САМАРСКИЙ
ГОСУДАРСТВЕННЫЙ
АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ
имени академика С.П.Королёва

В.М. Дуплякин

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЫБОРОЧНЫХ ДАННЫХ

САМАРА 2010

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
"САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С.П. КОРОЛЁВА"

В.М. Дуплякин

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЫБОРОЧНЫХ ДАННЫХ

Утверждено Редакционно-издательским советом университета в качестве учебного пособия

САМАРА

Издательство СГАУ

2010

УДК СГАУ: 519.2
ББК 22.171
Д 839

Рецензенты: канд. техн. наук, доц. Л.В. Коломиец,
канд. техн. наук, доц. Н.Я. Лищинский

Дуплякин В.М.

Д 839 Статистический анализ выборочных данных: учеб. пособие /
В.М.Дуплякин - Самара : Изд-во Самар. гос. аэрокосм. ун-та, 2010.
– 110 с.

ISBN 978-5-7883-0746-6

Пособие предназначается для изучения методов математической статистики применительно к решению наиболее распространённых задач обработки результатов выборочных наблюдений. Все рассматриваемые вопросы иллюстрируются примерами. Приведены необходимые для выполнения расчётов статистические таблицы. Пособие может использоваться не только с целью изучения статистических методов, но и как справочное руководство при курсовом и дипломном проектировании или при выполнении самостоятельных исследований студентами и аспирантами экономических, инженерных и других специальностей различных форм обучения.

Предназначено для студентов специальностей "Математические методы в экономике", "Менеджмент".

Разработано на кафедре математических методов в экономике Самарского аэрокосмического университета.

УДК СГАУ: 519.2
ББК 22.171

ISBN 978-5-7883-0746-6

© Самарский государственный
аэрокосмический университет, 2010

ВВЕДЕНИЕ

Статистические методы широко используются при решении разнообразных инженерных и экономических задач, при исследовании социальных вопросов и в менеджменте, в научных исследованиях в области механики, физики, химии, биологии, медицины, так как при изучении любого круга явлений, неизбежно наступает этап, когда требуется не только выявление основных закономерностей, но и анализ случайных отклонений от них, обусловленных погрешностями регистрации данных и нестабильностью условий наблюдения, а это возможно только на основе использования статистических методов.

Привлекательной стороной статистического подхода является развитие аппарата выборочного метода получения сведений об интересующих нас явлениях. Очевидная целесообразность использования выборок сравнительно небольшого числа элементов из всей совокупности данных элементов при оценке их свойств, независимо от физического содержания решаемых задач, является постоянным фактором, стимулирующим развитие статистики как раздела математики и постоянное расширение сферы её применения.

Статистический анализ выборочные данных представляет собой приложение математической статистики как раздела математики, предметом которого является разработка методов регистрации, описания и анализа статистических данных, получаемых в результате наблюдения массовых случайных явлений.

Основные задачи математической статистики:

1. Определение вероятностей событий.
2. Оценка числовых характеристик случайных величин.
3. Оценка параметров законов распределения.
4. Выявление законов распределения случайных величин.
5. Проверка статистических гипотез.
6. Выявление функциональных зависимостей между величинами в условиях использования ограниченных данных искажённых случайными отклонениями.

Настоящее пособие позволяет познакомиться с наиболее эффективными методами решения статистических задач, широко используемыми на практике. Пособие не заменяет учебник по курсу теории вероятностей и математической статистики, а является дополнением, которое даёт возможность освоить и самостоятельно решать различные задачи статистического анализа.

1. ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Теоретической основой математической статистики являются предельные теоремы теории вероятностей, которые охватывают различные формулировки закона больших чисел и центральной предельной теоремы.

Под законом больших чисел в теории вероятностей понимается ряд теорем, в каждой из которых для определённых условий устанавливается факт приближения средних характеристик к некоторым определённым постоянным при большом числе опытов.

Другая группа предельных теорем касается не приближения к средним значениям, а приближения к некоторым предельным законам распределения, все эти теоремы, так или иначе, представляют собой различные формулировки более общей как называемой центральной предельной теоремы, сформулированной и доказанной нашим знаменитым соотечественником, П.Л.Чебышевым*.

Предельные теоремы позволяют не только осуществлять научные прогнозы в области случайных явлений, но и оценить точность этих прогнозов.

Взаимосвязь различных интерпретаций предельных теорем теории вероятностей иллюстрируются схемой, приведенной на рисунке 1.1.

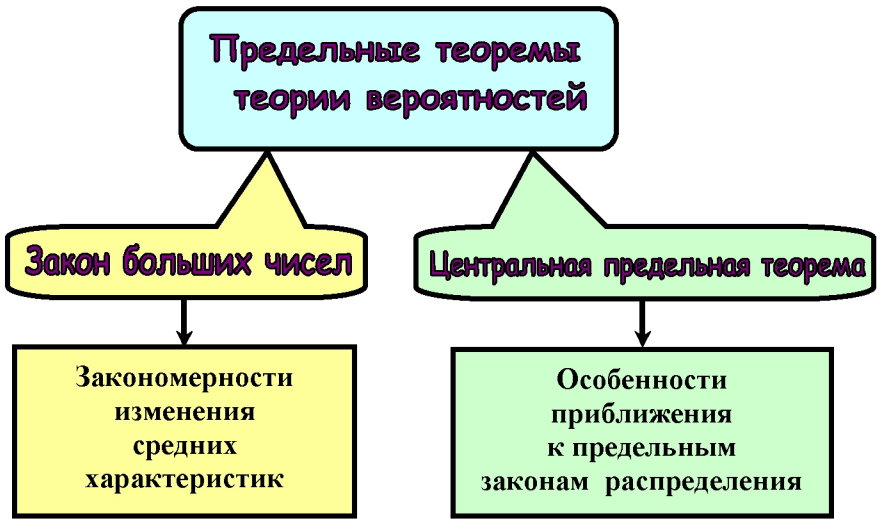

Рис. 1.1 – Взаимосвязь предельных теорем теории вероятностей

* Чебышёв Пафнутий Львович (1821-1894) – профессор Петербургского университета, считается одним из основоположников теории приближения функций. Значителен вклад Чебышева в теорию чисел и теорию вероятностей, а так же в механику. Член Петербургской, Берлинской и Болонской академий, Парижской Академии наук, член-корреспондент Лондонского Королевского общества, Шведской академии наук и др., всего 25 различных Академий и научных обществ. Чебышёв состоял почётным членом всех российских университетов.

Известный математик Шарль Эрмит заявил, что Чебышёв «является гордостью русской науки и одним из величайших математиков Европы», а профессор Стокгольмского университета Миттаг-Леффлер утверждал, что Чебышёв — гениальный математик и один из величайших аналистов всех времен.

1.1. Неравенство Чебышева

Одной из очень важных разновидностей "закона больших чисел" является неравенство П.Л.Чебышева, поскольку оно используется при доказательстве большинства предельных теорем в различных формулировках.

Допустим, что для случайной величины X известны её характеристики: m_x, D_x, тогда для любого положительного числа α выполняется неравенство Чебышева

$$P(|X - m_x| \geq \alpha) \leq \frac{D_x}{\alpha^2}. \qquad (1.1)$$

Неравенство Чебышева показывает, что для любого наперёд заданного положительного числа α, вероятность того, что любая случайная величина X отклонится от своего математического ожидания не меньше чем на α, ограничена сверху величиной $\frac{D_x}{\alpha^2}$.

Доказательство.

А) X – дискретная случайная величина.

Дискретная случайная величина задаётся своим статистическим рядом, например, в виде

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

Представим значения данной случайно величины отложенными на числовой оси, как это показано на рис.1.2.

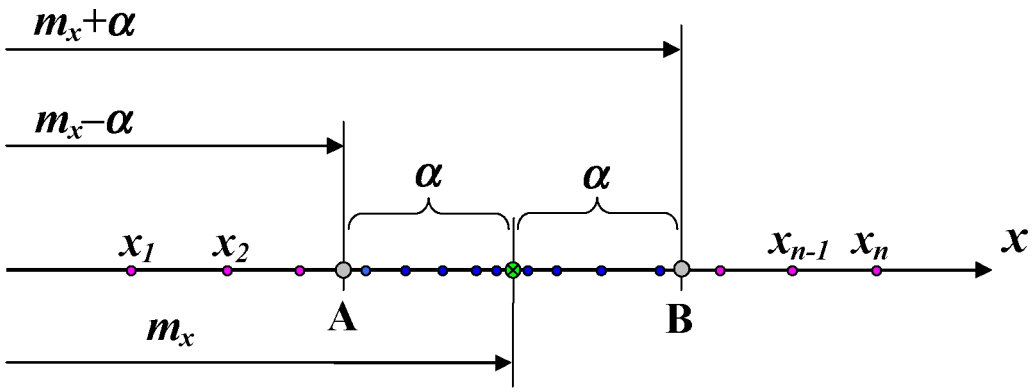

Рис. 1.2 – Распределение случайной величины на числовой оси

Очевидно, что левая часть неравенства Чебышева представляет собой вероятность того, что случайная величина X выходит за пределы отрезка **AB**, а именно

$$P(|X - m_x| \geq \alpha) = P\left(X \notin [A;B]\right). \tag{1.2}$$

Что бы найти эту вероятность, нужно просуммировать все вероятности для тех значений x_i, которые лежат вне отрезка **AB**

$$P(|X - m_x| \geq \alpha) = \sum_{|X - m_x| \geq \alpha} p_i \ . \tag{1.3}$$

Теперь обратимся к вычислению дисперсии

$$D_x = M\left[\left(X - m_x\right)^2\right] = \sum_{i=1}^{n}\left(x_i - m_x\right)^2 p_i = \sum_{i=1}^{n}\left|x_i - m_x\right|^2 p_i \ . \tag{1.4}$$

Т.к. все значения членов суммы неотрицательны, то эта сумма может только увеличиваться, если учитываются не все слагаемые, т.е.

$$D_x \geq \sum_{|x_i - m_x| \geq \alpha} \left|x_i - m_x\right|^2 p_i \ . \tag{1.5}$$

Заменим все слагаемые $\left|x_i - m_x\right|$ в последнем выражении на α. Очевидно, что от такой замены суммы может только уменьшиться, т.к. $\left|x_i - m_x\right| \geq \alpha$, поэтому предлагаемая замена усиливает предыдущее неравенство

$$D_x \geq \sum_{|x_i - m_x| \geq \alpha} \alpha^2 p_i = \alpha^2 \sum_{|x_i - m_x| \geq \alpha} p_i \ , \text{ а сумма представляет не что иное как } P\left(X \notin [A;B]\right). \tag{1.6}$$

$$\text{Следовательно:} \quad D_x \geq \alpha^2 P\left(|X - m_x| \geq \alpha\right). \tag{1.7}$$

Отсюда и следует неравенство Чебышева.

Б) $\boldsymbol{X}$ – непрерывная случайная величина

Доказательство аналогично предыдущему, но конечные суммы заменяются интегралами

$$P(|X - m_x| \geq \alpha) = \int_{|X - m_x| \geq \alpha} f(x)\,dx \ , \text{ где } f(x) - \text{плотность вероятности.} \tag{1.8}$$

Далее так же переходим к оценке величины дисперсии, используя поэтапное усиление неравенства

$$D_x = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x)\,dx = \int_{-\infty}^{+\infty} \left|x - m_x\right|^2 f(x)\,dx \geq \int_{|x - m_x| > \alpha} \left|x - m_x\right|^2 f(x)\,dx \ . \tag{1.9}$$

Поскольку в интересующем нас интервале интегрирования нижней границей модуля разности $|x-m_x|$ является заданная величина α, то можно выполнить дополнительное усиление предыдущего неравенства

$$D_x \geq \alpha^2 \int\limits_{|x-m_x|>\alpha} |x-m_x|^2 f(x)dx . \qquad (1.10)$$

Интеграл в данном неравенстве представляет собой вероятность выхода случайной величины X за пределы отрезка $AB = (m_x - \alpha;\ m_x + \alpha)$, поэтому

$$D_x \geq \alpha^2 P\left(|x-m_x| > \alpha\right). \qquad (1.11)$$

Что, собственно и требовалось доказать.

Пример. Оценить вероятность того, что случайная величина $\boldsymbol{X}$ с заданным математическим ожиданием m_x и средним квадратическим отклонением σ_x может отклониться от своего математического ожидание не менее чем на $3\sigma_x$.

Решение.

Воспользуемся неравенством Чебышева $P(|X-m_x| \geq \alpha) \leq \dfrac{D_x}{\alpha^2}$, положив $\alpha = 3\sigma_x$.

Подставляя, получим $P(|X-m_x| \geq 3\sigma_x) \leq \dfrac{D_x}{9(\sigma_x)^2} = \dfrac{1}{9} = 0,111(1)$.

Следует понимать, что это верхняя оценка искомой вероятности, которая справедлива для любого закона распределения, а на самом деле величина этой вероятности может быть ниже, например, для нормального закона распределения, данная задача приводит к так называемому правилу "трёх сигма", характеризующемуся следующей вероятностью

$$P(|X_{норм.распр.} - m_x| \geq 3\sigma_x) = 0,0028... \ .$$

1.2. Закон больших чисел (теорема П.Л.Чебышева)

Это теорема устанавливает связь между средним арифметическим наблюдаемых значений и их математическим ожиданием.

Предварительно рассмотрим вспомогательную задачу. Допустим, имеется случайная величина $\boldsymbol{X}$ и заданным математическим ожиданием m_x и дисперсией D_x. Над этой величиной производится n независимых опытов в постоянных условиях, после которых вычисляется среднее арифметическое всех имеющихся значений.

Требуется найти числовые характеристики этого среднего значения: математическое ожидание и дисперсию, а так же выяснить, как они изменяются с увеличением n.

Обозначим: X_1 – значение случайной величины X в первом опыте, X_2 – значение рассматриваемой величины во втором опыте и т.д.

Определим среднее арифметическое $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$. (1.12)

По теоремам о числовых характеристиках линейных функций статистически независимых аргументов имеем:

$m_y = \frac{1}{n}\sum_{i=1}^{n} m_{xi}$, т.к. $m_{xi} = m_x - const$ (опыты в постоянных условиях), то $m_y = m_x$. (1.13)

$D_y = \frac{1}{n^2}\sum_{i=1}^{n} D_{xi}$, т.к. $D_{xi} = D_x - const$ (постоянные условия опытов), то $D_y = \frac{1}{n} D_x$. (1.14)

Проведенный аналитический анализ показывает, что математическое ожидание среднего арифметического Y при независимых опытах в постоянных условиях не зависит от числа сделанных наблюдений n и равно математическому ожиданию наблюдаемой случайной величины X, а дисперсия среднего арифметического неограниченно убывает с увеличением числа опытов и при достаточно большом n может быть сколь угодно малой положительной величиной.

Теорема П.Л.Чебышева конкретизирует в точной количественной форме отмеченной свойство устойчивости среднего арифметического наблюдаемых значений случайной величины.

Теорема П.Л.Чебышева: "При достаточно большом числе независимых опытов в постоянных условиях среднее арифметическое наблюдаемых значений случайной величины сходится по вероятности к её математическому ожиданию".

Уточним смысл термина "сходимость по вероятности". Говорят, что случайная величина X сходится по вероятности к величине α, если при увеличении числа опытов n вероятность того, что величины X и α будут сколь угодно близки, неограниченно приближается к единице, а это значит, что при достаточно большом n удовлетворяется неравенство

$$P(|X - \alpha| < \varepsilon) > 1 - \delta, \quad (1.15)$$

где ε и δ – произвольные малые положительные числа.

Зачастую для наглядности сходимость величины X к величине α идентифицируется чисто символически как

$$X \xrightarrow[n]{P} \alpha$$

Что следует понимать следующим образом: величина X при увеличении числа опытов n сходится по вероятности к величине α.

Используя приведенную формализацию понятия сходимости по вероятности, можно записать теорему П.Л.Чебышева в виде соотношения

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_x\right| < \varepsilon\right) > 1 - \delta, \tag{1.16}$$

где ε и δ – произвольные малые положительные числа.

Рассмотрим доказательство приведенного неравенства и тем самым докажем теорему П.Л.Чебышева.

Сначала обратимся к статистическим характеристикам среднего арифметического $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ наблюдаемых значений случайной величины при независимых опытах в постоянных условиях, которые получены в предыдущем разделе в виде

$$m_y = \frac{1}{n}\sum_{i=1}^{n} m_{xi} \quad \text{и} \quad D_y = \frac{1}{n} D_x.$$

Применяя к случайной величине Y неравенство Чебышева и положив $\alpha = \varepsilon$, получим

$$P\left(\left|Y - m_y\right| \geq \varepsilon\right) \leq \frac{D_y}{\varepsilon^2} = \frac{D_x}{n \cdot \varepsilon^2}. \tag{1.17}$$

Как бы мало не было число ε, всегда можно взять n таким большим, что бы для произвольного положительно малого числа δ выполнялось неравенство

$$\frac{D_x}{n \cdot \varepsilon^2} < \delta. \tag{1.18}$$

Возвращаясь к неравенству Чебышева, получим

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_x\right| \geq \varepsilon\right) \leq \delta. \tag{1.19}$$

Переходя к противоположному событию, получим соотношение

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_x\right| < \varepsilon\right) > 1 - \delta, \tag{1.20}$$

которое является формализованной записью теоремы П.Л.Чебышева, следовательно, таким образом, доказана теорема Чебышева.

1.3. Обобщённая теорема Чебышева

Теорема П.Л.Чебышева обобщается на более сложный случай независимых опытов в переменных условиях, а именно когда закон распределения случайной величины X от опыта к опыту изменяется.

В этом случае мы имеем дело со средним арифметическим $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ случайных величин $X_1, X_2, \ldots, X_n$ с различными в каждом опыте математическими ожиданиями $m_{x1}, m_{x2}, \ldots, m_{xn}$ и дисперсиями $D_{x1}, D_{x2}, \ldots, D_{xn}$.

Кроме того предполагается, что все дисперсии ограничены сверху одним и тем же число L, т.е.

$$D_{xi} \le L \text{ при } i = 1, 2, \ldots, n. \tag{1.21}$$

В такой постановке обобщённая теорема Чебышева сводится к следующему: "При возрастании числа независимых опытов n в переменных условиях среднее арифметическое наблюдаемых значений величин $X_1, X_2, \ldots, X_n$ сходится по вероятности к среднему арифметическому их математических ожиданий".

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| < \varepsilon\right) > 1 - \delta, \text{ если } D_{xi} \le L \ (i = 1, 2, .. < n), \tag{1.22}$$

где ε и δ – произвольные малые положительные числа.

Доказательство.

Рассмотрим величину $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Найдём для этой величины математическое ожидание и дисперсию, пользуясь известными теоремами о числовых характеристиках

$$m_y = \frac{1}{n}\sum_{i=1}^{n} m_{xi} \quad \text{и} \quad D_y = \frac{1}{n^2}\sum_{i=1}^{n} D_{xi} \ . \tag{1.23}$$

Применим к величине Y неравенство Чебышева $P\left(\left|Y - m_y\right| \ge \varepsilon\right) \le \frac{D_y}{\varepsilon^2}$.

После подстановки получим

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| \ge \varepsilon\right) \le \frac{\sum_{i=1}^{n} D_{xi}}{n^2\varepsilon^2}. \tag{1.24}$$

При замене каждой дисперсии D_{xi} на гарантированно превосходящую величину L неравенство может только усиливаться, поэтому

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| \geq \varepsilon\right) \leq \frac{L}{n\,\varepsilon^2}. \qquad (1.25)$$

Как бы ни было мало ε, можно выбрать число опытов n настолько большим, чтобы выполнялось неравенство $\frac{L}{n\,\varepsilon^2} \leq \delta$, где δ – произвольное малое положительное число, тогда

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| \geq \varepsilon\right) \leq \delta. \qquad (1.26)$$

Переходя к противоположному событию с вероятностью $P = 1 - \delta$, получим

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| < \varepsilon\right) > 1 - \delta, \qquad (1.27)$$

что и представляет собой доказательство обобщённой теоремы Чебышева.

1.4. Теорема Маркова*

Эта теорема представляет собой наиболее общий случай закона больших чисел в следующей формулировке:

"Если имеются статистически зависимые* случайные величины $X_1, X_2, \ldots, X_n$ для совокупности которых при неограниченном увеличении числа опытов $n \to \infty$ соблюдается условие $\frac{1}{n^2}\mathrm{D}\left[\sum_{i=1}^{n} X_i\right] \to 0$, то среднее арифметическое наблюдаемых значений сходится по вероятности к среднему арифметическому их математических ожиданий".

---

* Марков Андрей Андреевич (1856-1922) – ближайший ученик П.Л.Чебышева, который внес особенно большой вклад в становление и развитие теории случайных процессов, специфическая разновидность которых вошла в современную математическую литературу как так называемые марковские процессы и до сих пор широко используется при исследовании и моделировании систем массового обслуживания.

* Напомним, что две случайные величины являются статистически зависимыми, если закон распределения одной из них зависит от того какое значение приобрела другая из рассматриваемых величин. Статистическая зависимость включает в себя как частный случай функциональную зависимость и поэтому является более широким понятием.

Формализованное оформление теоремы Маркова выглядит следующим образом

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| < \varepsilon\right) > 1-\delta, \text{ если при } n \to \infty \text{ имеем } \frac{1}{n^2}\mathrm{D}\left[\sum_{i=1}^{n} X_i\right] \to 0. \quad (1.28)$$

Доказательство.

Рассмотрим среднее арифметическое наблюдаемых значений $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Найдём для этой величины математическое ожидание и дисперсию, пользуясь известными

Очевидно, что $m_y = \frac{1}{n}\sum_{i=1}^{n} m_{xi}$ и $D_y = \frac{1}{n^2}\sum_{i=1}^{n} D_{xi}$.

Применим к величине Y неравенство Чебышева $P\left(\left|Y - m_y\right| \geq \varepsilon\right) \leq \frac{D_y}{\varepsilon^2}$. (1.29)

По условию теоремы при $n \to \infty$ неограниченно убывает величина $\frac{1}{n^2}\mathrm{D}\left[\sum_{i=1}^{n} X_i\right] \to 0$, т.е. неограниченно уменьшается дисперсия $D_y \to 0$, поэтому неравенство Чебышева можно представить в виде $P\left(\left|Y - m_y\right| \geq \varepsilon\right) \leq \delta$, где δ – произвольная малая положительная величина.

Переходя к противоположному событию, получим

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| < \varepsilon\right) > 1-\delta, \quad (1.30)$$

что и требовалось доказать.

1.5. Теорема Я. Бернулли*

Известная теорема Якоба Бернулли устанавливает связь между частотой появления наблюдаемого события и его вероятностью:

"При неограниченном увеличении числа независимых опытов в постоянных условиях, т.е. при , $n \to \infty$ частота события A, а именно p^* сходится по вероятности к его истинной вероятности p ".

* Якоб Бернулли (Jakob Bernoulli, 1654-1705) – знаменитый швейцарский математик.

Частота случайного события p^* здесь, как это принято в теории вероятностей, понимается в виде отношения числа опытов m благоприятствующих появлению данного события к общему числу опытов n, т.е.

$$p^* = \frac{m}{n}. \tag{1.31}$$

Формально теорема Я.Бернулли записывается в виде неравенства, которое обеспечивается при увеличении числа наблюдений ($n \to \infty$)

$$P(\left|p^* - p\right| < \varepsilon) > 1 - \delta \tag{1.32}$$

для произвольных малых положительных чисел ε и δ .

Доказательство.

Обозначим случайный результат $i-$го опыта как X_i. Эта дискретная случайная величина может принимать два значения: $X_{i1} = 0$, когда опыт не привёл к желаемому результату и $X_{i2} = 1$, если i – й опыт закончился с интересующим нас результатом.

Как и для всякой дискретной величина, её распределение характеризуется статистическим рядом, который в данном случае имеет вид

X_{ik}	0	1
p_k	$1-p$	p

Математическое ожидание величины X_i в отдельно взятом опыте вычисляется как

$$m_{x_i} = \sum_{k=1}^{2} x_{ik} \cdot p_k = p\,. \tag{1.33}$$

Частота представляет собой среднее арифметическое величин X_1, X_2, ..., X_i,..., X_n

$$p^* = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{1.34}$$

Согласно теореме П.Л.Чебышева при независимых опытах в постоянных условиях среднее арифметическое сходится по вероятности к математическому ожиданию, что в нашем случае записывается следующим образом

$$P(|p^* - p| < \varepsilon) > 1 - \delta\,. \tag{1.35}$$

Именно это и требовалось доказать.

Следует подчеркнуть особое значение теоремы Я.Бернулли для решения вероятностных задач, т.к. именно эта теорема обосновывает возможность оценки вероятностей интересующих нас на практике событий, исходя их обработки наблюдений независимых опытов в постоянных условиях.

1.6. Теорема Пуассона*

При изменяющихся условиях опыта так же имеет место устойчивость частоты появления события, что формулируется в тереме Пуассона:

"Если производится n независимых опытов и вероятность появления события A в $i-$м опыте равна p_i , то при увеличении числа опытов частота появления события A сходится по вероятности к среднему арифметическому вероятностей p_i ".

Формализованная запись теоремы Пуассона имеет следующий вид

$$P\left(\left|p^* - \frac{1}{n}\sum_{i=1}^{n} p_i\right| < \varepsilon\right) > 1 - \delta, \tag{1.36}$$

для любых малых положительных величин ε и δ, если $n \to \infty$.

Доказательство этой теоремы аналогично доказательству предыдущей теоремы, но при этом используется обобщённая теорема Чебышева, применяя которую необходимо убедиться в ограниченности дисперсий.

Обобщённая теорема Чебышева декларирует сходимость по вероятности среднего арифметического наблюдаемых значений величин $X_1, X_2, \ldots, X_n$ к среднему арифметическому их математических ожиданий при независимых опытах в переменных условиях и ограниченности дисперсий

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{xi}}{n}\right| < \varepsilon\right) > 1 - \delta, \quad \text{если } D_{xi} \le L \ \ (i = 1, 2, .. < n), \tag{1.37}$$

где ε и δ – произвольные малые положительные числа.

---

* Simeon Denis Poison (1781-1840) – выдающийся французский математик и физик, известное распределение которого и в настоящее время широко используется в теории массового обслуживания при моделировании потоков событий.

Применительно к частоте события на основании обобщённой теоремы Чебышева можно записать

$$P\left(\left|p^* - \frac{1}{n}\sum_{i=1}^{n} p_i\right| < \varepsilon\right) > 1 - \delta, \quad \text{если } D_{p_i} \le L \ \ (i = 1, 2, .. < n). \qquad (1.38)$$

Таким образом, что бы доказать сходимость частоты события к его вероятности, в данном случае нужно доказать ограниченность дисперсий статистических оценок вероятностей, т.е. убедиться в выполнении условия $D_{p_i} \le L \ \ (i = 1, 2, .. < n)$.

Обратимся к анализу дисперсии оценки вероятности $D_{p_i} = \mathrm{M}\left[(p_i - m_{p_i})^2\right]$.

Поскольку статистическая вероятность и её математическое ожидание могут изменяться в строго ограниченном интервале возможных значений

$$0 \le p_i \le 1, \quad 0 \le m_{p_i} \le 1,$$

то из этого следует ограничение $D_{p_i} = \mathrm{M}\left[(p_i - m_{p_i})^2\right] \le 0,25$.

Доказав ограниченность дисперсии статистических оценок вероятности, мы тем самым доказали теорему Пуассона, которая имеет принципиальное значение для практики, поскольку часто вероятностные методы используются для исследования явлений, которые физически не могут много раз повторяться в одинаковых условиях. При меняющихся условиях вероятности рассматриваемых событий сильно зависят этих условий. Устойчивость частот выражается здесь в том, что частота события приближается к средней вероятности.

2. ВЫБОРОЧНЫЕ ОЦЕНКИ И ИХ СВОЙСТВА

Числовые характеристики случайных величин, получаемые опытным путём, сами являются случайными величинами. Например, среднее арифметическое при увеличении числа опытов приближается к математическому ожиданию, но всё же здесь имеет место только лишь сходимость по вероятности, а обработка конкретной выборки приводит к получению статистических оценок, включающих в себя случайные отклонения от генеральных характеристик.

Любая из выборочных оценок случайна и при её использовании возможны ошибки, поэтому желательно выбрать такие оценки искомых характеристик, что бы эти ошибки были по возможности минимальными.

2.1. Требования к выборочным оценкам

Допустим, что нас интересует некоторая характеристика a для случайной величины $\boldsymbol{X}$, которая в реальной задаче приобретает конкретный смысл, например, математическое ожидание, дисперсия, какой-либо начальный или центральный момент с заданным порядком, вероятность некоторого события и т.п.

Выполнено n независимых опытов и получены значения рассматриваемой случайной величины

$$X_1, X_2, ..., X_i, ..., X_n \ .$$

Выборочная оценка a^* является функцией наблюдаемых случайных величин и поэтому сама является случайной величиной

$$a^* = a(X_1, X_2, ..., X_i, ..., X_n) .$$

Закон распределения величины a^* зависит от самого a, от наблюдаемых значений $X_1, X_2, ..., X_i, ..., X_n$, от числа опытов n и может быть найден известными методами теории вероятностей, что представляет сложную в математическом плане задачу, но для многих практически значимых случаев имеются известные решения.

Исходя из общей постановки задач математической статистики, к выборочным оценкам a^* предъявляются требования состоятельности, несмещённости и эффективности.

1. **Состоятельность** – сходимость по вероятности к истинному значению при увеличении числа опытов:

$$P(|a^* - a| < \varepsilon) > 1 - \delta, \quad \text{для любых малых положительных величин } \varepsilon \text{ и } \delta, \text{ если } n \to \infty. \tag{2.1}$$

2. **Несмещённость** – отсутствие систематическое ошибки (постоянное отклонение в одну сторону), что выражается в равенстве математического ожидания выборочной оценки истинному значению данной характеристики

$$\mathrm{M}\left[a^*\right] = a\,. \tag{2.2}$$

3. **Эффективность** – минимальное рассеивание (минимальная дисперсия)

$$\mathrm{D}\left[a^*\right] \to \min\,. \tag{2.3}$$

2.2. Свойства выборочных оценок математического ожидания

Рассмотрим свойства выборочных оценок математического ожидания при независимых опытах.

В качестве статистической оценки математического ожидания используем среднее арифметическое наблюдаемых значений случайной величины X

$$m_x^* = \frac{1}{n}\sum_{n-1}^{n} x_i\,. \tag{2.4}$$

Тем не менее, в этом следует убедиться, проверив состоятельность, несмещённость и эффективность предлагаемой статистической оценки дисперсии.

