, т.е.

$$\mathrm{M}(X + Y + \ldots + W) = \mathrm{M}(X) + \mathrm{M}(Y) + \ldots + \mathrm{M}(W).$$

5. Математическое ожидание произведения двух или нескольких взаимно независимых случайных величин равно произведению математических ожиданий этих величин, т.е.

$$\mathrm{M}(XY) = \mathrm{M}(X) \cdot \mathrm{M}(Y).$$

6. Математическое ожидание произведения случайной величины на постоянную C равно произведению математического ожидания случайной величины на постоянную C

$$\mathrm{M}(CX) = C \cdot \mathrm{M}(X).$$

Наряду с математическим ожиданием используют и другие числовые характеристики: ***медиана*** x_{med} делит распределение X на две равные части и определяется из условия $F(x_{med}) = 0{,}5$; ***мода*** x_{mod} – это макси-

мально часто встречающееся значение X и для непрерывно распределенной случайной величины является абсциссой точки максимума $f(x)$.

В симметричных распределениях все три числовые характеристики (математическое ожидание, медиана и мода) совпадают.

При наличии нескольких мод распределение называют мультимодальным.

Если математическое ожидание случайной величины дает нам ее «среднее значение» или точку на координатной прямой, вокруг которой «разбросаны» значения рассматриваемой случайной величины, то ***дисперсия*** характеризует «степень разброса» значений случайной величины около ее среднего.

Дисперсией $D(X)$ случайной величины X называется математическое ожидание квадрата отклонения значения случайной величины от ее математического ожидания, т.е.

$$D(X) = M([X - M(X)]^2). \qquad (1.7)$$

Дисперсию удобно вычислять по формуле:

$$D(X) = M(X^2) - [M(X)]^2 .$$

Для ***дискретной*** случайной величины $\boldsymbol{X}$ формула дает

$$D(X) = \sum_{i=1}^{k} (x_i)^2 p_i - [M(X)]^2 . \qquad \textbf{(1.7)*}$$

Для ***непрерывной*** случайной величины $\boldsymbol{X}$

$$D(X) = \int_{-\infty}^{\infty} \left(\boldsymbol{x - M(x)}\right)^2 \boldsymbol{f(x)dx}. \qquad \textbf{(1.7)**}$$

Дисперсия имеет размерность, равную квадрату размерности случайной величины.

Рассмотрим ***свойства дисперсии***:

1. Дисперсия постоянной величины всегда равна нулю:

$$D(C) = 0.$$

2. Постоянный множитель можно выносить за знак дисперсии, предварительно возведя его в квадрат:

$$D(CX) = C^2 D(X).$$

3. Дисперсия алгебраической суммы двух ***независимых*** случайных величин равна сумме их дисперсией:

$$D(X \pm Y) = D(X) + D(Y).$$

Положительный корень из дисперсии называется ***среднеквадратичным (стандартным) отклонением*** и обозначается $\sigma = \sqrt{D}(X)$ Среднее квадратичное отклонение имеет ту же размерность, что и случайная величина.

Случайная величина называется ***центрированной*,** если M*(X)* = 0 и ***стандартизированной***, если M*(X)* = 0 и $\sigma = 1$.

В общем случае свойства случайной величины могут характеризоваться различными начальными и центральными моментами.

Начальным моментом K–го порядка называется число α_K**,** определяемое формулой:

$$\alpha_K = \mathrm{M}(X^K) = \begin{cases} \sum x_i^K p_i & для \quad X - дискр., \\ \int\limits_{-\infty}^{+\infty} x^K f(x)\,\mathrm{d}\,x_i & для \quad X - непрер., \end{cases}$$

где $\mathrm{M}(X^K)$ – математическое ожидание *K*–й степени случайной величины X^K (соответственно – для случайных величин дискретного и непрерывного типа).

Центральным моментом K–го порядка называется число μ_K, определяемое формулой

$$\mu_K = \mathrm{M}[(X - \alpha)] = \begin{cases} \sum\limits_i (x_i - \alpha_1)^K p_i \quad , \\ \int\limits_{-\infty}^{+\infty} (x - \alpha_1)^K f(x)\,\mathrm{d}\,x_i . \end{cases}$$

Из определений моментов, в частности, следует:

$$\alpha_0 = \mu_0 = 1, \quad \alpha_1 = \mathrm{M}(X), \quad \mathrm{D}(X) = \mu_2 = \alpha_2 - \alpha_1^{\,2}.$$

Часто пользуются производными характеристиками от начальных и центральных моментов.

Коэффициентом вариации называется величина

$$V = \frac{\sigma}{\alpha_1} 100\ \%.$$

Коэффициент вариации – величина безразмерная, применяемая для сравнения степеней изменчивости случайных величин с разными единицами измерения.

Коэффициентом асимметрии **(или *скошенности*)** распределения называется величина

$$A = \frac{\mu_3}{\sigma^3}. \qquad (1.8)$$

Коэффициент асимметрии характеризует степень асимметрии распределения случайной величины относительно ее математического ожидания. Для симметричных распределений *A* = 0. Если пик графика

функции *f(x)* смещен в сторону малых значений («хвост» на графике функции *f(x)* справа), то $A > 0$. В противном случае $A < 0$ (см. рис.1.2).

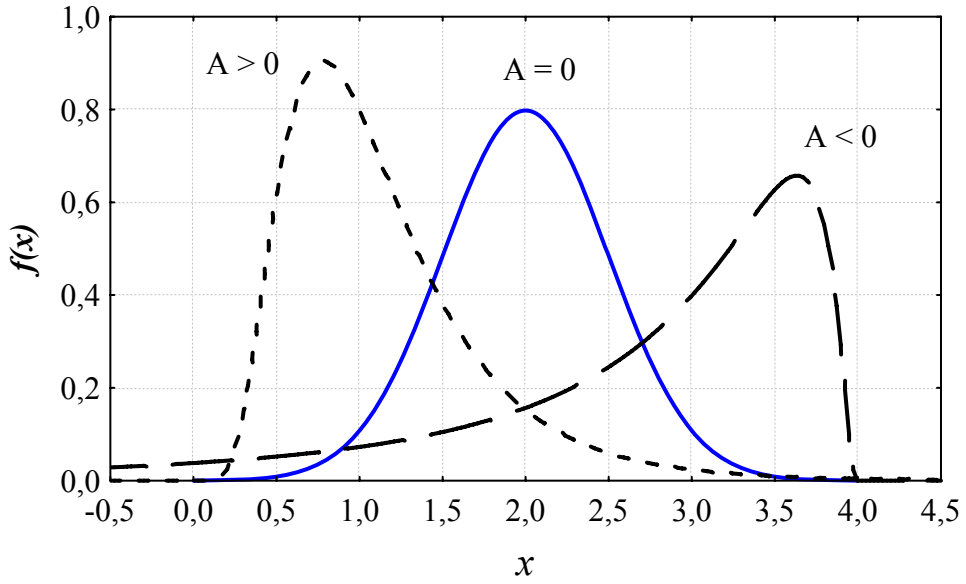


Рис.1.2. Зависимость графиков плотности вероятности ***f(x)*** *от коэффициента асимметрии A*

Коэффициентом эксцесса **(или** ***островершинности*****)** распределения называется величина

$$E = \frac{\mu_4}{\sigma^4} - 3. \qquad (1.9)$$

Коэффициент эксцесса является мерой остроты графика функции плотности распределения *f(x)* (рис. 1.3).

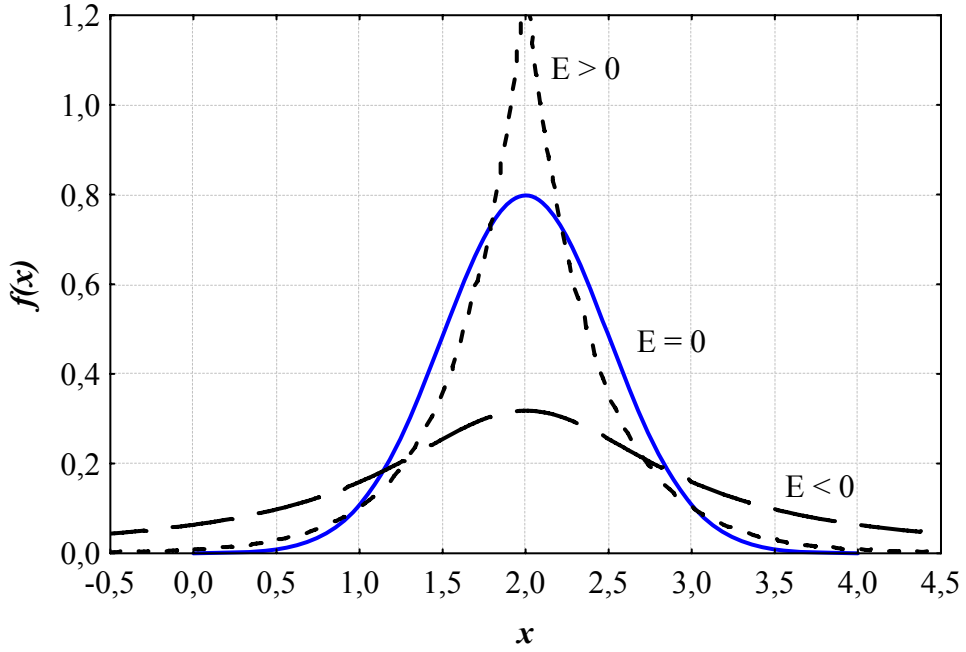


Рис.1.3. Зависимость графиков плотности вероятности симметричной f(x) от коэффициента эксцесса E

Квантилью порядка p распределения случайной величины X непрерывного типа называется действительное число t_p, удовлетворяющее уравнению

$$P(X < t_p) = p.$$

Значения $t_{0,75}$ и $t_{0,25}$ называются, соответственно, верхней и нижней ***квартилями***. Квартильный размах, равный разности верхней и нижней квартиляй, представляет собой интервал вокруг медианы, который содержит 50% значений ***X.***

Критической точкой порядка p распределения случайной величины ***X*** непрерывного типа называется действительное число k_p, удовлетворяющее уравнению

$$P(X \geq k_p) = p.$$

Квантиль и критическая точка одного и того же распределения связаны соотношением $k_p = t_{1-p.}$

1.1.4. Примеры законов распределения случайной величины

Рассмотрим некоторые важные для практики законы распределения случайных величин.

Непрерывная случайная величина имеет ***нормальное распределение*** с параметрами $a \in R$ и $\sigma > 0$, если плотность распределения вероятностей имеет вид:

$$f_N(x; a; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad (1.10)$$

где параметры a – математическое ожидание ($a = M[X]$), σ – среднее квадратичное отклонение X ($\sigma = +\sqrt{D[X]}$). В данном случае математическое ожидание a совпадает с медианой x_{med} и модой x_{mod}.

Если случайная величина распределена по закону $N(x;0;1)$, то она называется ***стандартизированной нормальной величиной***. Функция распределения для нее имеет вид: $F_N(x;0;1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

Графики плотности *f(x)* и функции *F(x)* распределения стандартизированной нормальной величины изображены на рис. 1.4.

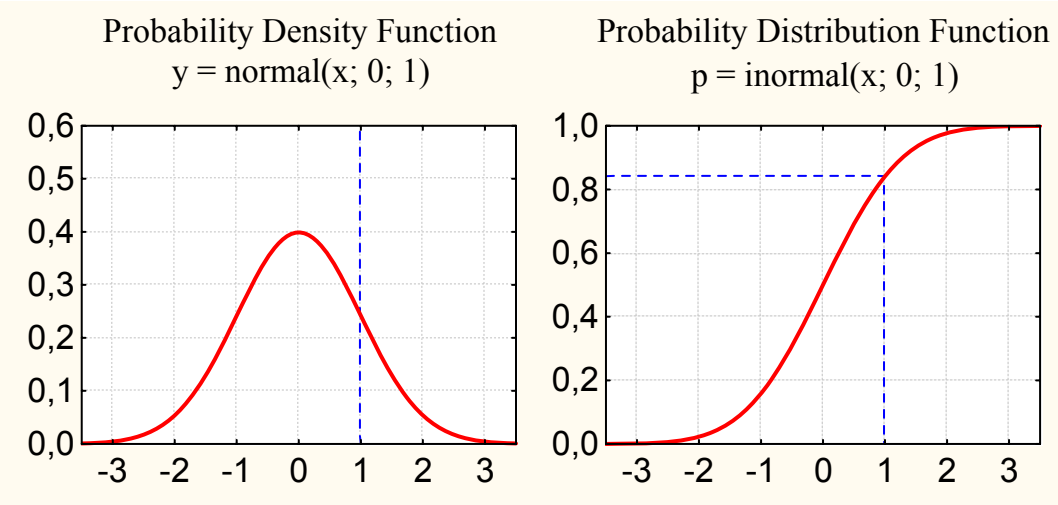


Рис. 1.4. Графики f(x) и F(x) по закону $N(x;0;1)$

На рис.1.5 изображены графики плотности *f(x)* нормального распределения при фиксированном $a = 2$ и разных σ. С уменьшением σ кривая *f(x)* сжимается, концентрируясь вокруг прямой $x = 2$.

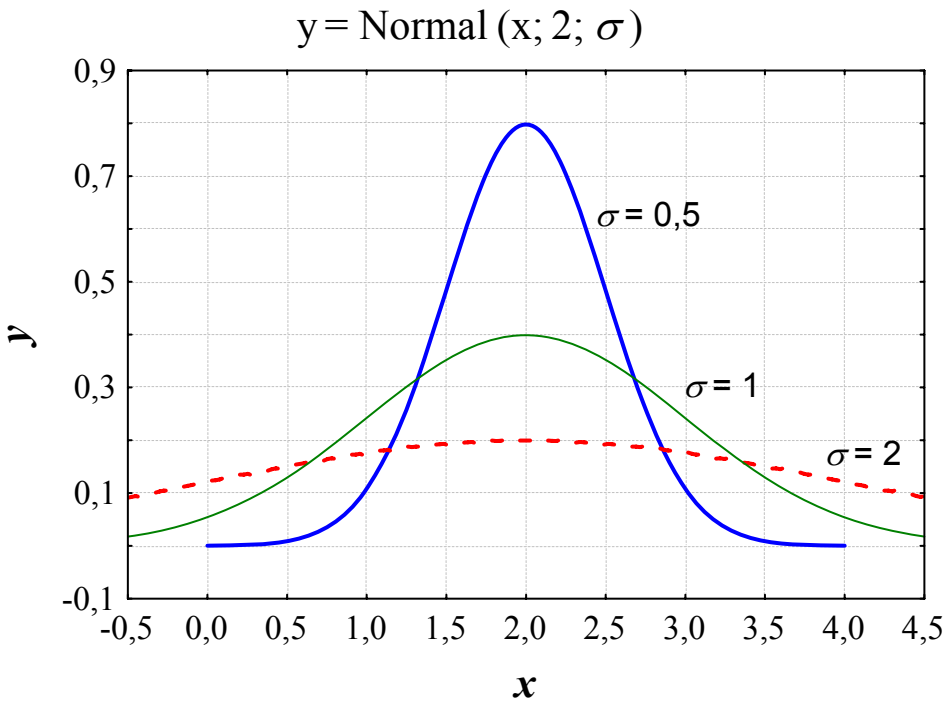


Рис. 1.5. Зависимость графиков плотности нормального распределения от стандартного отклонения σ

Центральные моменты нормального распределения удовлетворяют рекуррентному соотношению

$$\mu_{n+2} = (n+1)\sigma^2 \mu_n , \qquad n = 1, 2, \ldots ,$$

откуда, в частности, следует, что все центральные моменты нечетного порядка равны нулю, так как $\mu_1 = 0$ и, таким образом, $A = \frac{\mu_3}{\sigma^3} = 0$. С учетом $\mu_4 = 3\sigma^2, \mu_2 = 3\sigma^4$ имеем $E = \frac{\mu_4}{\sigma^4} - 3 = 0$. В этом смысле кривая плотности нормального распределения является эталонной ($A = 0$, $E = 0$), с которой сравнивают $f(x)$ других распределений при одинаковых M(X) и D(X). Причем, на фоне кривой плотности нормального распределения, график плотности распределения $f(x)$ деформирован (асимметричен) влево, если $A > 0$, и вправо, если $A < 0$; остроконечен (вытянут вверх), если $E > 0$, и тупоконечен, если $E < 0$ (рис.1.2 – 3).