2.2.1. **Состоятельность** – поскольку речь об оценке математического ожидания, декларируется следующим образом

$$P(|m_x^* - m_x| < \varepsilon) > 1 - \delta, \quad \text{для любых малых положительных величин } \varepsilon \text{ и } \delta. \tag{2.5}$$

Это свойство выборочных оценок математического ожидания фактически устанавливается теоремой Чебышева, в которой утверждается, что среднее арифметическое наблюдаемых значений при независимых опытах в постоянных условиях сходится по вероятности к математическому ожиданию

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_x\right| < \varepsilon\right) > 1 - \delta. \tag{2.6}$$

2.2.2. **Несмещённость** – $\mathrm{M}[m_x^*] = m_x\,.$ (2.7)

Для случайной величины, которая является линейной функцией других случайных величин, а именно $m_x^* = \frac{1}{n}\sum_{n-1}^{n} x_i$ по теореме о характеристиках случайных величин из теории вероятностей известно что $\mathrm{M}\left[\frac{1}{n}\sum_{n-1}^{n} x_i\right] = \frac{1}{n}\sum_{n-1}^{n} m_{xi}$, но, поскольку рассматриваются опыты в неизменных условиях, т.е. когда $m_{xi} - const$, то $m_{xi} = m_x$, то в итоге имеем доказательство свойства несмещённости выборочной оценки математического ожидания

$$M\left[\frac{1}{n}\sum_{n-1}^{n} x_i\right] = \frac{1}{n} n \cdot m_x = m_x . \qquad (2.8)$$

2.2.3. **Эффективность** – $D[m_x^*] \to \min$. (2.9)

Пользуясь теоремами о числовых характеристиках случайных величин и учитывая, что рассматриваются независимые опыты в постоянных условиях, приходим к следующим соотношениям

$$\mathrm{D}\left[\frac{1}{n}\sum_{n-1}^{n} x_i\right] = \frac{1}{n^2}\sum_{n-1}^{n} D_{xi} = \frac{1}{n^2} n \cdot D_x = \frac{D_x}{n} . \qquad (2.10)$$

Можно показать, что для нормального закона распределения приведенная статистическая оценка математического ожидания всегда будет минимальной, т.е. эффективной. Для других законов распределения это может быть и не так.

2.3. Свойства выборочных оценок дисперсии

На первый взгляд наиболее естественной оценкой, по аналогии с равновозможными дискретными случайными величинами, является статистическая дисперсия, вычисляемая как

$$D_x^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^2 , \text{ где } m_x^* = \frac{1}{n}\sum_{n-1}^{n} x_i . \qquad (2.11)$$

2.3.1. **Состоятельность** – т.к. речь идёт об оценке дисперсии, то её состоятельность формулируется следующим образом

$$P(\left|D_x^* - D_x\right| \geq \varepsilon) \leq \delta, \qquad (2.12)$$

для любых малых положительных величин ε и δ при неограниченном числе опытов.

Попробуем убедиться в состоятельности оценки дисперсии, обратившись к теореме Чебышева для некоторой величины $\boldsymbol{Y}$

$$P(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - m_y\right| \geq \varepsilon) \leq \delta. \qquad (2.13)$$

В качестве Y возьмём второй начальный момент $Y = \alpha_{2x}$. Выборочная оценка этого момента вычисляется как

$$\alpha^*_{2x} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 . \tag{2.14}$$

Как видно, выборочный второй начальный момент находится как среднее арифметическое квадратов отдельных значений рассматриваемой случайной величины, потому данная оценка сходится по вероятности к своему истинному значению по теореме Чебышева, т.е.

$$P\left(\left|\alpha^*_{2x} - \alpha_{2x}\right| \geq \varepsilon\right) \leq \delta, \tag{2.15}$$

для любых малых положительных величин ε и δ при неограниченном увеличении числа опытов.

Можно показать, что квадрат выборочного математического ожидания сходится к квадрату истинного значения математического ожидания, т.е.

$$P\left(\left|(m^*_x)^2 - (m_x)^2\right| \geq \varepsilon\right) \leq \delta, \tag{2.16}$$

для любых малых положительных величин ε и δ при неограниченном увеличении числа опытов.

Объединяя два предыдущих неравенства в одно и выполнив перегруппировку, и учитывая соотношения

$$D^*_x = \alpha^*_{2x} - (m^*_x)^2 \quad \text{и} \quad D_x = (\alpha_{2x} - (m_x)^2, \tag{2.17}$$

получим следующее неравенство показывающее сходимость статистической оценки дисперсии к своему истинному значению

$$P\left(\left|D^*_x - D_x\right| \geq \varepsilon\right) \leq \delta . \tag{2.18}$$

2.3.2. **Несмещённость** – а именно то, что $\mathrm{M}[D^*_x] = D_x$.

Выразим статистическую оценку дисперсии через оценки второго начального момента и математического ожидания

$$\begin{aligned} D^*_x = \alpha^*_{2x} - (m^*_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \\ = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\sum_{i=1}^{n} x_i - 2\frac{1}{n^2}\sum_{i<j} x_i x_j = \frac{n-1}{n^2}\sum_{i=1}^{n} x_i^2 - 2\frac{1}{n^2}\sum_{i<j} x_i x_j . \end{aligned} \tag{2.19}$$

Найдём математическое ожидание полученного выражения для оценки дисперсии

$$\mathrm{M}\left[D^*_x\right] = \frac{n-1}{n^2}\sum_{i=1}^{n}\mathrm{M}\left[x_i^2\right] - 2\frac{1}{n^2}\sum_{i<j}\mathrm{M}\left[x_i x_j\right]. \tag{2.20}$$

Перейдём к центрированным значениям $\mathring{x}_i = x_i - m_x$, так как дисперсия не зависит от того, в какой точке выбрать начало координат

$$\mathrm{M}\left[D^*_x\right] = \frac{n-1}{n^2}\sum_{i=1}^{n}\mathrm{M}\left[\mathring{x}_i^2\right] - 2\frac{1}{n^2}\sum_{i<j}\mathrm{M}\left[\mathring{x}_i \mathring{x}_j\right]. \tag{2.21}$$

Очевидно что $\sum_{i=1}^{n}\mathrm{M}\left[\mathring{x}_i^2\right] = n \cdot D_x$, $\quad \sum_{i<j}\mathrm{M}\left[\mathring{x}_i \mathring{x}_j\right] = K_{ij}$. (2.22, 2.23)

Будем считать, что рассматриваются независимые опыты, поэтому корреляционный момент равен нулю, т.е. $K_{ij} = 0$.

С учётом сделанных уточнений имеем

$$\mathrm{M}\left[D_x^*\right] = \frac{n-1}{n} D_x \ , \tag{2.24}$$

откуда видна смещённость используемой статистической оценки дисперсии.

Как следствие выявленной особенности, оценка дисперсии, получаемая по формуле

$$D_x^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^2 \tag{2.25}$$

или, что даёт тот же результат, но несколько удобнее для вычислений

$$D_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(m_x^*\right)^2 \tag{2.26}$$

называется смещённой оценкой дисперсии.

Устранение смещения оценки дисперсии не вызывает затруднений и выполняется следующим образом, позволяя получать несмещённые выборочные оценки по формулам

$$D_x^* = \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^2$$

или (2.27)

$$D_x^* = \frac{n-1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(m_x^*\right)^2\right].$$

Таким образом, связь между несмещёнными оценками дисперсии и соответствующими смещёнными дисперсиями устанавливается в виде соотношения

$$D^*_{несмещ}(x) = \frac{n-1}{n} D^*_{смещ}(x). \tag{2.28}$$

Оценим насколько велика может быть поправка дисперсии, устраняющая её смещение. Обратимся к данным приведенным ниже в таблице для выборок различного объёма.

Таблица 2.1 – Поправка на смещение дисперсии

n	$\frac{n}{n-1}$
10	1,111
30	1,034
100	1,010
200	1,005

Очевидно, что поправку на смещение дисперсии следует использовать только для средних выборок, а для представительных выборок она не имеет значения.

2.3.3. **Эффективность** $-\mathrm{D}\left[D^*\right] \to \min$. (2.29)

Можно показать, но это выходит за рамки данного пособия, что в общем случае выборочные оценки дисперсии не эффективны.

В случае нормального закона распределения наблюдается так называемая асимптотическая эффективность, т.е. при неограниченном увеличении числа опытов статистическая оценка дисперсии приближается к минимальной, становясь тем самым эффективной.

2.4. Свойства выборочных оценок вероятности случайного события

Рассмотрим статистическую оценку вероятности по частоте появления события при ограниченном числе независимых опытов в постоянных условиях

$$p^* = \frac{m}{n}, \tag{2.30}$$

где n – объём выборочных данных, m – число опытов, в которых реализовалось интересующее нас событие, или как ещё говорят по-другому – число опытов благоприятствующих рассматриваемому событию.

Представим число появлений данного события в виде суммы дискретных случайных величин X_i, каждая из которых может принимать только два значения, т.е. 0 или 1 с вероятностями, приведенными в статистическом ряду

x_{ik}	0	1
p_k	$1-p$	p

Попутно заметим, что статистические характеристики бинарной случайной величины X_i определяются, как это известно, из теории вероятностей, следующим образом

$$m_{x_i} = p, \quad D_{x_i} = p(1-p)\ . \tag{2.31, 2.32}$$

2.4.1. **Состоятельность** – это свойство статистической оценки вероятности являющееся следствием теоремы Я.Бернулли, в соответствии с которой частота события сходится по вероятности к вероятности данного события при увеличении числа опытов, что в формализованном виде описывается соотношением

$P(\left|p^* - p\right| \geq \varepsilon) \leq \delta$, для любых малых положительных величин ε и δ.

2.4.2. **Несмещённость** – $\mathrm{M}[p^*] = p$. (2.33)

Найдём математическое ожидание статистической оценки вероятности как математическое ожидание суммы случайных бинарных величин, используя известные теоремы теории вероятностей о числовых характеристиках линейных функций случайных величин

$$\mathrm{M}\left[p^*\right] = \mathrm{M}\left[\frac{m}{n}\right] = \frac{1}{n}\mathrm{M}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} m_{x_i}\right] = \frac{1}{n} n \cdot p = p. \quad (2.34)$$

Представленные выкладки доказывают несмещённость статистической оценки вероятности по частоте регистрации события.

2.4.3. **Эффективность** – $\mathrm{D}\left[p^*\right] \to \min$. (2.35)

Найдём дисперсию частоты наблюдаемых событий

$$\mathrm{D}\left[p^*\right] = \mathrm{D}\left[\frac{m}{n}\right] = \frac{1}{n^2}\mathrm{M}\left[\sum_{i=1}^{n} D_{x_i}\right] = \frac{1}{n^2} p \cdot (1-p) = \frac{1}{n} p \cdot (1-p). \quad (2.36)$$

Можно показать, что такая дисперсия является минимально возможной и, следовательно, оценка $p^* = \frac{m}{n}$ является эффективной.

Проведенный анализ свойств выборочной оценки вероятности по частоте показал, что эта оценка отвечает требованиям состоятельности, несмещённости и эффективности, т.е. всем тем основным требованиям, которым должны удовлетворять статистические оценки характеристики случайных величин.

3. ОБРАБОТКА ОПЫТОВ

3.1. Простая статистическая совокупность. Статистический ряд. Гистограмма

Предположим, что изучается некоторая случайная величина X, закон распределения которой неизвестен. Требуется решить одну из задач математической статистики, например, определить численные значения параметров закона её распределения, используя данные наблюдений за рассматриваемой величиной, или проверить гипотезу о том, что величина X подчиняется тому или иному закону распределения.

Для решения этой задачи над случайной величиной X производится ряд независимых опытов (наблюдений), в результате которых величина X принимает определённые значения. Эти данные представляют собой первичный статистический материал, подлежащий последующей обработке и анализу.

Совокупность зарегистрированных опытных значений случайной величины называется простой статистической совокупностью или простым статистическим рядом, а так же выборкой.

Наблюдаемые значения случайных величин протоколируются, как это показано в качестве примера, приведенного в таблице 3.1.

Таблица 3.1 – Регистрация статистических данных

Номер опыта (наблюдение) i	Значение случайной величины x_i
1	120
2	110
3	105
…	…
…	…
n	150

Исходя из целей решения выбранной статистической задачи, необходимо определиться с необходимым числом опытов или наблюдений n. Этому вопросу мы уделим в данном пособии особое внимание (см. раздел 8), однако сразу же отметим, что по объёму

зарегистрированных данных различают малые выборки $n = 6 - 12$, средние выборки $n = 20 - 60$ и представительные (репрезентативные) выборки объёмом $n = 200 - 600$.

Приведенные границы малых, средних и представительных выборок достаточно условны и при этом очевидно, что чем больше объём выборки, тем достовернее решение соответствующей статистической задачи.

Менее очевидно то, что желая повысить достоверность получаемых оценок случайных величин и используя для этого, например, выборку объёмом $n_2 = 70$ вместо выборки меньшего объёма $n_1 = 60$ мы не получим ощутимого увеличения точности оценок, т.к. для этого нужно значительно увеличить объём данных, переходя к представительным выборкам, имеющим $n_2 = 200 - 600$.

Малые выборки, из-за низкой достоверности результатов их обработки, в практических приложениях математической статистики используются в исключительно редких случаях, когда объём выборки затруднительно увеличить по каким-либо соображениям.

Основным инструментом практической статистики является обработка средних выборок.

Используя представительные выборки, мы, конечно, увеличиваем достоверность получаемых решений статистических задач, но при этом следует отметить, что в отличие от средних выборок, где в качестве исходных данных выступает простая статистическая совокупность, здесь это представление исходных данных теряет наглядность, т.к. простая статистическая совокупность при числе наблюдений в несколько сотен перестаёт быть удобной и наглядной, т.к. становится громоздкой и необозримой.

Для придания компактности и наглядности данным представительной выборки выполняется предварительная обработка, в результате которой строится статистический ряд. Что бы получить статистический ряд необходимо разделить весь диапазон опытных данных $[x_{\min}; x_{\max}]$ на интервалы или "разряды".

Число разрядов можно определить, например, по формуле Стерджесса (Herbert Sturges, 1926), результат применения которой округляется в большую сторону

$$k = 1 + \log_2 N \quad \text{или в десятичных логарифмах} \quad k = 1 + 3{,}322 \cdot \lg N, \qquad (3.1)$$

где N – объём выборки.

Можно задать число разрядов экспертным путём назначив $k = 12 - 20$, т.к. такое число разрядов обеспечивает достаточную точность последующих вычислений.

Выбрав число разрядов k, найдем ширину разрядного интервала, считая её постоянной величиной

$$h = \frac{x_{\max} - x_{\min}}{k}. \tag{3.2}$$

Далее находим границы разрядных интервалов

$$J_i = \left(x_i;\ x_{i+1} = x_i + h\right),\ x_1 = x_{\min},\ \ i = 1, 2, \ldots, k. \tag{3.3}$$

В заключение предварительной обработки представительной выборки, найдём частоту попадания опытных данных в соответствующие интервалы, используя число попаданий в каждый из разрядов m_i найденное для данной выборки

$$p_i^* = \frac{m_i}{N},\ i = 1, 2, \ldots, k. \tag{3.4}$$

Очевидно, что $\sum_{i=1}^{k} p_i^* = 1.$

Собственно статистическим рядом называется таблица, отражающая полученные результаты предварительной обработки представительной выборки, пример структуры оформления которой представлен в виде таблицы 3. 2.

При последующем анализе границы интервалов используются редко, обычно требуются численные значения "представителей" разрядов, которые вычисляются следующим образом

$$x_i^* = \frac{x_i + x_{i+1}}{2},\ \ i = 1, 2, \ldots, k. \tag{3.5}$$

Таблица 3.2 – Статистический ряд

i	1	2	…	k
$J_i = \left(x_i;\ x_{i+1}\right)$	$\left(x_1;\ x_2\right)$	$\left(x_2;\ x_3\right)$	…	$\left(x_k;\ x_{k+1}\right)$
$p_i^* = \frac{m_i}{N}$	p_1^*	p_2^*	…	p_k^*

При построении статистических рядов возникают вопросы, один из которых состоит в том, что непонятно как поступать, если какое-то из наблюдаемых значений попадает точно на границу смежных разрядов – это маловероятно, но возможно.

Здесь можно поступать по-разному, например, для каждого такого значения прибавлять к накопленным частотам смежных разрядов по 1/2, как бы разделив эту точку поровну между соседними разрядами. Поскольку таких значений в принципе не может быть много, а выборка представительная, в которой число опытов составляет сотни наблюдений, то вполне подойдёт любое формализованное правило учёта таких данных, например, смещение этих точек в левый или в правый ближайший разряд – на результатах решения статистических задач практически это не отразится.

Часто также возникает вопрос о целесообразности использования неравномерного разбиения разрядных интервалов. При ограниченном и заведомо малом числе разрядов, например, $k = 6 - 8$ такой подход может повысить точность последующего использования полученного статистического ряда, но реализация неравномерного разбиения зависит от вида распределения исследуемой случайной величины, а его ещё предстоит определить. Поэтому вопрос решается чисто технически – увеличивается число разрядов, а разрядные интервалы принимаются одинаковыми.

Графическое изображение статистического ряда называется гистограммой, схематичное построение гистограммы, приведено на рис.3.1. Очевидно, что при увеличении числа наблюдений и при соответствующем увеличении числа разрядов гистограмма начнёт приближаться к плотности вероятностей рассматриваемой случайной величины.

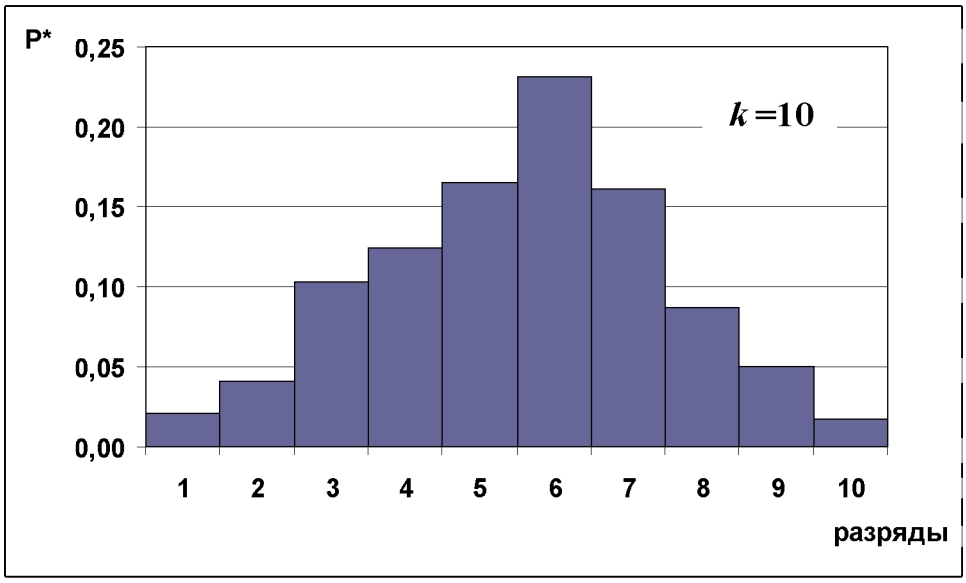

Рис. 3.1 – Гистограмма случайной величины

3.2. Числовые характеристики статистического распределения

Конечно, наиболее общей характеристикой случайной величина является закон распределения в той или иной форме, а именно в виде функции распределения или в виде плотности вероятности. Закон распределения содержит всю полноту информации о данной случайной величине с точки зрения математики, а точнее, теории вероятностей и ещё точнее, с позиций математической статистики.

Однако выявление закона распределения зачастую и не требуется, что имеет место во многих приложениях математической статистики к решению практических задач, когда достаточно ограничиться только получением оценок характеристик распределения.

Отметим некоторое примечание о возможной неоднозначности терминологии в математической статистике: так например синонимами выборочная оценка какой-либо характеристики являются широко распространённые синонимы: статистическая, опытная оценка.

Для генеральной оценки синонимами являются синонимами: истинная, теоретическая оценка.

3.2.1. Выборочное математическое ожидание

А. Средняя выборка $m_x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$. (3.6)

Б. Представительная выборка $m_x^* = \sum_{i=1}^{k} x_i \cdot p_i^*$. (3.7)

3.2.2. Выборочная дисперсия и среднего квадратического отклонения

А. Средняя выборка $D_x^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - m_x^*)^2$. (3.8)

Б. Представительная выборка $D_x^* = \sum_{i=1}^{k} (x_i - m_x)^2 \cdot p_i^*$. (3.9)

Среднее квадратическое отклонение

$$s_x^* = \sqrt{D_x^*}\,. \tag{3.10}$$

3.2.3. Выборочные начальные моменты

А. Средняя выборка $\alpha_s^*(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i)^s$. (3.11)

Б. Представительная выборка $\alpha_s^*(x) = \sum_{i=1}^{k} (x_i)^s \cdot p_i^*$. (3.12)

3.2.4. Выборочные центральные моменты

А. Средняя выборка $\mu_s^*(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - m_x^*)^s$. (3.13)

Б. Представительная выборка $\mu_s^*(x) = \sum_{i=1}^{k} (x_i - m_x)^s \cdot p_i^*$. (3.14)

Свойства выборочных моментов такие же, как у одноимённых моментов, соответствующих генеральной совокупности наблюдаемых данных, так, например

$$\mu_1^*(x) = 0;$$
$$\mu_2^*(x) = D_x^* = \alpha_2^*(x) - (m_x^*)^2 \text{ и т.п.}$$

Приведенные формулы для статистических оценок характеристик распределения случайных величин можно получить, используя формулы для аналогичных характеристик дискретных случайных величин и записывая их для частного случая равной вероятности наблюдаемых значений, что на самом деле вовсе не обязательно. Поэтому такой формальный подход имеет недостатки, которые, как показывает исследование данного вопроса, в наибольшей степени проявляются при обработке средних выборок, к чему мы вернёмся в данном пособии несколько позднее.

Примечание. Выборочные моменты выше 4-го порядка редко используются в практике статистического анализа, т.к. при постоянном объёме выборок с ростом порядка вычисляемых моментов значительно снижается достоверность получаемых статистических оценок. Поэтому обеспечение достаточной точности этих оценок требует увеличения объёма выборки до таких значений, которые обычно трудно реализовать по чисто техническим причинам.

3.3. Выравнивание статистических рядов

Задача выравнивания статистических рядов заключается в том, что бы подобрать теоретическую кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение.

Очевидно, что критерий наилучшего подбора может быть различным, а решение данный задачи не является однозначным.

Обратимся к причинам вызывающим постановку задачи выравнивания статистических рядов. Дело здесь в том, что на практике мы имеем всегда ограниченное число наблюдений (выборка), поэтому статистическое (выборочное) распределение в большей или меньшей мере имеет случайные черты, которые маскируют распределение генеральной совокупности рассматриваемых данных.

Наиболее распространённым методом выравнивания статистических рядов является метод моментов.

Рассмотрим применение данного метода в тех случаях, когда случайная величина X распределена в генеральной совокупности по нормальному закону с функцией распределения, описываемой выражением

$$F(x) = \frac{1}{s_x\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-m_x)^2}{2s_x^2}\right) dx, \qquad (3.15)$$

параметры которого, а именно m_x и s_x, мы не знаем, кроме того мы не знаем вообще, какой вид имеет закон распределения генеральной совокупности интересующих нас данных.

В нашем распоряжении только лишь ограниченная выборка данных, статистическая обработка которых позволила получить выборочные оценки характеристик m_x^* и s_x^*.

Подставив выборочные оценки моментов распределения m_x^* и s_x^* в выражение предполагаемого нами закона распределения, получим выравнивающую функцию распределения

$$F^*(x) = \frac{1}{s_x^*\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-m_x^*)^2}{2(s_x^*)^2}\right) dx. \qquad (3.16)$$

Очевидно, что выравнивающая функция распределения и исходные выборочные данные имеют одинаковые моменты распределения, в этом и заключается идея метода моментов.

Убедимся в возможностях выравнивания статистических функций распределения, обратившись к иллюстрациям на рисунках 3.2, 3.3 и 3.4, где приведены результаты статистической обработки трёх выборок из одной генеральной совокупности. Выборки различаются объёмом, который составляет в этих примерах n_1=15, n_2=40 и n_3=100.

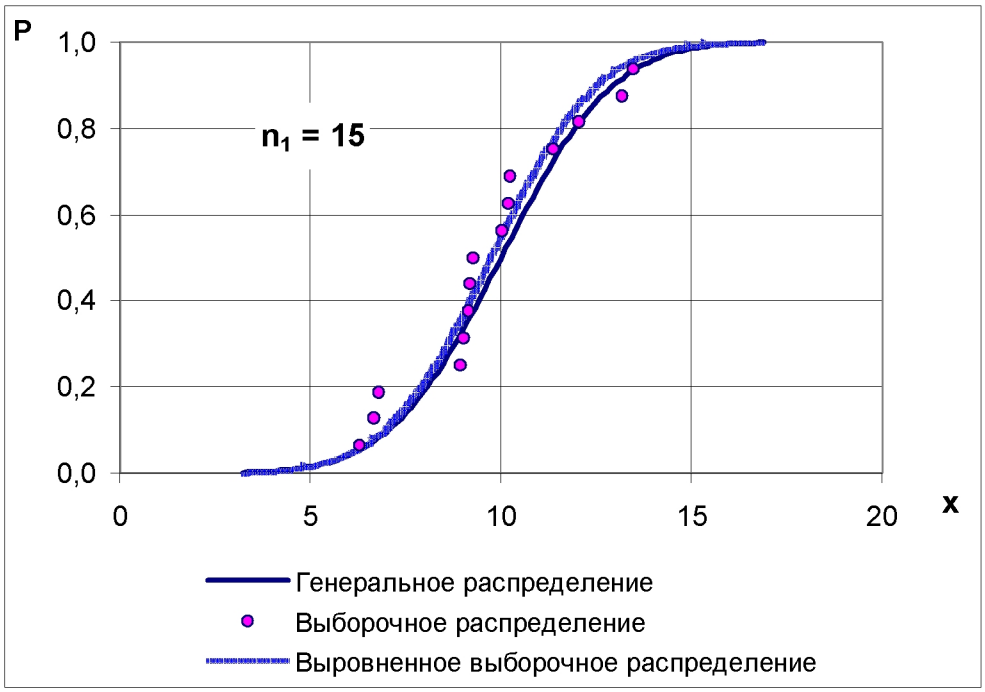

Рис. 3.2 - Выравнивание исходных данных нормальным законом распределения, n=15

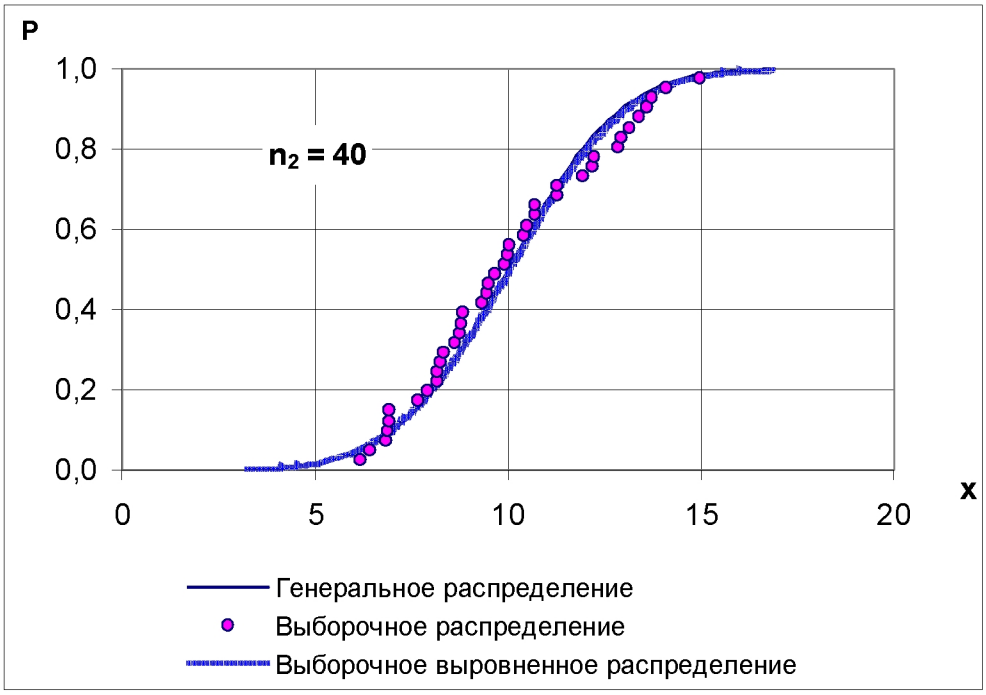

Рис. 3.3 - Выравнивание исходных данных нормальным законом распределения, n=40

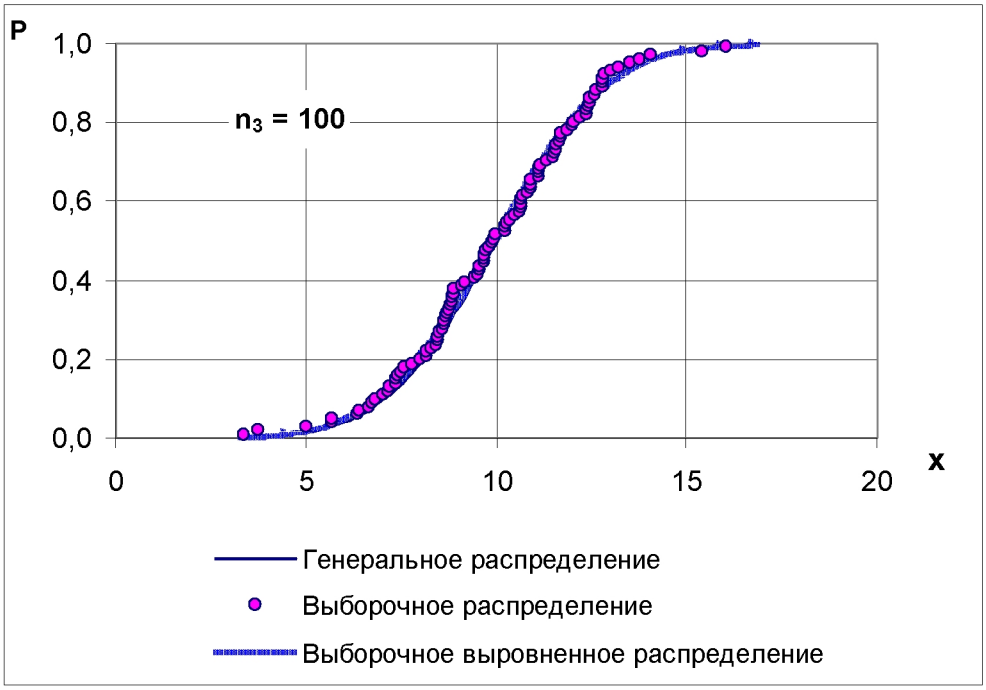

Рис. 3.4 - Выравнивание исходных данных нормальным законом распределения, n=100

Как видно из приведенных примеров, увеличение объёма выборочных данных приводит к сближению выборочной функции распределения к генеральному распределению, что вполне предсказуемо. Но даже при наименьшем объёме выборки из числа рассмотренных процедура выравнивания даёт убедительный результат, особенная ценность которого заключается в воспроизведении "хвостов" распределения, куда статистические данные не попадают. Последнее имеет большое практическое значение, так как именно "хвосты", т.е. участки функции распределения в области наименьших и наибольших возможных значений представляют наибольший интерес для практического использования.

Рассмотрим особенности применения метода моментов для выравнивания статистических данных, которые в генеральной совокупности подчиняются закону распределения с равномерной плотностью вероятности

$$f(x) = \begin{cases} \dfrac{1}{\alpha - \beta} & \text{при } \alpha \le x \le \beta; \\ 0 & \text{при } x < \alpha \text{ или } x > \beta. \end{cases} \qquad (3.17)$$

Для справки отметим, что интегральная функция равномерного закона распределения имеет вид

$$F(x) = \frac{x - \alpha}{\beta - \alpha}, \text{ если } 0 \le x \le 1;$$
$$F(x) = 0, \text{ если } x < 0; \qquad (3.18)$$
$$F(x) = 1, \text{ если } x > 0.$$

Из теории вероятностей известно, что моменты равномерного закона распределения связаны следующими соотношениями с его явными параметрами

$$m_x = \frac{\alpha + \beta}{2}; \quad D_x = \frac{(\beta - \alpha)^2}{12}. \qquad (3.19,\ 3.20)$$

Воспользуемся результатами статистической обработки имеющихся выборочных данных m_x^* и $D_x^* = (s_x^*)^2$, приравняв между собой следующие выражения

$$\frac{\alpha^* + \beta^*}{2} = m_x^*; \quad \frac{(\beta^* - \alpha^*)^2}{12} = D_x^*, \qquad (3.21,\ 3.22)$$

откуда найдём параметры выравнивающей функции равномерного закона распределения

$$\alpha^* = m_x^* - \sqrt{3D_x^*}; \quad \beta^* = m_x^* + \sqrt{3D_x^*}. \qquad (3.23,\ 3.24)$$

Рассмотрим выравнивание равномерным законом распределения на конкретных примерах иллюстрируемых рисунками 3.5, 3.6 и 3.7, где выборки получены из одной и той же генеральной совокупности.

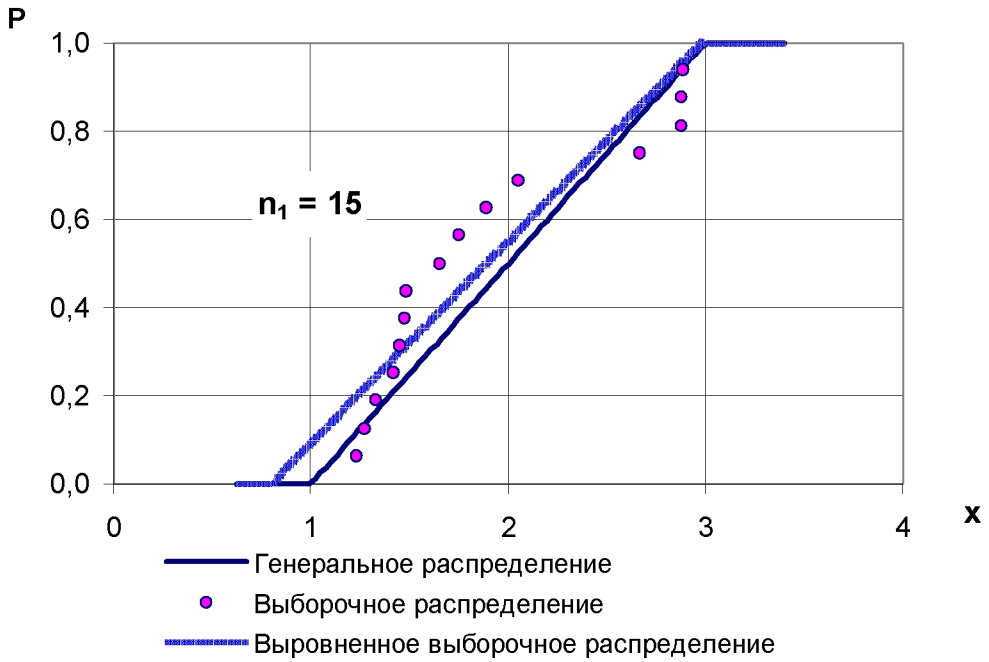

Рис. 3.5 - Выравнивание исходных данных законом равномерной плотности, n=15

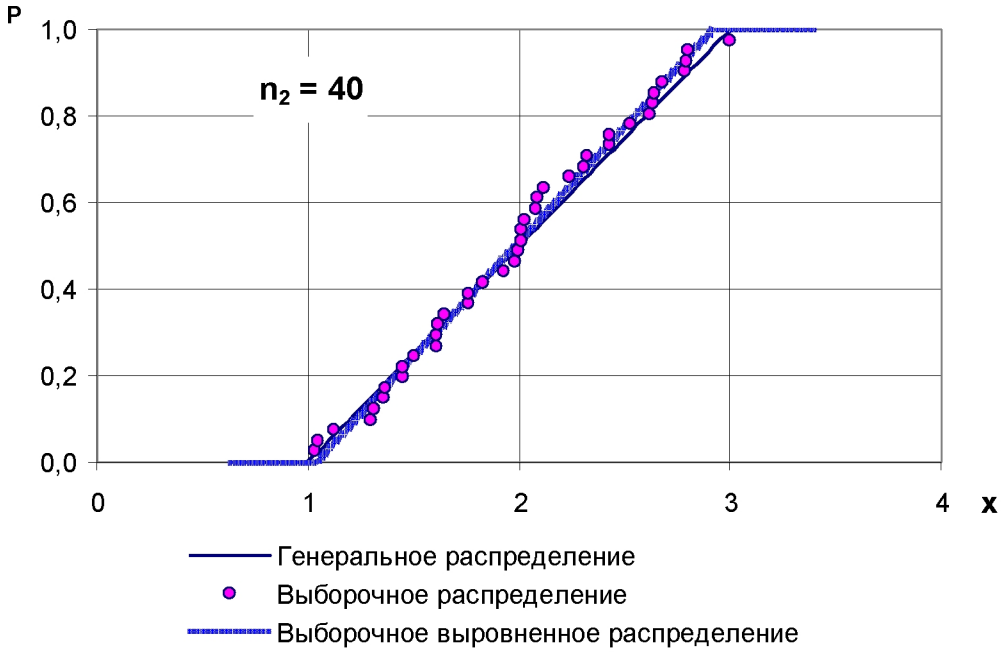

Рис. 3.6 - Выравнивание исходных данных законом равномерной плотности, n=40

Рис. 3.7 - Выравнивание исходных данных законом равномерной плотности, n=100

Следует отметить, что в рассматриваемом примере особенно для наименее представительной выборки объёмом n=15

$$\alpha^* \neq x_{\min} \quad \text{и} \quad \beta^* \neq x_{\max} \ .$$

Поэтому выравнивание равномерного закона распределения с использованием наблюдаемых наибольших и наименьших значений, а именно

$$\alpha^* = x_{\min} \quad \text{и} \quad \beta^* = x_{\max} \qquad (3.25,\ 3.26)$$

можно рекомендовать только при достаточно большом числе наблюдений.

В приведенных на рисунках 3.5, 3.6 и 3.7 примерах нетрудно заметить, что выборочная выравнивающая функция распределения намного лучше соответствует функции распределения генеральной совокупности данных, чем статистическая функция распределения, представленная отдельными точками.

4. ПРЕДВАРИТЕЛЬНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

4.1. Оценка математических ожиданий и средних квадратических отклонений

Одной из наиболее распространённых задач предварительного статистического анализа является оценка значений математического ожидания и дисперсии или среднего квадратического отклонения генеральной совокупности по результатам обработки выборки элементов. Эта задача имеет большое практическое значение по многим причинам, одной из которых является то, что эти параметры однозначно определяют нормальный закон распределения.

Методика расчёта математических ожиданий и средних квадратических отклонений зависит от объёма используемой выборки элементов. Различают средние и представительные выборки. Средними считаются выборки, объём которых составляет n = 12 – 30 элементов.

Представительные выборки имеют объём n = 200 – 300 элементов.

Математическое ожидание характеризует центр рассеивания случайных величин

$$m_x = \frac{\sum_{i=1}^{N} x_i}{N}$$, где N – объём генеральной совокупности. (4.1)

Дисперсия является мерой рассеивания случайных величин относительно математического ожидания

$$D_x^* = \frac{\sum_{i=1}^{n} (x_i - m_x)^2}{N} \quad . \qquad (4.2)$$

Использование дисперсии при решении практических задач зачастую неудобно из–за размерности этой величины. Например, если значения исследуемой величина имеют размерность кг, то размерность дисперсии соответствует кг^2 , что лишено смысла. Более удобной, в этом смысле является мера рассеивания в виде среднего квадратического отклонения, которое связано с дисперсией следующим ниже соотношением и имеет размерность исследуемой величины

$$s_x = \sqrt{D_x} \quad . \tag{4.3}$$

4.1.1. Оценка математических ожиданий и средних квадратических отклонений для средних выборок

Представим каждый результат наблюдения x_i графически на числовой оси, как это показано на рис. 4.1.

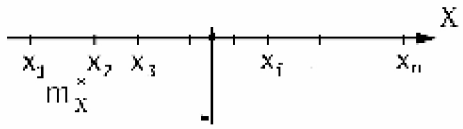

Рис. 4.1 – К вычислению статистической оценки математического ожидания

Статистическая оценка математического ожидания находится по формуле

$$m_x^* = \frac{\sum_{i=1}^{n} x_i}{n} \quad . \tag{4.4}$$

Статистическая оценка дисперсии определяется следующим образом

$$D_x^* = \left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(m_x^*\right)^2 \right] \cdot \frac{n}{n-1} \quad . \tag{4.5}$$

Следует отметить, что множитель n/(n–1) в формуле (4.5) устраняет смещение статистической оценки дисперсии, которое увеличивается при уменьшении объёма выборки.

Пример. Рассмотрим определение статистических оценок математического ожидания и среднего квадратического отклонения наблюдений, представленных в таблице 4.1.

Таблица 4.1 Обработка выборочных данных

i	x_i^*	$(x_i^*)^2$	i	x_i^*	$(x_i^*)^2$
1	10,42	108,63	11	15,81	249,98
2	11,49	132,09	12	16,24	263,66
3	12,22	149,34	13	16,69	278,67
4	12,8	163,92	14	17,2	295,74
5	13,31	177,06	15	17,78	316,11
6	13,76	189,41	16	18,51	342,51
7	14,19	201,33	17	19,58	383,28
8	14,6	213,12			
9	15	225	Σ	255	3927,04
10	15,4	237,2			

В таблице приведены данные 17 наблюдений. Оценка математического ожидания по формуле (4.4) производится следующим образом

$$m_x^* = \frac{255,00}{17} = 15,00 \quad .$$

Оценка дисперсии по формуле (4.5) имеет значение

$$D_x^* = \left[\frac{3927,04}{17} - 15,00^2 \right] \cdot \frac{17}{17-1} = 6,38 \quad .$$

В итоге получаем статистическую оценку среднего квадратического отклонения

$$s_x^* = \sqrt{6,38} = 2,53 \quad .$$

4.1.2. Оценка математических ожиданий и средних квадратических отклонений для представительных выборок

Оценка математического ожидания и среднего квадратическкского отклонения в случае представительной выборки начинается с предварительной обработки исходных данных. Для этого они распределяются по разрядам, как это показано на рис. 4.2, где x_1, x_2, … , x_{k+1} представляют собой границы разрядов.

Число разрядов k обычно выбирается в пределах 12 – 30, а сами разряды принимаются одинаковыми по ширине. Увеличение числа разрядов сверх указанного не даёт повышения точности статистических оценок и поэтому нецелесообразно. Также, как правило, не имеет смысла введение разрядов переменной ширины.

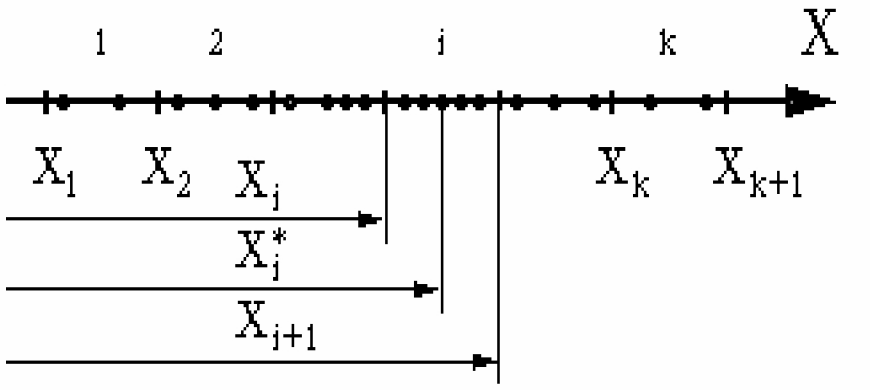

Рис. 4.2. - Выделение разрядов представительной выборки

Сначала для каждого разряда следует вычислить частоту появления в разряде

$$p_i^* = \frac{n_i}{N}, \quad \text{где} \quad N = \sum_{i=1}^{k} n_i \ . \tag{4.6}$$

После этого вычисляются характерные представители разрядов

$$x_i^* = \frac{x_i + x_{i+1}}{2} \ , \ i = 1,\ 2\ ,\ \ldots\ , k \ . \tag{4.7}$$

Оценка математического ожидания определяется по формуле $m_x^* = \sum_{i=1}^{k} x_i^* p_i^*$. (4.8)

Статистическая оценка дисперсии находится следующим образом

$$D_x^* = \left[\sum_{i=1}^{k} \left(x_i^* \right)^2 p_i^* \right] - \left(m_x^* \right)^2 \ . \tag{4.9}$$

Пример. Рассмотрим оценку математического ожидания и среднего квадратического отклонения результатов наблюдений, которые зарегистрированы в 15 разрядах и представлены в таблице 4.2.

Таблица 4.2 – Оценка выборочных характеристик

Исходные данные				Результаты вычислений			
i	x_i	x_{i+1}	ni	$p_i^*=n_i/N$	x_i^*	$x_i^* \cdot p_i^*$	$(x_i^*)^2 \cdot p_i^*$
1	30,00	31,00	2	0,0096	30,50	0,29	8,94
2	31,00	32,00	4	0,0192	31,50	0,61	19,08
3	32,00	33,00	7	0,0337	32,50	1,09	35,55
4	33,00	34,00	11	0,0529	33,50	1,77	59,35
5	34,00	35,00	19	0,0913	34,50	3,15	108,72
6	35,00	36,00	27	0,1298	35,50	4,61	163,59
7	36,00	37,00	33	0,1587	36,50	5,79	211,37
8	37,00	38,00	32	0,1538	37,50	5,77	216,35
9	38,00	39,00	28	0,1346	38,50	5,18	199,53
10	39,00	40,00	17	0,0817	39,50	3,23	127,52
11	40,00	41,00	12	0,0577	40,50	2,34	94,63
12	41,00	42,00	6	0,0288	41,50	1,20	49,68
13	42,00	43,00	5	0,0240	42,50	1,02	43,42
14	43,00	44,00	3	0,0144	43,50	0,63	27,29
15	44,00	45,00	2	0,0096	44,50	0,43	19,04
Σ			208	1,0000	Σ	37,11	1384,07

Суммирование числа наблюдений по всем разрядам даёт общее число опытов N=208.

Заполнение вычисляемой части таблицы определяется заголовками соответствующих столбцов. Поразрядное суммирование частот даёт 1,0000 , что используется для контроля правильности вычислений.

Суммирование элементов столбца $x_i^* p_i^*$ позволяет получить оценку математического ожидания

$$m_x^* = 37{,}106 \ .$$

Статистическая оценка дисперсии даёт следующий результат

$$D_x^* = 1384{,}07 - (37{,}106)^2 = 7{,}229 \ .$$

Статистическая оценка среднего квадратического отклонения имеет значение

$$s_x^* = \sqrt{7{,}229} = 2{,}689 \ .$$

4.2. Построение статистических функций распределения на нормально–вероятностной бумаге

Статистическая функция распределения строится по результатам наблюдений, она имеет следующее теоретическое определение

$$F^*(x) = P^*(X < x) \ , \tag{4.10}$$

т.е. для каждого значения аргумента эта функция даёт вероятность того, что рассматриваемая случайная величина будет меньше заданного значения аргумента.

Смысл функции распределения можно проиллюстрировать с помощью графика, представленного на рисунке 4.3, из которого, например, видно, что случайная величина, для которой построен этот график, может принимать значения меньше чем 15,2 с вероятностью

$$P = P\,(X<15{,}2) = F\,(15{,}2) = 0{,}82 \ .$$

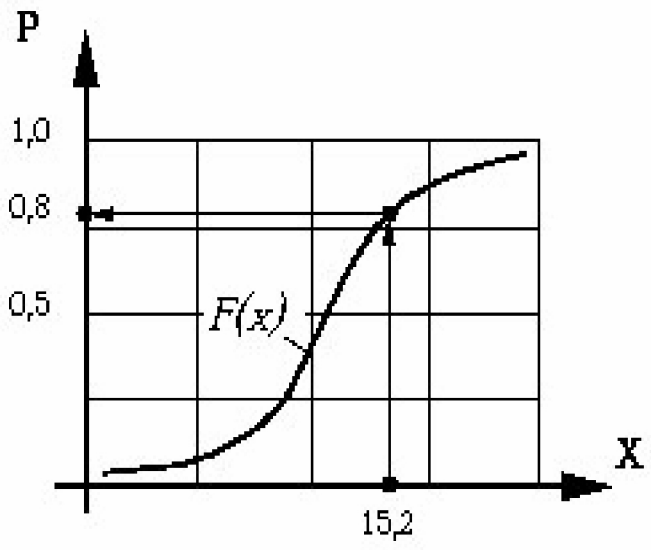

Рис. 4.3 – Пример использования функции распределения

Функция распределения, являясь одной из форм закона распределения, представляет собой наиболее общую характеристику рассматриваемой случайной величины, которая содержит в себе информацию обо всех её статистических свойствах.

Оценка функции распределения является весьма распространённой на практике процедурой предварительного статистического анализа так, как её значения используются в качестве вероятностей событий различного физического содержания.

Методика построения статистических функций распределения зависит от объёма используемых выборок элементов. Здесь, как и в предыдущем разделе, различают средние выборки объёмом $n = 12 – 30$ элементов и большие или представительные выборки, которые имеют объём $n = 200 – 300$ элементов.

Для предварительного анализа статистической функции распределения удобно пользоваться её графическим представлением. Поскольку вследствие центральной предельной теоремы подавляющее большинство случайных величин, характеризующих реальные явления, подчиняется нормальному закону распределения, то первой задачей предварительного статистического анализа является сопоставление полученной статистической функции распределения с функцией нормального закона, описываемого зависимостью

$$F(x) = \frac{1}{s_x\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-m_x)^2}{2s_x^2}} dx \ , \qquad (4.11)$$

где m_x – математическое ожидание, s_x – среднее квадратическое отклонение.

Непосредственно построив графики нормального закона распределения и статистической функции распределения, затруднительно судить об их взаимном соответствии, что видно, например, из рисунка 4.4.

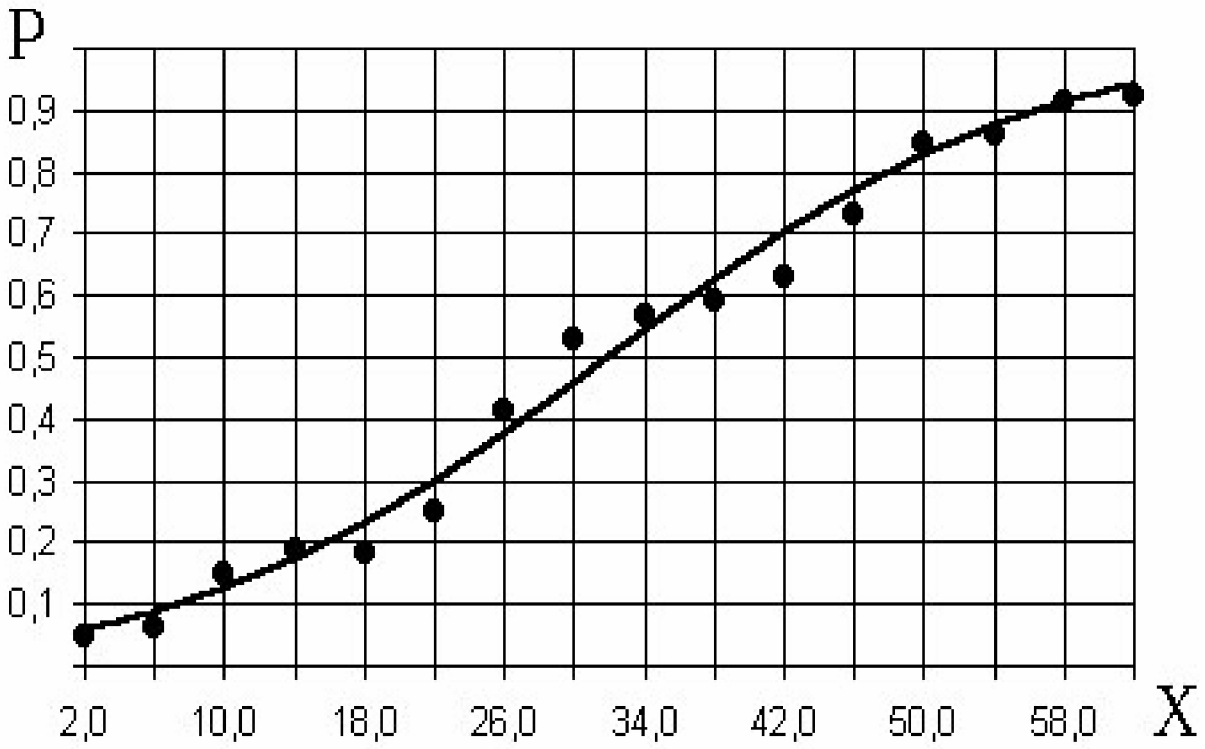

Рис. 4.4 - Построение функции распределения в линейных координатах

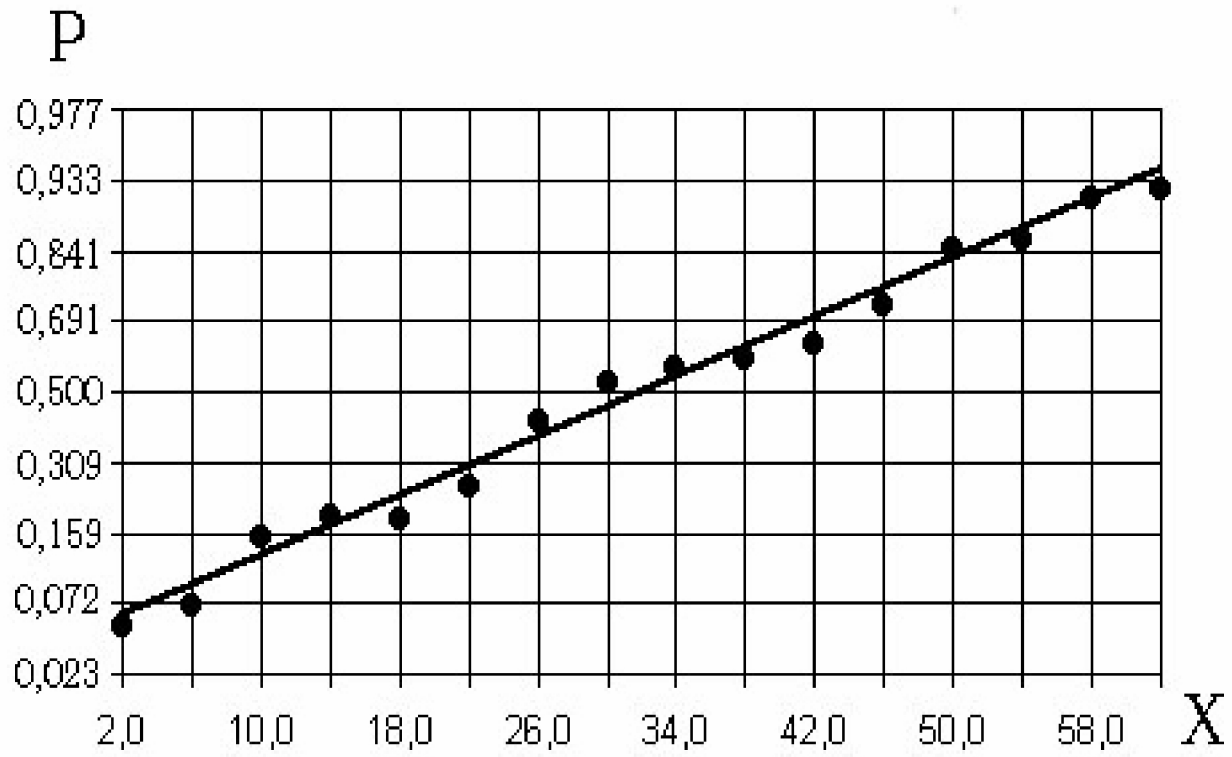

Рис. 4.5 - Построение функции распределения с неравномерным масштабом координаты ***P***

Задачу предварительной визуальной оценки соответствия нормальному закону распределения значительно облегчает использование нормально–вероятностной бумаги, на которой за счет неравномерного масштаба оси ординат график нормальной функции распределения представляется в виде прямой линии, как показано на рис. 4.5 для тех же данных, которые использовались при построении графика на рис. 4.4.

Чтобы неравномерно разметить ось ординат, на которой откладываются вероятности P, вводится вспомогательная вертикальная ось Up , изображаемая справа.

Ось Up имеет равномерную разбивку и на ней откладывается квантиль нормального распределения, определяемый по следующей формуле

$$u_p = \frac{x - m_x}{s_x} \quad , \tag{4.12}$$

где m_x – математическое ожидание, s_x – среднее квадратическое отклонение.

Значения u_p можно взять из таблицы нормального распределения П.1, которая представлена в приложении. Для этого следуем иметь в виду, что табличному значению F (x)=P соответствует квантиль u_p=x, если P >0,5 и u_p= –x, если P<0,5.

Откладывая значения функции нормального распределения в координатах x и Up , мы всегда получим прямую линию, что используется для разграфления вероятностных бумаг.

Чтобы сделать шкалу квантилей существенно положительной обычно увеличивают их значения, вводя новое обозначение

$$U_p + 5 = u_p + 5 \quad . \tag{4.13}$$

Пример использования нормально–вероятностной бумаги показан на рисунке 4.6, а сама нормально–вероятностная бумага представлена в приложении на рисунке П.8.

Рассмотрим построение функции нормального закона распределения на нормально–вероятностной бумаге с заданными значениями математического ожидания и среднего квадратического отклонения m_x и s_x показанное на рисунке 4.6.

Сначала воспользуемся значением математического ожидания и обозначим положение точки M(m_x, 5), воспользовавшись правой осью ординат U_p+5. Очевидно, что эта же точка может быть построена с использованием правой оси ординат P, тогда она имеет координаты M(m_x, 50).

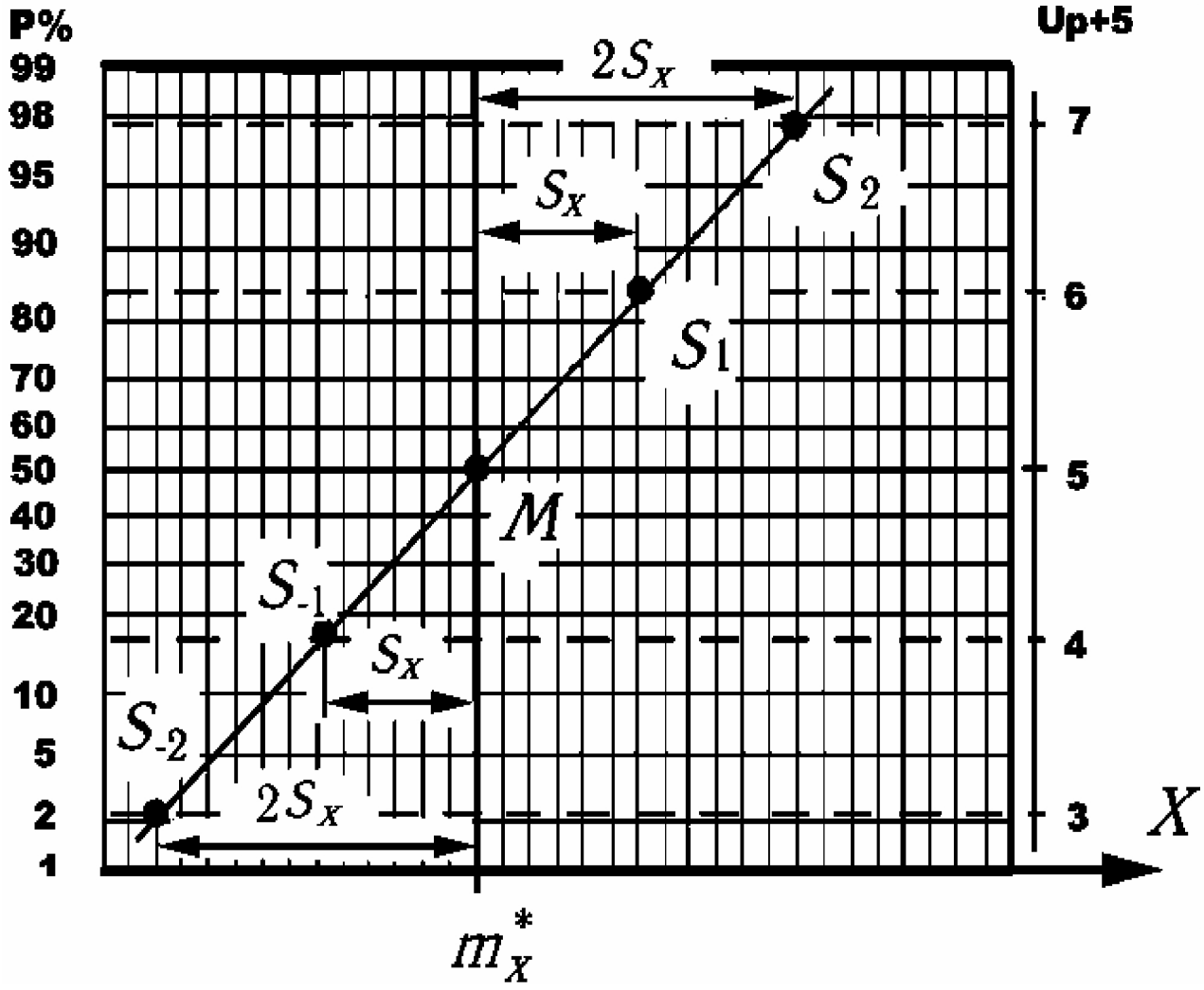

Рис. 4.6 – Использование нормально-вероятностной бумаги

Из выражения (4.12) следует, что перемещение от точки M вправо или влево на величину кратную среднему квадратическому отклонению, т.е. $m_x \pm s_x$, $m_x \pm 2s_x$ и т.д. соответствует приращению по шкале квантилей ±1, ±2. Отсюда ясно, как построить точки S_1, S_{-1}, S_2, S_{-2}.