Значения функции $p = F(x)$ и обратной к ней можно вычислить с помощью калькулятора распределения вероятности (рис.1.6).

Рис.1.6. Калькулятор нормального распределения вероятности. Площадь затемненной области равна 0,841345 = p = F(1)

Нормальное распределение непрерывной случайной величины имеет очень широкое распространение в случайных явлениях природы, так как ему подчиняется распределение случайной величины, представ-

ленной в виде суммы слабо зависимых случайных величин, сравнимых по порядку их влияния на рассеивание суммы. На практике очень часто встречаются случайные величины, образующиеся именно в результате суммирования многих случайных слагаемых, сравнимых по степени своего влияния на рассеивания суммы.

Непрерывная случайная величина X имеет ***логнормальное распределение***, если Y = lnX подчинен нормальному закону $N(y;a;\sigma)$, то есть если плотность распределения вероятностей имеет вид (см рис.1.7):

$$f_{LN}(x;\mu;\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ \mu=\ln a,\quad 0<x<\infty. \qquad (1.11)$$

Числовые характеристики логнормального распределение:

$$M[x]=ae^{\sigma^2/2},$$

$$x_{med}=a,\ x_{mod}=ae^{-\sigma^2},$$

$$D[x]=a^2e^{\sigma^2}(e^{\sigma^2}-1), \qquad (1.11)^*$$

$$A=(e^{\sigma^2}+2)\sqrt{(e^{\sigma^2}-1)},$$

$$E=(e^{\sigma^2}-1)(e^{3\sigma^2}+3e^{2\sigma^2}+6e^{\sigma^2}+6).$$

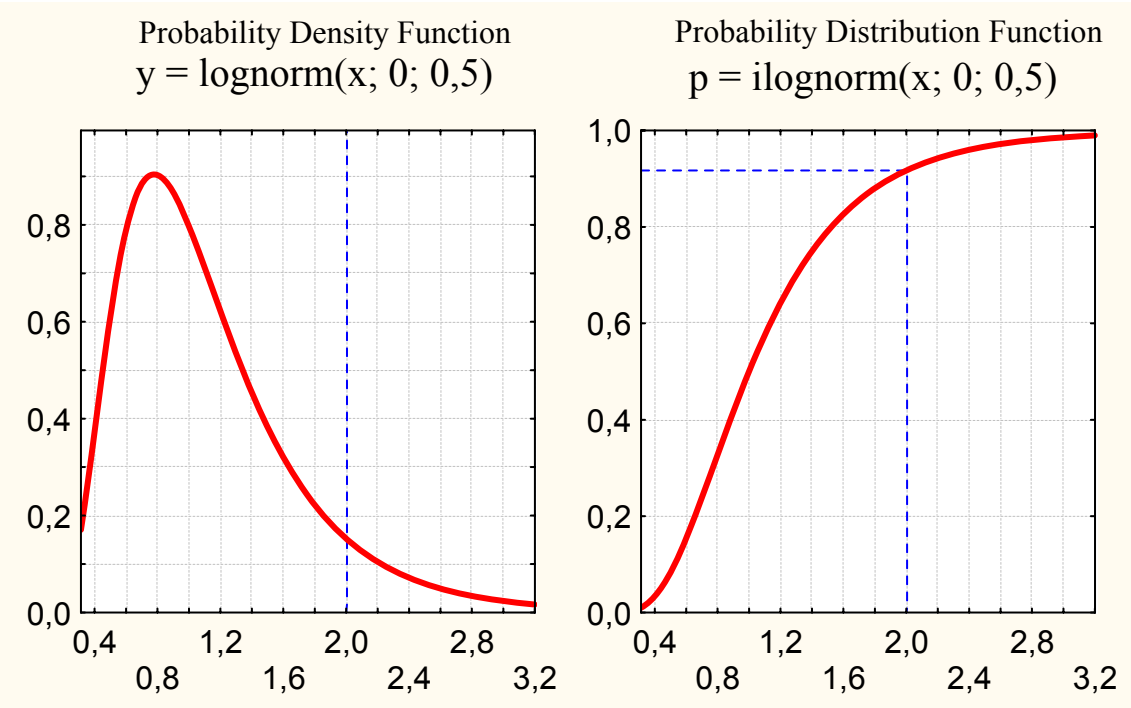


Рис. 1.7. Графики f(x) и F(x) логнормального распределения

Значения функции логнормального распределения $p = F(x)$ и обратной к ней можно вычислить с помощью калькулятора распределения вероятности (см. рис.1.8).

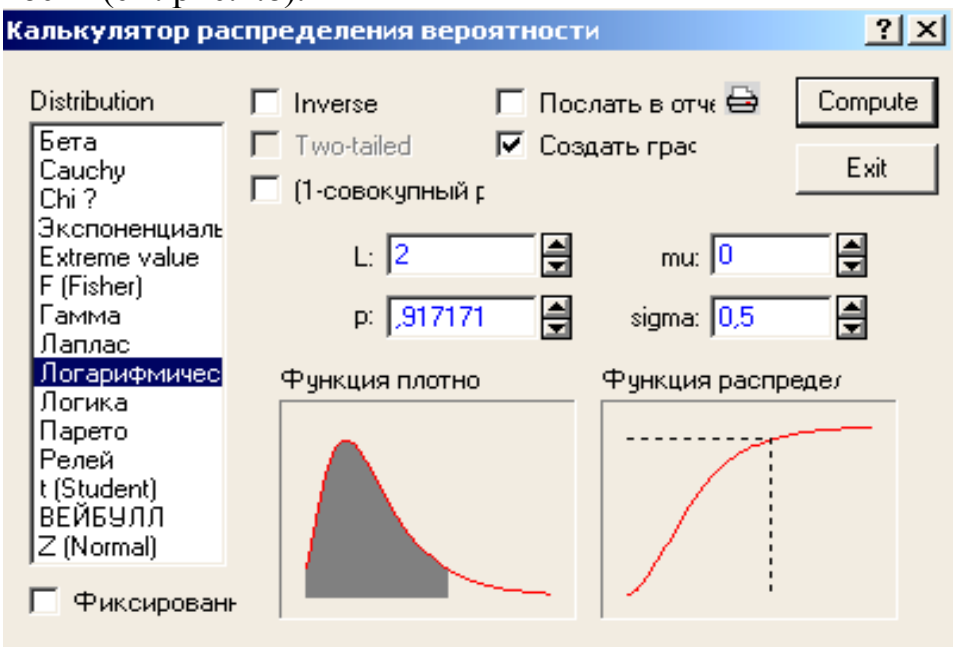


Рис.1.8. Калькулятор логнормального распределения вероятности. Площадь затемненной области равна 0,917171 = p = F(2)

Из определения логнормального распределения следует, что если случайная величина Y распределена нормально, то $X = e^Y$ распределена логнормально. Таким образом, логнормальному распределению подчиняется распределение случайной величины, представленной в виде произведения слабозависимых случайных величин, сравнимых по порядку их влияния.

Непрерывная случайная величина X имеет ***хи–квадрат распределение*** с m–степенями свободы, если она представима в виде суммы квадратов m величин, распределенных по нормальному закону $N(1;0)$, то есть если плотность распределения вероятностей имеет вид (рис. 1.9):

$$f_{Ch}(x;m) = \frac{1}{2^{m/2}\Gamma(m/2)} e^{-\frac{x}{2}} x^{\frac{m}{2}-1}, \quad 0 < x < \infty, \qquad (1.12)$$

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

$$\Gamma\left(\frac{2n+1}{2}\right) = \frac{1}{2^n}(2n-1)!!\sqrt{\pi} \quad \text{и} \quad \Gamma(n+1) = n! \quad \text{для} \quad n = \overline{0,\infty}.$$

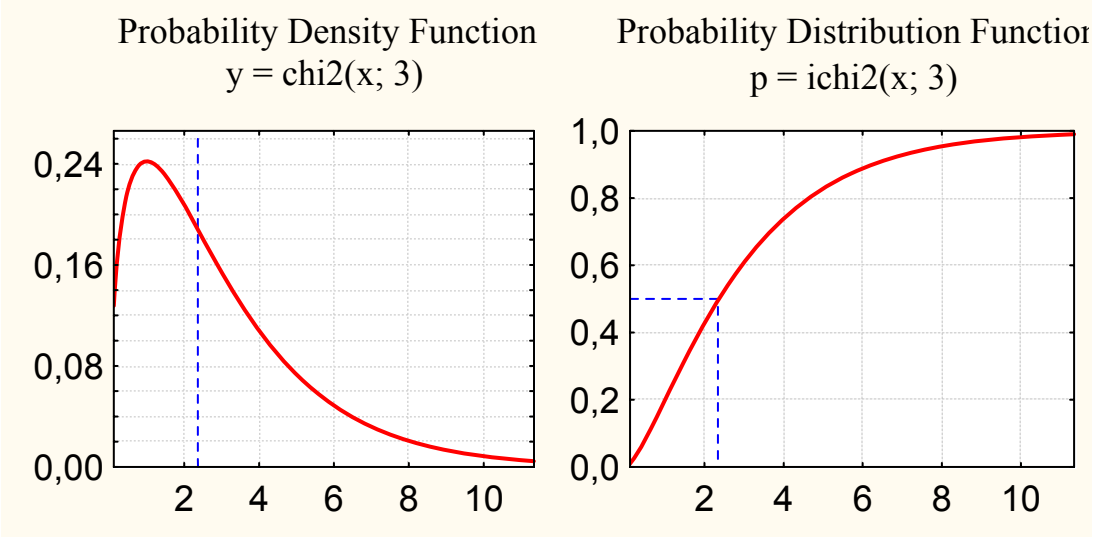


Рис. 1.9. Графики f(x) и F(x) хи– квадрат распределения

Числовые характеристики хи–квадрат распределение:

$$M[x] = m\ ,\ x_{mod} = m - 2\ ,\ D[x] = 2m\ ,$$

$$A = 2^{3/2} / \sqrt{m}\ ,\ E = 12 / m\ .$$

График плотности хи–квадрат распределения асимметричен (скошен влево, так как $A > 0$), островершинен ($E > 0$) и $x_{mod} < M[x]$,

Зависимость графиков плотности хи–квадрат распределения от m представлена на рис. 1.10.

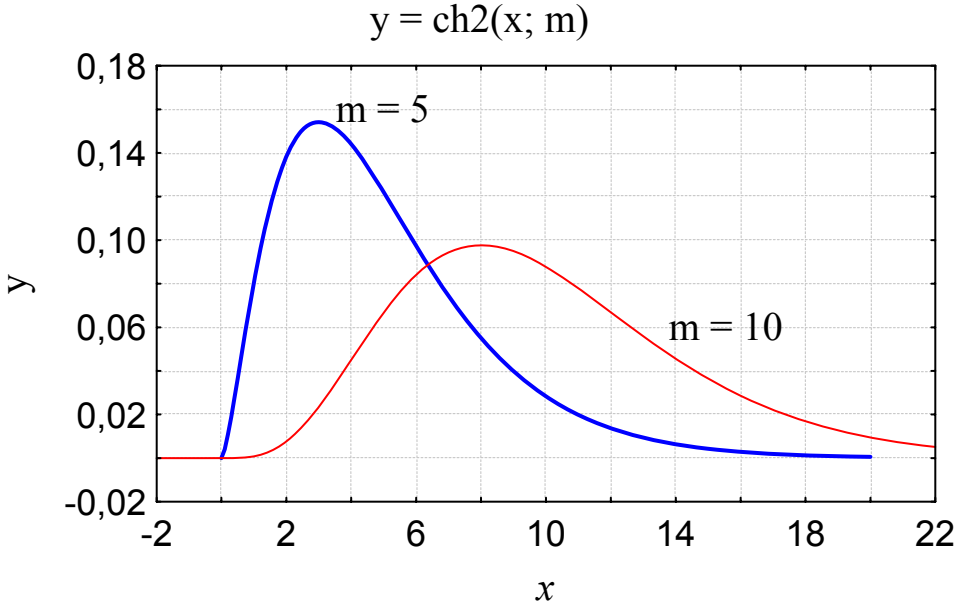


Рис. 1.10. Зависимость графиков f(x) хи–квадрат распределения от m

Значения функции хи–квадрат распределения p = *F(x)* и обратной к ней можно вычислить с помощью калькулятора распределения вероятности (рис. 1.11).