Функция нормального закона распределения будет изображаться на рисунке 4.6 в виде прямой линии, проходящей через точки M, S_1, S_{-1}, S_2, S_{-2}. При проведении прямой линии достаточно любых двух точек, но для контроля правильности построения желательно использовать три точки, например, M, S_1, S_{-1}, или M, S_{-2}, S_2.

4.2.1. Средняя выборка

Статистическая функция распределения для средней выборки строится по точкам на основании следующей зависимости

$$F^*(x_i) = P_i^* = \frac{i}{n+1} \ , \qquad (4.14)$$

где i – номер наблюдения, полученный после сортировки результатов наблюдения по возрастанию числовых значений, n – общее число наблюдений.

Пример. Рассмотрим построение статистической функции распределения для средней выборки, которая в отсортированном по возрастанию порядке представлена в таблице 4.3, где также приведены результаты вычисления статистических вероятностей P_i^* и промежуточные данные, необходимые для оценки математического ожидания и среднего квадратического отклонения по формулам (4.4, 4.5, 4.3).

Таблица 4.3 - Обработка выборочных данных

i	x_i^*	$(x_i^*)2$	P_i^*	i	x_i^*	$(x_i^*)2$	P_i^*
1	4,74	22,47	0,063	9	6,15	37,82	0,563
2	5,02	25,2	0,125	10	6,31	39,82	0,625
3	5,27	27,77	0,188	11	6,42	41,22	0,688
4	5,45	29,7	0,25	12	6,54	42,77	0,75
5	5,56	30,91	0,313	13	6,78	45,97	0,813
6	5,67	32,15	0,375	14	6,92	47,89	0,875
7	5,9	34,81	0,438	15	7,22	52,13	0,938
8	6,05	36,6	0,5	Σ	90	547,23	

Вычисления по ранее приведенным формулам дают следующие результаты

$$m_x^* = \frac{90,00}{15} = 6,00 \ , \qquad s_x^* = \sqrt{\frac{15}{15-1}\left[\frac{547,23}{15} - (6,00)^2\right]} = 0,719 \ .$$

Статистическая функция распределения $F^*(x_i)=P^*_i$, построенная по данным таблицы 4.3, представлена на рисунке 4.7 в виде ломаной линии, а прямая линия соответствует нормальному закону распределения с математическим ожиданием $m_x=6,00$ и средним квадратическим отклонением $s_x=0,719$. График нормального закона распределения построен на рисунке 4.7 по точкам S_{-2}, S_{-1}, M, S_1, S_2 .

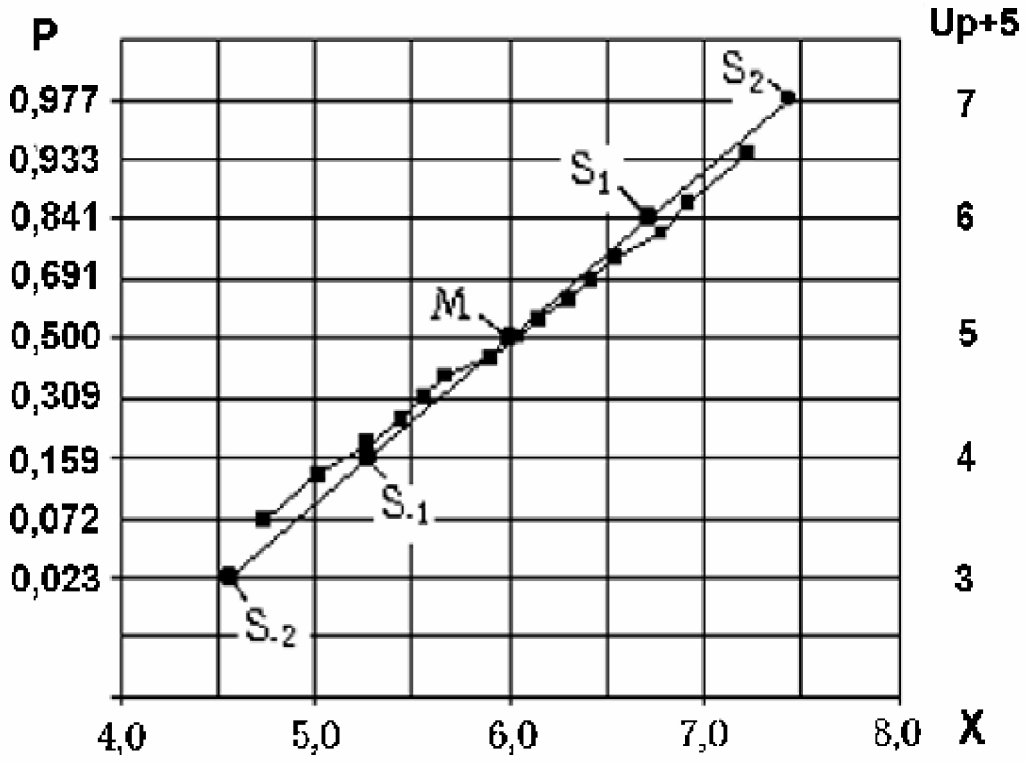

Рис. 4.7 – Построение функции распределения на нормально-вероятностной бумаге

Положение точек S_{-2}, S_{-1}, M, S_1, S_2 в координатах X, U_p+5 определяется следующим образом

$S_{-2}(6,00 - 2* 0,719;\ 3,0)= S_{-2}(4,562;\ 3,0)$,

$S_{-1}\ (6,00 - 0,719;\ 4,0)= S_{-1}(5,281;\ 4,0)$,

$M(6,00;\ 5,0)$,

$S_1(6,00 + 0,719;\ 6,0)= S_1(6,719;\ 6,0)$,

$S_2(6,00 + 2*0,719;\ 7,0)= S_2(7,438;\ 7,0)$.

4.2.2. Представительная выборка

Построение статистической функции распределения для представительной выборки требует предварительной обработки исходных данных. Сначала результаты опытов распределяются по разрядам, как это было показано на рис. 4.2, где x_1 , x_2, … , x_{k+1} представляют собой границы разрядов.

Число разрядов k обычно принимается в пределах 12 – 30. Увеличение числа разрядов сверх указанного не приводит к повышению точности статистических оценок. Также не имеет смысла введение разрядов переменной ширины.

Частота появления рассматриваемого события в разряде определяется как

$$p_i^* = \frac{n_i}{N}, \quad \text{где} \quad N = \sum_{i=1}^{k} n_i \quad . \tag{4.15}$$

Значения статистической функции распределения определяются суммированием частот во всех разрядах, начиная с первого и кончая рассматриваемым разрядом

$$F^*(x_{i+1}) = P_i^* = \sum_{j=1}^{i} p_j^* \quad , \; i = 1,\ 2,\ ...,\ k \quad . \tag{4.16}$$

Следует отметить, что значения статистической функции распределения, получаемые по формуле (1.16) соответствуют правой границе рассматриваемого разряда.

Пример. Рассмотрим построение статистической функции распределения для представительной выборки, содержащей результаты 240 опытов, зафиксированных в 15 разрядах, представленные в таблице 4.4, где также приведены результаты вычисления статистических вероятностей P_i^* и промежуточные данные, необходимые для оценки математического ожидания и среднего квадратического отклонения по формулам (4.15, 4.16, 4.7, 4.8, 4.9, 4.3).

Статистические оценки математического ожидания и среднего квадратического отклонения, полученные по методике, изложенной в разделе 4.1.2, имеют следующие значения

$$m_x^* = 34,72 \quad , \; s_x^* = \sqrt{1230,84 - (34,72)^2} = 5,02 \quad .$$

На рисунке 4.8 точки вида ■, соответствуют значениям статистической функция распределения $F^*(xi+1)=P^*_i$, построенной по данным таблицы 4.4. Прямая линия является изображением нормального закона распределения с математическим ожиданием $m_x=34,72$ и средним квадратическим отклонением $s_x=5,02$.

Таблица 4.4 – Обработка выборочных данных

i	x_i	x_{i+1}	N_i	$p_i^*=n_i/N$	x_i^*	$x_i^* \cdot p_i^*$	$(x_i^*)^2 \cdot p_i^*$	$P_i^*=\Sigma p_j$
1	20,0	22,0	1	0,0062	21,0	0,13	2,74	0,0062
2	22,0	24,0	3	0,0125	23,0	0,29	6,60	0,0187
3	24,0	26,0	6	0,0249	25,0	0,62	15,59	0,0436
4	26,0	28,0	11	0,0457	27,0	1,23	33,34	0,0894
5	28,0	30,0	19	0,0790	29,0	2,29	66,44	0,1684
6	30,0	32,0	28	0,1164	31,0	3,61	111,89	0,2848
7	32,0	34,0	38	0,1580	33,0	5,21	172,07	0,4428
8	34,0	36,0	39	0,1622	35,0	5,68	198,65	0,6050
9	36,0	38,0	35	0,1455	37,0	5,38	199,23	0,7505
10	38,0	40,0	25	0,1040	39,0	4,05	158,11	0,8545
11	40,0	42,0	17	0,0707	41,0	2,90	118,83	0,9252
12	42,0	44,0	10	0,0416	43,0	1,79	76,88	0,9667
13	44,0	46,0	5	0,0208	45,0	0,94	42,10	0,9875
14	46,0	48,0	2	0,0083	47,0	0,39	18,37	0,9958
15	48,0	50,0	1	0,0042	49,0	0,20	9,98	1,0000
Σ			240	1,0000	Σ	34,72	1230,84	

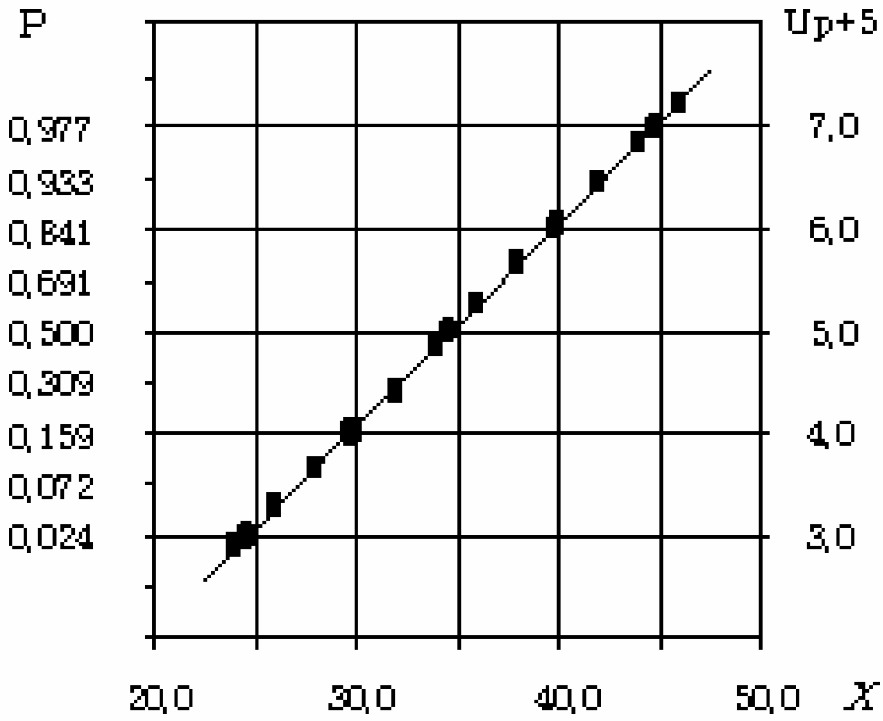

Рис. 4.8 – Выравнивание статистической функции распределения

График нормального закона распределения построен по точкам S_{-2}, S_{-1}, M, S_1, S_2, обозначенным как • , положение которых в координатах X, U_p+5 определяется следующим образом:

$S_{-2}(34,72-2*5,02;\ 3,0) = S_{-2}(24,68;\ 3,0)$,

$S_{-1}(34,72-5,02;\ 3,0) = S_{-1}(29,70;\ 4,0)$,

M(34,72; 5,0),

$S_1(34,72+5,02;\ 3,0) = S_1(39,74;\ 6,0)$,

$S_2(34,72+2*5,02;\ 7,0) = S_2(44,76;\ 7,0)$.

5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

5.1. Проверка гипотезы нормальности статистической функции распределения

Статистические функции распределения, являясь одной из форм закона распределения, содержат всю информацию, которая необходима для оценки любых числовых характеристик исследуемых случайных величин. Функция распределения определяется соотношением

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

В предыдущем разделе подробно рассматривалось построение этих функций на основе средних и представительных выборок. При статистическом анализе используемых данных весьма важно решить вопрос о соответствии полученной статистической функции распределения какому–либо из известных законов распределения.

Поскольку наибольший практический интерес представляет нормальный закон распределения, то поэтому в настоящем пособии рассматривается вопрос определения вероятности соответствия наблюдаемых результатов нормальному закону распределения.

Поставленный вопрос методически может быть решён разными средствами. В настоящем пособии демонстрируется использование критерия согласия Колмогорова–Смирнова для средних выборок и критерия согласия Пирсона для представительных выборок.

Предварительно познакомимся с принципиальной стороной применения критериев согласия. Допустим, что статистическая функция распределения F*(x) заменяется или, как ещё говорят, выравнивается с помощью теоретической функции распределения F(x). Этого можно добиться, если в качестве параметров теоретической функции использовать оценки соответствующих параметров, полученные при обработке опытных данных. Так, если определены оценки математического ожидания m_x^* и среднего квадратического отклонения s_x^* (см. раздел 3.3), то функция нормального закона распределения, "выравнивающая" статистическое распределение определяется выражением

$$F(x) = \frac{1}{s_x^* \sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-m_x^*)^2}{2(s_x^*)^2}} dx \, .$$

Общим для статистической функции распределения и для "выравнивающей" теоретической функции будет то, что у них совпадают значения математического ожидания и средних квадратических отклонений. Если построить графики этих функций, то они не будут совпадать, как это, например, показано на рис. 5.1.

При увеличении числа наблюдений статистическая функция распределения приближается к соответствующему теоретическому распределению.

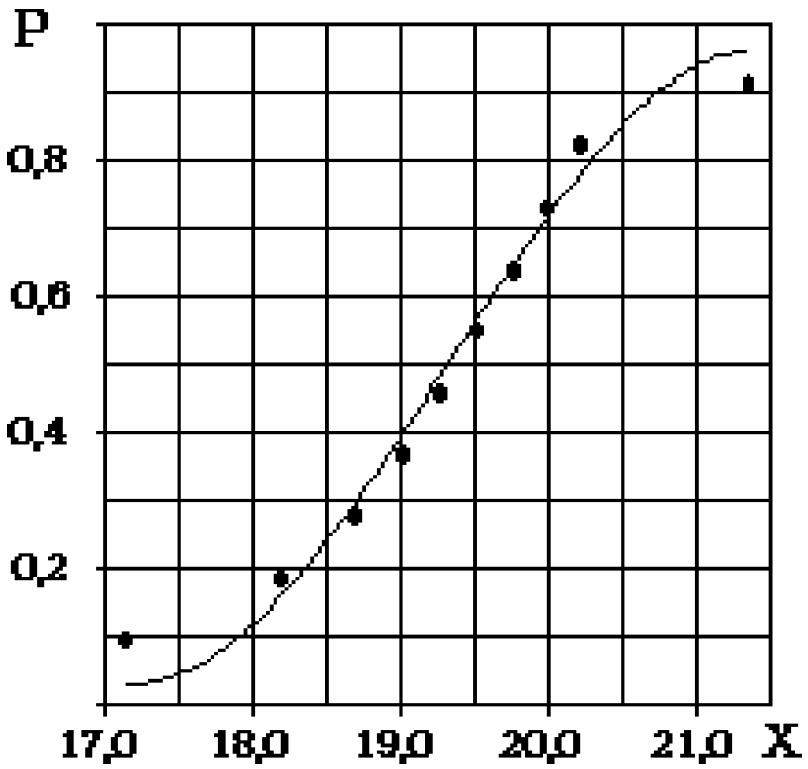

Рис. 5.1 – К проверке нормальности распределения

Несовпадение функций F(x) и F*(x) может иметь различные причины. Во–первых, статистическая функция F*(x) строится с использованием выборки данных из некоторой генеральной совокупности. Так как элементы выборки не представляют всех элементов генеральной совокупности, то оценки характеристик m_x^* , s_x^* и сама функция F*(x) имеют случайные отклонения от истинных значений m_x, s_x, F(x) для генеральной совокупности.

Во–вторых, получение и регистрация выборочных данных x_1, x_2, … , x_n на физическом уровне (определение размеров, взвешивание и т.п.) всегда сопровождается случайными отклонениями, погрешностями и ошибками.

В этой связи встаёт вопрос о том, насколько случайны отклонения статистической функции от теоретической функции распределения. Возможно, что это отклонение имеет неслучайные причины, обусловленные использованием несоответствующего теоретического закона распределения. Для ответа на такой вопрос служат так называемые "критерии согласия".

Применение критерием согласия заключается в следующем. Допустим, предстоит проверить гипотезу H о том, что случайная величина X распределена по определённому, например, по нормальному закону. Может быть, это отклонение имеет неслучайные причины, связанные с тем, что для выравнивания статистических данных используется неподходящий закон распределения. Для ответа на такой вопрос служат так называемые "критерии согласия".

Применение критериев согласия заключается в следующем.

Для оценки вероятности реализации гипотезы H введём величину U, характеризующую степень расхождения теоретического и статистического распределений. Эта величина может быть выбрана различным образом, два варианта её выбора будут рассмотрены ниже в следующих далее разделах.

Очевидно, что величина U является случайной величиной, закон распределения которой зависит от закона распределения исследуемой величины и от общего числа опытов. Если этот закон известен, то можно вычислить доверительную вероятность

$$\beta = P(U < u) \ ,$$

которая равна вероятности того, что за счёт только случайных отклонений расхождение между статистическим и теоретическим распределением будет больше, чем это имело место в проведенных опытах.

Перейдём к вероятности противоположного события

$$\alpha = 1 - \beta \ ,$$

которая оценивает уровень значимости полученных расхождений и является вероятностью того, что наблюдаемое расхождение имеет неслучайные причины. Вероятность α называется уровнем значимости расхождений.

Получив расчетным путём достаточно высокие значения доверительной вероятности (β>0,8), можно принять гипотезу о приемлемости выбранного теоретического закона распределения. Напротив, если большое значение имеет уровень значимости расхождений (α>0,8), то следует отвергнуть обсуждаемую гипотезу как малоправдоподобную.

Какие именно значения доверительных вероятностей или уровней значимости отклонений считать достаточно большими, зависит от физического содержания проводимых опытов и от принимаемой меры ответственности использования полученных результатов.

Получив недостаточно высокие значения доверительных вероятностей при оценке приемлемости гипотезы о нормальности закона распределения рассматриваемой случайной величины, зачастую при выполнении последующего статистического анализа пользуются методами, которые специально разработаны для нормального закона распределения. То

есть малоправдоподобная гипотеза продолжает использоваться при выполнении последующих вычислений. В этих случаях получаемые результаты будут приближёнными.

Высокие значения доверительной вероятности β позволяют при последующем анализе отказаться от статистической функции распределения и воспользоваться выравнивающей теоретической функцией распределения, которая очевидно будет больше соответствовать генеральной совокупности значений рассматриваемой величины. Выбранная теоретическая функция распределения имеет плавный характер и может быть успешно продолжена на основании известных зависимостей, как в область малых, так и в область больших вероятностей, до которых не доходит статистическая функция распределения вследствие ограниченности объёма используемой выборки элементов.

5.1.1. Проверка гипотезы нормальности статистической функции распределения для средних выборок

А.Н.Колмогоров и Н.В.Смирнов предложили использовать максимальное значение модуля разности между статистической функцией распределения F*(x) и соответствующей выравнивающей теоретической функцией распределения F(x) в качестве меры расхождения

$$D = \max\left|F^*(x_i) - F(x_i)\right| \ , \ i = 1,2,...,n \ . \qquad (5.1)$$

А.Н.Колмогоров и Н.В.Смирнов доказали, что какова бы ни была функция распределения непрерывной случайной величины X , при увеличении числа независимых опытов имеет место следующее соотношение

$$\beta = P(\lambda) = P(D\sqrt{n} \geq \lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2} \ . \qquad (5.2)$$

Значения вероятности P(λ), подсчитанные по формуле (5.2) в работе [1] , представлены в приложении в таблице П.2.

Пример. Рассмотрим среднюю выборку объёмом n=20 элементов, приведенную в таблице 5.1. Результаты опытов прошли предварительную сортировку в возрастающем порядке.

В таблице приведены также результаты промежуточных вычислений, которые необходимы для определения статистических оценок математического ожидания и среднего квадратического отклонения (подробнее смотри раздел 4.1.1).

$$m_x^* = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{632,23}{20} = 31,61 \ .$$

$$s_x^* = \sqrt{\frac{n}{n-1}\left[\frac{\sum_{i=1}^{n}(x_i)^2}{n} - (m_x^*)^2\right]} = \sqrt{\frac{20}{20-1}\left[\frac{20182,46}{20} - (31,61)^2\right]} = 3,218 \ .$$

Таблица 5.1 – Обработка выборочных данных

i	**x_i***	**$(x_i^*)^2$**	**P_i***	**P_i**	**$\lvert P_i - P_i^* \rvert$**
1	26,46	700,13	0,0476	0,0547	0,0071
2	28,20	795,24	0,0952	0,1445	0,0493
3	28,70	823,69	0,1429	0,1828	0,0399
4	29,01	841,58	0,1905	0,2094	0,0189
5	29,28	857,32	0,2381	0,2344	0,0037
6	29,53	872,02	0,2857	0,2589	0,0269
7	29,77	886,25	0,3333	0,2836	0,0498
8	30,00	900,00	0,3810	0,3082	0,0727
9	30,22	913,25	0,4286	0,3327	0,0959
10	30,46	927,81	0,4762	0,3602	0,1160
11	30,71	943,10	0,5238	0,3897	0,1341
12	31,20	973,44	0,5714	0,4491	0,1223
13	32,00	1024,00	0,6190	0,5481	0,0710
14	32,50	1056,25	0,6667	0,6088	0,0579
15	33,42	1116,90	0,7143	0,7130	0,0013
16	34,30	1176,49	0,7619	0,7983	0,0364
17	34,90	1218,01	0,8095	0,8466	0,0371
18	35,25	1242,56	0,8571	0,8709	0,0138
19	37,15	1380,12	0,9048	0,9574	0,0526
20	39,17	1534,29	0,9524	0,9906	0,0382
Σ	632,23	20182,46	--	--	--

Значения статистической функции распределения определяются соотношением

$$F^*(x_i) = P_i^* = \frac{i}{n+1} \quad , \tag{5.3}$$

где i – порядковый номер результата опыта x_i в отсортированном по возрастанию списке, n–общее число опытов.

Результаты вычисления статистической функции распределения представлены в таблице 5.1, а полученные значения функции изображены точками на рисунке 5.2.

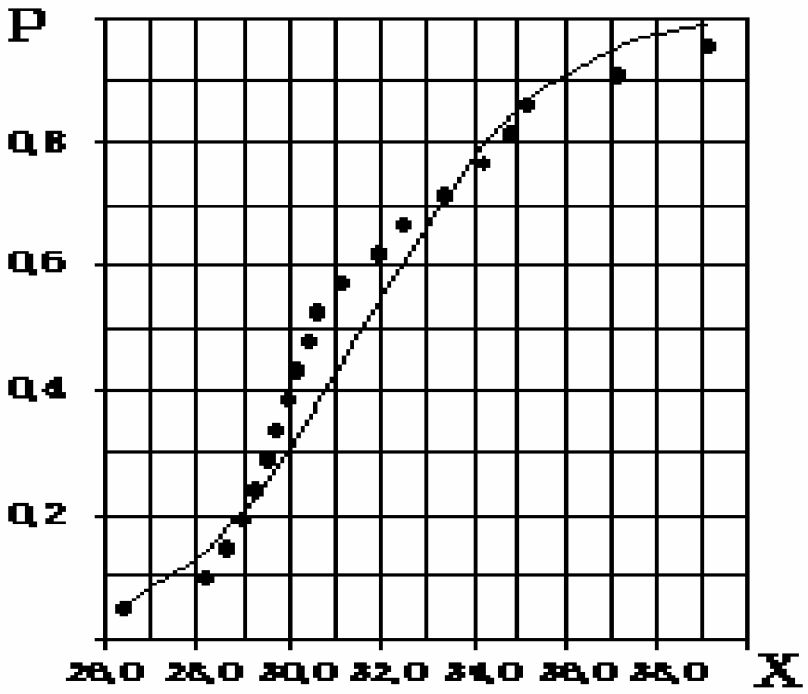

Рис. 5.2 – Проверка нормальности статистического распределения

Выравнивающая теоретическая функция распределения строится с помощью таблицы нормального распределения П.1 из приложения. Чтобы воспользоваться этой таблицей следует вычислить нормализованные значения аргументов по формуле

$$t_i = \frac{x_i - m_x^*}{s_x^*} \ , \ i = 1,\ 2,\ ...,\ n \ . \tag{5.4}$$

После этого нужно воспользоваться значениями функции нормального распределения, взятыми из таблицы П.1, которые будут равны значениям статистической функции распределения

$$P_i = F(x_i) \qquad \text{при } x_i = t_i \ , \qquad i = 1,2,...,n \ . \tag{5.5}$$

Следует обратить внимание на отсутствие в таблице отрицательных аргументов. В этом случае следует воспользоваться соотношением $F(-t_i)=1-F(t_i)$.

Результаты определения теоретической функции распределения P_i представлены в таблице 5.1 и изображены на рисунке 5.2 в виде сплошной линии.

Модули разности статистической и теоретической функций распределения $\Delta_i=|P_i-P_i^*|$ представлены в таблице 5.1, откуда видно, что максимальное расхождение составляет

$$D = \max\left|P^*_i - P_i\right| = 0{,}1341 .$$

При таком расхождении найдём значение параметра

$$\lambda = D\sqrt{n} = 0{,}5999 \quad .$$

По таблице П.2. определим значение доверительной вероятности

$$\beta = P(\lambda) = 0{,}864 \quad .$$

Соответствующий уровень значимости расхождений составляет

$$\alpha = 1 - \beta = 0{,}136 \quad .$$

Сравнительно высокая доверительная вероятность $\beta=0{,}864$ показывает, что для представленных в таблице 2.1 данных можно с большой уверенностью использовать нормальный закон распределения. На первый взгляд это противоречит картине, изображённой на рис. 5.2, где видно существенное различие значений $F(x_i)$ и $F^*(x_i)$. Однако, здесь следует иметь в виду, что для построения статистической функции распределения $F^*(x)$ использовалась средняя выборка объёмом $n=20$ элементов.

Если, например, наблюдаемое расхождение $D=0{,}1341$ имело бы место при числе опытов $n=100$, то получим $\lambda=1{,}341$, что соответствует доверительной вероятности $\beta=0{,}0551$, а это говорит о совершенной непригодности нормального закона распределения при числе опытов $n=100$, так как в этом случае уровень значимости расхождений составил бы $\alpha=0{,}9449$.

5.1.2. Проверка гипотезы нормальности статистической функции распределения для представительных выборок

При наличии представительной выборки для оценки нормальности статистической функции распределения часто используется критерий К.Пирсона, который здесь даёт несколько лучшие результаты, чем критерий Колмогорова–Смирнова [2].

В качестве меры расхождения распределений Пирсон предложил использовать величину

$$U = N\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} \quad , \qquad (5.6)$$

где k – число разрядов, используемое для предварительного анализа при вычислении статистических оценок математического ожидания m_x^* и среднего квадратического отклонения s_x^* (подробнее смотри раздел 4.1.2),

$N = \sum_{i=1}^{k} n_i$ – общее число опытов, n_i – число опытов, зафиксированное в i–м разряде,

$p_i^* = \frac{n_i}{N}$ – частота появления в i–м разряде (статистическая оценка вероятности),

p_i – вероятность появления события в данном разряде в соответствии с выбранным теоретическим законом распределения.

К.Пирсон показал, что величина U имеет распределение Пирсона, называемое часто распределением хи-квадрат χ^2, которое зависит от числа "степеней свободы"

$$r = k - s,$$

где k – число разрядов, s – число связей, определяемое при использовании в качестве теоретического закона нормального закона распределения как s = 3 , откуда

$$r = k - 3. \qquad (5.7)$$

Значения распределения Пирсона $\chi 2$ представлены в приложении в таблице П.3. в виде $\chi^2 = \chi^2(p, r)$.

Имея вычисленное значение U, найдём в таблице П.3 ближайшее к нему число $\chi_2 = U$ в столбце с конкретным значением r. В выбранной строке слева находится значение вероятности p, которое равно значению доверительной вероятности приемлемости нормального закона распределения β. Пример использование таблицы П.3 схематично показан на рисунке 5.3.