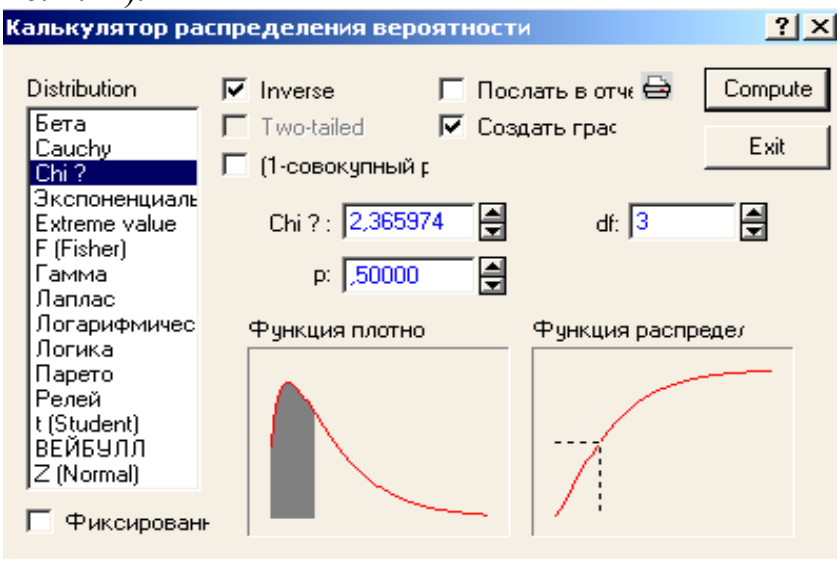


Рис.1.11. Калькулятор хи–квадрат распределения вероятности. Площадь затемненной области равна 0,5 = p = F(2,365974)

Непрерывная случайная величина X имеет ***t–распределение Стьюдента*** с m степенями свободы, если плотность распределения вероятностей имеет вид (рис. 1.12):

$$f_t(x;m)=\frac{1}{\sqrt{\pi m}}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}},\quad -\infty<x<\infty. \tag{1.13}$$

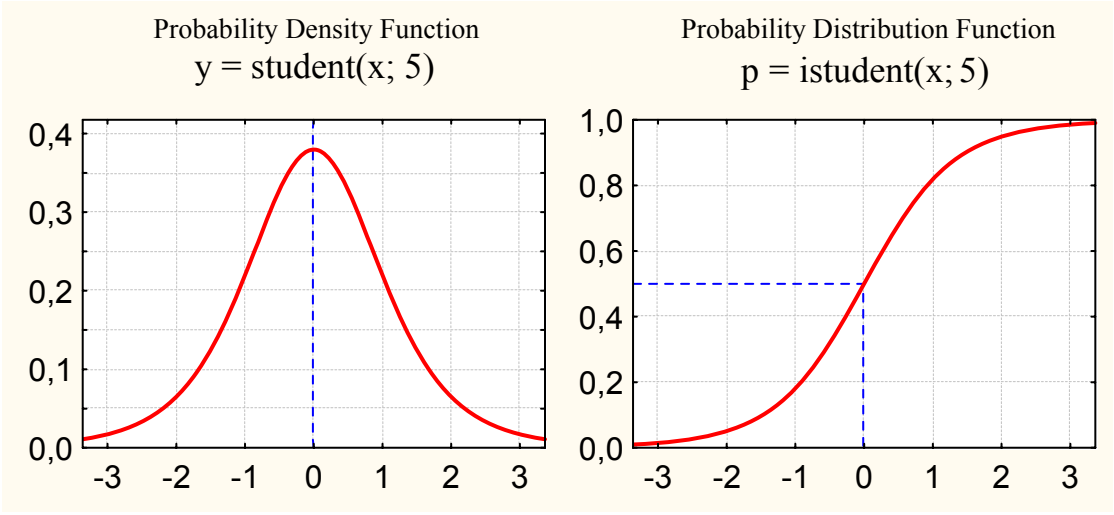


Рис. 1.12. Графики f(x) и F(x) закона t–распределения

Числовые характеристики $t-$ распределения:

$$M[x] = x_{med} = x_{mod} = 0\,,\ D[x] = \frac{m}{m-2}\,,$$

$$A = 0\,,\ E = \frac{6}{m-4}\,.$$

При больших степенях свободы (m > 30) t–распределение практически совпадает с нормальным распределением $N(x;0;1)$.

Значения функции $t-$распределения p = $F(x)$ и обратной к ней можно вычислить с помощью калькулятора распределения вероятности (см. рис.1.1).

Непрерывная случайная величина X имеет ***F–распределение Фишера,*** если она представима в виде отношения двух случайных величин, распределенных по закону хи–квадрат со степенями свободы ν и ω, и плотность распределения вероятностей имеет вид (рис.1.13):

$$f_F(x;\nu;\omega) = \frac{\Gamma\left(\frac{\nu+\omega}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\omega}{2}\right)}\left(\frac{\nu}{\omega}\right)^{\nu/2} x^{\frac{\nu}{2}-1}\left(1+\frac{\nu}{\omega}x\right)^{-\frac{\nu+\omega}{2}}, \quad 0 < x < \infty. \qquad (1.14)$$

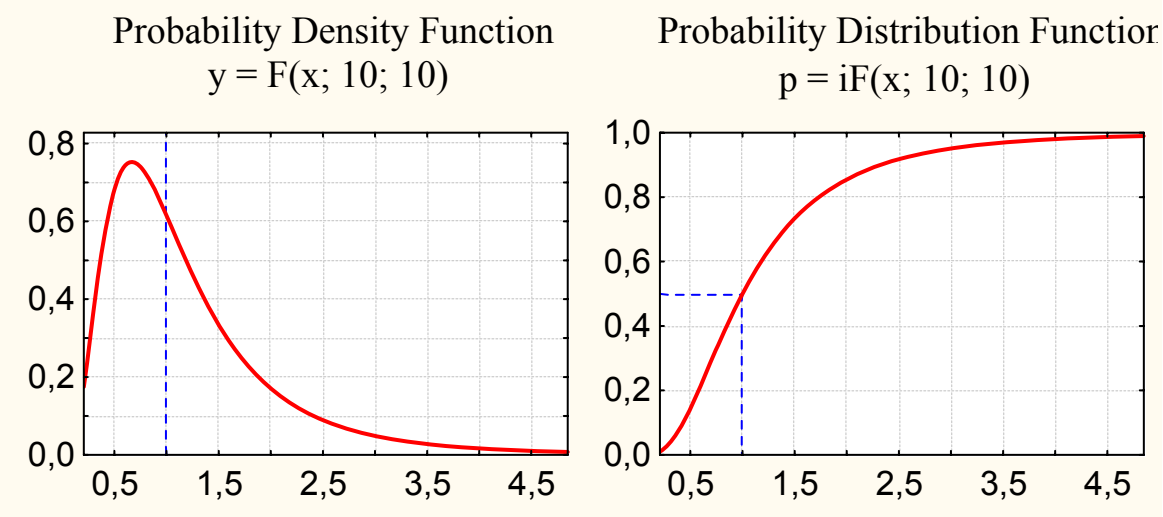


Рис. 1.13. Графики f(x) и F(x) закона F – распределения Фишера

Значения функции $F-$распределения Фишера p = $F(x)$ и обратной к ней можно вычислить с помощью калькулятора распределения вероятности (рис. 1.14).

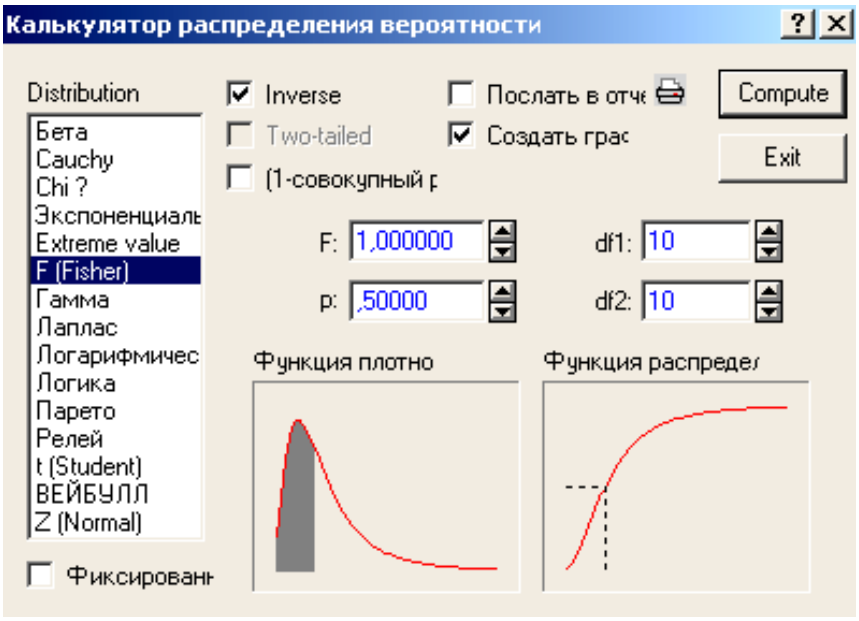


Рис.1.14. Калькулятор F–распределения вероятности. Площадь затемненной области равна 0,5 = p = F(1)

Зависимость графиков плотности $f_F(x;\nu;\omega)$ *F*–распределения Фишера от параметров ν и ω представлена на рис.1.15.

$$y_1 = F(x;5;5)$$
$$y_2 = F(x;20;20)$$
$$y_3 = F(x;100;100)$$

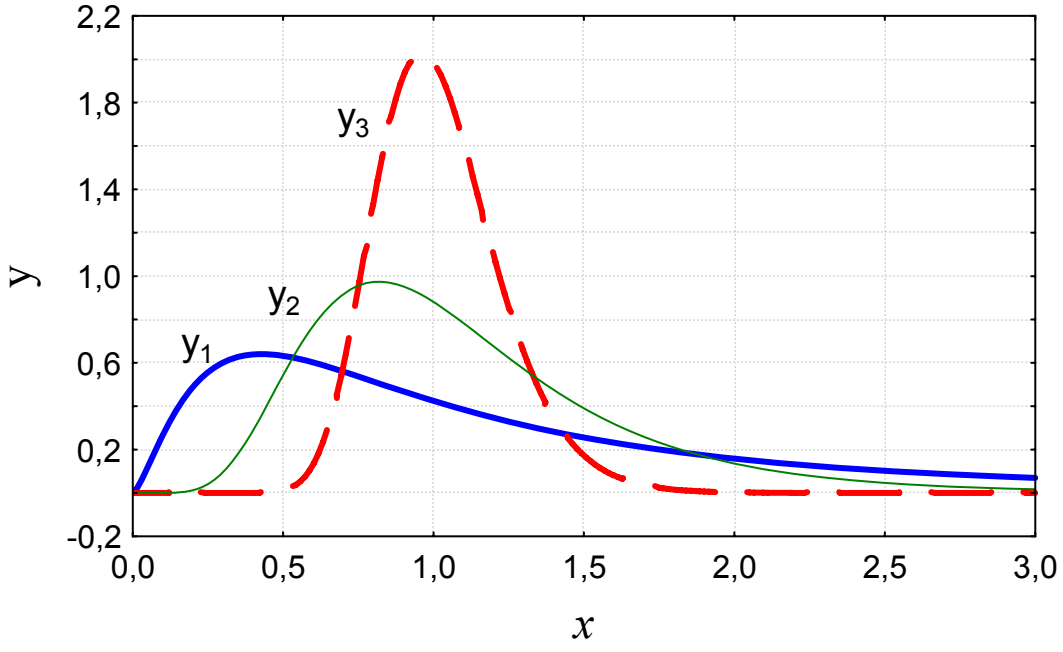


Рис.1.15. Зависимость графиков плотности F–распределения от ν и ω

1.1.5. Система двух случайных величин

Функцией распределения двумерной случайной величины (системы двух случайных величин $\{X,Y\}$) называется неубывающая функция двух действительных переменных, определяемая как вероятность совместного выполнения двух неравенств:

$$F_{X,Y}(x, y) = P\{X < x,\ Y < y\},$$

и удовлетворяющая следующим свойствам:

$$\lim_{x\to-\infty} F_{X,Y}(x, y) = \lim_{y\to-\infty} F_{X,Y}(x, y) = 0,\ \lim_{x,y\to+\infty} F_{X,Y}(x, y) = 1,$$

$$\lim_{x\to+\infty} F_{X,Y}(x, y) = F_Y(y),\ \lim_{y\to+\infty} F_{X,Y}(x, y) = F_X(x).$$

Систему двух случайных величин называют ***непрерывно распределенной***, если их функция распределения непрерывна на всей плоскости и существует такая неотрицательная интегрируемая функция $f_{X,Y}(x,y)$, называемая ***плотностью распределения вероятностей*** $\{X,Y\}$, что

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} d\xi \int_{-\infty}^{y} f_{X,Y}(\xi,\eta)d\eta,\qquad f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y},$$

$$\int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} f_{X,Y}(\xi,\eta)d\eta = 1,\qquad f_{X,Y}(x, y) \rho 0.$$

Плотности распределения вероятностей по каждой переменной выражаются в виде

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,\eta)d\eta,\qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(\xi,y)d\xi.$$

Тогда $F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi$, $\qquad F_Y(y) = \int_{-\infty}^{y} f_Y(\eta)d\eta.$

Для системы двух случайных величин {X,Y} вводятся числовые характеристики – ***моменты порядка K + S***.

Начальный момент порядка K + S

$$\alpha_{K,S} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^K y^S f_{X,Y}(x, y)dxdy.$$

В частности, $(m_X, m_Y) = (\alpha_{1,0}, \alpha_{0,1})$ называется ***математическим ожиданием*** $\{X,Y\}$ или центром рассеивания.

Центральный момент порядка K + S

$$\mu_{K,S} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-m_X)^K (y-m_Y)^S f_{X,Y}(x,y)dxdy.$$

В частности, $\mu_{2,0} = D_X$, $\mu_{0,2} = D_Y$ – ***дисперсии*** , $\mu_{1,1} = K_{XY}$ – ***ковариация (корреляционный момент)***. Нормированная ковариация

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} \tag{1.15}$$

называется ***коэффициентом корреляции*** системы двух случайных величин. Здесь $\sigma_X = \sqrt{D_X}$, $\sigma_Y = \sqrt{D_Y}$ – среднеквадратичные отклонения. Корреляционный коэффициент удовлетворяет условию $|\rho_{XY}| \le 1$ и определяет степень линейной зависимости между X и Y.

Систему двух случайных величин $\{X,Y\}$ называют ***дискретно распределенной***, если множество возможных значений $\{x_i, y_j\}$ счетное и задана соответствующая каждой паре вероятность $p_{ij} = P\{X = x_i , Y = y_j\}$, удовлетворяющая условию

$$\sum_i \sum_j p_{ij} = 1 ,$$

где суммирование распространяется на все возможные значения индексов *i* и *j*. В случае конечного числа возможных значений строят таблицу распределения системы двух случайных величин {X,Y}.

Таблица 1.2

Матрица распределения дискретной двумерной случайной величины (системы двух случайных величин $\{X,Y\}$

Y	X			
	x_1	x_2	$\propto$	x_k
y_1	p_{11}	p_{21}	$\propto$	p_{k1}
y_2	p_{12}	p_{22}	$\propto$	p_{k2}
∂	∂	∂	∂	∂
y_m	p_{1m}	p_{2m}	$\propto$	p_{km}

Одномерные законы распределения отдельных компонент выражаются формулами

$$p_i = \mathrm{P}\{\mathrm{X} = x_i\} = \sum_j p_{ij} , \quad p_j = \mathrm{P}\{\mathrm{Y} = y_j\} = \sum_i p_{ij} .$$

Начальный момент порядка K + S

$$\alpha_{K,S} = \sum_i \sum_j x_i^K y_j^S p_{ij} .$$

Центральный момент порядка K + S

$$\mu_{K,S} = \sum_i \sum_j (x_i - m_X)^K (y_j - m_Y)^S p_{ij} .$$

В частности,

$$m_X = \sum_i \sum_j x_i\, p_{ij} = \sum_i x_i p_i \,,$$

$$m_Y = \sum_i \sum_j y_j\, p_{ij} = \sum_j y_j p_j \,,$$

$$D_X = \sum_i \sum_j (x_i - m_X)^2\, p_{ij} = \sum_i (x_i - m_X)^2 p_i = \sum_i x_i^2 p_i - m_X^2 \,,$$

$$D_Y = \sum_i \sum_j (y_j - m_Y)^2\, p_{ij} = \sum_j (y_j - m_Y)^2 p_j = \sum_j y_j^2 p_j - m_Y^2 \,,$$

$$K_{XY} = \sum_i \sum_j (x_i - m_X)\ (y_j - m_Y)\ p_{ij} = \sum_i \sum_j x_i y_j p_{ij} - m_X m_Y \quad .$$

1.2. Элементы математической статистики

Теоретической базой математической статистики является теория вероятностей, изучающая вероятностные закономерности массовых однородных случайных событий. Теория вероятностей изучает математические модели случайных явлений, при этом сама математическая модель остаётся заданной. В практических задачах характеристики математической модели, как правило, неизвестны, но имеются некоторые экспериментальные данные о событии или случайной величине. Требуется на основании этих данных построить подходящую теоретико-вероятностную модель изучаемого явления. Это и является задачей математической статистики − обширного раздела современной математики.