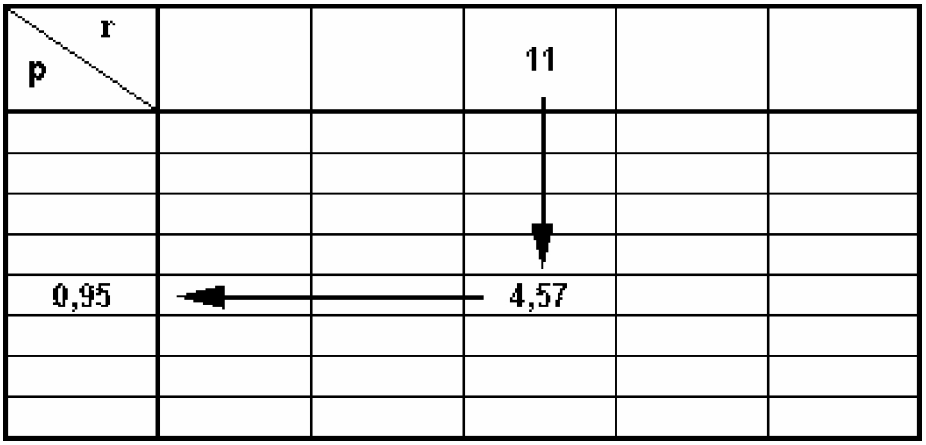

Рис. 5.3 – Схема определения вероятности β

Получив значение β, найдём уровень значимости расхождений как

$$\alpha = 1 - \beta.$$

Следует отметить, что в отличие от критерия согласия Колмогорова–Смирнова критерий согласия Пирсона можно использовать для любых теоретических законов распределения, но при этом следует учесть возможные отличия в определении числа связей s и соответствующим образом вычислить значения теоретических вероятностей p_i.

Пример. Рассмотрим оценку приемлемости нормального закона распределения для представительной выборки имеющей объём N=255 элементов, результаты предварительной обработки которой в виде числа опытов n_i , распределённых в 15 разрядах с фиксированными границами x_i , x_{i+1} при i=1, 2, … 15 приведены в таблице 5.2.

Таблица 5.2

i	x_i	x_{i+1}	n_i	p_i*	P_i*	t_i	t_{i+1}	$F(t_i)$	$F(t_{i+1})$	p_i	N·pi	u_i
1	35	37	1	0,0039	0,0039	–3,089	–2,702	0,0010	0,0034	0,0024	0,621	0,2311
2	37	39	2	0,0078	0,0118	–2,702	–2,316	0,0034	0,0103	0,0068	1,744	0,0373
3	39	41	4	0,0157	0,0275	–2,316	–1,929	0,0103	0,0269	0,0166	4,228	0,0123
4	41	43	8	0,0314	0,0588	–1,929	–1,542	0,0269	0,0615	0,0347	8,836	0,0791
5	43	45	18	0,0706	0,1294	–1,542	–1,155	0,0615	0,1240	0,0625	15,929	0,2693
6	45	47	28	0,1098	0,2392	–1,155	–0,768	0,1240	0,2211	0,0971	24,768	0,4216
7	47	49	23	0,0902	0,3294	–0,768	–0,381	0,2211	0,3514	0,1303	33,221	3,1448
8	49	51	40	0,1569	0,4863	–0,381	0,005	0,3514	0,5021	0,1507	38,435	0,0637
9	51	53	43	0,1686	0,6549	0,005	0,392	0,5021	0,6525	0,1504	38,357	0,5619
10	53	55	32	0,1255	0,7804	0,392	0,779	0,6525	0,7820	0,1295	33,019	0,0315
11	55	57	28	0,1098	0,8902	0,779	1,165	0,7820	0,8782	0,0962	24,518	0,4943
12	57	59	12	0,0471	0,9373	1,165	1,552	0,8782	0,9398	0,0616	15,704	0,8738
13	59	61	9	0,0353	0,9725	1,552	1,939	0,9398	0,9738	0,0340	8,676	0,0121
14	61	63	5	0,0196	0,9922	1,939	2,326	0,9738	0,9900	0,0162	4,134	0,1811
15	63	65	2	0,0078	1,0000	2,326	2,713	0,9900	0,9967	0,0067	1,699	0,0531
$N=\Sigma n_i = 255$											$U=\Sigma u_i=$	7,316

Статистические оценки математического ожидания и среднего квадратического отклонения в соответствии с методикой раздела 1.2 вычисляются по следующим формулам

$$m_x^* = \sum_{i=1}^{15} x_i^* p_i^* ,$$

$$s_x^* = \sqrt{\sum_{i=1}^{15} x_i^* p_i^* - (m_x^*)^2} ,$$

где $x_i^* = \frac{x_i + x_{i+1}}{2}$, $p_i^* = \frac{n_i}{N}$, $N = \sum_{i=1}^{15} n_i = 255.$

В результате вычислений получены значения статистических оценок

$m_x^* = 50{,}97$ и $s_x^* = 5{,}170$.

Статистическая функция распределения определяется суммированием частот появления в данном и предшествующих разрядах

$$P_i^* = F^*(x_{i+1}) = \sum_{j=1}^{i} p_j^* \quad . \tag{5.8}$$

Значения статистической функции распределения представлены в таблице 5.2 и изображены в виде отдельных точек на рис. 5.4. При этом следует иметь в виду, что полученные значения статистической функции на графике функции откладываются на правых границах соответствующих интервалов.

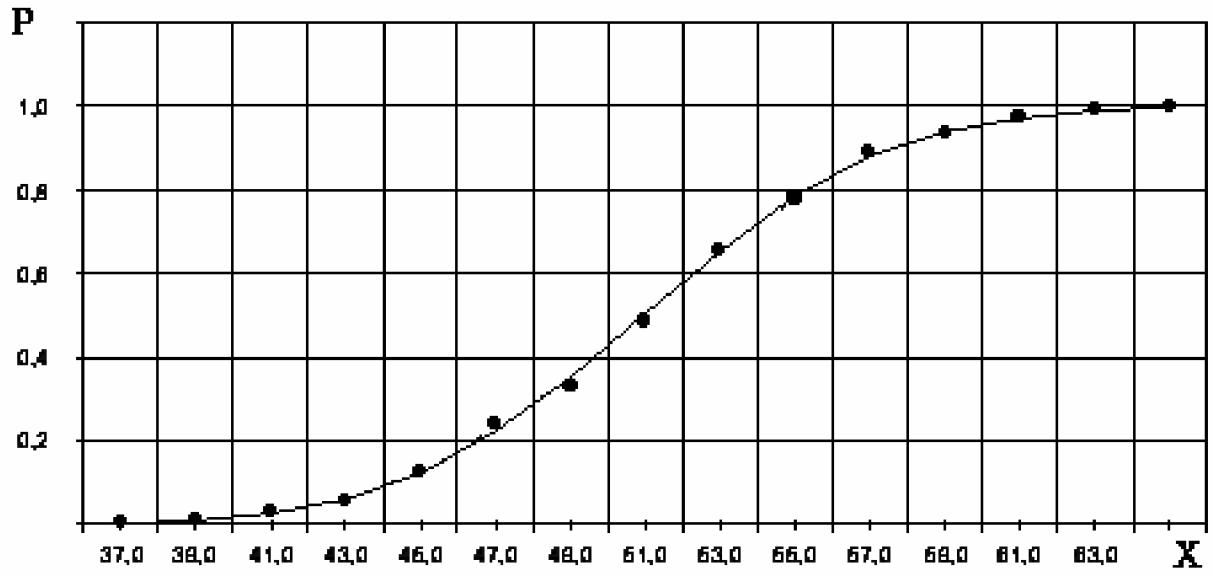

Рис. 5.4 – Выравнивание представительной выборки

После этого оценим вероятности попадания случайных величин, распределённых по выбранному теоретическому нормальному закону, в каждый из имеющихся разрядов

$$p_i = F(t_{i+1}) - F(t_i)\ , \quad \text{где } t_i = \frac{x_i - m_x^*}{s_x^*}\ . \tag{5.9}$$

Значения $F(t_i)$ берём из таблицы П.1 нормального закона, помещённой в приложении. При этом для отрицательных значений аргумента, которых нет в таблице, следует воспользоваться соотношением

$$F(-t_i) = 1 - F(t_i)\ . \tag{5.10}$$

Откладывая для всех значений x_i соответствующие значения $F(t_i)$, построим теоретическую функцию нормального закона распределения $F(x_i)$, которая показана на рис.5.4 сплошной линией.

Перейдём к непосредственному вычислению критерия согласия. Формула (5.6) неудобна для практического использования, поэтому приведём её к более удобному виду

$$U = \sum_{i=1}^{k} u_i = N\sum_{i=1}^{k} \frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{k} \frac{(Np_i^* - Np_i)^2}{Np_i} = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}\ .$$

Вычисленные промежуточные результаты p_i^*, t_i, $F(t_i)$, p_i, $N\cdot p_i$, u_i представлены в таблице 5.2.

Суммируя все полученные значения u_i, определим значение критерия согласия

$$U=7{,}316\ .$$

По таблице П.3 для числа степеней свободы $r=15-3=12$, интерполируя, находим значение доверительной вероятности

$$\beta = 0{,}885\ .$$

Отсюда имеем следующее значение уровня значимости расхождений

$$\alpha = 1 - \beta = 0{,}115\ .$$

Полученное значение доверительной вероятности $\beta=0{,}885$ является довольно высоким, а уровень значимости расхождений низкий, поэтому гипотеза о приемлемости нормального закона распределения для рассматриваемых представительных выборочных данных может быть принята с большой уверенностью.

5.2. Оценка значимости расхождений статистических оценок

Допустим, что получены результаты предварительной статистической обработки двух выборок n_1, m_{x1}^*, D_{x1}^* и n_2, m_{x2}^*, D_{x2}^* , где n_i – число опытов, m_{xi}^*– статистическая оценка математического ожидания, D_{xi}^*– статистическая оценка дисперсии, i =1,2 – номер выборки. Вычисление оценок m_x^*, D_x^* подробно рассматривалось в разделе 4.

Очевидно, что в общем случае $m_{x1}^* \neq m_{x2}^*$ и $D_{x1}^* \neq D_{x2}^*$. Эти различия числовых значений статистических характеристик с одной стороны обусловлены случайным выбором элементов из генеральной совокупности. При этом, сколько бы мы не получали выборок, каждая из них будет иметь свои значения статистических характеристик, отличные от характеристик других выборок из той же генеральной совокупности. С другой стороны, наблюдаемое различие может иметь неслучайный характер из-за того, что при формировании выборок были использованы различные генеральные совокупности, например выборки 3 и 4 на рис. 5.5. В таком случае различие статистических характеристик имеет неслучайную причину. Обе схемы появления расхождений числовых значений статистических характеристик представлены графически на рисунке 5.5.

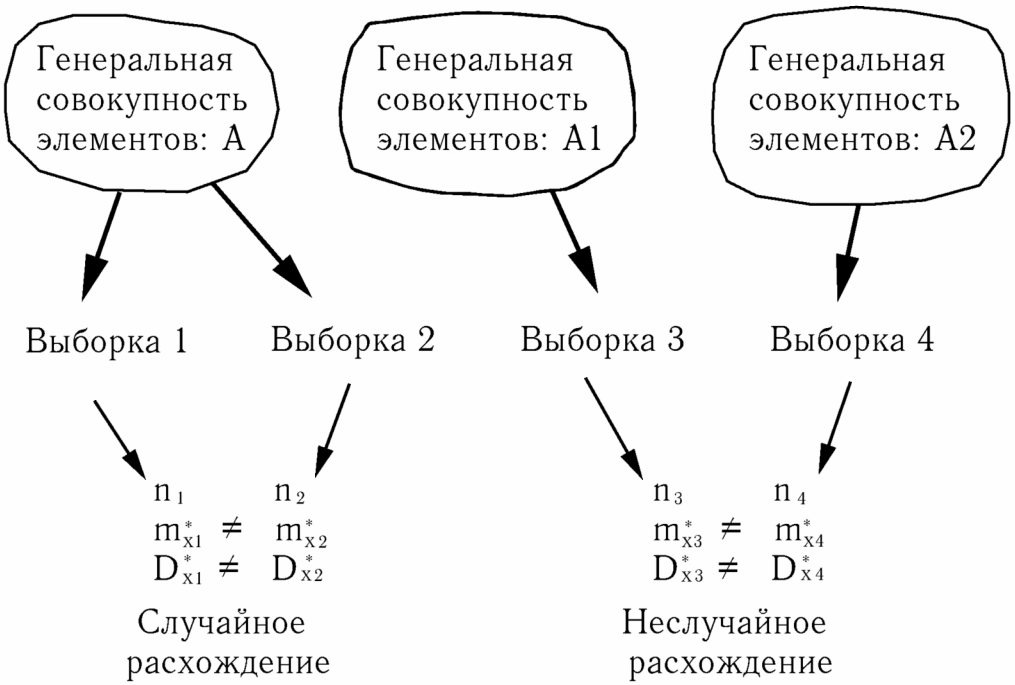

Рис. 5.5 - Схема взаимодействия выборок с генеральными совокупностями

Определённую роль в расхождении статистических характеристик играет несовпадение объёмов выборок, так как в общем случае $n_1 \neq n_2$. Однако, и при одинаковых по объёму выборках наблюдается расхождение их характеристик.

Статистика не может идентифицировать физическую причину наблюдаемых расхождений, однако в статистике развиты методы, позволяющие ответить на вопрос о том, являются ли эти расхождения случайными или они вызваны неслучайной причиной. Ниже будет рассмотрено, как это делается при оценке различий статистических оценок математических ожиданий и дисперсий.

5.2.1. Оценка расхождений средних значений

Воспользуемся результатами предварительной статистической обработки двух выборок n_1, m_{x1}^*, D_{x1}^* и n_2, m_{x2}^*, D_{x2}^*, где n_i – число опытов, m_{xi}^* – статистическая оценка математического ожидания, D_{xi}^*– статистическая оценка дисперсии, i =1,2 – номер выборки.

Рассмотрим оценку значимости расхождений средних значений

$$s = \left| m_{x1}^* - m_{x2}^* \right| . \tag{5.11}$$

Сначала вычислим дисперсию разности математических ожиданий [1]

$$D_s = \frac{(n_1 - 1)D_{x1} + (n_2 - 1)D_{x2}}{(n_1 - 1) + (n_2 - 1)} . \tag{5.12}$$

Так как истинные значения дисперсий D_{x1} и D_{x2} неизвестны, то вместо них в формулу (5.9) подставим их статистические оценки D_{x1}^*, D_{x2}^*.

После этого вычислим статистику Стьюдента

$$t_\beta = \frac{s}{\sqrt{D_s}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} . \tag{5.13}$$

Случайная величина t_β имеет закон распределения с плотностью вероятности, которая определяется выражением [1]

$$S_{n-1}(t) = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})} \left(1 + \frac{t^2}{n-1}\right)^{-n/2}, \tag{5.14}$$

где $n - 1 = n_1 + n_2 - 2$,

$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ - гамма - функция.

Уровень значимости расхождений или вероятность того, что расхождения средних значений имеют неслучайную причину, определяется как

$$\beta = 1 - 2\int_0^{t_\beta} S_{n-1}(t)dt \ . \tag{5.15}$$

Значения верхнего предела t_β интеграла (5.15), заимствованные в работе [3], представлены в таблице П.4 в приложении.

Следует отметить, что распределение вероятности значимости расхождений в виде выражения (5.13) с использованием распределения Стьюдента представляет собой точное решение только при нормальном распределении случайных величин в рассматриваемых выборках.

Если же распределение этих величин отличается от нормального, то распределение статистики (5.13) всё равно оказывается близким к распределению вида (5.15).

Зная доверительную вероятность, оценим уровень значимости расхождений

$$\alpha = 1 - \beta, \tag{5.16}$$

который представляет собой вероятность того, что наблюдаемое расхождение имеет случайный характер и рассматриваемые выборки принадлежат одной генеральной совокупности.

Пример 1. Рассмотрим известные опыты по определению веса определённого объёма азота, выполненные в 1903 году [1]. Исследовались две выборки данных, первая выборка – результаты получения азота из азотистых соединений, вторая выборка – результаты получения азота из воздуха. Условия всех опытов (температура, давление) идентичны. Результаты опытов представлены в таблице 5.3.

Статистические характеристики выборок имеют следующие значения:

m_{x1}*=2,29947, D_{x1}*=0,0000019022, n1=8;

m_{x2}*=2,31016, D_{x2}*=0,000000021456, n2 =10.

Найдем расхождение средних значений:

$$s = 2{,}31016 - 2{,}29947 = 0{,}01069.$$

Наблюдаемое расхождение средних значений в относительных единицах составляет около 0,5%, то есть невелико и может быть объяснено неточным взвешиванием, т.е. случайными погрешностями измерений.

С другой стороны, можно предположить, что использовались различные способы получения азота, и именно это могло привести к появлению неслучайных причин наблюдаемых расхождений.

Таблица 5.3 Опытные данные

Номер Опыта	Вес азота в граммах	
	1. Получение из соединений	2. Получение из воздуха
1	2,30143	2,31017
2	2,29890	2,30986
3	2,29816	2,31010
4	2,30182	2,31010
5	2,29869	2,31024
6	2,29940	2,31010
7	2,29849	2,31028
8	2,29889	2,31035
9	–	2,31026
10	–	2,31024

Для выяснения того, какая из этих гипотез более правдоподобна, вычислим дисперсию расхождения средних значений по формуле (5.12)

$$D_s = \frac{7 \cdot 0,0000019022 + 9 \cdot 0,000000021456}{16} = 0,0000008443 \text{ , откуда } \sqrt{D_s} = 0,00092 \text{ .}$$

Статистика Стьюдента, вычисленная по формуле (5.13) имеет значение

$$t_\beta = \frac{0,01069}{0,00092}\sqrt{\frac{80}{18}} = 24,5 \text{ .}$$

По таблице распределения Стьюдента П.4 из приложения для n–1=8+10–2=16 убеждаемся, что вероятность неслучайного расхождения средних значений составляет $\beta > 0,999$, так как значению n–1=16 соответствует максимальное табличное значение t_β=4,01 при β=0,999.

Следовательно, вероятность гипотезы случайного характера расхождений средних значений в рассматриваемых опытах невелика

$$\alpha < 1 - 0,999 = 0,001.$$

Выполненные расчёты убедительно показывают, что из двух рассматриваемых гипотез следует выбрать гипотезу о неслучайном характере полученных расхождений средних значений. Именно эти вычисления в своё время явились обоснованием ряда экспериментальных исследований, которые привели к открытию газа аргона, в то время неизвестного.

Пример 2. Рассмотрим ещё один пример оценки значимости расхождений средних значений. В таблице 5.4 представлены результаты предварительной статистической обработки оценок двух групп студентов по высшей математике.

Группы не отличаются по своему составу, что подтверждается незначительным различием дисперсий, однако методики преподавания данной дисциплины в этих группах были различными.

Таблица 5.4 Статистические данные

№ группы	Численность группы	Среднее значение оценок	Дисперсия оценок
1	18	4,520	1,250
2	22	4,560	1,270

Встаёт вопрос, является ли различие средних значений полученных оценок свидетельством различной эффективности используемых методик преподавания?

Для ответа на поставленный вопрос выясним значимость расхождений средних значений и для этого сначала вычислим дисперсию разности средних значений по формуле

$$D_s = \frac{17 \cdot 1,25 + 21 \cdot 1,27}{38} = 1,261, \quad \text{откуда} \quad \sqrt{D_s} = 1,123 . \qquad (5.17)$$

Статистика Стьюдента, по формуле (5.13) имеет значение

$$t_\beta = \frac{4,56 - 4,52}{1,123} \sqrt{\frac{22 \cdot 18}{22 + 18}} = 0,112 .$$

Интерполируя в таблице распределения Стьюдента П.4 из приложения для n–1=18+22–2=38, находим вероятность неслучайного расхождения средних значений

$$\beta = 0,089 .$$

Вероятность гипотезы случайного характера расхождений средних значений в рассматриваемом примере составляет

$$\alpha = 1 - 0,089 = 0,911.$$

Расчёт убедительно показывает, что здесь следует выбрать гипотезу о случайном характере полученных расхождений средних значений, то есть считать, что применяемые методики преподавания в данном случае не отличаются по своей эффективности.

5.2.2. Оценка расхождений дисперсий

Рассмотрим две выборки: n_1, m_{x1}^*, D_{x1}^* и n_2, m_{x2}^*, D_{x2}^*.

Для оценки расхождений дисперсий воспользуемся статистикой Р.Фишера

$$F = \frac{D_{x1}^*}{D_{x2}^*}, \quad \text{где} \quad D_{x1}^* > D_{x2}^* \quad . \tag{5.18}$$

Плотность вероятности распределения этой величины при нормальном распределении элементов выборки определяется соотношением [1]

$$f(F) = \frac{f_1^{\frac{f_1}{2}} f_2^{\frac{f_2}{2}} \Gamma(\frac{f_1 + f_2}{2})}{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})} \cdot \frac{F^{\frac{f_1-2}{2}}}{(f_2 + f_1 F)^{\frac{f_1+f_2}{2}}}, \tag{5.19}$$

где $f_1 = n_1 - 1$; $f_2 = n_2 - 1$; $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ - гамма-функция.

Если элементы выборок распределены по закону, отличающемуся от нормального, то использование распределения Р.Фишера даёт приближённые, но достаточно точные оценки.

Для практического использования распределения Р.Фишера разработаны таблицы, соответствующие заданным уровням значимости расхождений (смотри таблицу П.5 в приложении). Таблицы заимствованы в работе [1] и даны для четырёх уровней значимости расхождений β=0,01, 0,05, 0,10 и 0,25 в виде F=F (f_1, f_2, β=const), где $f_1=n_1 - 1$, $f_2=n_2 - 1$, причём f_1 всегда соответствует большей по величине дисперсии.

Если расчётное значение $F_{расч}$, определяемое по формуле (5.18), меньше табличного значения $F_{табл}$ для определённого уровня значимости расхождений β, то это значит, что фактический уровень значимости расхождений меньше табличного. При этом вероятность случайного характера расхождений рассматриваемых дисперсий будет больше чем

$$\alpha = 1 - \beta .$$

Пример. Два предприятия выпускают идентичную продукцию, например, майонез. В лаборатории произведен выборочный контроль продукции по параметру $\boldsymbol{X}$ (жирность). Эта величина в соответствии с установленными нормами должна находиться в заданном интервале $4{,}55 < \boldsymbol{X} < 4{,}58$.

Можно ли утверждать, что у этих предприятий существенное различие качества продукции, если воспользоваться результатами предварительной обработки показателя $\boldsymbol{X}$, которые приведены в таблице 5.5 и где $\boldsymbol{n}$ – число проверенных банок майонеза.

Таблица 5.5 - Показатели жирности

Характеристики	**Фирма** А	**Фирма** Б
n	18	15
$\mathbf{m^*_x}$	4,57	4,56
$\mathbf{D^*_x}$	0,0295	0,0139

При решении данной задачи нельзя основываться на сравнении средних значений, так как они фактически заданы техническими условиями, поэтому следует обратиться к сравнению оценок дисперсий, которые здесь значительно отличаются между собой.

Вычисленное значение статистики Р.Фишера в данной задаче составляет

$$F_{\text{расч}} = \frac{0,0295}{0,0139} = 2,12 \ .$$

Из имеющихся таблиц распределения Р.Фишера следует выбрать таблицу с ближайшим по величине значением F. Очевидно, что для $f_1 = 18 - 1 = 17$ и $f_2 = 15 - 1 = 14$, таких таблиц две: при $\beta = 0,90$ имеем $F_{\text{табл}} = 1,988$, а при $\beta = 0,95$ находим $F_{\text{табл}} = 2,428$.

Выполним линейную интерполяцию

$$\beta = 0,90 + \frac{0,95-0,90}{2,428-1,988} \times (2,12-1,988) = 0,915.$$

Мы получили вероятность того, что имеется неслучайная причина наблюдаемых расхождений (в данном случае это расхождение дисперсий жирности). Вычислим вероятность противоположного события, т.е. вероятность того, что различия имеют чисто случайный характер

$$\alpha = 1 - 0,915 = 0,085.$$

Сравнивая полученные вероятности, делаем вывод о том, что наблюдаемые расхождения имеют неслучайный характер. Следовательно фирмы А и Б выпускают продукцию одного наименования, но с существенно различным качеством.

6. ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

Любые статистические характеристики случайных величин, определяемые в результате обработки выборки элементов из некоторой генеральной совокупности, в силу ограниченного объёма выборок также являются случайными величинами. Повторяя определение статистических характеристик для новых выборок из общей генеральной совокупности, каждый раз будем получать новые числовые значения характеристик. При этом возникает вопрос, как, ориентируясь на полученные значения статистических характеристик, определить их истинные значения. Для ответа на такой вопрос используется построение доверительного интервала.

Рассмотрим общее понятие доверительного интервала. Допустим, что в результате обработки выборки элементов получена статистическая оценка параметра **a***. Это может быть статистическая оценка математического ожидания, дисперсии, вероятности события или оценка какой-либо другой характеристики. Изобразим полученный результат графически, как это показано на рисунке 6.1.

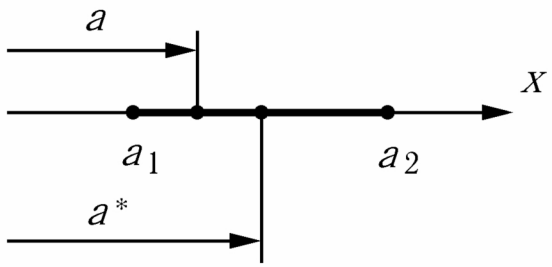

Рис. 6.1 – К построению доверительного интервала

Зная значение **a***, построим некоторый интервал $J_\beta(a) = (a_1; a_2)$, который с заданной вероятностью β будет содержать истинное значение параметра **a**, соответствующее всей генеральной совокупности рассматриваемых элементов.

Вероятность β называется доверительной вероятностью и соответствует условию

$$\beta = P(a_1 \le a \le a_2). \qquad (6.1)$$

В зависимости от физического содержания решаемых задач в расчётной практике используются значения доверительных вероятностей: 0,8 , 0,9 , 0,95 , 0,99 , 0,999.

Интервал $J_\beta(a) = (a_1; a_2)$, построенный относительно имеющейся статистической оценки параметра **a*** и "накрывающий" с заданной вероятностью β истинное значение **a** называется доверительным интервалом.

Условие (3.1) не является достаточным для однозначного определения доверительного интервала, поэтому оно дополняется условием равновозможности выхода истинного значения случайной величины направо и налево за пределы доверительного интервала, а именно

$$P(a < a_1) = P(a > a_2) = (1-\beta)/2 . \qquad (6.2)$$

Вероятность выхода истинного значения случайной величины за пределы построенного интервала как направо так налево равна вероятности противоположного события

$$\alpha = 1 - \beta . \qquad (6.3)$$

Ниже подробно рассматриваются методики построения доверительного интервала для математического ожидания, дисперсии и вероятности события.

6.1 Доверительный интервал математического ожидания

Доверительный интервал математического ожидания $J_\beta(m)=(m_1; m_2)$, располагается симметрично относительно оценки математического ожидания m*, как это показано на рисунке 6.2.

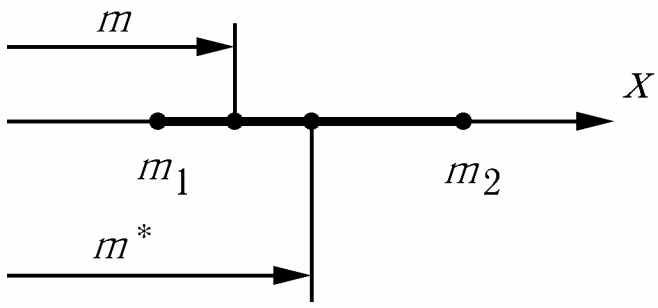

Рис. 6.2 – Доверительный интервал математического ожидания

Границы доверительного интервала в данном случае определяются следующим образом [3]

$$m_1 = m^* - \varepsilon_\beta, \quad m_2 = m^* + \varepsilon_\beta. \qquad (6.4)$$

Отклонение математического ожидания от статистической оценки определяется как

$$\varepsilon_\beta = t_\beta \sqrt{\frac{D^*}{n}} . \qquad (6.5)$$

Параметр t_β связан с доверительной вероятностью β выражением

$$\beta = 2\int_0^{t_\beta} S_{n-1}(t)dt \ , \qquad (6.6)$$

где $S_{n-1}(t)$ – плотность распределения Стьюдента с числом степеней свободы $n-1$, при определении которого n – число опытов.

Значения параметра t_β можно взять из таблицы распределения Стьюдента П.4 в приложении, однако лучше использовать более удобную таблицу П.6.

Представленная методика даёт точные результаты при нормальном законе распределения случайной величины. В противном случае этой методикой также можно пользоваться, но тогда значения границ доверительного интервала будут приближёнными.

Пример. Предварительная статистическая обработка выборки объёмом n=20 элементов дала следующие оценки математического ожидания и среднего квадратического отклонения (см. раздел 4.1.1)

$$m^*=20,15 \ ; \ S^*=1,52 \ .$$

Необходимо построить доверительный интервал для математического ожидания с доверительной вероятностью $\beta = 0,9$.

Сначала по таблице распределения Стьюдента П.6 для $n-1 = 19$ и $\beta = 0,9$ находим $t_\beta = 1,729$.

По формуле (6.5) получаем $\varepsilon_\beta = 1,729\dfrac{1,52}{\sqrt{20}} = 0,5877$.

Значения границ доверительного интервала вычисляем как

$$m_1 = 20,15 - 0,5877 = 19,5623; \qquad m_2 = 20,15 + 0,5877 = 20,7377 \ .$$

Выполнив округление, окончательно представим доверительный интервал в виде

$$J_\beta(m) = (\ 19,56 \ ; \ 20,74) \ .$$

6.2 Доверительный интервал дисперсии

Доверительный интервал дисперсии $J_\beta(D)=(D_1;\ D_2)$, располагается несимметрично относительно статистической оценки дисперсии D*, как это показано на рисунке 6.3.

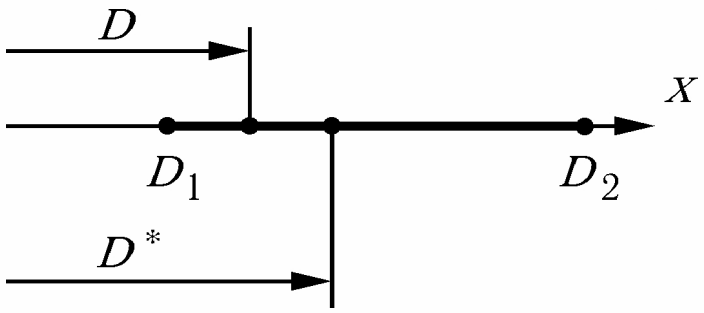

Рис. 6.3 – Доверительный интервал дисперсии

Границы доверительного интервала дисперсии находятся следующим образом [3]

$$D_1 = \frac{D^*(n-1)}{\chi_1^2}, \qquad D_2 = \frac{D^*(n-1)}{\chi_2^2} . \tag{6.7}$$

Значения знаменателей в формулах (6.7) определяются по таблицам распределения Пирсона в виде $\chi^2 = \chi^2(P, r)$, принимая

$$r = n - 1, \qquad P_1 = (1 - \beta)/2, \qquad P_2 = 1 - P_1 , \tag{6.8}$$

где n – число опытов, β – принятое значение доверительной вероятности.

Для определения значений $\chi^2_{1,2}$ удобно пользоваться таблицей П.7, которая представлена в приложении и специально разработана для построения доверительного интервала дисперсии.

Рассмотренная методика даёт точные результаты, когда величина, для которой оценивается дисперсия, распределена по нормальному закону. В других случаях этой методикой тоже можно пользоваться, но результаты построения доверительного интервала будут приближёнными.

Очевидно, что, определив границы доверительного интервала дисперсии $(D_1;\ D_2)$, можно построить доверительный интервал среднего квадратического отклонения

$$J_\beta(S) = (S_1;\quad S_2) = (\sqrt{D_1};\quad \sqrt{D_2}) \ . \tag{6.9}$$

Пример. При статистической обработке выборки объёмом $n = 20$ элементов получена оценка среднего квадратического отклонения $S^* = 1{,}52$.

Необходимо построить доверительный интервал для среднего квадратического отклонения с доверительной вероятностью $\beta=0{,}9$.

По формулам (3.8) вычислим вероятности

$$P_1 = (1 - 0{,}9)/2 = 0{,}05 , \qquad P_2 = 1 - P1 = 0{,}95.$$

В таблице П.7 для $r=20-1=19$ найдём параметры распределения Пирсона:

$$\chi^2_1 = 30{,}144 ; \quad \chi^2_2 = 10{,}117 .$$

Далее по формулам (3.7) вычислим границы доверительного интервала дисперсии

$$D_1 = \frac{1{,}52^2 (20-1)}{30{,}144} = 1{,}456 , \qquad D_2 = \frac{1{,}52^2 (20-1)}{10{,}177} = 4{,}339 .$$

Воспользовавшись формулой (6.9) запишем доверительный интервал среднего квадратического отклонения в виде

$$J_\beta(S) = (\sqrt{1{,}456}; \ \sqrt{4{,}339}) = (1{,}207 \ ; \ 2{,}083) .$$

6.3 Доверительный интервал вероятности наблюдаемых событий

Вероятности различных событий необходимые для решения практических задач чаще всего определяются опытным путём, основываясь на теореме Я.Бернулли (см. раздел 1.5). В этой теореме утверждается, что частота события при независимых опытах сходится по вероятности к вероятности данного события.

Исходя из этого, статистическая вероятность приравнивается частоте события

$$p^* = \frac{m}{n} , \tag{6.10}$$

где n – общее число независимых опытов, m – число опытов, в которых реализовалось данное событие.

Сколь велико бы ни было число опытов **n** , всё равно статистическая вероятность p* ведёт себя как случайная величина. В соответствии с законом больших чисел при увеличении числа опытов распределение частоты события приближается к нормальному.

Так как частота события m/n представляет собой среднее арифметическое числа положительных опытов, то для построения доверительного интервала вероятности можно воспользоваться методикой построения доверительного интервала для математического ожидания (см. раздел 6.1). Однако даже при нормальном законе распределения рассматриваемой величины, построенный таким образом интервал будет приближённым.

Точное решение рассматриваемой задачи основывается на рассмотрении биномиального распределения [4], в соответствии, с которым вероятность появления события ровно m раз при n независимых опытах определяется зависимостью

$$P_n^m = C_n^m p^m q^{n-m} \quad , \tag{6.11}$$

где $C_n^m = \dfrac{n!}{m!(n-m)!}$ - число сочетаний из n элементов по m элементам,

p – вероятность появления события в отдельном опыте,

q = (1 – p) – вероятность противоположного события.

Предположим, что вероятность p известна. Найдём интервал (p_1 , p_2) в который с вероятностью $\beta = 1 - \alpha$ попадает частота события p* , как показано на рисунке 6.4.

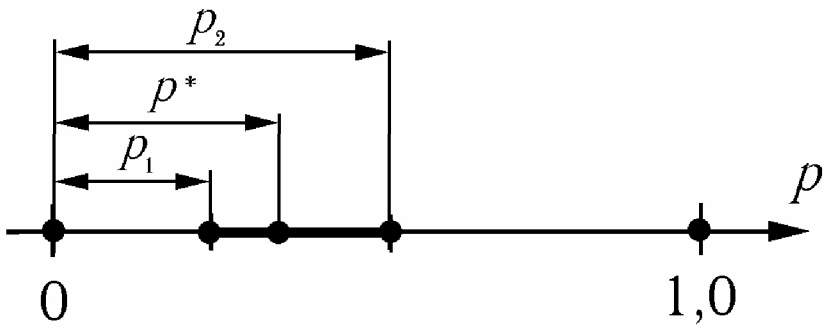

Рис. 6.4 – Доверительный интервал вероятности

Вероятность попадания частоты события p* левее p_1 и правее p_2 выражается соотношениями

$$\sum_{m=0}^{k} C_n^m p_1^m (1-p_1)^{n-m} = \frac{\alpha}{2} \quad , \tag{6.12}$$

$$\sum_{m=k-1}^{n} C_n^m p_2^m (1-p_2)^{n-m} = \frac{\alpha}{2} \quad , \tag{6.13}$$

где $k = np^*$ - число появлений события.

Решая уравнения (6.12 – 6.13) относительно p_1 и p_2 , получим доверительный интервал для вероятности рассматриваемого события $J_\beta(p)=(p_1; p_2)$.

На рисунках 6.5 и 6.6 представлены графически решения уравнений (6.12 – 6.13) для доверительных вероятностей 0,80 и 0,90, заимствованные в работе [4]. По оси абсцисс отложена частота события p*, а по оси ординат – вероятности p_1 и p_2. Чтобы найти доверительный интервал, нужно на оси абсцисс отложить полученное значение частоты p*, провести через эту точку прямую, параллельную оси ординат и отметить ординаты точек пересечения с кривыми, соответствующими заданному числу опытов.

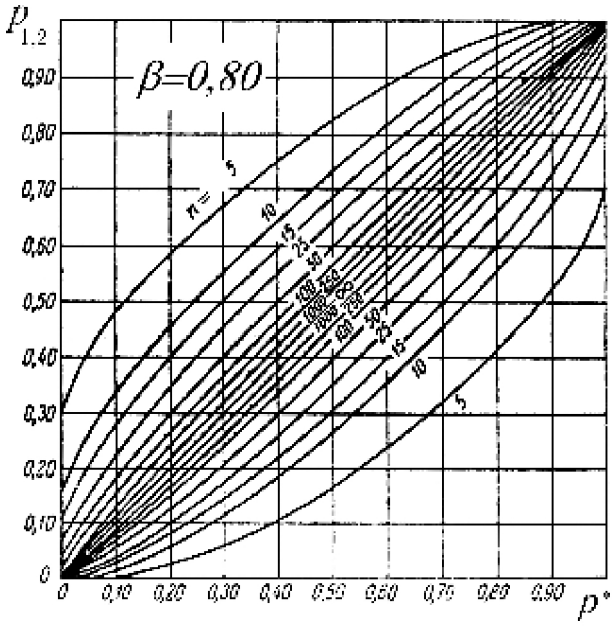

Рис. 6.5 – Границы доверительных интервалов вероятности $\beta = 0,80$

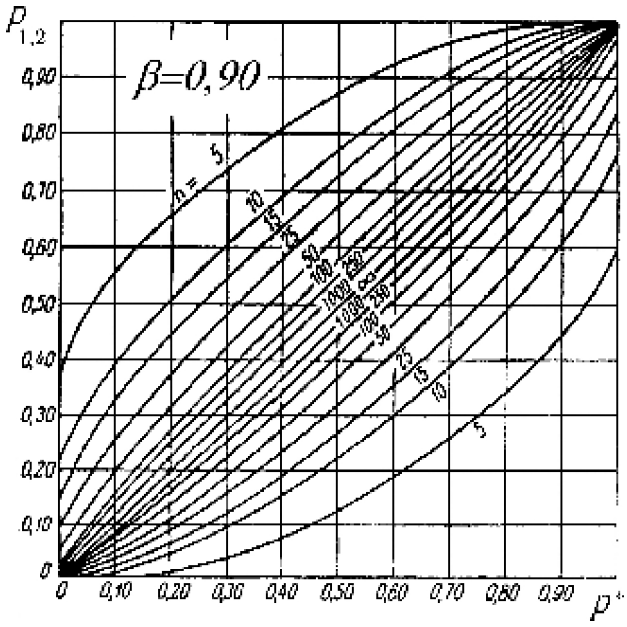

Рис. 6.6 – Границы доверительных интервалов вероятности $\beta = 0,90$

Пример. При выполнении опытов, число которых составляет n = 50, положительные результаты дали m = 15 опытов. Требуется определить доверительный интервал вероятности рассматриваемого события, задавшись доверительной вероятностью β = 0,90.

Частота данного события составляет

$$p^* = \frac{m}{n} = \frac{15}{50} = 0{,}30 \ .$$

По графикам рисунка 3.6 находим границы доверительного интервала

$$J_\beta(p) = (p_1; p_2) = (0{,}19 \ ; \ 0{,}42) \ .$$

Точное решение с использованием биномиального распределения, а именно по формулам (6.12, 6.13), даёт следующие значения доверительного интервала

$$J_\beta(p) = (0{,}1949 \ ; \ 0{,}4237) \ .$$

Рассмотрим также приближённое построение доверительного интервала, воспользовавшись методикой построения доверительных интервалов для математического ожидания, изложенной в разделе 6.1.

Сначала по таблице П.6 для n –1 = 49 и β = 0,9, интерполируя, находим $t_\beta = 1{,}677$.

Дисперсия частоты события может быть оценена следующим образом

$$D_p = p^*(1 - p^*) = 0{,}30 \cdot (1 - 0{,}30) = 0{,}21 \ .$$

По формуле (6.5) получаем

$$\varepsilon_\beta = 1{,}677\sqrt{\frac{0{,}21}{50}} = 0{,}109 \ .$$

Значения границ доверительного интервала вычисляем как

$$p_1 = 0{,}30 - 0{,}109 = 0{,}191; \qquad p_2 = 0{,}30 + 0{,}109 = 0{,}409 \ .$$

Представим полученный приближённый доверительный интервал в виде

$$J_\beta(p) \approx (0{,}191 \ ; \ 0{,}409) \ .$$

Сравнивая приближённый и точный доверительные интервалы, можно убедиться в достаточно высокой точности приближённого решения, однако следует иметь в виду, что такое соответствие имеет место при соблюдении условия

$$np^* > 10 \quad \text{и} \quad n(1 - p^*) > 10 \ ,$$

которое, как можно убедиться, в данном примере выполняется.

6.4 Доверительный интервал вероятности редких событий

Использование частоты события при оценке вероятности события имеет специфические особенности, если сама вероятность события очень мала, то есть рассматриваемое событие является редким. В таком случае, проводя ограниченное число опытов n, получим нулевую частоту события $p^* = m/n$, так как $m = 0$. На самом деле вероятность этого события, хотя и мала, но отличается от нуля, то есть $p > 0$.

Построение доверительного интервала позволяет оценить возможное значение вероятности данного редкого события. При этом, очевидно, что левая граница доверительного интервала равно нулю, то есть $p_1 = 0$ или

$$J_\beta(p) = (p_1; p_2) = (0 ; p_2) . \qquad (6.14)$$

Графически пример доверительного интервала редкого события схематично показан на рисунке 6.7.

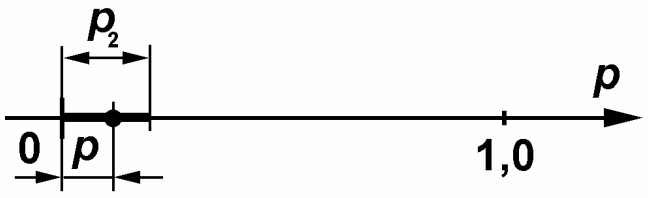

Рис. 6.7 – Доверительный интервал редкого события

Приближённые методы построения доверительного интервала вероятности в данном случае неприменимы, так как они дают здесь существенную ошибку.

Использование биномиального закона распределения позволяет получить точное решение, но его применение затруднительно в вычислительном плане.

Рассмотрим более простое и точное для рассматриваемого частного случая решение, представленное в работе [3]. Допустим, что проведено n опытов и ни в одном из них не обнаружено интересующее нас событие. Требуется построить доверительный интервал вероятности ненаблюдаемого в опытах события.

Так как левая граница доверительного интервала $p_1=0$ известна, то остаётся определить значение p_2 при заданной вероятности β.

Обратимся к вероятности противоположного события $(1-\beta)$, которое заключатся в том, что каждый опыт даёт положительный результат. В отдельном опыте положительный результат появляется с вероятностью $(1-p_2)$, если в качестве вероятности события взять правую границу доверительного интервала $p = p_2$. По теореме умножения вероятностей независимых событий получим

$$1-\beta = (1-p_2)\cdot(1-p_2)\cdot\ldots\cdot(1-p_2) = (1-p_2)\cdot n\,.$$

Из последнего выражения имеем

$$p_2 = 1-\sqrt[n]{1-\beta}\quad . \tag{6.15}$$

Пример. Выполнено $n = 100$ опытов, в которых рассматриваемое событие не появилось ни разу. Требуется определить доверительный интервал данного редкого события с доверительной вероятностью $\beta = 0{,}90$.

По формуле (6.15) получаем

$$p_2 = 1-\sqrt[100]{1-0{,}9} = 0{,}023\quad .$$

Доверительный интервал строится следующим образом $J_\beta(p) = (0\ ;\ 0{,}023)$.

Какое практическое значение имеет полученный результат?

Допустим, что предполагается провести $n = 1000$ опытов. С доверительной вероятностью 0,90 можно ожидать минимальное число результативных опытов $n_{min} = 0$ и максимальное возможное число результативных опытов $n_{max} = 0{,}023\cdot 1000 = 23$. Поэтому очевидно, что выполнив 1000 опытов, мы не можем рассчитывать более чем на 23 положительных результата.

7. РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ представляет собой процедуру определения зависимости между двумя переменными X и Y с последующей проверкой её адекватности или, по-другому, с проверкой соответствия используемым данным. В качестве исходных данных рассматривается ряд наблюдений за системой, работу которой характеризуют параметры X и Y. Результаты наблюдений представляют собой два массива чисел [X] и [Y] . Элементы массивов x_i , y_i (i = 1,…, n, где n – число наблюдений) с одинаковыми индексами соответствуют одному и тому же наблюдению.

Следует отличать цель регрессионного анализа от аппроксимации данных, поскольку при аппроксимации желательно, что бы аппроксимирующая зависимость в наибольшей степени соответствовала наблюдаемым данным. В регрессионном анализе предполагается, что данные наблюдений всегда содержат в себе случайные погрешности и поэтому нет необходимости точного их воспроизведения, а необходимо найти зависимость, которая наилучшим образом воспроизводит закономерность, скрывающуюся за этими данными.

7.1. Линейный регрессионный анализ

При линейном регрессионном анализе зависимость между исследуемыми переменными принимается в виде

$$y = a + b \cdot x \quad . \tag{7.1}$$

Для определения параметров a и b, входящих в уравнение регрессии используется метод наименьших квадратов. В соответствии с этим методом регрессионная зависимость определяется так, что бы сумма квадратов отклонений вычисляемых значений y_{pi} от полученных опытным путём y_i была минимальной.

Минимизируемая сумма квадратов отклонений записывается как

$$S = \sum_{i=1}^{n} (y_{p\,i} - y_i)^2 = \sum_{i=1}^{n} (a + b \cdot x_i - y_i)^2 \to \min \quad . \tag{7.2}$$

Условие минимума может быть представлено следующим образом

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2\sum_{i=1}^{n}(a + b \cdot x_i - y_i) = 2(n \cdot a + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0 \ , \\ \frac{\partial S}{\partial b} &= 2\sum_{i=1}^{n}(a + b \cdot x_i - y_i) \cdot x_i = 2(a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i \cdot x_i) = 0 \ . \end{aligned} \tag{7.3}$$

Поделив на число опытов n, и введя новые обозначения, запишем условие минимума в виде системы уравнений

$$\begin{aligned} &a + b \cdot m_x^* - m_y^* = 0 \ , \\ &a \cdot m_x^* + b \cdot \alpha_2^*(x) - \alpha_2^*(x,y) = 0 \ . \end{aligned} \tag{7.4}$$

Здесь введены обозначения

$$\begin{aligned} &m_x^* = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \ , \quad m_y^* = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \ ; \\ &\alpha_2^*(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 \ , \ \alpha_2^*(x,y) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i \ . \end{aligned} \tag{7.5}$$

Полученные значения m^*_{x} и m^*_{y} представляют собой оценки математических ожиданий переменных x и y, $\alpha_2^*(x)$ – статистический второй начальный момент величины x, $\alpha_2^*(x,y)$ – статистический второй начальный смешанный момент.

Дисперсии переменных x и y вычисляются по формулам

$$\begin{aligned} D_x^* &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_x^*)^2 = \frac{1}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} (x_i)^2 - (m_x^*)^2 \right] = \frac{n}{n-1} \left(\alpha_2^*(x) - (m_x^*)^2 \right) , \\ D_y^* &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - m_y^*)^2 = \frac{1}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i)^2 - (m_y^*)^2 \right] = \frac{n}{n-1} \left(\alpha_2^*(y) - (m_y^*)^2 \right) . \end{aligned}$$

(7.6)

Средние квадратические отклонения определяются как

$$\sigma_x^* = \sqrt{D_x^*} , \ \sigma_y^* = \sqrt{D_y^*} \ . \tag{7.7}$$

Близость взаимосвязи рассматриваемых переменных к линейной зависимости характеризует корреляционный момент

$$K_{xy}^* = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - m_y^*) \cdot (x_i - m_x^*) = \frac{n}{n-1} \cdot \left(\alpha_2^*(x,y) - m_y^* \cdot m_x^* \right) \ . \tag{7.8}$$

Коэффициент корреляции так же характеризует близость к линейной зависимости, но эта величина позволяет оценить именно корреляцию, исключив влияние рассеивания исследуемых величин, что имеет место у корреляционного момента

$$r_{xy}^* = K_{xy}^* / (\sigma_x^* \cdot \sigma_y^*) \ . \tag{7.9}$$

Следует отметить, что коэффициент корреляции имеет ограниченную область изменения, которая для любых величин $\boldsymbol{X}$ и $\boldsymbol{Y}$ составляет

$$-1 \le r_{xy}^* \le +1 \ .$$

При $r_{xy} = -1$ величины y и x связаны убывающей линейной зависимостью, $r_{xy} = 1$ свидетельствует о возрастающей линейной зависимости, для независимых величин $r_{xy} = 0$. Промежуточные значения корреляционного коэффициента указывают на статистическую взаимосвязь исследуемых величин.

Решая приведенную выше систему уравнений (7.3) с использованием введённых обозначений, получим формулы для определения коэффициентов уравнения регрессии

$$b = \frac{\alpha_2^*(x,y) - m_y^* \cdot m_x^*}{\alpha_2^*(x) - (m_x^*)^2} \quad , \qquad a = m_y^* - b \cdot m_x^* \quad . \tag{7.10}$$

7.2. Значимость выборочной корреляции

Выполняя регрессионный анализ, необходимо убедиться в склонности исследуемых величин к линейному взаимодействию, характеристикой которого является выборочный коэффициент корреляции.

Численное значение коэффициента корреляции, приближающееся по абсолютной величине к единице, свидетельствует о выраженной в статистическом смысле склонности к линейному взаимодействию рассматриваемых случайных величин, таких, как $\boldsymbol{X}$ и $\boldsymbol{Y}$ в нашем случае. Однако, используя выборочные данные, мы получаем статистическую оценку коэффициента корреляции, которая сама по себе является случайной величиной из-за ограниченности объёма выборочных данных. Поэтому встаёт вопрос о том, будет ли действительно значимой корреляция рассматриваемых величин в их генеральной совокупности и есть ли смысл в поиске регрессии между исследуемыми величинами.

Для ответа на этот вопрос проверяется значимость гипотезы склонности к линейному взаимодействию с использованием статистики Стьюдента в виде

$$t_\beta = \frac{\left| r_{xy}^* \right|}{\sqrt{1 - (r_{xy}^*)^2}} \sqrt{n-2} .$$

Используя таблицы распределения Стьюдента, найдём значение вероятности реализации гипотезы о наличии корреляции в генеральной совокупности данных

$$P(\mathrm{H1}) = \beta .$$

Вероятность противоположного события представляет собой вероятность того, что в генеральной совокупности нет корреляции между рассматриваемыми случайными величинами

$$P(\mathrm{H0}) = \alpha = 1 - \beta .$$

7.3. Оценка адекватности линейной регрессии

Получив уравнение регрессии, необходимо убедиться в том, насколько удовлетворительно оно соответствует имеющимся результатам наблюдений. Для этого используется оценка надёжности уравнения регрессии.

Под надёжностью здесь понимается вероятность того, что отклонения опытных данных от соответствующих вычисленных значений имеют чисто случайный характер. Чем ближе надёжность H к единице, тем с большей уверенностью можно использовать полученную регрессию. Если рассматривалось несколько видов регрессионных зависимостей, то предпочтение следует отдать регрессии с наибольшей надёжностью.

7.3.1. Критерий Фишера. Надёжности регрессии

В качестве статистического критерия надёжности регрессии наиболее часто используется критерий Фишера в виде

$$F = \frac{D_{ад} \cdot (n-2)}{D_{ост}} \le F_{табл}(f_1, f_2, \alpha) \quad . \tag{7.11}$$

Здесь $f_1 = 1$, $f_2 = n-2$, α - уровень значимости расхождений.

Дисперсия адекватности и остаточная дисперсия, входящие в критерий Фишера, определяются следующим образом:

$$\begin{aligned} D_{ад} &= \frac{1}{n}\sum_{i=1}^{n}(y_{p\,i} - m_y)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{p\,i})^2 - m_y^2 = \alpha_{2p} - m_y^2 \quad , \\ D_{ост} &= D_y - D_{ад} \quad . \end{aligned} \tag{7.12}$$

Для численной оценки дисперсии адекватности необходимо вычислить соответствующий начальный момент

$$\alpha_{2p} = \frac{1}{n}\sum_{i=1}^{n}(a + b \cdot x_i)^2 \quad . \tag{7.13}$$

Значения $F_{табл}$ берутся из известных таблиц Фишера для заданного уровня значимости расхождений α.

Надёжность регрессии связана с уровнем значимости расхождений соотношением

$$H = 1 - \alpha \quad . \tag{7.14}$$

В учебных целях вместо таблиц Фишера допускается использовать график надёжности регрессии, который получен с использованием критерия Фишера и представлен на рисунке 7.1. Этот график даёт удовлетворительные по точности значения при числе наблюдений n = 8–12. Если F>10, то принимаем H = 0,99.

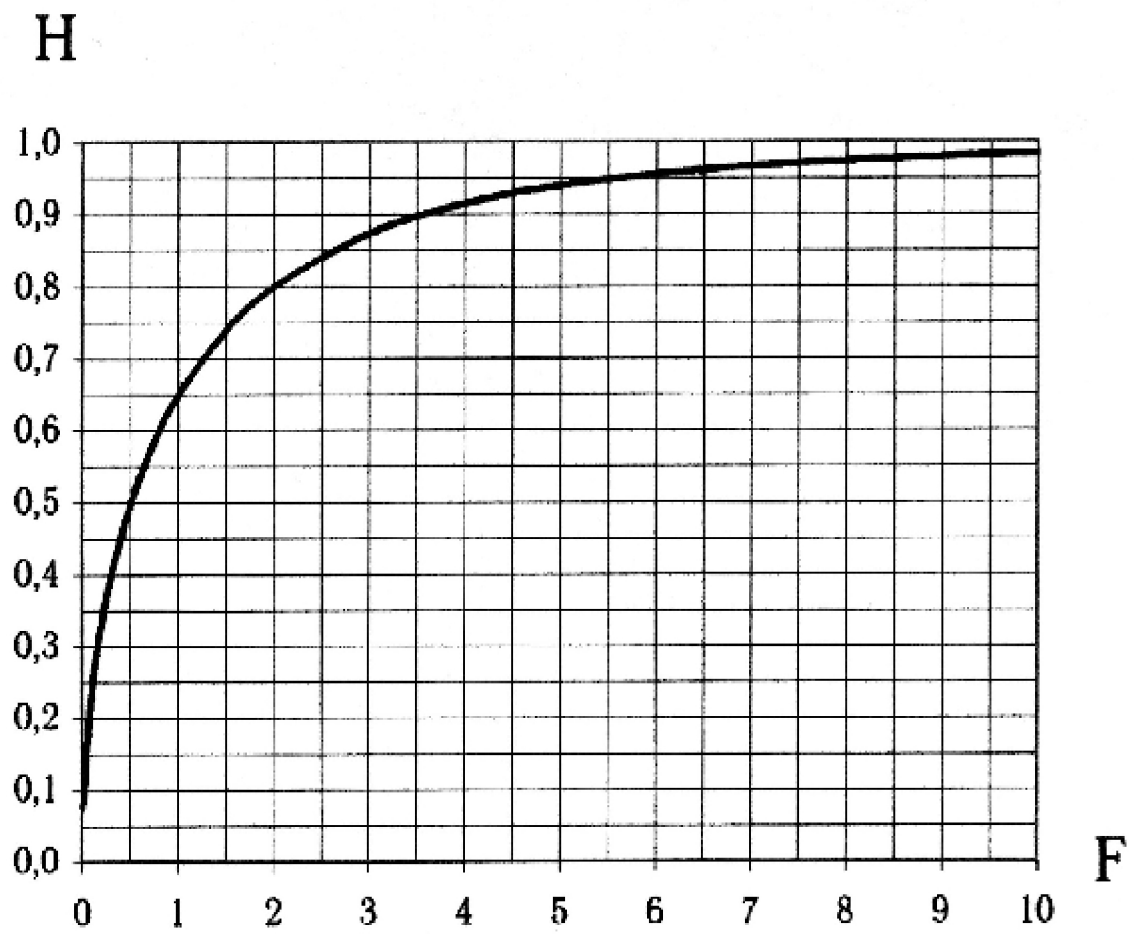

Рис. 7.1. Надёжность регрессии (n = 8–12)

7.3.2. Коэффициент детерминации

В качестве количественной меры адекватности обычно используется коэффициент детерминации

$$R^2 = \frac{D_{адекв}}{D_y^*}. \qquad (7.15)$$

Полученное значение коэффициента детерминации показывает, на то какая часть дисперсии рассматриваемой переменной $\boldsymbol{Y}$ "объясняется" найденной регрессией.

Дисперсия адекватности и остаточная дисперсия подсчитываются по формулам (7.12), из которых видно, что $D_{адекв} \le D_y$, следовательно, при любых исходных данным мы будем иметь $0 \le R^2 \le 1$.

Чем ближе значения коэффициента детерминации к единице, тем адекватнее рассматриваемая линейная регрессионная модель.

Пример.

Установим линейную регрессионную зависимость между переменными X и Y для конкретных данных парных наблюдений, приведенных в таблице 7.1.

Таблица 7.1 - Исходные данные парных наблюдений

i	1	2	3	4	5	6	7	8
x(i)	1,3310	2,5688	4,0063	5,1643	6,8280	8,1590	9,4102	10,6480
y(i)	3,7268	3,7268	3,3275	3,3275	3,9930	5,0578	5,0578	4,9247

Расчётные значения статических характеристик соответствующие рассматриваемым данным и полученные с использованием формул 7.5, 7.6 и 7.7 представлены в таблице 7.2.

Таблица 7.2 - Числовые характеристики парных наблюдений

Исследуемая величина	Расчётные значения			
	m	α_2	D	σ
X	6,0145	45,7767	10,975	3,313
Y	4,1427	17,6602	0,5691	0,7544

Корреляционный момент и коэффициент корреляции, характеризующие близость взаимосвязи рассматриваемых величин к линейной зависимости, вычисленные с использованием формул 7.8 и 7.9, приведены в таблице 7.3.

Таблица 7.3 - Характеристики корреляции

Наименование величины	K_{xy}	$\alpha_2(x, y)$	r_{xy}
Численное значение	2,0404	26,7016	0,8164

Выполняя корреляционный анализ, в рассматриваемом примере получим

$$t_\beta = \frac{|0,1864|}{\sqrt{1-(0,1864)^2}}\sqrt{8-2} = 3,4630.$$

Воспользовавшись таблицей распределения Стьюдента из приложения П.4 и применяя линейную интерполяцию, найдём значение вероятности реализации гипотезы о наличии корреляции в генеральной совокупности данных

$$P(\text{H1}) = \beta = 0,989.$$

Вероятность противоположного события представляет собой вероятность того, что в генеральной совокупности нет корреляции между рассматриваемыми случайными величинами

$$P(\text{H0}) = \alpha = 1-\beta = 0,011.$$

Сравнивая значения полученных вероятностей рассматриваемых гипотез, убеждаемся в возможности существенной корреляции исследуемых переменных X и Y за пределами имеющейся выборки, т.е. в генеральной совокупности. Как следствие из выполненного анализа корреляции, можно утверждать, что в данном случае велика вероятность регрессионной зависимости между рассматриваемыми случайными величинами.

Воспользовавшись данными расчётов, которые представлены в таблицах 7.2 и 7.3, получим следующие значения коэффициентов уравнения линейной регрессии:

$$b = 0,1859; \quad a = 3,0246 .$$

Очевидно, что уравнение линейной регрессии будет иметь следующий вид

$$y = 3,0246 + 0,1859 \cdot x .$$

Для наглядного представления соответствия исходных данных и полученного уравнения регрессии построен график, приведенный на рисунке 7.2. Прямая линия, соответствующая найденному уравнению регрессии проведена через две точки, вычисленные координаты, которых приведены в таблице 7.4.