Методы математической статистики широко применяются в различных отраслях естествознания.

Все это предопределяет необходимость овладения методами математической статистики как инструментом статистического анализа и прогнозирования естественно−научных явлений и процессов.

1.2.1. Выборочный метод

Полный набор всех возможных N значений дискретной случайной величины X называют ***генеральной совокупностью***. Часть генеральной совокупности из n элементов, отобранных случайным образом, называется ***выборкой***. При этом число n называют ***объемом выборки***. Различают выборки малого объема ($n < 30$) и большого объема ($n > 30$) .

В начале на основе результатов эксперимента строят ***простой статистический ряд*** – таблицу, состоящую из двух строк, в первой – порядковый номер измерения, во второй – его результат.

Таблица 1.3

Простой статистический ряд случайной величины X

i	1	2	…	n
x_i	x_1	x_2	…	x_n

Для визуальной оценки распределения случайной величины производят группировку данных. Вначале x_i располагают в порядке возрастания, затем интервал наблюдаемых значений случайной величины разбивают на k последовательных непересекающихся частичных интервалов $\tilde{x}_0 \div \tilde{x}_1$, $\tilde{x}_1 \div \tilde{x}_2$, … $\tilde{x}_j \div \tilde{x}_{j+1}$, …, $\tilde{x}_{k-1} \div \tilde{x}_k$, далее подсчитывают ***частоты*** n_j – количество x_i, попавших в j – тый интервал. Полученный таким образом ***группированный статистический ряд*** отражают таблицами следующего вида.

Таблица 1.4

Группированный статистический ряд частот n_j случайной величины X

$\tilde{x}_{j-1} \div \tilde{x}_j$	$\tilde{x}_0 \div \tilde{x}_1$	$\tilde{x}_1 \div \tilde{x}_2$	···	$\tilde{x}_{k-1} \div \tilde{x}_k$
n_j	n_1	n_2	. . .	n_k

или, подсчитывая ***относительные частоты*** (1.1)*, $p_j = \dfrac{n_j}{n}$.

Таблица 1.5

Группированный статистический ряд относительных частот p_j случайной величины X

$\tilde{x}_{j-1} \div \tilde{x}_j$	$\tilde{x}_0 \div \tilde{x}_1$	$\tilde{x}_1 \div \tilde{x}_2$	···	$\tilde{x}_{k-1} \div \tilde{x}_k$
p_j	p_1	p_2	. . .	p_k

или, определяя середину j–ого интервала $\overline{x}_j = \tilde{x}_j - 0.5\Delta_j$, где $\Delta_j = \tilde{x}_j - \tilde{x}_{j-1}$ – длина j–ого интервала, получим ряд распределения.

Таблица 1.5*

Группированный статистический ряд относительных частот p_j случайной величины X (с указанием середин интервалов)

$\overline{x}_j$	$\overline{x}_1$	$\overline{x}_2$	. . .	$\overline{x}_k$
p_j	p_1	p_2	. . .	p_k

При этом частоты p_j **удовлетворяют условию** $\sum_{j=1}^{k} p_j = 1$.

Деля частоту p_j на длину соответствующего интервала Δ_j, получим таблицу ***плотностей частоты*** f_j. Откладывая по оси абсцисс интервалы $\tilde{x}_{j-1} \div \tilde{x}_j$ и надстраивая на каждом интервале, как на основании, прямоугольник высотой f_j и площадью p_j, получим ступенчатую фигуру – ***гистограмму*** частот – статистический аналог кривой плотности распределения. Еще более точной оценкой кривой плотности распределения является ***полигон*** частот – ломаная, отрезки которой соединяют точки ($\overline{x}_j$, f_j). В итоге ряд распределения принимает следующий вид.

Таблица 1.6

Группированный статистический ряд плотностей частот f_j *случайной величины* X

$\overline{x}_j$	$\overline{x}_1$	$\overline{x}_2$	. . .	$\overline{x}_k$
f_j	f_1	f_2	. . .	f_k

Другим способом представления эмпирического закона распределения являются накопленные частоты $\sum_{i=1}^{j} n_i$ (или $\sum_{i=1}^{j} p_i$ – накопленные относительные частоты).

Таблица 1.7

Группированный статистический ряд накопленных частот $\sum_{i=1}^{j} n_i$ *или накопленных относительных частот* $\sum_{i=1}^{j} p_i$ *случайной величины* X

$\tilde{x}_{j-1} \div \tilde{x}_j$	$\tilde{x}_0 \div \tilde{x}_1$	$\tilde{x}_1 \div \tilde{x}_2$	. . .	$\tilde{x}_{k-1} \div \tilde{x}_k$
$\sum_{i=1}^{j} n_i$	n_1	n_1+n_2	. . .	$\sum_{i=1}^{k} n_i$
$\sum_{i=1}^{j} p_i$	p_1	p_1+p_2	. . .	1

Накопленные относительные частоты порождают ***эмпирическую функцию распределения*** – оценку функции распределения дискретной случайной величины $\boldsymbol{X}$, вычисляемую по формуле (1.2)*

$$F(x) = \sum_{x_i < x} p_i$$

и являющуюся разрывной ступенчатой, равной нулю (левее наименьшего наблюдаемого значения), испытывающей скачок величиной p_j при переходе через левую границу j–ого интервала и в итоге достигающей единицы правее наибольшего наблюдаемого значения.

Система STATISTICA позволяет по выборке микроэлемента (прил. 1) построить таблицу частот n_j, p_j, $\sum_{i=1}^{j} n_i$ и $\sum_{i=1}^{j} p_i$ (например, для $La_И$ – табл. 1.8), а также нарисовать гистограммы частот (рис.1.16).

Таблица 1.8

Частоты распределения содержания $La_И$

j	Интервалы	n_j	$\sum_{i=1}^{j} n_i$	p_j, %	$\sum_{i=1}^{j} p_i$,%
1	10,0<x<=15,0	1	1	1,0989	1,0989
2	15,0<x<=20,0	8	9	8,7912	9,8901
3	20,0<x<=25,0	19	28	20,8791	30,7692
4	25,0<x<=30,0	35	63	38,4615	69,2308
5	30,0<x<=35,0	25	88	27,4725	96,7033
6	35,0<x<=40,0	2	90	2,1978	98,9011
7	40,0<x<=45,0	1	91	1,0989	100,0000
	∑	91		100,0000	

Гистограмма частот n_j и $\sum_{i=1}^{j} n_i$ дана на рис.1.16.

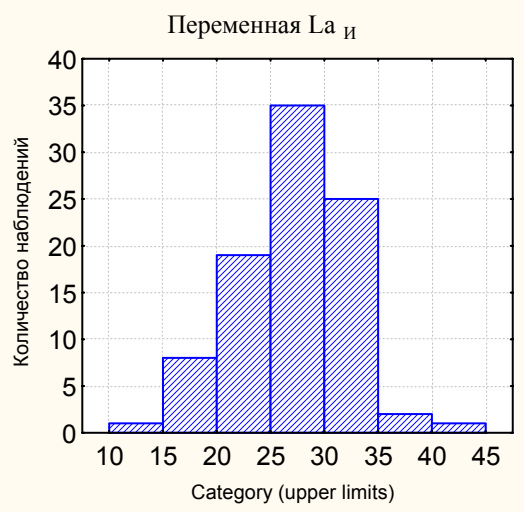


Рис. 1.16. Гистограмма частот $\mathbf{n_j}$ *распределения* $La_И$

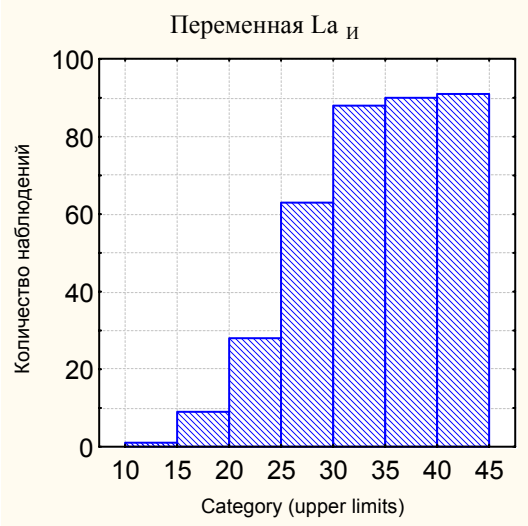


Рис. 1.16. Гистограмма накопленных частот распределения* $La_И$

Характеристики случайной величины, построенные на основании выборочных данных, называются ***выборочными*** или ***точечными оценками***. Свойства случайной величины могут характеризоваться различными начальными и центральными моментами, вычисляемыми в случае дискретной случайной величины по следующим формулам :

Начальный момент порядка K

$$\alpha_K = \sum_i^{=K} x_i \; p_i$$

Центральный момент порядка K

$$\mu_K = \sum_i \left(\overset{=}{x_i} - \alpha_1 \right)^K \; p_i \; .$$

Важнейшие из них – ***математическое ожидание*** $M(X) = m_X$ и ***дисперсия*** $D(X) = \sigma^2 (X)$, где через σ обозначено ***стандартное отклонение***– являются частными случаями моментов:

$$\overline{x} = \alpha_1 \; , \; \overline{D} = \mu_2 \; , \; \overline{\sigma} = \sqrt{\overline{D}} \; . \qquad (1.16)$$

Выделяют также несмещенную выборочную дисперсию

$$s^2 = \frac{n}{n-1} \overline{D} \; . \qquad (1.17)$$

Если ***выборочное*** математическое ожидание случайной величины дает нам ее «среднее» значение или точку на координатной прямой , вокруг которой «разбросаны» значения рассматриваемой случайной величины, то ***выборочная*** дисперсия характеризует «степень разброса» значений случайной величины X.

Используются также оценки коэффициента асимметрии (1.8) $A = \frac{\mu_3}{s^3}$ и коэффициента эксцесса (1.9) $E = \frac{\mu_4}{s^4} - 3$, как степени отклонения полигона частот от плотности нормального распределения непрерывной случайной величины, для которой они равны нулю.

Система STATISTICA позволяет по выборке микроэлемента вычислить точечные оценки, например, для $La_И$.

Таблица 1.9

Выборочные числовые характеристики распределения содержания $La_И$

	n	$\overline{x}$	s	A	E
$La_И$	91	27,02527	5,039656	-0,205849	0,082287

Выборочные числовые характеристики или точечные оценки случайной величины – приближенные значения параметров распределения. Чтобы охарактеризовать погрешность этих значений, нужно указать граничные значения, за которые не выходит оцениваемый параметр. Поскольку все расчёты производятся на основании случайных результатов опыта, то и граничные значения также являются случайными величинами. Таким образом, речь идёт о построении интервала со случайными границами, который с заданной вероятностью содержал бы неизвестное значение параметра распределения.

Для определения погрешности полученных значений используют ***интервальные оценки***, применяя понятие «***доверительного интервала***» – интервала, внутри которого параметр, как ожидается, найдется с некоторой доверительной вероятностью (надежностью) β. Иногда вместо β используют величину α, *равную* $1-\beta$ и называемую уровнем значимости.

Рассмотрим нахождение доверительного интервала для математического ожидания $\boldsymbol{m_x}$ нормально распределенной случайной величины. Ширина 2ε такого интервала ($\overline{x}$ –ε, $\overline{x}$ +ε), обладающего симметрией относительно $\overline{x}$ –выборочного значения m_x, находится из условия

$$P(|x - \overline{x}| < \varepsilon) = \beta ,$$

причем сама вероятность $P(|x - \overline{x}| < \varepsilon)$ определяется законом распределения Стьюдента (1.13) со степенью свободы $k = n-1$, если дисперсия не известна, а лишь подсчитано ее несмещенное значение s^2; $\beta = F_t(x_\beta; k)$. По заданным β и k калькулятор распределения вероятности распределения Стьюдента (рис. 1.1, где p = β, $t = x_\beta$ и df = k) по-

зволяет найти соответствующее значение x_β . Из условия $x_\beta = \dfrac{\varepsilon\sqrt{n}}{s}$ можно найти $\varepsilon = \dfrac{x_\beta s}{\sqrt{n}}$. В результате можно построить доверительный интервал $(\bar{x} - \varepsilon,\ \bar{x} + \varepsilon)$, содержащий параметр m_x с вероятностью β.

В случае с $La_И$ при $\beta = 0{,}95$ имеем x_β = 1,986675 (рис. 1.17), ε = 1,04955 и доверительный интервал (25,9757; 28,0748) , содержащий параметр m_x с вероятностью (надежностью) $\beta = 0{,}95$**.** Иными словами, погрешность вычисления математического ожидания по приближенному значению $\bar{x}$ = 27,02527 не превышает ε = 1,04956 при уровне значимости $\alpha = 0{,}05$. Величина $\dfrac{s}{\sqrt{n}}$ = 0,5283 называется ***стандартной ошибкой*** X и равна ε при $x_\beta = 1$, чему соответствует не очень высокая надежность $\beta = 0{,}68$ при $k = 90$. Практически доверительный интервал можно построить с помощью точечных оценок распределения содержания $La_И$, аналогично табл.1.9 (рис.1.17).

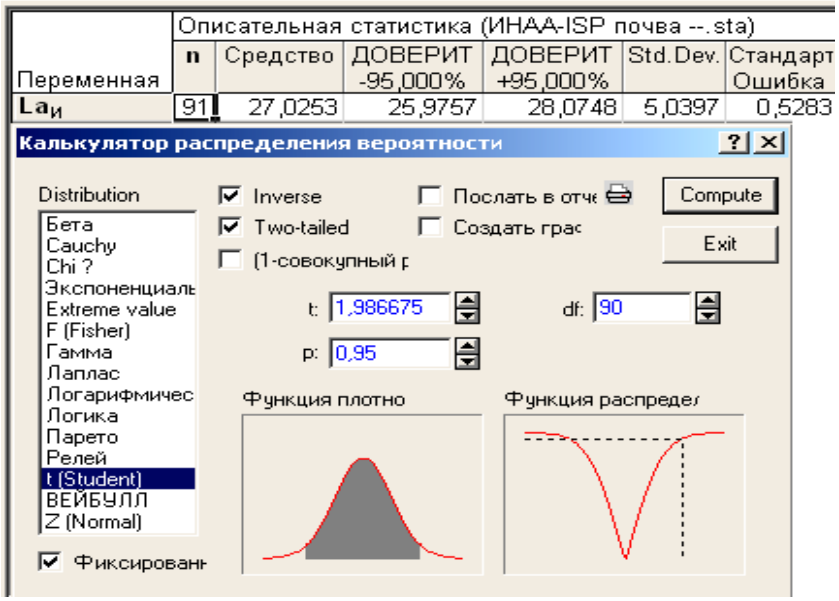


Рис. 1.17. Доверительный интервал для m_x распределения $La_И$

1.2.2. Корреляционно–регрессионный анализ

Для многих явлений в природе типичны случайные зависимости. Случайные величины находятся в корреляционной зависимости, если каждому значению одной из них соответствует некоторое распределе-

ние другой, что математически отражается в уравнении регрессии одной случайной величины на другую.

По результатам эксперимента сначала оформляется таблица наблюдений системы дискретных случайных величин (*X, Y*) – ***матрица распределения*** – прямоугольная таблица, в которой записаны наблюдаемые значения для *X: {x1, x2, … , xk }* , для Y*: {y1, y2, … , ym }* и соответствующая каждой паре $\{x_i, y_j\}$ вероятность $p_{ij} = P\{X = x_i , Y = y_j\}$, удовлетворяющая условию $\sum_i \sum_j p_{ij} = 1$. При этом система двух случайных величин (*X, Y*) характеризуется набором начальных и центральных моментов (п. 1.1.5).

В общем случае Y и *X* связаны вероятностной зависимостью, справедливой лишь в среднем, так как при фиксированном значении *X = x* зависимая переменная *Y* имеет случайный разброс (столбец значений) из-за ошибок измерения, влияние неучтенных факторов или других причин. Таким образом, фиксированному значению $X = x_i$ соответствует усредненное значение $Y_{x_i} = M[Y/X = x_i]$ – условное математическое ожидание, вычисляемое по формуле

$$Y_{x_i} = \tilde{y}_i = \frac{1}{p_i} \sum_{j=1}^{m} y_j p_{ij} \qquad (1.18)$$

В итоге исходная таблица $\{x_i, y_j\}$ эквивалентна таблице $\{x_i, \tilde{\boldsymbol{y}}_{\boldsymbol{i}}\}$.

Условное математическое ожидание $Y_x = M[Y/X = x]$ называется ***регрессией*** *Y* на *X* , график зависимости $Y_x (x)$ называется линией регрессии. Аналогично определяется регрессия *X* на *Y*.

Таблица 1.10

Регрессионная матрица распределения двумерной случайной величины

x_i	x_1	x_2	…	x_k
$\tilde{\boldsymbol{y}}_{\boldsymbol{i}}$	$\tilde{\boldsymbol{y}}_1$	$\tilde{\boldsymbol{y}}_2$	…	$\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$
p_i	p_1	p_2	…	p_k

Рассмотрим модель линейной по параметрам регрессии Y на X, находящей линейную комбинацию $Y_x(x) = f(x) = \sum_{j=1}^{n_\beta} \beta_j f_j(x)$ базисных функций f_j **,** которая лучше всего в смысле метода наименьших квадратов аппроксимирует массив $\{x_i , \tilde{y}_i\}$. В этом случае результаты наблюдений представляются в виде

$$\tilde{y}_i = f(x_i) + \varepsilon_i ,$$

где ε_I – случайные некоррелированные ошибки наблюдений в предположении, что $M[\varepsilon_i] = 0$, $D[\varepsilon_i] = M[\varepsilon_i^2] = \sigma_i^2$. Таким образом, при выбранных базисных функциях f_j оценки $\overline{\beta}_j$ коэффициентов β_j определяются из условия

$$\varepsilon(\beta_j) = \sum_i \varepsilon_i^2 p_i = \sum_i [\tilde{y}_i - f(x_i)]^2 p_i = \min .$$

Качество аппроксимации результатов наблюдений регрессивной моделью определяется остаточной дисперсией $s^2 = \dfrac{\varepsilon}{k - n_\beta}$ (n_β – число оцениваемых параметров β_j), которую можно использовать для сравнительного анализа нескольких регрессивных моделей.

Рассмотрим простую линейную регрессию, которая считается выполненной, $f(x) = \sum_{j=1}^{2} \beta_j x^{j-1} = \beta_1 + \beta_2 x$, если найдем оценки коэффициентов β_1 и β_2 из условия минимизации выражения $\sum_i [\tilde{y}_i - \beta_1 - \beta_2 x_i]^2 p_i$:

$$\overline{\beta}_1 + \overline{\beta}_2 \sum_{i=1}^{k} x_i p_i = \sum_{i=1}^{k} \tilde{y}_i p_i ;$$

$$\overline{\beta}_1 \sum_{i=1}^{k} x_i p_i + \overline{\beta}_2 \sum_{i=1}^{k} x_i^2 p_i = \sum_{i=1}^{k} \tilde{y}_i x_i p_i .$$

В этом случае $\overline{\beta}_1$ и $\overline{\beta}_2$ можно выразить через точечные оценки числовых характеристик системы дискретных случайных величин:

$$f(x) = \overline{y} + \overline{r}_{xy} \frac{\overline{\sigma}_y}{\overline{\sigma}_x} (x - \overline{x}) ,$$

где $\overline{x} = \sum_{i=1}^{k} x_i p_i$ – оценка m_x по массиву $\{x_i\}$,

$\overline{y} = \sum_{j=1}^{m} y_j p_j$ – оценка m_y по массиву $\{y_k\}$,

$\bar{\sigma}_x^2 = \sum_i (x_i - \bar{x})^2 p_i = \bar{D}_x$ – оценка D_x по массиву $\{x_i\}$,

$\bar{\sigma}_y^2 = \sum_j (y_j - \bar{y})^2 p_j = \bar{D}_y$ – оценка D_y по массиву $\{y_k\}$,

$\bar{K}_{xy} = \sum_{i,j} (x_i - \bar{x})(y_j - \bar{y}) p_{ij}$ – оценка **ковариации** по $\{x_i, y_k\}$,

$\bar{r}_{xy} = \dfrac{\bar{K}_{xy}}{\bar{\sigma}_x \bar{\sigma}_y}$ – выборочный **коэффициент корреляции Пирсона**, значение которого по модулю равно единице в случае линейной зависимости Y и X .

Таким образом, $|\bar{r}_{xy}|$ характеризует степень тесноты линейной зависимости между Y и X, проявляющейся в том, что при возрастании одной случайной величины другая проявляет тенденцию также возрастать (в этом случае $\bar{r}_{xy} > 0$) или убывать (в таком случае $\bar{r}_{xy} < 0$). В первом случае говорят, что **Y** и **X** связаны положительной корреляцией, а во втором – корреляция отрицательна. При этом зависимость тем ближе к линейному закону, чем $|\bar{r}_{xy}|$ ближе к единице слева. Если $\bar{r}_{xy} = 0$, то это означает только отсутствие линейной связи между **Y** и **X**, любой другой вид связи может при этом присутствовать.

Аналогично рассматривается регрессия $f(x) = \beta_1 + \beta_2 x + \beta_3 x^2$, которая лучше всего аппроксимирует массив $\{x_i, \tilde{y}_i\}$ в смысле метода наименьших квадратов, то есть определяющая коэффициенты β_j из условия $\varepsilon = \sum_i [\tilde{y}_i - f(x_i)]^2 \tilde{p}_i = \min$, где $\tilde{y}_i$ вычисляются по формуле (1.18).

Таким образом, наряду с прямой линейной регрессии строятся кривые полиномиальных регрессий, построенных методом наименьших квадратов и аппроксимированных полиномами порядка *M:*

$$\hat{Y}(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_M x^M .$$

Оценка значимости регрессии (качество уравнения регрессии) проверяется с помощью F-критерия (Справочник …, 1987; Боровиков, 2001), имеющего F-распределение (1.14) со степенями свободы *M* и *k – M–1*.

$$F = \frac{R^2}{1-R^2}\frac{k-M-1}{M}, \qquad R^2 = \frac{\sum_{i=1}^{k}\left(\hat{Y}(x_i)-\overline{Y}\right)^2 p_i}{\sum_{i=1}^{k}\left(\tilde{y}_i-\overline{Y}\right)^2 p_i}, \qquad \overline{Y} = \sum_{i=1}^{k}\tilde{y}_i p_i ,$$

Если уравнение регрессии служит для прогнозирования, то для повышения надежности рекомендуется добиться путем подбора соответствующего уравнения регрессии выполнения соотношения

$$F > 4F_{0,5;\, M,\ n-M-1}.$$

Степень адекватности регрессионной модели можно оценить, например, скорректированным коэффициентом детерминации

$$\hat{R}^2 = 1-\left(1-R^2\right)\frac{k-1}{k-M-1},$$

лежащим в пределах от 0 до 1. Он измеряет качество построенной регрессии: чем ближе коэффициент детерминации к 1, тем лучше регрессия «объясняет» зависимость в данных.

1.2.3. Проверка статистических гипотез

Во многих случаях результаты наблюдений используются для проверки предположений (гипотез) относительно тех или иных свойств распределения случайной величины. В частности, такого рода задачи возникают при сравнении методов обработки по определенным измеряемым признакам и т. д.

К основным задачам математической статистики относится статистическая проверка гипотез о законах распределения и о параметрах распределения случайной величины. При исследовании различных случайных величин на определённом его этапе появляется возможность выдвинуть ту или иную гипотезу о свойствах изучаемой величины, например, сделать предположение о законе распределения её, или, если закон распределения известен, но неизвестны его параметры, то сделать предположение о их величине. Наиболее правдоподобную по каким-то соображениям гипотезу называют нулевой (основной) и обозначают H_0. Наряду с основной гипотезой рассматривают другую (альтернативную) гипотезу H_1 , противоречащую основной. Выдвинутая нулевая гипотеза нуждается в дальнейшей проверке. При этом могут быть допущены ошибки двух видов:

- ✓ ошибка первого рода – отвергнута правильная гипотеза;
- ✓ ошибка второго рода – принята неправильная гипотеза.

Вероятность совершить ошибку первого рода (вероятность отвергнуть правильную гипотезу) обычно обозначают α и называют ***уровнем значимости.*** Случайную величину Z , служащую для проверки гипотезы, называют ***критерием.*** Совокупность значений критерия, при которых нулевую гипотезу отвергают, называют ***критической областью***. Граничные точки критической области z_{kp} называют ***критическими точками***. Различают три вида критической области:

- правосторонняя, определяемая неравенством $Z > z_{kp} > 0$;
- левосторонняя, определяемая неравенством $Z < z_{kp} < 0$;
- двусторонняя, определяемая неравенством $Z < z_1 < z_2 < Z$.

Например, если критические точки симметричны относительно нуля, то двусторонняя критическая область имеет вид $|Z| > z_{kp} > 0$.

При отыскании критической области задаются уровнем значимости α и ищут критические точки, исходя из требования, чтобы вероятность того, что критерий Z примет значения, лежащие в критической области, была равна принятому уровню значимости. В результате получаем:

- для правосторонней критической области $P\ (Z > z_{kp}) = \alpha$;
- для левосторонней критической области $P\ (Z < z_{kp}) = \alpha$;
- для двусторонней симметричной области $P\ (Z > z_{kp}) = \alpha/2$.

Основной принцип статистической проверки гипотез заключается в следующем. Если наблюдаемое значение критерия $Z_{набл}$, вычисленное по данным выборки, принадлежит критической области, то гипотезу отвергают; если наблюдаемое значение не принадлежит критической области, то нет оснований отвергать гипотезу.

Для многих критериев Z с учетом законов их распределения калькулятор распределения вероятности системы STATISTICA позволяет по α найти критические точки z_{kp} и наоборот (определить уровень значимости α значения критерия Z). Степень значимости отличия сравниваемых законов распределения или параметров распределения качественно определяется по уровню значимости (Боровиков, 2003): не значимые ($\alpha \geq 0{,}100$), слабо значимые ($0{,}100 > \alpha \geq 0{,}050$), статистически значимые ($0{,}050 > \alpha \geq 0{,}010$), сильно значимые ($0{,}010 > \alpha \geq 0{,}001$), высоко значимые ($0{,}001 > \alpha$).

Рассмотрим ***проверку гипотезы о законе распределении***.

Пусть дана выборка наблюдений случайной величины X: $\{x_1, x_2, \ldots, x_n\}$. Проверяется гипотеза H_0 , утверждающая, что X имеет функцию распределения $F(x)$ или плотность распределения $f(x)$. По выборке наблюдений находят оценки неизвестных параметров (если таковые есть) предполагаемого закона распределения случайной величины X.

Далее, интервал Ω возможных значений случайной величины X разбивается на k непересекающихся подинтервалов $\Omega_i = (a_i, b_i)$, $i = 1,2,...,k$ *n*. Число k определяется с учетом эмпирической формулы $k = 1+4lg(n)$. Пусть n_i – число элементов выборки, принадлежащих подинтервалу Ω_i. Очевидно, что $\sum_{i=1}^{k} n_i = n$. Используя предполагаемый закон распределения случайной величины X, находят вероятности p_i того, что значение X принадлежит подинтервалу Ω_i:

$$p_i = P(X \in \Omega_i) = \int_{\Omega_i} f(x)dx = F(b_i) - F(a_i), \quad \sum_{i=1}^{k} p_i = 1.$$

Далее вычисляют статистическое значение критерия по формуле

$$\chi^2 = \sum_{i=1}^{k} \chi_i^{\,2} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}. \qquad (1.19)$$

По теореме Пирсона величина χ^2 должна быть распределена (при условии $\min_i \{np_i\} \geq 5$) по закону (1.12) χ^2 (рис.1.9–10) с $(k – L – 1)$–степенями свободы, где L – число неизвестных параметров распределения, оцениваемых по выборке, а сам критерий проверки гипотезы о соответствии эмпирического распределения теоретическому закону носит название критерия Пирсона.

При заданном уровне значимости α гипотезу о распределении X по закону $F(x)$ отвергают, если $\chi^2 > \chi^2_{кр}$ и нет оснований отвергать, если $\chi^2 < \chi^2_{кр}$, где $\chi^2_{кр}$ определяется с помощью калькулятора распределения вероятности (см. рис.1.11) по закону χ^2 с $(k – L – 1)$–степенями свободы так, чтобы $P(\chi^2 > \chi^2_{кр}) = \alpha$.

С учетом чувствительности критерия Пирсона к разбиению выборки на k интервалов можно $\alpha(k)$ вычислять, например, по интерполяционной формуле Лагранжа для N целочисленных точек k_j, ближайших к значению k, *равному* $1+4lg(n)$.

$$\alpha(k) = \sum_{j=1}^{N} \alpha_j \frac{(k-k_0)...(k-k_{j-1})(k-k_{j+1})...(k-k_N)}{(k_j-k_0)...(k_j-k_{j-1})(k_j-k_{j+1})...(k_j-k_N)}, \qquad (1.20)$$

где $\alpha_j = \alpha(k_j)$ вычисляется системой STATISTICA (см. п. 2.2.1).