Таблица 7.4 - Данные к построению графика регрессии

x_i	0,00	12,00
$y_i = a + b \cdot x_i$	3,025	5,255

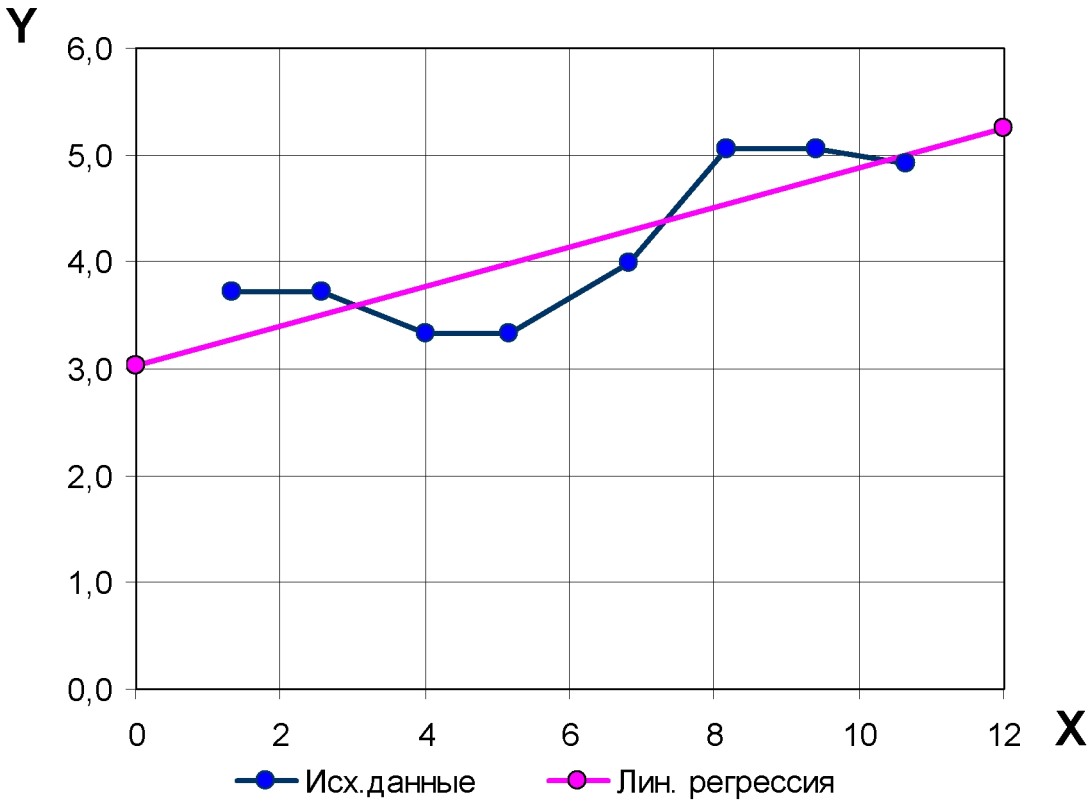

Рис.7.2. - Регрессионная зависимость

Дисперсия адекватности и остаточная дисперсия, необходимые для последующей оценки соответствия полученного уравнения регрессии основной зависимости между исходными данными, вычисленные по формулам 7.12 имеют следующие значения

$$D_{ад}=0,3793 \;\; ; \;\; D_{ост}=0,1898 \; .$$

В рассматриваемой задаче численное значение критерия Фишера для оценки адекватности найденной регрессии определяется как

$$F=\frac{0,3793\cdot(8-2)}{0,1898}=11,99.$$

По графику критерия Фишера на рисунке 7.1, приближённо определяем, что поскольку F>10, то поэтому надёжность найденной линейной регрессии составляет H = 0,99.

Полученный высокий уровень надёжности линейной регрессии указывает на то, что найденная с использованием имеющихся результатов парных наблюдений линейная зависимость в данном случае адекватно отражает основную зависимость между рассматриваемыми переменными.

Коэффициент детерминации имеет следующее значение

$$R^2 = \frac{D_{адекв}}{D_y} = 0{,}667 \ .$$

Найденное значение коэффициента детерминации показывает, что только 67% дисперсии рассматриваемой переменной ***Y*** "объясняется" найденной регрессией, что может приводить к существенным расхождениям между расчётными и наблюдаемыми значениями ***Y***.

Таким образом, переходя к выводам о проверке адекватности, можно утверждать, что полученное уравнение линейной регрессии адекватно отражает взаимосвязь рассматриваемых величин, что подтверждается статистическим критерием Фишера, однако количественный критерий в виде коэффициента детерминации имеет невысокое значение, которое говорит о невысоких прогностических свойствах найденного тренда.

8. ПЛАНИРОВАНИЕ ОБЪЁМА ВЫБОРОК

Использование выборок для оценки характеристик генеральных совокупностей элементов широко применяется при решении большого круга задач. В этой связи можно рассматривать две ситуации. Во–первых, отсутствие предварительного планирования объёма выборок на основе статистических методов. Вторая ситуация заключается в предварительной статистической оценке числа опытов, обеспечивающих заданную точность определения исследуемых характеристик. Отметим что, соответствующий раздел статистики за последнее время выделился в самостоятельное развитое научное направление, которое называется "Планирование эксперимента".

В зависимости от того, к какой из рассматриваемых ситуаций относится получаемая статистическая оценка рассматриваемой величины, различают активный статистический анализ (объём выборки предварительно рассчитывается статистическими методами), и в противном случае – пассивный статистический анализ, т.е. обработка имеющихся данных, собранных без обеспечения их представительности.

По существу в рассматриваемой задаче, задавая погрешность оценки некоторой случайной величины, мы тем самым определяем доверительный интервал, в котором с доверительной вероятностью должно находиться истинное значение рассматриваемой величины β , то есть

$$\beta = P\,(a^* - \Delta a < \mathbf{a} < a^* + \Delta a)\,, \qquad (8.1)$$

где a^* – статистическая оценка, получаемая обработкой выборки,

$\mathbf{a}$ – истинное значение рассматриваемой величины,

Δa – допустимая погрешность оценки данной характеристики.

При планировании объёма выборок необходимо определять число опытов **n**, обеспечивающее выполнение соотношения (8.1). Очевидно, что задачи планирования объёма выборок являются обратными по отношению к рассматриваемым в предыдущем разделе задачам определения доверительных интервалов.

8.1. Планирование оценивания математического ожидания

Обратившись к построению доверительного интервала для математического ожидания, рассмотренному в разделе 6.1, на основании формул (6.4) и (6.5) запишем выражение для погрешности оценки математического ожидания в виде

$$\Delta m_x = t_\beta \sqrt{\frac{D_x^*}{n}} \quad , \tag{8.2}$$

где: n – объём выборки элементов при независимых опытах,

D_x^* – статистическая оценка дисперсии исследуемой случайной величины,

t_β – параметр распределения Стьюдента, соответствующий заданному значению доверительной вероятности β (смотри соотношение (6.6) в разделе 6), определяемый по таблице П.6 в приложении.

Из формулы (8.2) непосредственно следует, что $n = \frac{t_\beta^2}{(\Delta m_x)^2} D_x^*$. (8.3)

Решение уравнения (8.3) осложняется тем, что правая часть этого уравнения также зависит от числа опытов n, так как $t_\beta = t\,(\beta, n-1)$.

Преобразуем выражение (8.3) к виду

$$n \cdot A^2 - t_\beta^2 = 0 \quad , \text{ где } A = \frac{\Delta m_x^*}{\sqrt{D_x^*}} \quad . \tag{8.4}$$

Значения параметра A, вычисленные для различных n и β представлены графически на рисунке 8.1.

Использование графиков на рисунке 8.1 позволяет приближённо определить необходимое число опытов для оценки математического ожидания с заданной погрешностью. Для этого приняв значение допустимой погрешности Δm_x^* и используя значение статистической дисперсии D_x^* , определим параметр A как

$$A = \frac{\Delta m_x^*}{\sqrt{D_x^*}} \quad . \tag{8.5}$$

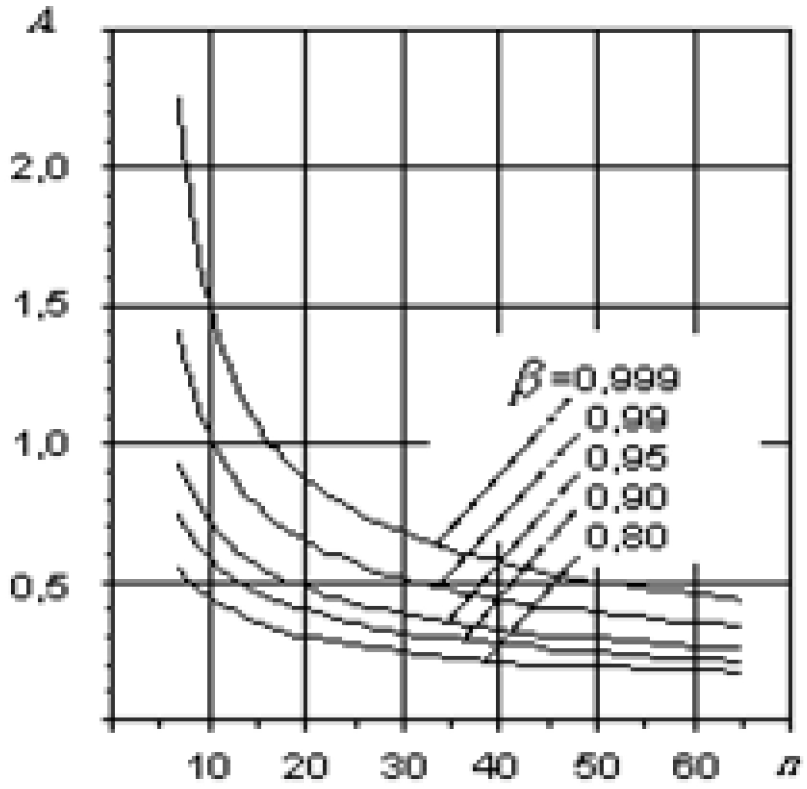

Рис. 8.1 – К планированию оценки математического ожидания

Далее, отложив полученное значение на оси ординат, смещаемся вправо до кривой соответствующей заданной доверительной вероятности и спускаемся на ось абсцисс, где и отметим необходимое число опытов.

Чтобы воспользоваться рассмотренной методикой, необходимо знать оценку дисперсии D_x^*. Приближённо D_x^* определяют следующими способами [2]:

- Используется значение Dx*, полученное для предыдущих выборок, если аналогичные исследования уже проводились.
- Применяется так называемое правило "трёх сигма", из которого следует, что если известно максимальное значение исследуемой случайной величины x_{max} и минимальное значение x_{min}, то для случайной величины, распределённой по нормальному закону, можно приближённо принять

$$D_x^* = (\frac{x_{max} - x_{min}}{6})^2 \quad .$$

Последовательное уточнение необходимого числа опытов, при котором сначала зададимся небольшим числом опытов, например, $n_1 = 10$ и выполнив их, вычислим оценки m_{x1}^* и D_{x1}^* (смотри разделы 4.1.1 и 4.1.2). По приведенной выше методике оценим необходимое число опытов n.

Если n > n1, то выполняется ещё некоторое число опытов, например, Δn = 10 и тогда с учётом уже выполненных опытов имеем выборку n2 = n1+Δn, из которой получаем оценки m_{x2}^* и D_{x2}^* и заново определяем потребное число опытов.

Увеличение объёма выборки и проведение дополнительных опытов проводится до тех пор, пока число выполненных опытов не станет меньше потребного.

Пример. На предприятии с конвейерным производством осуществляется выборочный контроль качества выпускаемой продукции. При оценке предыдущей выборки были получены оценки математического ожидания и среднего квадратического отклонения

$$m_x^* = 35{,}10 \text{ и } S_x^* = 4{,}17.$$

Требуется определить объём (число элементов) следующей выборки, которое позволит оценить математическое ожидание исследуемой величины с абсолютной погрешностью ±2,50 при доверительной вероятности β = 0,95.

Сначала по формуле (8.5), учитывая $D_x = S_{x2}$, определим значение параметра A

$$A = \frac{2{,}5}{4{,}17} = 0{,}60 \ .$$

Затем, выбрав на рисунке 8.1 кривую для вероятности 0,95 и отметив A = 0,60, находим необходимое число опытов n = 14 .

8.2. Планирование оценивания дисперсии

Определение объёма выборки **n** , необходимого для оценки дисперсии с необходимой точностью основывается на рассмотрении доверительного интервала дисперсии, который представляется (смотри раздел 6.2) в виде

$$J_\beta(D)=(D_1; D_2) \ .$$

Если рассматривается случайная величина с нормальным законом распределения, то ближайшая к оценке дисперсии граница доверительного интервала определяется из распределения Пирсона выражением

$$D_1 = \frac{D^*(n-1)}{\chi_1^2} \ . \qquad (8.6)$$

Задаваемое значение абсолютной погрешности дисперсии ΔD связано с левой границей доверительного интервала соотношением $D_1 = D^* - \Delta D$. Отсюда с учётом формулы (8.6) получим

$$\frac{D^*(n-1)}{\chi_1^2} = D^* - \Delta D \quad . \tag{8.7}$$

Вводя относительную погрешность $\varepsilon = \Delta D / D^*$, из последнего соотношения получим

$$n = (1 - \varepsilon)\chi^2_1 + 1 \; . \tag{8.8}$$

Значения χ^2_1 берутся по таблицам Пирсона П.7 в зависимости от $r = (n - 1)$

и $p_1 = (1 - \beta)/2$, β – где доверительная вероятность.

Достаточно сложное решение уравнения (8.8) относительно числа опытов n представлено графически на рисунке 8.2.

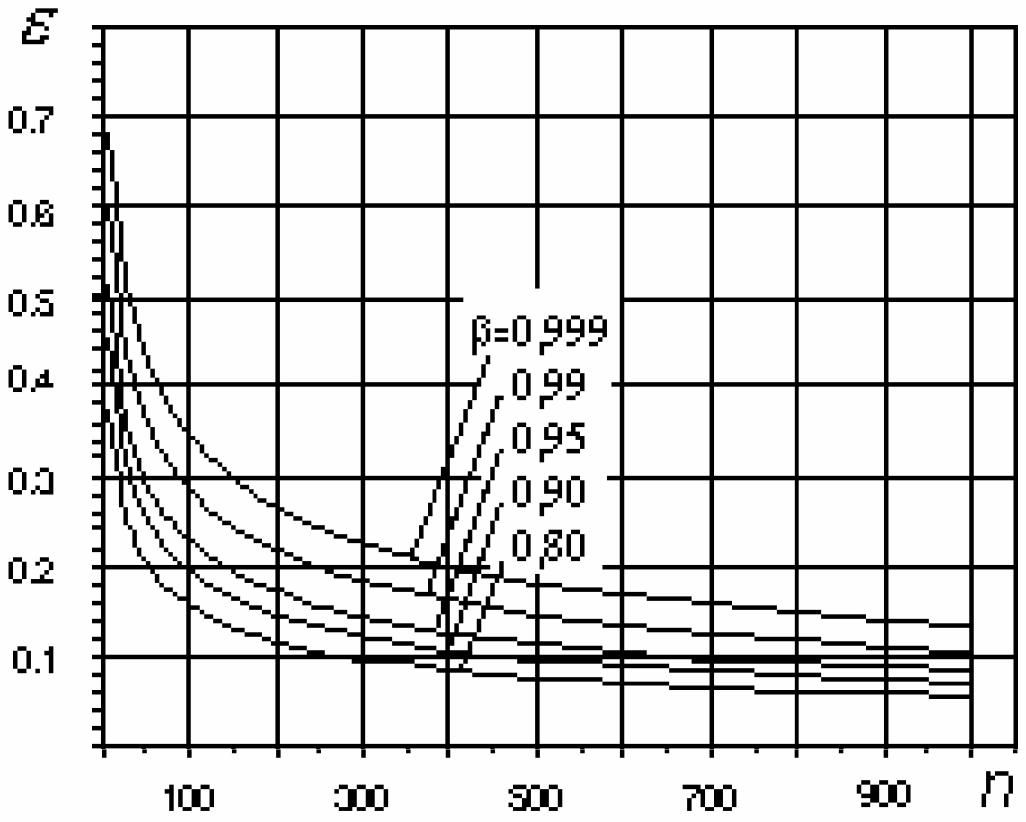

Рис. 8.2 – К планированию оценки дисперсии

Из приведенных рассуждений видно, что, определяя объём выборки необходимый для оценки дисперсии, получаем, однозначное решение, задаваясь относительной, а не абсолютной погрешностью.

Если рассматриваемая случайная величина, для которой планируется определение дисперсии, распределена по закону, отличному от нормального, то приведенная методика может приводить к существенным ошибкам.

Возвратившись к абсолютной погрешности дисперсии ΔD , отметим, что её планирование намного сложнее. При необходимости контроля этой величины можно предложить следующее решение.

Задавшись некоторой относительной погрешностью ε, найдём потребное число опытов n. Затем, после выполнения этих опытов, определим статистическую оценку D_x^* (смотри раздел 4.1.1) и на основании формулы (8.7) вычислим абсолютную погрешность

$$\Delta D = D_x^*(1-\frac{n-1}{\chi_1^2}) \quad , \tag{8.9}$$

выбрав значение χ^2_1 по таблицам Пирсона П.7 в зависимости от $r = (n-1)$ и $p_1 = (1-\beta)/2$.

Убедившись, что полученное значение абсолютной погрешности ΔD нас удовлетворяет, считаем выполненное число опытов достаточным. В противном случае проводятся дополнительные опыты.

Пример. Необходимо определить число опытов, которое позволит получить оценку дисперсии случайной величины, подчиняющейся нормальному закону распределения, с относительной погрешностью 15% при доверительной вероятности 0,90.

Воспользовавшись графиком на рисунке 8.2, получим $n = 200$. Из того же графика видно, что, при планировании вдвое меньшего числа опытов $n = 100$, с принятой доверительной вероятностью 0,90 возможное значение относительной погрешности увеличивается до 20%.

Сопоставляя полученный результат с приведенным ранее примером планирования оценки математического ожидания можно убедиться, что, выполнив определённое число опытов и получив статистические оценки математического ожидания и дисперсии, мы всегда получаем оценку дисперсии с гораздо меньшей точностью по сравнению с точностью оценки математического ожидания для этой же выборки.

8.3. Планирование оценивания вероятности наблюдаемых событий

Обратимся к приближённому построению доверительного интервала вероятности наблюдаемых событий $J_\beta(p) = (p_1; p_2)$, рассмотренному в разделе 6.3.

Границы доверительного интервала определяются с помощью распределения Стьюдента в виде

$$p_{1,2} = p^* \bullet t_\beta \sqrt{\frac{p^*(1-p^*)}{n}} \quad , \quad \text{где} \tag{8.10}$$

p* – частота появления события (статистическая вероятность),

t_β – параметр распределения Стьюдента, определяемый по таблице П.4 для заданной доверительной вероятности β и числа опытов n–1.

Погрешность оценки вероятности на основании формулы (8.10) определяется как

$$\varepsilon = t_\beta \sqrt{\frac{p^*(1-p^*)}{n}} \quad . \tag{8.11}$$

Решение уравнения (8.11) относительно числа опытов n весьма затруднительно, так как от n зависит параметр t_β. Поэтому обратимся к приближённому решению, записав уравнение (8.11) в виде

$$\begin{gathered} n - A \cdot t_\beta^2 = 0 \quad , \\ \text{где} \quad A = p^*(1-p^*)/\varepsilon^2 \ . \end{gathered} \tag{8.12}$$

Решение уравнения (8.12) для заданных значений параметра A представлено графически на рисунке 8.3.

Чтобы воспользоваться графиками на рисунке 8.3, следует для заданной погрешности ε и принятой величины частоты события p* вычислить значение параметра A. Затем, непосредственно, из графика для соответствующей доверительной вероятности определяется необходимое число опытов n.

Очевидным неудобством в данном случае является необходимость предварительной оценки частоты события p*.

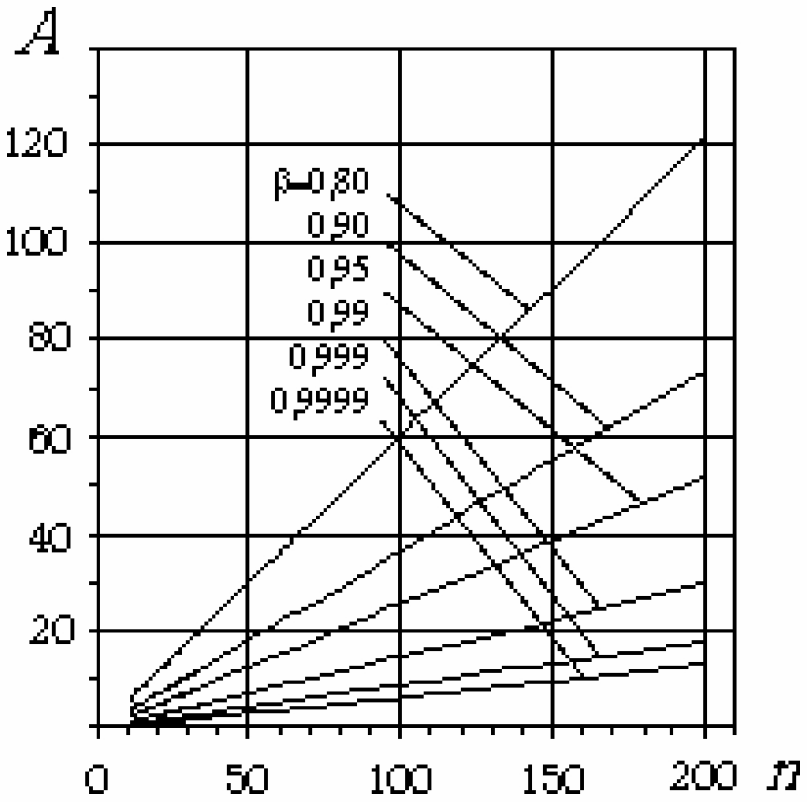

Рис. 8.3 – К планированию оценивания вероятности

Отклонение распределения рассматриваемой случайной величины от нормального закона распределения в данной задаче большого значения не имеет, так как частота события по закону больших чисел имеет распределение, приближающееся при увеличении числа опытов к нормальному закону.

Пример. При выполнении 100 опытов получено 25 благоприятных исходов рассматриваемого события, поэтому полученная частота события принимается в качестве его вероятности

$$p^* = \tfrac{25}{100} = 0,25 \quad .$$

Можно ли с доверительной вероятностью β=0,90 утверждать, что полученная оценка имеет абсолютную погрешность менее 0,10 ?

По формуле (8.12) вычислим $A = 0,25 \cdot (1 - 0,25) / 0,1^2 = 18,75$.

На рисунке 8.3 для β = 0,90 находим n = 55. Потребное число опытов оказалось меньше 100, поэтому полученная оценка вероятности обладает желаемой точностью.

8.4. Планирование оценивания вероятности редких событий

Допустим, что планируется наблюдение редких событий, которые при проведении опытов могут ни разу не реализоваться. В разделе 8.4 рассматривалось построение доверительного интервала для такого случая и показано, что его левая граница равна нулю, то есть $J_\beta(p)=(0\ ;\ p_2)$.

Отсутствие результата при проведении заданного числа опытов в данном случае позволяет назначить верхнее возможное значение вероятности события p_2.

Рассмотрим обратную задачу, определения необходимого числа безрезультативных опытов, проведение которых позволяет назначить верхнюю границу вероятности редкого события с доверительной вероятностью β. Непосредственно из формулы (6.15) следует решение этой задачи в виде

$$n = \frac{\lg(1-\beta)}{\lg(1-p_2)} \quad . \tag{8.13}$$

Пример. Требуется определить, сколько безрезультативных опытов следует запланировать, чтобы с вероятностью 0,95 назначить верхнее значение вероятности рассматриваемого события равное 0,05.

По формуле (8.13) получаем

$$n = \frac{\lg(1-0{,}95)}{\lg(1-0{,}05)} = 58{,}4 \approx 58 \quad .$$

Смысл этого результата заключается в том, что если при выполнении 58 опытов рассматриваемое событие не реализуется ни разу, то с вероятностью 0,95 можно принять максимальное возможное значение вероятности данного события равное 0,05.

ПРИЛОЖЕНИЕ

Таблица П.1.

Нормальное распределение $F(x) = P(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{t^2}{2})dt$

x	F (x)	x	F (x)	x	F (x)	x	F (x)
0,00	0,50000	0,76	0,77637	1,52	0,93574	2,28	0,98870
0,02	0,50798	0,78	0,78230	1,54	0,93822	2,30	0,98928
0,04	0,51595	0,80	0,78814	1,56	0,94062	2,32	0,98983
0,06	0,52392	0,82	0,79389	1,58	0,94295	2,34	0,99036
0,08	0,53188	0,84	0,79955	1,60	0,94520	2,36	0,99086
0,10	0,53983	0,86	0,80511	1,62	0,94738	2,38	0,99134
0,12	0,54776	0,88	0,81057	1,64	0,94950	2,40	0,99180
0,14	0,55567	0,90	0,81594	1,66	0,95154	2,42	0,99224
0,16	0,56356	0,92	0,82121	1,68	0,95352	2,44	0,99266
0,18	0,57142	0,94	0,82639	1,70	0,95543	2,46	0,99305
0,20	0,57926	0,96	0,83147	1,72	0,95728	2,48	0,99343
0,22	0,58706	0,98	0,83646	1,74	0,95907	2,50	0,99379
0,24	0,59483	1,00	0,84134	1,76	0,96080	2,52	0,99413
0,26	0,60257	1,02	0,84614	1,78	0,96246	2,54	0,99446
0,28	0,61026	1,04	0,85083	1,80	0,96407	2,56	0,99477
0,30	0,61791	1,06	0,85543	1,82	0,96562	2,58	0,99506
0,32	0,62552	1,08	0,85993	1,84	0,96712	2,60	0,99534
0,34	0,63307	1,10	0,86433	1,86	0,96856	2,62	0,99560
0,36	0,64058	1,12	0,86864	1,88	0,96995	2,64	0,99585
0,38	0,64803	1,14	0,87286	1,90	0,97128	2,66	0,99609
0,40	0,65542	1,16	0,87698	1,92	0,97257	2,68	0,99632
0,42	0,66276	1,18	0,88100	1,94	0,97381	2,70	0,99653
0,44	0,67003	1,20	0,88493	1,96	0,97500	2,72	0,99674
0,46	0,67724	1,22	0,88877	1,98	0,97615	2,74	0,99693
0,48	0,68439	1,24	0,89251	2,00	0,97725	2,76	0,99711
0,50	0,69146	1,26	0,89617	2,02	0,97831	2,78	0,99728
0,52	0,69847	1,28	0,89973	2,04	0,97932	2,80	0,99744
0,54	0,70540	1,30	0,90320	2,06	0,98030	2,82	0,99760
0,56	0,71226	1,32	0,90658	2,08	0,98124	2,84	0,99774
0,58	0,71904	1,34	0,90988	2,10	0,98214	2,86	0,99788
0,60	0,72575	1,36	0,91308	2,12	0,98300	2,88	0,99801
0,62	0,73237	1,38	0,91621	2,14	0,98382	2,90	0,99813
0,64	0,73891	1,40	0,91924	2,16	0,98461	2,92	0,99825
0,66	0,74537	1,42	0,92220	2,18	0,98537	2,94	0,99836
0,68	0,75175	1,44	0,92507	2,20	0,98610	2,96	0,99846
0,70	0,75804	1,46	0,92785	2,22	0,98679	2,98	0,99856
0,72	0,76424	1,48	0,93056	2,24	0,98745	3,00	0,99865
0,74	0,77035	1,50	0,93319	2,26	0,98809	3,20	0,99931

Примечание: для отрицательных аргументов $F(-|x|) = 1 - F(+|x|)$.

Таблица П.2.