Наряду с критерием Пирсона, основанным на сравнении эмпирических и теоретических частот, применяется также критерий Колмогорова–Смирнова, основанный на сравнении накопленных частот. В слу-

чае критерия Колмогорова–Смирнова уровень значимости $\alpha_{K\text{-}S}$ рассчитывался приближенно (для $0{,}01 < \alpha < 0{,}2$ и $n > 10$) по формуле (Большев и др., 1983).

$$\alpha_{K-S} \approx 2\exp\left[\sqrt{\frac{1}{2}+\left(1+\frac{9}{2}n\right)^2 - 18n^2\left(d+\frac{1}{6n}\right)^2} - \frac{9}{2}n - 1\right]. \qquad (1.21)$$

Здесь D–статистическое значение критерия Колмогорова–Смирнова (Поллард, 1982), вычисляемое по формуле $d = \max d_j = \max\left|F_j^* - F_j\right|$, где $F_j^* = \sum_{i=1}^{j}\frac{n_i}{n}$ – выборочная функция распределения (накопленные частоты), вычисленная с учетом найденных выше частот n_i, а $F_j = \sum_{i=1}^{j} p_i$ - теоретическая функция распределения, вычисленная с учетом найденных выше p_i (Поллард, 1982).

В качестве критерия соответствия эмпирического распределения нормальному теоретическому используют также отношения коэффициентов асимметрии A и эксцесса E к их стандартным отклонениям $\sigma_{\tilde{A}}$ и $\sigma_{\tilde{E}}$, соответственно: $\tilde{t}_1 = \frac{A}{\sigma_A}$, $\tilde{t}_2 = \frac{E}{\sigma_E}$ (1.22)

Если эти отношения по абсолютной величине превышают 3, то гипотеза о нормальном распределении отвергается.

Рассмотрим гипотезы о параметрах нормального или логнормального распределения. Пусть имеются две серии опытов, регистрирующие значения некоторой случайной величины и определяющие две выборки объемом n_X и n_Y.

Рассмотрим ***сравнение двух дисперсий.***

Рассмотрим тестирование гипотезы H_0 о равенстве дисперсий $D_X = D_Y$ при неизвестных математических ожиданиях. Пусть даны две случайные величины X и Y, распределенные по нормальному закону. По данным выборок объёмом n_X и n_Y соответственно подсчитаны исправленные выборочные дисперсии s_x^2 и s_y^2. Требуется при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что $D_X = D_Y$. Такая задача возникает при сравнении точности двух приборов, при сравнении различных методов измерений. Обычно выборочные дисперсии оказываются различными. Возникает вопрос: существенно или нет они различаются? Если различие незначимо, то имеет место нулевая гипотеза, следовательно, методы имеют одинаковую точность, а

различие эмпирических дисперсий объясняется случайными причинами, в частности, случайным отбором объектов выборки.

По данным выборок объёмом n_X и n_Y вычисляют $F_{набл}$, как отношение большей дисперсии к меньшей.

$$F_{набл} = \frac{s_Б^2}{s_M^2} \ . \qquad (1.23)$$

Критическая область строится в зависимости от конкурирующей гипотезы $\boldsymbol{H_1}$ следующим образом. С помощью калькулятора распределения вероятности (рис. 1.14) по закону распределения Фишера по заданному уровню значимости α и вычисленным степеням свободы k_1 и k_2 находят $F_{kp}(\alpha, k_1, k_2)$ для $H_1: D_X > D_Y$ или $F_{kp}(\alpha/2, k_1, k_2)$ для $H_1: D_X \neq D_Y$. Если $F_{набл} > F_{kp}$, то H_0 отвергают, а при $F_{набл} < F_{kp}$ нет оснований отвергать H_0.

Величина F удовлетворяет распределению (1.14) Фишера (рис. 1.13–15) со степенями свободы k_1, определенной разностью объема выборки с большей дисперсией и единицы, и k_2, определенной разностью объема выборки с меньшей дисперсией и единицы.

Рассмотрим ***сравнение математических ожиданий.***

Для проверки подобия выборок (соответствия их распределению одной и той же случайной величины) рассмотрим вопрос о значимости расхождения между выборочными значениями математических ожиданий $\overline{x}$ и $\overline{y}$: выдвинем в качестве H_0 равенство математических ожиданий $m_X = m_Y$. Тестирование такой гипотезы основано на нормальном (1.10) распределении (рис. 1.4–5) в случае большого объема выборок ($n > 30$), когда дисперсии считаются известными, и на распределении (1.13) Стьюдента (рис. 1.12) в случае малых выборок ($n < 30$), когда дисперсии считаются неизвестными.

Рассмотрим первый случай. Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : m_X = m_Y$ о равенстве математических ожиданий двух больших нормальных выборок *с известными дисперсиями* D_X и D_Y, надо вычислить наблюдаемое значение критерия

$$Z_{набл} = \frac{\overline{x} - \overline{y}}{\sqrt{D_X / n_X + D_Y / n_Y}} . \qquad (1.24)$$

Далее следует построить критическую область в зависимости от конкурирующей гипотезы следующим образом:

При конкурирующей гипотезе $H_1: m_X \neq m_Y$ (двусторонняя критическая область) или $H_1: m_X > m_Y$ ($m_X < m_Y$) (односторонняя критическая

область) с помощью калькулятора распределения вероятности (см. рис.1.6) по нормальному закону найти критическую точку z_{kp} .

Если $|\ Z_{набл}\ | < z_{kp}$, то нет оснований отвергать нулевую гипотезу;

если $|\ Z_{набл}\ | > z_{kp}$, то нулевую гипотезу отвергают.

Рассмотрим второй случай. Пусть имеются две выборки объемов n_X и n_Y, на основании которых подсчитаны выборочными значениями математических ожиданий $\overline{x}$ и $\overline{y}$ и исправленные выборочные дисперсии s_x^2 и s_y^2. Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 ($m_X = m_Y$) о равенстве математических ожиданий двух малых нормальных выборок ***с неизвестными дисперсиями*** D_X и D_Y , надо предварительно проверить гипотезу о равенстве дисперсий (1.23) по подсчитанным исправленным выборочным дисперсиям s_x^2 и s_y^2. Если не будет оснований отвергать гипотезу о равенстве дисперсий, то есть дисперсии хотя и неизвестны, но предполагаются одинаковыми, то надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{\overline{x} - \overline{y}}{\sqrt{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}} \sqrt{\frac{n_X n_Y (n_X + n_Y - 2)}{n_X + n_Y}} . \qquad (1.25)$$

Затем построить критическую область в зависимости от конкурирующей гипотезы следующим образом.

При конкурирующей гипотезе H_1: $m_X \neq m_Y$ (двусторонняя критическая область) или H_1: $m_X > m_Y$ ($m_X < m_Y$) (односторонняя критическая область) с помощью калькулятора распределения вероятности (1.13) Стьюдента (см. рис.1.1) по заданному уровню значимости и числу степеней свободы $k = n_X + n_Y - 2$ найти критическую точку t_{kp}.

Если $|\ T_{набл}\ | < t_{kp}$, то нет оснований отвергать нулевую гипотезу;

если $|\ T_{набл}\ | > t_{kp}$, то нулевую гипотезу отвергают.

Вернемся ко второму случаю и рассмотрим далее второй вариант, когда гипотеза о равенстве дисперсий (1.23) отвергается. Пусть имеются две выборки объемов $\boldsymbol{n_X}$ и $\boldsymbol{n_Y}$, на основании которых подсчитаны выборочные значения математических ожиданий $\overline{x}$ и $\overline{y}$ и исправленные выборочные дисперсии s_x^2 и s_y^2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 : $m_X = m_Y$ о равенстве математических ожиданий двух малых нормальных выборок ***с неизвестными дисперсиями*** D_X и D_Y, надо предварительно проверить гипотезу о равенстве дисперсий (1.23) по подсчитанным исправленным выбороч-

ным дисперсиям s_x^2 и s_y^2. Пусть гипотеза о равенстве дисперсий отвергается, то есть дисперсии хотя и неизвестны, но предполагаются разными. Тестирование такой гипотезы H_0: $m_X = m_Y$ основано на распределении (1.13) Стьюдента с числом степеней свободы k:

$$k = \frac{(s_x^2/n_x + s_y^2/n_y)^2}{\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{(s_y^2/n_y)^2}{n_y - 1}}.$$

В этом случае вычисляют наблюдаемое значение критерия по формуле

$$T_{набл} = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2/n_X + s_y^2/n_Y}}. \qquad (1.25)^*$$

Затем строят критическую область в зависимости от конкурирующей гипотезы следующим образом.

При конкурирующей гипотезе H_1: $m_X \neq m_Y$ (двусторонняя критическая область) или H_1: $m_X > m_Y$ ($m_X < m_Y$) (односторонняя критическая область) с помощью калькулятора распределения вероятности (1.13) Стьюдента (рис. 1.1) по заданному уровню значимости и числу степеней свободы k найти критическую точку t_{kp}.

Если $|\ T_{набл}\ | < t_{kp}$, то нет оснований отвергать нулевую гипотезу;

если $|\ T_{набл}\ | > t_{kp}$, то нулевую гипотезу отвергают.

В случае логнормальной (1.11) модели (рис. 1.7) рекомендуется использовать критерий Родионова (уровень значимости α_R). В начале при заданном уровне значимости α предварительно проверяется гипотеза о равенстве дисперсий по F–критерию Фишера (1.23). Если не будет оснований отвергать гипотезу о равенстве дисперсий, то далее надо рассчитать значение критерия Стьюдента

$$T = \frac{\overline{\ln x} - \overline{\ln y}}{\sqrt{(n_x - 1)s_{\ln x}^2 + (n_y - 1)s_{\ln y}^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}, \qquad (1.26)$$

а критическое значение $T_{кр} = T(\alpha, k)$ найти с помощью калькулятора распределения (1.13) вероятности Стьюдента (см. рис.1.1) по заданному уровню значимости α и числу степеней свободы $k = n_x + n_y - 2$. Если гипотеза о равенстве дисперсий отвергается, то тестирование основной гипотезы основано на нормальном (1.10) распределении $N(z, 0, 1)$ случайной величины Z :

$$Z = \frac{\left|\overline{\ln x} - \overline{\ln y}\right| + \frac{1}{2}\left|s^2_{\ln x} - s^2_{\ln y}\right|}{\sqrt{s^2_{\ln x}/n_x + s^2_{\ln y}/n_y + (s^4_{\ln x}/(n_x - 1) + s^4_{\ln y}/(n_y - 1))/2}}. \qquad (1.26)^*$$

Неопределенность с законом распределения приводит к ***непараметрическим критериям***, которые особенно полезны для малых выборок.

Рассмотрим в качестве примера *U*–критерий Манна–Уитни для проверки гипотезы H_0 об однородности двух выборок, представляющий непараметрическую альтернативу *t*–критерию Стьюдента для независимых выборок. *U*–критерий Манна–Уитни предполагает, что все значения по обеим выборкам случайных величин *X* и *Y* объемов *n* и *m*, соответственно, ранжируются, то есть записываются в один ряд в порядке возрастания. После этого каждый элемент выборок характеризуется рангом – порядковым номером каждого элемента выборок в общем ранжированном ряду из обеих выборок. Наблюдаемое значение критерия *U* вычисляется по формуле

$$U = W - \frac{1}{2}m(m+1) = \sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij},$$

где *W*–значение критерия Уилкоксона, численно равное сумме рангов элементов второй выборки (объема ***m***) в общем ранжированном ряду,

$$\delta_{ij} = \begin{cases} 1, & если \quad X_i < Y_j, \\ 0, & в \quad противном \quad случае. \end{cases}$$

Таким образом, критерий *U* считает общее число тех случаев, в которых элементы второй выборки превосходят элементы первой выборки.

Распределение случайной величины *U* асимптотически нормально с параметрами $M[U] = nm/2$ и $D[U] = nm(n+m+1)/12$, чем и пользуются на практике, если $min\{n,m\} > 25$, для определения критического значения $U_{кр}(\alpha, n, m)$, соответствующего заданному уровню значимости α. Для случаев, когда *n* и *m* < 25, пользуются специальными таблицами (Каждан и др., 1990; Большев и др., 1983).

Проверка гипотезы о равенстве средних, определенных по двум выборкам объемов n_1 и n_2, с помощью **X–** критерия Ван–дер–Вардена начинается с того, что все значения по обеим выборкам ранжируются, то есть записываются в один ряд в порядке возрастания. X–критерий представляет собой величину

$$X = \sum_{i=1}^{n_2} \psi\left(\frac{i}{n_1 + n_2 + 1}\right),$$

где i – порядковый номер каждого значения второй выборки в общем ряду; ψ–функция, обратная функции нормального распределения, вычисляется с помощью калькулятора распределения вероятности по нормальному закону (рис. 1.6).

Вычисленное значение критерия X сравнивается с $X_{кр}$, определенным по специальным таблицам для заданного уровня значимости и объемов выборок (Каждан и др., 1990; Большев и др., 1983). Если $/X/ > X_{кр}$, то гипотеза о равенстве выборочных средних отвергается.

При этом следует учитывать особенности применения непараметрических критериев, например, ранговый X–критерий Ван–дер–Вардена (Каждан и др., 1990) рекомендуется применять, если предполагается, что наблюдения близко следуют нормальному закону (Вероятность …, 1999). Статистическим U–критерием Манна–Уитни (Боровиков В.П., 2003) для проверки гипотезы об однородности двух выборок X и Y объемов n_x и n_y следует пользоваться на практике, если только $\min\{n_x, n_y\} > 25$ (Вероятность …, 1999). Критерии серий Вальда–Вольфовица предполагает, что рассматриваемые переменные являются непрерывными и измерены в порядковой шкале (Боровиков В.П., 2003). Заметим, что двухвыборочный критерий Колмогорова–Смирнова (уровень значимости α_{2K-S}), основанный на сравнении эмпирических функций распределения двух выборок и проверяющий гипотезу однородности двух выборок (Боровиков В.П., 2003), является чувствительным как к различию в положении двух выборок, так и к различию общих форм распределений двух выборок (в частности, различия в рассеянии, асимметрии и т. д.).

Рассмотрим ***гипотезу о значимости выборочного коэффициента корреляции.***

Пусть дана нормально распределенная система дискретных случайных величин (X, Y) – совокупность n пар наблюдаемых значений $\{x_i, y_j\}$, характеризуемая, в частности, выборочным коэффициентом корреляции r Пирсона (п. 1.2.2), который оказался отличным от нуля. При этом возникает необходимость при заданном уровне значимости проверить нулевую гипотезу H_0: $r = 0$ при альтернативной H_1: $r \neq 0$ (двусторонняя критическая область). Если нулевая гипотеза отвергается, то выборочный коэффициент корреляции значимо отличается от нуля, а X и Y коррелированны. В качестве критерия проверки нулевой гипотезы H_0: $r = 0$ принимается случайная величина

$$\hat{T}_{набл} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad (1.27)$$

а критическое значение $T_{кр} = T(\alpha, k)$ находится с помощью калькулятора распределения (1.13) вероятности Стьюдента (рис. 1.1) по заданному уровню значимости α и числу степеней свободы $k = n - 2$.

Гипотеза H_0: $r = 0$ отвергается, т.е. выборочный коэффициент корреляции значимо отличается от ноля или иными словами X и Y коррелированны, если $\left|\hat{T}_{набл}\right| > T_{кр}$.

Если выборки малы или распределения существенно отличаются от нормального закона, то для проверки гипотезы о наличии корреляционной связи можно использовать непараметрический аналог коэффициента корреляции r Пирсона – ранговый коэффициент корреляции R Спирмена, вычисляемый аналогично r заменой наблюдаемых значений случайных величин их рангами (порядковыми номерами наблюдаемых значений в объединенной выборке, записанной в порядке возрастания). Значимость коэффициента корреляции R Спирмена проверяется аналогично значимости коэффициента корреляции r Пирсона.

1.2.4. Особенности применения статистического анализа эколого-геохимической информации в случае малых выборок

Эколого-геохимическая оценка состояния окружающей среды часто проводится с использованием небольшого объема выборки. Основным фактором в данном случае являются дорогостоящие методы анализа. Рассмотрим возможность применения методов статистической обработки при небольшом объеме выборок для сопоставления результатов исследований химического состава солевых образований из посуды населенных пунктов Томской и Челябинской областей.

При статистическом моделировании предполагается, что выборочная совокупность удовлетворяет требованиям массовости (объем выборки $n > 30$), однородности (измерения выполнены одинаковым способом), случайности (непредсказуемость результата единичного выборочного измерения, объективность отбора проб) и независимости (независимость результата каждого измерения от времени и места измерения). В ходе выполнения эколого-геохимических исследований возникают ситуации, когда требования математической статистики не могут быть приняты безоговорочно. Так, например, в силу дороговизны метода анализа приходится мириться с нарушением первого требования, т. е. использовать малые выборки. В этом случае применение статистиче-

ских методов должно базироваться на всестороннем анализе характера решаемой задачи, выборе наиболее эффективных статистических методов обработки измерений, методов статистических оценок и статистических критериев, менее чувствительных к объему выборки или учитывающих особенность малого объема выборки.

Статистический анализ эколого-геохимической информации проводят поэтапно:

I. ***Проверка гипотезы о законе распределения,*** применяя совокупность всесторонних способов:

1. Использование опыта геохимической практики (Каждан и др., 1990). Так, например, элементы с высокой концентрацией распределены по нормальному закону, а элементы с низкой концентрацией распределены по логарифмически нормальному закону.

2. Графический способ придает выборке наглядную форму, позволяющую выдвигать и проверять гипотезы о законе распределения. К сожалению, малый объем выборки не позволяет использовать гистограммы для проверки гипотезы о соответствии выборочных данных теоретическому закону. Особенностью графического способа в случае малого объема выборки является не построение гистограммы, а сравнение выборочных плотностей частот f_i^*, вычисленных по частотам n_i делением их на n и на длину i – го интервала, с теоретической кривой плотности распределения. В геохимической практике большое значение имеет нормальный закон распределения. Таково распределение $N(X,\mu,\sigma)$ химических элементов с высокой концентрацией X, где μ и σ – математическое ожидание и среднее квадратичное отклонение случайной величины X. Для химических элементов с низкой концентрацией X следует проверить гипотезу о распределении случайной величины X по логарифмически нормальному закону, т. е. гипотезу о распределение случайной величины lnX по нормальному закону $N(lnX, \mu_L, \sigma_L)$, где μ_L и σ_L – математическое ожидание и среднее квадратичное отклонение случайной величины lnX. Вначале по выборочным данным вычисляют точечные несмещенные оценки математического ожидания и стандартного отклонения случайной величины (a и s – для μ и σ или a_L и s_L – для μ_L и σ_L), затем рассматривают интервал $(a-3s\ ,\ a+3s)$ для случайной величины X или $(a_L-3s_L\ ,\ a_L+3s_L)$ для случайной величины lnX, в котором находится абсолютное большинство выборочных значений (≈ 99,73 %) нормально распределенной случайной величины. Данный интервал разбивают на k неравномерных интервалов, где число k определяется с учетом эмпирической формулы $k = 1+ 4lgn$ (Шестаков, 1988). Затем производят последовательную парную группировку элементов выборки по

принципу наименьшего расстояния, когда два соседних ближайших элемента выборки объединяют в группу, усредняя их значения для определения центра группы и т. д., пока не останется *k* групп. Внутренними границами интервалов выбирают значения средних арифметических центров соседних групп. Проиллюстрируем такой подход на примере выборки (А) содержания *X* химического элемента *Sc* в солевых отложениях населенного пункта Аргаяш Челябинской области (табл. 1.11).

Таблица 1.11

Содержание X химического элемента Sc в солевых отложениях из посуды с. Аргаяш Челябинской области

Sc,	Номер пробы						
мг/кг	1	2	3	4	5	6	7
X	0,26	0,27	0,29	0,38	0,38	0,52	1,2
lnX	1,347	1,309	1,238	0,968	0,968	0,654	0,182

Вычисляя точечные несмещенные оценки по формулам (1.16–17)

$$a = \frac{1}{n}\sum_{i=1}^{n} x_i , \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - a)^2$$

при $n = 7$, получим $a \approx 0{,}47$ и s $\approx 0{,}33$. В данном случае $k = 1 + 4\ lg7 \approx 4{,}38$, т. е. $4 < k < 5$. В связи с этим рассмотрим два варианта дробления интервала выборочных значений $(a - 3s\ ,\ a + 3s) = (-0{,}53;\ 1{,}47)$ на $k = 4$ и 5 интервалов.

Построенные группированные распределения позволяют рассчитать выборочные плотности частот f_i^* по соответствующим частотам n_i, деля n_i на n и на длину i –го интервала.

Сравнение взаимного расположение эмпирических плотностей частот в координатах (f_i^*, x_i) с теоретическими кривыми плотностей распределения по нормальному (1.10) $f_N(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2s^2}}$ и логнормальному (1.11) $f_{LN}(x) = \frac{1}{s_L\sqrt{2\pi}\,x} e^{-\frac{(\ln x - a_L)^2}{2s_L^2}}$ закону отражено на рис. 1.18.

Таким образом, согласно графическому способу, можно предположить, что данная выборка (А) распределена скорее по логнормальному закону, чем по нормальному.

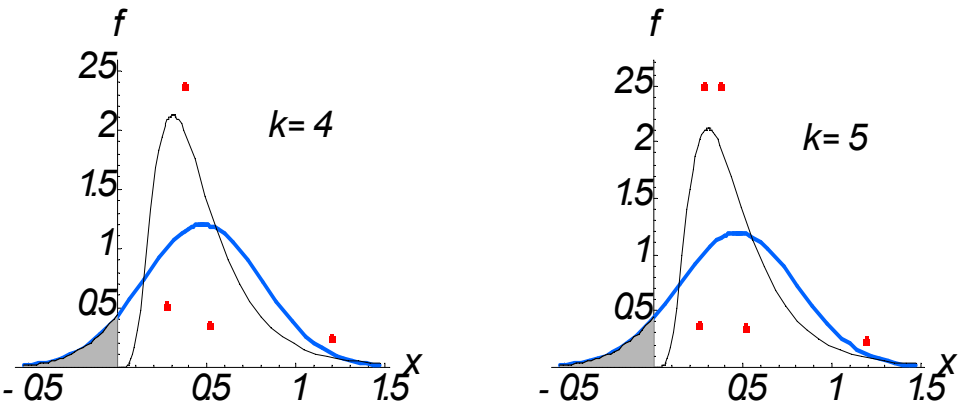


Рис. 1.18. Взаимное расположение точечных эмпирических плотностей частот для случаев k = 4 и 5 в координатах (f_i^*, x_i) с теоретическими кривыми плотностей распределения по нормальному $f_N(x)$ (толстая кривая) и логнормальному $f_{LN}(x)$ (тонкая кривая) закону

Гипотезу о распределении выборки (А) по нормальному закону можно отвергнуть на том основании, что в этом случае соответствующий нормальный закон N (x; 0,47; 0,33) приводит к возможности принять случайной величиной X отрицательное значение с вероятностью 0.08, что превышает принятый здесь уровень значимости α = 0,05. На рис.1 затемненной заливкой выделены области отрицательных значений выборки (А).

Более строгим способом является аналитические критерии, рассматриваемые ниже.

3. Аналитические способы сравнения числовых характеристик (Справочник …, 1987). В качестве критерия соответствия эмпирического распределения нормальному теоретическому используют отношения выборочных показателей асимметрии A и эксцесса E за вычетом их смещений m_A и m_E , соответственно, к их стандартным отклонениям σ_A и σ_E (1.22): $t_1 = \dfrac{A - m_A}{\sigma_A}$ и $t_2 = \dfrac{E - m_E}{\sigma_E}$. Если эти отношения по абсолютной величине превышают 3, то гипотеза о нормальном распределении отвергается. Для нормального распределения вероятность того, что выборочное значение этих отношений будет отличаться от математического ожидания больше, чем на 3 стандартных отклонения, очень мала (≤0,0027). Обычно ограничиваются асимптотическими оценками стандартных отклонений показателей асимметрии и эксцесса (Каждан и др., 1990; Родионов, 1981; Шарапов, 1965 и Большев и др., 1983):

$$\sigma_{A_\infty} \approx \sqrt{6/n}, \quad \sigma_{E_\infty} \approx \sqrt{24/n}, \quad m_{A_\infty} = 0 \quad , \quad m_{E_\infty} = 0. \qquad (1.28)$$

$$\overline{x}=\frac{1}{n}\sum_i x_i,\quad \mu_k=\frac{1}{n}\sum_i\left(x_i-\overline{x}\right)^k,\quad s^2=\frac{n}{n-1}\mu_2,\quad A_\infty=\frac{\mu_3}{s^3},\quad E_\infty=\frac{\mu_4}{s^4}-3$$

Особенностью применения этого критерия в случае малых ***n*** является использование более точных оценок стандартных отклонений показателей асимметрии и эксцесса (Справочник …, 1987; Большев и др., 1983; Пустыльник, 1968 и Крамер, 1975). Для точечных оценок показателей асимметрии *A* и эксцесса *E* оценки стандартных отклонений показателей асимметрии и эксцесса имеют вид:

$$\sigma_A=\sqrt{\frac{6(n-2)}{(n+1)(n+3)}},\quad \sigma_E=\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}},\quad m_A=0\quad,\quad m_E=-\frac{6}{n+1},$$

$$\overline{x}=\frac{1}{n}\sum_i x_i,\quad \mu_k=\frac{1}{n}\sum_i\left(x_i-\overline{x}\right)^k,\quad s^2=\frac{n}{n-1}\mu_2,\quad A=\frac{\mu_3}{s^3},\quad E=\frac{\mu_4}{s^4}-3. \quad (1.29)$$

При этом сама оценка *E* является смещенной. В работе (Миллер, 1965) наряду со смещенными оценками приведены также и несмещенные $\tilde{\boldsymbol{A}},\tilde{\boldsymbol{E}}$ **:**

$$\tilde{A}=\frac{\sqrt{n(n-1)}}{n-2}\frac{\mu_3}{{\mu_2}^{1.5}},\quad \tilde{E}=\frac{(n-1)(n+1)}{(n-2)(n-3)}\left(\frac{\mu_4}{{\mu_2}^2}-3+\frac{6}{n+1}\right),\quad m_{\tilde{A}}=0,\quad m_{\tilde{E}}=0\quad,$$

$$\sigma_{\tilde{A}}=\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}},\quad \sigma_{\tilde{E}}=\sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}. \quad (1.30)$$

Из вида смещенных и несмещенных оценок показателей асимметрии и эксцесса следуют формулы, связывающие несмещенные оценки $\tilde{\boldsymbol{A}},\tilde{\boldsymbol{E}}$ со смещенными A и E :

$$\tilde{A}=\frac{n^2}{(n-1)(n-2)}A,\quad \tilde{E}=\frac{n^2(n+1)}{(n-1)(n-2)(n-3)}\left(E+3-3\frac{(n-1)^3}{(n-1)n^2}\right).$$

В случае малого объема выборки *n* смещенные и несмещенные оценки могут отличаться в несколько раз. Так, например, в случае выборки (А) вычисления по формулам (2) дают для смещенных оценок $A \approx 1{,}39$ и $E \approx 0{,}30$, а по формулам (3) для несмещенных оценок $\tilde{\boldsymbol{A}} \approx 2{,}27$ и $\tilde{\boldsymbol{E}} \approx 5{,}39$. Сравнение смещенных и несмещенных оценок показателей асимметрии, эксцесса и критерия соответствия эмпирического распределения нормальному теоретическому $N(x; 0{,}47; 0{,}33)$ отражено в табл. 1.12.

Как видно из табл. 1.12, различие в критериях t_1 и t_2, вычисляемых по формулам (1.29) для смещенных оценок и по формулам (1.30) для несмещенных оценок, является существенным настолько, что приводит к разным выводам относительно соответствия эмпирического распреде-

ления нормальному теоретическому (для несмещенных оценок $t_2 = 3{,}396 > 3$, т.е. распределение выборки (А) существенно отличаются от нормального закона).

Таблица 1.12

Проверка гипотезы о нормальном законе распределения Sc по смещенным и несмещенным оценкам показателей асимметрии A и эксцесса E

	Расчетные формулы						
	(1.28)	(1.29)	(1.30)	(1.28)	(1.29)	(1.30)	
A	1,391	1,391	2,272	0,303	0,303	5,389	E
σ_A	0,926	0,612	0,793	1,852	0,661	1,587	σ_E
t_1	1,50	2,27	2,86	0,16	1,59	3,40	t_2

4. Аналитические способы сравнения законов распределения (Каждан и др., 1990; Справочник …, 1987; Родионов, 1981; Шестаков, 1988; Большев и др., 1983; Пустыльник, 1968 и Поллард, 1982). Применение наиболее распространенного критерия проверки гипотезы о соответствии эмпирического распределения теоретическому закону – критерию Пирсона (1.19) предполагает разделение выборочных данных на k интервалов. Число k определяется с учетом эмпирической формулы $k = 1 + 4\ lgn$ и требования критерия Пирсона $k > 3$, при этом предполагается, что в каждом интервале содержится не менее трех значений случайной величины (Шестаков, 1988). Нарушение последнего требования в случае малого объема выборки n делает критерий Пирсона чувствительным к способам группировки выборки на k интервалов и порождают так называемые ошибки I (отвергается правильная гипотеза) и II (не отвергается неправильная гипотеза) родов.

В отличие от критерия Пирсона, основанного на сравнении эмпирических и теоретических частот, критерий Колмогорова–Смирнова (1.21) основан на сравнении накопленных частот.

В случае выборки (А) для варианта $k = 4$ промежуточные вычисления приведены в табл. 1.13, где, как принято в случае сравнения с нормальным законом распределения, границы крайних интервалов расширены до бесконечности (∞).

В результате имеем $\chi^2 \approx 2{,}47 < 3{,}84 \approx \chi^2_{0,05;1}$, $D \approx 0{,}24 < 0{,}48 \approx D_{0,05;7}$. Таким образом, согласно критериям Пирсона и Колмогорова–Смирнова нет оснований отвергать гипотезу о том, что данный вариант выборки (А) распределен по нормальному закону. Аналогичным образом можно проверить гипотезу о распределении данного варианта вы-

борки (А) по логнормальному закону, т.е. распределение логарифмов выборки (А) по нормальному закону.

Таблица 1.13

Проверка гипотезы о нормальном законе распределения Sc по критериям Пирсона и Колмогорова–Смирнова

k	№ группы	Центры групп	$c_i \div b_i$	n_i	$\boldsymbol{p_i}$	χ_i^2	$\boldsymbol{F_i}$	$\boldsymbol{D_i}$
4	1	0,2775	$-\infty \div 0{,}33$	3	0,33	0,20	0,33	0,10
	2	0,38	$0{,}33 \div 0{,}45$	2	0,14	1,06	0,47	0,24
	3	0,52	$0{,}45 \div 0{,}86$	1	0,41	1,19	0,88	0,02
	4	1,2	$0{,}86 \div \infty$	1	0,12	0,02	1,00	0,00

Для исследования характера ошибок при проверке гипотезы о законе распределения (*N* – нормальный, *LN* – логнормальный) в случае малого объема выборки *n* применим критерии (1.19) и (1.21) к разным вариантам группировки выборки (А), наряду с только что рассмотренным вариантом группировки выборки (А), т.е. вариантом частот (3211), также к вариантам частот (2311) и (1411) при $k = 4$, а при $k = 5$ к варианту (21211), и к аналогичным ему (12211) и (11311). Результаты исследования приведены в табл. 1.14.

Таблица 1.14

Зависимость критериев Пирсона и Колмогорова–Смирнова от способов группировки выборки (А)

k	Варианты групп	χ^2	χ^2	d	d
		N	*LN*	*N*	*LN*
4	(3211)	2,47	1,12	0,24	0,15
	(2311)	5,3	2,1	0,26	0,18
	(1411)	10,5	4,6	0,27	0,21
5	(21211)	3,2	1,1	0,24	0,15
	(12211)	8,3	2,8	0,24	0,15
	(11311)	7,8	2,8	0,26	0,18

Как следует из табл. 1.14, с учетом критических значений $\boldsymbol{D_{0,05;7}} \approx 0{,}48$ и $\boldsymbol{\lambda_{0,05}} \approx 1{,}36$, согласно критерию Колмогорова–Смирнова независимо от способов группировки нет оснований отвергать обе гипотезы (о

соответствии выборки (А) нормальному *N* и логнормальному *LN* законам), что соответствует ошибке II рода.

При использовании χ^2 - критерия Пирсона для $k = 4$ с учетом критического значения $\chi^2_{0,05;1} \approx 3{,}84$ в случае варианта (3211) обе гипотезы (о соответствии выборки (А) нормальному *N* и логнормальному *LN* закону) не отвергались (ошибка II рода), в случае варианта (2311) отвергалась гипотеза о нормальном *N* законе и не было оснований отвергать гипотезу о логнормальном *LN* законе (правильное решение), в случае варианта (1411) отвергались обе гипотезы (ошибка I рода).

При использовании χ^2 –критерия для $k = 5$ с учетом критического значения $\chi^2_{0,05;2} \approx 6{,}0$ в случае варианта (21211) обе гипотезы не отвергались (ошибка II рода). В случае вариантов (12211) и (11311) отвергалась гипотеза о нормальном *N* законе и не было оснований отвергать гипотезу о логнормальном *LN* законе (правильное решение). Таким образом, учитывая чувствительность критериев Пирсона и Колмогорова–Смирнова к способам группировки в случае малого объема выборки *n,* можно ограничиться в случае малого объема выборки *n* более скромным выводом о том, что в случае выборки (А) логнормальный закон предпочтительней, так как независимо от способа группировки все статистические значения критериев для *N* меньше соответствующих значений критериев для *LN,* а ведь именно значение статистического критерия характеризует меру отличия выборочного и теоретического законов распределения случайной величины.

II. ***Проверка гипотезы о равенстве средних*** на основе выбранного закона распределения:

1. Использование параметрических критериев (Каждан, 1990; Родионов, 1983). В случае нормальной модели используют критерии Фишера (1.23) и Стьюдента (1.24–25).В случае логнормальной модели рекомендуется использовать критерий Родионова (1.26).

2. **Использование непараметрических критериев** (Каждан и др., 1990; Большев и др., 1983). Неопределенность с законом распределения приводит к ***непараметрическим критериям***, которые особенно полезны для малых выборок.

Непараметрическими альтернативами критериям Стьюдента и Родионова являются, например, критерии серий Вальда–Вольфовица, *U*–критерием Манна–Уитни, двухвыборочный критерий Колмогорова–Смирнова для независиимых выборок или критерий знаков, критерий Вилкоксона для зависимых выборок (Боровиков В.П., 2003).

При этом следует учитывать особенности применения непараметрических критериев (п. 1.2.3).

Применим статистический анализ для установления значимости отличия средних по урану значений по четырем населенным пунктам Томской области (табл. 1.15).

Таблица 1.15

Содержание урана в солевых отложениях из посуды жителей населенных пунктов юга Томской области

Населенный пункт	№ пробы	U, мг/кг	Среднее, мг/кг
с. Новониколаевка	1	1,3	3,16
с. Новониколаевка	2	0,2	
с. Новониколаевка	3	0,2	
с. Новониколаевка	4	4,4	
с. Новониколаевка	5	9,7	
п. Комсомольск	1	1,6	0,48
п. Комсомольск	2	0,2	
п. Комсомольск	3	0,2	
п. Комсомольск	4	0,2	
п. Комсомольск	5	0,2	
с. Семёновка	1	0,2	5,70
с. Семёновка	2	3,3	
с. Семёновка	3	12,0	
с. Семёновка	4	12,0	
с. Семёновка	5	1,0	
с. Коломинские гривы	1	0,2	0,30
с. Коломинские гривы	2	0,2	
с. Коломинские гривы	3	0,2	
с. Коломинские гривы	4	0,2	
с. Коломинские гривы	5	0,7	

Рассмотрим, например, статистический анализ U по двум населенным пунктам юга Томской области: с. Новониколаевка (Н) и с. Семеновка (С). Гипотезу о распределении соответствующих выборок (Н) и (С) по нормальному закону можно отвергнуть только на том основании, что в этом случае соответствующие нормальные законы N (x_H; 3,16; 4,04) и N(x_C; 5,70; 5,86) приводят к возможности принять случайной величиной X отрицательное значение с вероятностями 0.22 и 0.17 соответственно. На рис. 1.19 затемненной заливкой выделены области отрицательных значений выборок (Н) и (С).

Учитывая принадлежность U к элементам с низкой концентрацией, проверим гипотезу о распределении U по логнормальному закону. Графический способ проверки гипотезы о распределении выборок (Н) и (С) по логнормальному закону проиллюстрирован на рис 1.20.

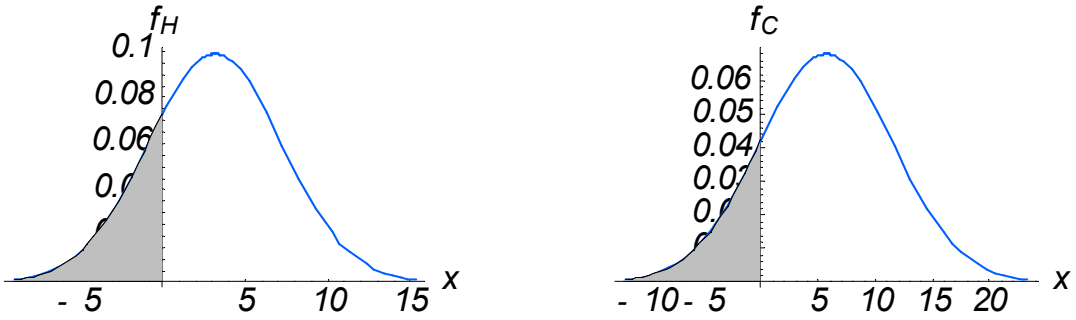


Рис. 1.19. Графики функций f(x) плотности нормального распределения $N(x_H; 3{,}16; 4{,}04)$ и $N(x_C; 5{,}70; 5{,}86)$, на фоне которых затемненной заливкой выделены области, в которых $P(X < 0) = 0{,}22$ и $0{,}17$

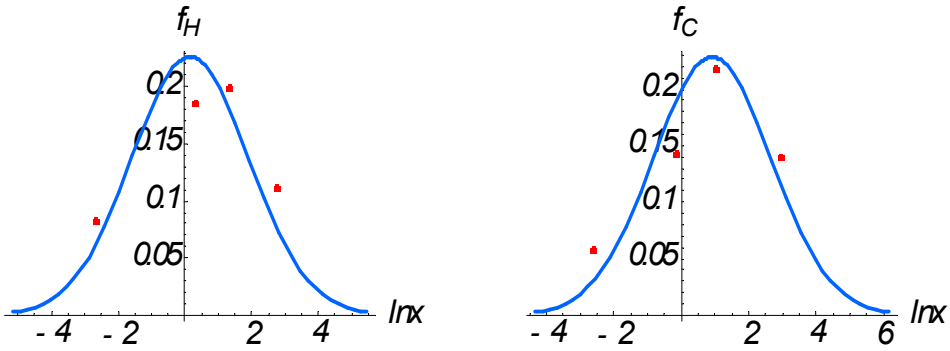


Рис. 1.20. Эмпирические плотности частот в координатах (f_i^*, lnx_i) на фоне теоретических кривых f(lnx) по нормальному закону N(lnx; 0,16; 1,77) для выборки (Н) и N(lnx; 0,91; 1,75) для выборки (С)

Применение критерия Пирсона для выборки (Н) и (С) дает χ^2_H = 0,085 и χ^2_C = 0,141 соответственно, что меньше $\chi^2_{\alpha,k-3}$ = 3,84 при α = *0,05* и k = *4*. Т. е. нет оснований отвергать гипотезу о соответствии выборок (Н) и (С) логнормальному распределению.

Критерии, основанные на сравнении числовых характеристик в случае выборок (Н) и (С), дают отношения выборочных показателей асимметрии и эксцесса к их стандартным отклонениям, равные 1,27 и 2,25 для (Н) и 0,43 и 1,71 для (С). Во всех случаях эти отношения меньше 3, что также подтверждает выше сформулированный вывод.

Таким образом, по совокупности приведенных критериев нет оснований отвергать гипотезу о соответствии выборочных данных (Н) и (С) логнормальному распределению для уровня значимости $\alpha = 0,05$.

Использование критерия Родионова в случае выборок (Н) и (С) дает по формуле (1.23) $F_{набл} = 1,02 < 6,4 = F(0,05; 4; 4) = F_{кр}$, т. е нет оснований отвергать гипотезу о равенстве дисперсий, и, согласно (1.26), $T = 0,68 < 1,86 = T(0,05; 8) = T_{кр}$, т. е. расхождения между выборочными значениями математических ожиданий m_H и m_C не являются значимыми.

В случае малых выборок (n < 10) точечную несмещенную оценку среднего квадратичного отклонения **s** можно заменить выборочным размахом ω (Шарапов, 1965):

$$s \approx \frac{\omega}{\beta(n)}, \quad \omega = (\ln x)_{\max} - (\ln x)_{\min}, \quad напр., \beta(5) \approx 2.326 .$$

При этом сам критерий для выборок равного объема принимает очень простой вид (Миллер, 1965):

$$\tau = \frac{a_1 - a_2}{\omega_1 + \omega_2} . \qquad (1.31)$$

В случае выборок (Н) и (С) критерий (1.31) дает $\tau_{набл} = 0,09 < 0,246 = \tau_{кр}$, т. е. нет оснований отвергать гипотезу о равенстве математических ожиданий m_H и m_C .

Аналогичным образом проведено сравнение других пар выборок табл. 1.7. Результаты сравнения выборки (Н) с выборками Км (с. Комсомольск) и Кл (с. Коломинские Гривы) приведены в табл. 1.16.

Таблица 1.16

Проверка гипотез о равенстве средних содержаний U в солевых отложениях из посуды населенных пунктов юга Томской области по критерию Родионова (сравнение m_H с m_C, $m_{Кл}$ и $m_{Км}$)

	a	*s*	$F_{набл} / F_{кр}$	$T / T_{кр}$	$\tau_{набл} / \tau_{кр}$
Н	0,16	1,77			
С	0,911	1,75	1,02 / 6.39	0,68 / 1,86	0,09 / 0,25
Кл	–1,91	0,87	4,12 / 6.39	2,35 / 1,86	0,36 / 0,25
Км	–1,75	1,24	2,03 / 6.39	1,98 / 1,86	0,29 / 0,25

Как следует из табл. 1.16, критерий проверки гипотезы о равенстве средних по формулам (1.26) и (1.31) приводит к выводу о значимом различии средних значений как по выборкам (Н) и (Км), так и по выборкам (Н) и (Кл).

Единообразие критерия проверки гипотезы о равенстве средних при одинаковых объемах рассмотренных выборок позволяет использовать формулу (1.26) в координатах (s,a) или формулу (1.31) в координатах (ω,a) для геометрической иллюстрации полученных результатов.

Приравнивая выражение T–критерия по формуле (1.26) к $T_{кр}$ или выражение τ–критерия по формуле (1.31) к $\tau_{кр}$, можно получить уравнения граничных линий (толстые линии на рис. 1.21 – ветви гиперболы в координатах (*a, s)* или прямые линии в координатах (*a, ω)*) критической области (затемненная область на рис. 1.21) в координатах (*a, s)* или в координатах (*a, ω)* по отношению к оценкам числовых характеристик выборки (Н). Точками указаны соответствующие пары оценок числовых характеристик прочих выборок, тонкими линиями соответствующие изолинии.

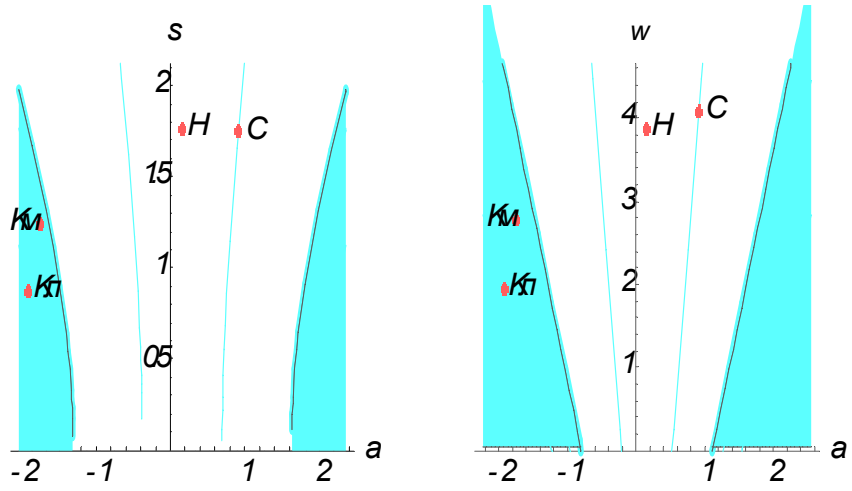


Рис.1.21. Графическая иллюстрация значимости различий выборок (С), (Км) и (Кл) по отношению к выборке (Н) в координатах (s, a) и (ω, a). Критическая для (Н) область затемнена.

Как видно из рис. 1.21, оценки числовых характеристик выборок (Км) и (Кл) попали в критическую область по отношению к (Н), а оценки (С) – нет. Причем, близость расположения точек (С), (Км) и (Кл) по отношению к граничной линии позволяет судить о запасе прочности значимости отличий этих выборок по отношению к выборке (Н).