Критерий согласия Колмогорова. Значения $\beta=P(\lambda)$

λ	Сотые доли λ									
	0	1	2	3	4	5	6	7	8	9
0, 3	0,99999	0,99998	0,99995	0,99991	0,99983	0,9997	0,99949	0,99917	0,99872	0,99807
0, 4	0,99719	0,99603	0,99452	0,99262	0,99027	0,9874	0,98400	0,97998	0,97532	0,96998
0, 5	0,96394	0,95719	0,94969	0,94147	0,93250	0,9228	0,91242	0,90134	0,88960	0,87724
0, 6	0,86428	0,85077	0,83678	0,82225	0,80732	0,7920	0,77636	0,76042	0,74422	0,72781
0, 7	0,71124	0,69453	0,67774	0,66089	0,64402	0,6271	0,61036	0,59363	0,57700	0,56050
0, 8	0,54414	0,52796	0,51197	0,49619	0,48063	0,4653	0,45026	0,43545	0,42093	0,40668
0, 9	0,39273	0,37907	0,36571	0,35266	0,33992	0,3274	0,31536	0,30356	0,29206	0,28087
1, 0	0,27000	0,25943	0,24917	0,23922	0,22957	0,2202	0,21114	0,20236	0,19387	0,18566
1, 1	0,17772	0,17005	0,16264	0,15550	0,14861	0,1419	0,13556	0,12939	0,12345	0,11774
1, 2	0,11225	0,10697	0,10190	0,09703	0,09235	0,0878	0,08357	0,07944	0,07550	0,07171
1, 3	0,06809	0,06463	0,06132	0,05815	0,05513	0,0522	0,04949	0,04686	0,04435	0,04196
1, 4	0,03968	0,03751	0,03545	0,03348	0,03162	0,0298	0,02815	0,02655	0,02503	0,02359
1, 5	0,02222	0,02092	0,01969	0,01852	0,01742	0,0163	0,01539	0,01446	0,01357	0,01274
1, 6	0,01195	0,01121	0,01051	0,00985	0,00922	0,0086	0,00808	0,00756	0,00707	0,00661
1. 7	0,00618	0,00577	0,00539	0,00503	0,00469	0,0043	0,00408	0,00380	0,00354	0,00330
1, 8	0,00307	0,00285	0,00265	0,00247	0,00229	0,0021	0,00198	0,00186	0,00170	0,00158
1, 9	0,00146	0,00136	0,00126	0,00116	0,00108	0,0010	0,00092	0,00085	0,00079	0,00073
2, 0	0,00067	0,00062	0,00057	0,00053	0,00048	0,0004	0,00041	0,00038	0,00035	0,00032
2, 1	0,00030	0,00027	0,00025	0,00023	0,00021	0,0001	0,00018	0,00016	0,00015	0,00014
2, 2	0,00013	0,00011	0,00010	0,00010	0,00009	0,0000	0,00007	0,00007	0,00006	0,00006
2. 3	0,00005	0,00005	0,00004	0,00004	0,00004	0,0000	0,00003	0,00003	0,00002	0,00002
2, 4	0,00002	0,00002	0,00002	0,00001	0,00001	0,0000	0,00001	0,00001	0,00001	0,00001

Распределение Пирсона χ^2 Таблица П.3

p	r												
	7	8	9	10	11	12	13	14	15	16	17	18	25
0,0001	29,881	31,827	33,725	35,557	37,365	39,131	40,873	42,575	44,260	45,926	47,559	49,185	60,136
0,0010	24,321	26,124	27,877	29,588	31,264	32,909	34,527	36,124	37,698	39,252	40,791	42,312	52,619
0,01	18,475	20,090	21,666	23,209	24,725	26,217	27,688	29,141	30,578	32,000	33,409	34,805	44,314
0,05	14,067	15,507	16,919	18,307	19,675	21,026	22,362	23,685	24,996	26,296	27,587	28,869	37,652
0,10	12,017	13,362	14,684	15,987	17,275	18,549	19,812	21,064	22,307	23,542	24,769	25,989	34,382
0,15	10,748	12,027	13,288	14,534	15,767	16,989	18,202	19,406	20,603	21,793	22,977	24,155	32,282
0,20	9,803	11,030	12,242	13,442	14,631	15,812	16,985	18,151	19,311	20,465	21,615	22,760	30,675
0,25	9,037	10,219	11,389	12,549	13,701	14,845	15,984	17,117	18,245	19,369	20,489	21,605	29,339
0,30	8,383	9,524	10,656	11,781	12,899	14,011	15,119	16,222	17,322	18,418	19,511	20,601	28,172
0,35	7,806	8,909	10,006	11,097	12,184	13,266	14,345	15,421	16,494	17,565	18,633	19,699	27,118
0,40	7,283	8,351	9,414	10,473	11,530	12,584	13,636	14,685	15,733	16,780	17,824	18,868	26,143
0,45	6,800	7,833	8,863	9,892	10,920	11,946	12,972	13,996	15,020	16,042	17,065	18,086	25,222
0,50	6,346	7,344	8,343	9,342	10,341	11,340	12,340	13,339	14,339	15,338	16,338	17,338	24,337
0,55	5,913	6,877	7,843	8,812	9,783	10,755	11,729	12,703	13,679	14,656	15,633	16,611	23,472
0,60	5,493	6,423	7,357	8,295	9,237	10,182	11,129	12,078	13,030	13,983	14,937	15,893	22,616
0,65	5,082	5,975	6,876	7,783	8,695	9,612	10,532	11,455	12,381	13,310	14,241	15,174	21,752
0,70	4,671	5,527	6,393	7,267	8,148	9,034	9,926	10,821	11,721	12,624	13,531	14,440	20,867
0,70	4,671	5,527	6,393	7,267	8,148	9,034	9,926	10,821	11,721	12,624	13,531	14,440	20,867
0,80	3,822	4,594	5,380	6,179	6,989	7,807	8,634	9,467	10,307	11,152	12,002	12,857	18,940
0,85	3,358	4,078	4,817	5,570	6,336	7,114	7,901	8,696	9,499	10,309	11,125	11,946	17,818
0,90	2,833	3,490	4,168	4,865	5,578	6,304	7,041	7,790	8,547	9,312	10,085	10,865	16,473
0,95	2,167	2,733	3,325	3,940	4,575	5,226	5,892	6,571	7,261	7,962	8,672	9,390	14,611
0,99	1,239	1,647	2,088	2,558	3,053	3,571	4,107	4,660	5,229	5,812	6,408	7,015	11,524
0,9990	0,599	0,857	1,152	1,479	1,834	2,214	2,617	3,041	3,483	3,942	4,416	4,905	8,649
0,9999	0,300	0,463	0,661	0,889	1,145	1,428	1,734	2,060	2,408	2,774	3,156	3,556	6,709

Распределение Стьюдента. Значения t_β, удовлетворяющие равенству $\beta = 2\int_0^{t_\beta} S_{n-1}(t)dt$ Таблица П.4

n-1	β													
	0,001	**0,01**	**0,05**	**0,10**	**0,20**	**0,30**	**0,40**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**0,99**	**0,999**
7	0,0012989	0,012988	0,064989	0,130293	0,263167	0,40154	0,54911	0,71114	0,89603	1,11916	1,41492	1,89458	3,49948	5,40807
8	0,0012926	0,012931	0,064700	0,129708	0,261921	0,39947	0,54593	0,70639	0,88889	1,10815	1,39682	1,85955	3,35538	5,04137
9	0,0012886	0,012886	0,064477	0,129253	0,260956	0,39787	0,54348	0,70272	0,88340	1,09972	1,38303	1,83311	3,24984	4,78089
10	0,0012847	0,012850	0,064298	0,128890	0,26018	0,39659	0,54153	0,69981	0,87906	1,09306	1,37218	1,81246	3,16926	4,58676
11	0,0012824	0,012822	0,064152	0,128594	0,25956	0,39555	0,53994	0,69744	0,87553	1,08767	1,36343	1,79588	3,10582	4,43688
12	0,0012795	0,012797	0,064031	0,128347	0,25903	0,39469	0,53862	0,69548	0,87261	1,08321	1,35622	1,78229	3,05454	4,31784
13	0,0012778	0,012777	0,063928	0,128139	0,25859	0,39396	0,53750	0,69383	0,87015	1,07947	1,35017	1,77093	3,01228	4,22093
14	0,0012756	0,012759	0,063840	0,127961	0,25821	0,39333	0,53655	0,69242	0,86805	1,07628	1,34503	1,76131	2,97685	4,14031
15	0,0012744	0,012744	0,063764	0,127806	0,25788	0,39279	0,53573	0,69120	0,86624	1,07353	1,34061	1,75305	2,94673	4,07279
16	0,0012733	0,012731	0,063698	0,127671	0,25760	0,39232	0,53501	0,69013	0,86467	1,07114	1,33676	1,74588	2,92079	4,01487
17	0,0012722	0,012719	0,063639	0,127552	0,25735	0,39190	0,53438	0,68919	0,86328	1,06903	1,33338	1,73961	2,89823	3,96511
18	0,0012710	0,012709	0,063587	0,127446	0,25712	0,39153	0,53382	0,68836	0,86205	1,06717	1,33039	1,73406	2,87844	3,92174
19	0,0012699	0,012699	0,063540	0,127352	0,25692	0,39120	0,53331	0,68762	0,86095	1,06551	1,32773	1,72913	2,86094	3,88332
20	0,0012687	0,012691	0,063499	0,127267	0,25674	0,39091	0,53286	0,68695	0,85996	1,06402	1,32534	1,72472	2,84534	3,84956
21	0,0012682	0,012683	0,063461	0,127190	0,25658	0,39064	0,53246	0,68635	0,85907	1,06267	1,32319	1,72074	2,83137	3,81930
22	0,0012676	0,012677	0,063426	0,127120	0,25643	0,39039	0,53209	0,68581	0,85827	1,06145	1,32124	1,71714	2,81876	3,79223
23	0,0012670	0,012670	0,063395	0,127056	0,25630	0,39017	0,53175	0,68531	0,85753	1,06034	1,31946	1,71387	2,80734	3,76764
24	0,0012665	0,012665	0,063366	0,126998	0,25617	0,38997	0,53144	0,68485	0,85686	1,05932	1,31784	1,71088	2,79695	3,74537
25	0,0012659	0,012660	0,063339	0,126944	0,25606	0,38978	0,53115	0,68443	0,85624	1,05838	1,31635	1,70814	2,78744	3,72514
26	0,0012653	0,012654	0,063315	0,126895	0,25595	0,38961	0,53089	0,68404	0,85567	1,05752	1,31497	1,70562	2,77872	3,70666
27	0,0012653	0,012650	0,063292	0,126849	0,25586	0,38945	0,53065	0,68369	0,85514	1,05673	1,31370	1,70329	2,77068	3,68949
28	0,0012642	0,012646	0,063271	0,126806	0,25577	0,38930	0,53042	0,68335	0,85465	1,05599	1,31253	1,70113	2,76326	3,67392
29	0,0012642	0,012642	0,063252	0,126767	0,25568	0,38916	0,53021	0,68304	0,85419	1,05530	1,31143	1,69913	2,75639	3,65952
30	0,0012636	0,012639	0,063234	0,126730	0,25561	0,38903	0,53002	0,68276	0,85377	1,05466	1,31042	1,69726	2,74998	3,64598

Таблица П.4 (продолжение)

Распределение Стьюдента. Значения t_β, удовлетворяющие равенству $\beta = 2\int_0^{t_\beta} S_{n\text{-}1}(t)dt$

n-1	β													
	0,001	**0,01**	**0,05**	**0,10**	**0,20**	**0,30**	**0,40**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**0,99**	**0,999**
31	0,0012636	0,012635	0,063217	0,126695	0,25553	0,38891	0,52984	0,68249	0,85337	1,05406	1,30946	1,69552	2,74404	3,63347
32	0,0012631	0,012632	0,063201	0,126662	0,25546	0,38880	0,52966	0,68223	0,85300	1,05350	1,30857	1,69389	2,73849	3,62183
33	0,0012631	0,012629	0,063185	0,126632	0,25540	0,38869	0,52950	0,68200	0,85265	1,05298	1,30774	1,69236	2,73329	3,61091
34	0,0012625	0,012626	0,063171	0,126603	0,25534	0,38859	0,52935	0,68177	0,85232	1,05249	1,30695	1,69092	2,72839	3,60073
35	0,0012619	0,012623	0,063158	0,126577	0,25528	0,38850	0,52921	0,68156	0,85201	1,05202	1,30621	1,68957	2,72381	3,59112
36	0,0012619	0,012621	0,063145	0,126551	0,25523	0,38841	0,52908	0,68137	0,85172	1,05158	1,30551	1,68830	2,71948	3,58210
37	0,0012619	0,012618	0,063134	0,126527	0,25518	0,38833	0,52895	0,68118	0,85144	1,05116	1,30485	1,68709	2,71541	3,57366
38	0,0012619	0,012616	0,063122	0,126504	0,25513	0,38825	0,52883	0,68100	0,85118	1,05077	1,30423	1,68595	2,71157	3,56566
39	0,0012614	0,012614	0,063112	0,126482	0,25508	0,38817	0,52871	0,68083	0,85093	1,05040	1,30364	1,68488	2,70791	3,55809
40	0,0012614	0,012612	0,063101	0,126462	0,25504	0,38810	0,52861	0,68067	0,85070	1,05005	1,30308	1,68385	2,70446	3,55096
41	0,0012608	0,012610	0,063092	0,126443	0,25500	0,38803	0,52850	0,68052	0,85048	1,04971	1,30254	1,68288	2,70118	3,54426
42	0,0012608	0,012608	0,063083	0,126423	0,25496	0,38797	0,52840	0,68038	0,85026	1,04939	1,30203	1,68195	2,69807	3,53772
44	0,0012608	0,012605	0,063065	0,126389	0,25488	0,38785	0,52822	0,68011	0,84987	1,04879	1,30109	1,68023	2,69229	3,52578
46	0,0012602	0,012602	0,063050	0,126357	0,25482	0,38774	0,52805	0,67986	0,84951	1,04825	1,30023	1,67866	2,68701	3,51487
48	0,0012597	0,012599	0,063035	0,126328	0,25476	0,38763	0,52790	0,67964	0,84917	1,04775	1,29944	1,67722	2,68221	3,50497
50	0,0012597	0,012597	0,063022	0,126302	0,25470	0,38754	0,52776	0,67943	0,84887	1,04729	1,29871	1,67591	2,67779	3,49595
52	0,0012597	0,012594	0,063010	0,126277	0,25465	0,38746	0,52763	0,67924	0,84859	1,04687	1,29804	1,67469	2,67373	3,48766
54	0,0012591	0,012592	0,062999	0,126254	0,25460	0,38738	0,52751	0,67906	0,84833	1,04648	1,29743	1,67357	2,66999	3,47995
56	0,0012591	0,012590	0,062988	0,126232	0,25455	0,38730	0,52740	0,67890	0,84809	1,04612	1,29685	1,67252	2,66651	3,47296
58	0,0012585	0,012587	0,062979	0,126213	0,25451	0,38723	0,52729	0,67874	0,84786	1,04578	1,29632	1,67155	2,66329	3,46627
60	0,0012585	0,012586	0,062970	0,126195	0,25447	0,38717	0,52720	0,67860	0,84765	1,04547	1,29582	1,67065	2,66027	3,46015

Таблица П.5

Распределение Р.Фишера (уровень значимости расхождений $\beta = 0{,}01$)

f_2	f_1																				
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
10	4,849	4,772	4,706	4,650	4,601	4,558	4,520	4,487	4,457	4,430	4,405	4,383	4,363	4,344	4,327	4,311	4,296	4,283	4,270	4,258	4,247
11	4,539	4,462	4,397	4,342	4,293	4,251	4,213	4,180	4,150	4,123	4,099	4,077	4,057	4,038	4,021	4,005	3,990	3,977	3,964	3,952	3,941
12	4,296	4,220	4,155	4,100	4,052	4,010	3,972	3,939	3,910	3,883	3,858	3,836	3,816	3,798	3,780	3,765	3,750	3,736	3,724	3,712	3,701
13	4,100	4,025	3,960	3,905	3,857	3,815	3,778	3,745	3,716	3,689	3,665	3,643	3,622	3,604	3,587	3,571	3,556	3,543	3,530	3,518	3,507
14	3,939	3,864	3,800	3,745	3,698	3,656	3,619	3,586	3,556	3,529	3,505	3,483	3,463	3,444	3,427	3,412	3,397	3,383	3,371	3,359	3,348
15	3,805	3,730	3,666	3,612	3,564	3,522	3,485	3,452	3,423	3,396	3,372	3,350	3,330	3,311	3,294	3,278	3,264	3,250	3,237	3,225	3,214
16	3,691	3,616	3,553	3,498	3,451	3,409	3,372	3,339	3,310	3,283	3,259	3,237	3,216	3,198	3,181	3,165	3,150	3,137	3,124	3,112	3,101
17	3,593	3,518	3,455	3,401	3,353	3,312	3,275	3,242	3,212	3,186	3,162	3,139	3,119	3,101	3,083	3,068	3,053	3,039	3,026	3,014	3,003
18	3,508	3,434	3,371	3,316	3,269	3,227	3,190	3,158	3,128	3,101	3,077	3,055	3,035	3,016	2,999	2,983	2,968	2,955	2,942	2,930	2,919
19	3,434	3,360	3,297	3,242	3,195	3,153	3,116	3,084	3,054	3,027	3,003	2,981	2,961	2,942	2,925	2,909	2,894	2,880	2,868	2,855	2,844
20	3,368	3,294	3,231	3,177	3,130	3,088	3,051	3,018	2,989	2,962	2,938	2,916	2,895	2,877	2,859	2,843	2,829	2,815	2,802	2,790	2,778
21	3,310	3,236	3,173	3,119	3,072	3,030	2,993	2,960	2,931	2,904	2,880	2,857	2,837	2,818	2,801	2,785	2,770	2,756	2,743	2,731	2,720
22	3,258	3,184	3,121	3,067	3,019	2,978	2,941	2,908	2,879	2,852	2,827	2,805	2,785	2,766	2,749	2,733	2,718	2,704	2,691	2,679	2,667
23	3,211	3,137	3,074	3,020	2,973	2,931	2,894	2,861	2,832	2,805	2,780	2,758	2,738	2,719	2,702	2,686	2,671	2,657	2,644	2,632	2,620
24	3,168	3,094	3,032	2,977	2,930	2,889	2,852	2,819	2,789	2,762	2,738	2,716	2,695	2,676	2,659	2,643	2,628	2,614	2,601	2,589	2,577
25	3,129	3,056	2,993	2,939	2,892	2,850	2,813	2,780	2,751	2,724	2,699	2,677	2,657	2,638	2,620	2,604	2,589	2,575	2,562	2,550	2,538
26	3,094	3,021	2,958	2,904	2,857	2,815	2,778	2,745	2,715	2,688	2,664	2,642	2,621	2,602	2,585	2,569	2,554	2,540	2,526	2,514	2,503
27	3,062	2,988	2,926	2,872	2,824	2,783	2,746	2,713	2,683	2,656	2,632	2,609	2,589	2,570	2,552	2,536	2,521	2,507	2,494	2,481	2,470
28	3,032	2,959	2,896	2,842	2,795	2,753	2,716	2,683	2,653	2,626	2,602	2,579	2,559	2,540	2,522	2,506	2,491	2,477	2,464	2,451	2,440
29	3,005	2,931	2,868	2,814	2,767	2,726	2,689	2,656	2,626	2,599	2,574	2,552	2,531	2,512	2,495	2,478	2,463	2,449	2,436	2,423	2,412
30	2,979	2,906	2,843	2,789	2,742	2,700	2,663	2,630	2,600	2,573	2,549	2,526	2,506	2,487	2,469	2,453	2,437	2,423	2,410	2,398	2,386

Таблица П-5 (продолжение)

Распределение Р.Фишера (уровень значимости расхождений $\beta = 0{,}05$)

f_2	f_1																				
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
10	2,978	2,943	2,913	2,887	2,865	2,845	2,828	2,812	2,798	2,785	2,774	2,764	2,754	2,745	2,737	2,730	2,723	2,716	2,710	2,705	2,700
11	2,854	2,818	2,788	2,761	2,739	2,719	2,701	2,685	2,671	2,658	2,646	2,636	2,626	2,617	2,609	2,601	2,594	2,588	2,582	2,576	2,570
12	2,753	2,717	2,687	2,660	2,637	2,617	2,599	2,583	2,568	2,555	2,544	2,533	2,523	2,514	2,505	2,498	2,491	2,484	2,478	2,472	2,466
13	2,671	2,635	2,604	2,577	2,554	2,533	2,515	2,499	2,484	2,471	2,459	2,448	2,438	2,429	2,420	2,412	2,405	2,398	2,392	2,386	2,380
14	2,602	2,565	2,534	2,507	2,484	2,463	2,445	2,428	2,413	2,400	2,388	2,377	2,367	2,357	2,349	2,341	2,333	2,326	2,320	2,314	2,308
15	2,544	2,507	2,475	2,448	2,424	2,403	2,385	2,368	2,353	2,340	2,328	2,316	2,306	2,297	2,288	2,280	2,272	2,265	2,259	2,253	2,247
16	2,494	2,456	2,425	2,397	2,373	2,352	2,333	2,317	2,302	2,288	2,276	2,264	2,254	2,244	2,235	2,227	2,220	2,212	2,206	2,200	2,194
17	2,450	2,413	2,381	2,353	2,329	2,308	2,289	2,272	2,257	2,243	2,230	2,219	2,208	2,199	2,190	2,181	2,174	2,167	2,160	2,154	2,148
18	2,412	2,374	2,342	2,314	2,290	2,269	2,250	2,233	2,217	2,203	2,191	2,179	2,168	2,159	2,150	2,141	2,134	2,126	2,119	2,113	2,107
19	2,378	2,340	2,308	2,280	2,256	2,234	2,215	2,198	2,182	2,168	2,155	2,144	2,133	2,123	2,114	2,106	2,098	2,090	2,084	2,077	2,071
20	2,348	2,310	2,278	2,250	2,225	2,203	2,184	2,167	2,151	2,137	2,124	2,112	2,102	2,092	2,082	2,074	2,066	2,059	2,052	2,045	2,039
21	2,321	2,283	2,250	2,222	2,197	2,176	2,156	2,139	2,123	2,109	2,096	2,084	2,073	2,063	2,054	2,045	2,037	2,030	2,023	2,016	2,010
22	2,297	2,259	2,226	2,198	2,173	2,151	2,131	2,114	2,098	2,084	2,071	2,059	2,048	2,038	2,028	2,020	2,012	2,004	1,997	1,990	1,984
23	2,275	2,236	2,204	2,175	2,150	2,128	2,109	2,091	2,075	2,061	2,048	2,036	2,025	2,014	2,005	1,996	1,988	1,981	1,973	1,967	1,961
24	2,255	2,216	2,183	2,155	2,130	2,108	2,088	2,070	2,054	2,040	2,027	2,015	2,003	1,993	1,984	1,975	1,967	1,959	1,952	1,945	1,939
25	2,236	2,198	2,165	2,136	2,111	2,089	2,069	2,051	2,035	2,021	2,007	1,995	1,984	1,974	1,964	1,955	1,947	1,939	1,932	1,926	1,919
26	2,220	2,181	2,148	2,119	2,094	2,072	2,052	2,034	2,018	2,003	1,990	1,978	1,966	1,956	1,946	1,938	1,929	1,921	1,914	1,907	1,901
27	2,204	2,166	2,132	2,103	2,078	2,056	2,036	2,018	2,002	1,987	1,974	1,961	1,950	1,940	1,930	1,921	1,913	1,905	1,898	1,891	1,884
28	2,190	2,151	2,118	2,089	2,064	2,041	2,021	2,003	1,987	1,972	1,959	1,946	1,935	1,924	1,915	1,906	1,897	1,889	1,882	1,875	1,869
29	2,177	2,138	2,104	2,075	2,050	2,027	2,007	1,989	1,973	1,958	1,945	1,932	1,921	1,910	1,901	1,891	1,883	1,875	1,868	1,861	1,854
30	2,165	2,126	2,092	2,063	2,037	2,015	1,995	1,976	1,960	1,945	1,932	1,919	1,908	1,897	1,887	1,878	1,870	1,862	1,854	1,847	1,841

Таблица П.5 (продолжение)

Распределение Р.Фишера (уровень значимости расхождений $\beta = 0{,}10$)

f_2	f_1																				
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
10	2,323	2,302	2,284	2,269	2,255	2,244	2,233	2,224	2,215	2,208	2,201	2,194	2,189	2,183	2,178	2,174	2,170	2,166	2,162	2,159	2,155
11	2,248	2,227	2,209	2,193	2,179	2,167	2,156	2,147	2,138	2,130	2,123	2,117	2,111	2,105	2,100	2,095	2,091	2,087	2,083	2,080	2,076
12	2,188	2,166	2,147	2,131	2,117	2,105	2,094	2,084	2,075	2,067	2,060	2,053	2,047	2,041	2,036	2,031	2,027	2,022	2,019	2,015	2,011
13	2,138	2,116	2,097	2,080	2,066	2,053	2,042	2,032	2,023	2,014	2,007	2,000	1,994	1,988	1,983	1,978	1,973	1,969	1,965	1,961	1,958
14	2,095	2,073	2,054	2,037	2,022	2,010	1,998	1,988	1,978	1,970	1,962	1,955	1,949	1,943	1,938	1,933	1,928	1,923	1,919	1,916	1,912
15	2,059	2,037	2,017	2,000	1,985	1,972	1,961	1,950	1,941	1,932	1,924	1,917	1,911	1,905	1,899	1,894	1,889	1,885	1,880	1,876	1,873
16	2,028	2,005	1,985	1,968	1,953	1,940	1,928	1,917	1,908	1,899	1,891	1,884	1,877	1,871	1,866	1,860	1,855	1,851	1,847	1,843	1,839
17	2,001	1,978	1,958	1,940	1,925	1,912	1,900	1,889	1,879	1,870	1,862	1,855	1,848	1,842	1,836	1,831	1,826	1,821	1,817	1,813	1,809
18	1,977	1,954	1,933	1,916	1,900	1,887	1,875	1,864	1,854	1,845	1,837	1,829	1,823	1,816	1,810	1,805	1,800	1,795	1,791	1,787	1,783
19	1,956	1,932	1,912	1,894	1,878	1,865	1,852	1,841	1,831	1,822	1,814	1,807	1,800	1,793	1,787	1,782	1,777	1,772	1,767	1,763	1,759
20	1,937	1,913	1,892	1,875	1,859	1,845	1,833	1,821	1,811	1,802	1,794	1,786	1,779	1,773	1,767	1,761	1,756	1,751	1,746	1,742	1,738
21	1,920	1,896	1,875	1,857	1,841	1,827	1,815	1,803	1,793	1,784	1,776	1,768	1,761	1,754	1,748	1,742	1,737	1,732	1,728	1,723	1,719
22	1,904	1,880	1,859	1,841	1,825	1,811	1,798	1,787	1,777	1,768	1,759	1,751	1,744	1,737	1,731	1,726	1,720	1,715	1,711	1,706	1,702
23	1,890	1,866	1,845	1,827	1,811	1,796	1,784	1,772	1,762	1,753	1,744	1,736	1,729	1,722	1,716	1,710	1,705	1,700	1,695	1,691	1,686
24	1,877	1,853	1,832	1,814	1,797	1,783	1,770	1,759	1,748	1,739	1,730	1,722	1,715	1,708	1,702	1,696	1,691	1,686	1,681	1,676	1,672
25	1,866	1,841	1,820	1,802	1,785	1,771	1,758	1,746	1,736	1,726	1,718	1,710	1,702	1,695	1,689	1,683	1,678	1,672	1,668	1,663	1,659
26	1,855	1,830	1,809	1,790	1,774	1,760	1,747	1,735	1,724	1,715	1,706	1,698	1,690	1,683	1,677	1,671	1,666	1,660	1,656	1,651	1,647
27	1,845	1,820	1,799	1,780	1,764	1,749	1,736	1,724	1,714	1,704	1,695	1,687	1,680	1,673	1,666	1,660	1,655	1,649	1,645	1,640	1,636
28	1,836	1,811	1,790	1,771	1,754	1,740	1,726	1,715	1,704	1,694	1,685	1,677	1,669	1,662	1,656	1,650	1,644	1,639	1,634	1,630	1,625
29	1,827	1,802	1,781	1,762	1,745	1,731	1,717	1,705	1,695	1,685	1,676	1,668	1,660	1,653	1,647	1,640	1,635	1,630	1,625	1,620	1,616
30	1,819	1,794	1,773	1,754	1,737	1,722	1,709	1,697	1,686	1,676	1,667	1,659	1,651	1,644	1,638	1,632	1,626	1,621	1,616	1,611	1,606

Таблица П-5 (продолжение)

Распределение Р.Фишера (уровень значимости расхождений $\beta = 0{,}25$)

f_2	f_1																				
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
10	1,551	1,547	1,543	1,540	1,537	1,534	1,531	1,529	1,527	1,525	1,523	1,522	1,520	1,519	1,518	1,517	1,516	1,515	1,514	1,513	1,512
11	1,523	1,518	1,514	1,510	1,507	1,504	1,501	1,499	1,497	1,495	1,493	1,491	1,490	1,488	1,487	1,486	1,485	1,483	1,482	1,481	1,481
12	1,500	1,495	1,490	1,486	1,483	1,480	1,477	1,474	1,472	1,470	1,468	1,466	1,464	1,463	1,461	1,460	1,459	1,458	1,456	1,455	1,454
13	1,480	1,475	1,470	1,466	1,462	1,459	1,456	1,453	1,451	1,449	1,447	1,445	1,443	1,441	1,440	1,438	1,437	1,436	1,435	1,433	1,432
14	1,463	1,458	1,453	1,449	1,445	1,441	1,438	1,435	1,433	1,431	1,428	1,426	1,425	1,423	1,421	1,420	1,418	1,417	1,416	1,415	1,414
15	1,449	1,443	1,438	1,434	1,430	1,426	1,423	1,420	1,417	1,415	1,413	1,411	1,409	1,407	1,405	1,404	1,402	1,401	1,400	1,398	1,397
16	1,437	1,431	1,426	1,421	1,417	1,413	1,410	1,407	1,404	1,401	1,399	1,397	1,395	1,393	1,391	1,390	1,388	1,387	1,385	1,384	1,383
17	1,426	1,420	1,414	1,409	1,405	1,401	1,398	1,395	1,392	1,389	1,387	1,385	1,383	1,381	1,379	1,377	1,376	1,374	1,373	1,372	1,370
18	1,416	1,410	1,404	1,399	1,395	1,391	1,388	1,384	1,381	1,379	1,376	1,374	1,372	1,370	1,368	1,366	1,365	1,363	1,362	1,360	1,359
19	1,407	1,401	1,395	1,390	1,386	1,382	1,378	1,375	1,372	1,369	1,367	1,364	1,362	1,360	1,358	1,356	1,355	1,353	1,352	1,350	1,349
20	1,399	1,393	1,387	1,382	1,378	1,374	1,370	1,367	1,363	1,361	1,358	1,356	1,353	1,351	1,349	1,348	1,346	1,344	1,343	1,341	1,340
21	1,392	1,386	1,380	1,375	1,370	1,366	1,362	1,359	1,356	1,353	1,350	1,348	1,345	1,343	1,341	1,340	1,338	1,336	1,335	1,333	1,332
22	1,386	1,379	1,374	1,368	1,364	1,359	1,355	1,352	1,349	1,346	1,343	1,341	1,338	1,336	1,334	1,332	1,330	1,329	1,327	1,326	1,324
23	1,380	1,374	1,368	1,362	1,357	1,353	1,349	1,346	1,342	1,339	1,337	1,334	1,332	1,330	1,327	1,326	1,324	1,322	1,321	1,319	1,318
24	1,375	1,368	1,362	1,357	1,352	1,347	1,343	1,340	1,337	1,333	1,331	1,328	1,326	1,323	1,321	1,319	1,318	1,316	1,314	1,313	1,311
25	1,370	1,363	1,357	1,352	1,347	1,342	1,338	1,335	1,331	1,328	1,325	1,323	1,320	1,318	1,316	1,314	1,312	1,310	1,309	1,307	1,306
26	1,366	1,359	1,352	1,347	1,342	1,337	1,333	1,330	1,326	1,323	1,320	1,318	1,315	1,313	1,311	1,309	1,307	1,305	1,303	1,302	1,300
27	1,361	1,354	1,348	1,342	1,337	1,333	1,329	1,325	1,322	1,318	1,315	1,313	1,310	1,308	1,306	1,304	1,302	1,300	1,298	1,297	1,295
28	1,358	1,350	1,344	1,338	1,333	1,329	1,325	1,321	1,317	1,314	1,311	1,308	1,306	1,304	1,301	1,299	1,297	1,295	1,294	1,292	1,291
29	1,354	1,347	1,340	1,335	1,330	1,325	1,321	1,317	1,313	1,310	1,307	1,304	1,302	1,299	1,297	1,295	1,293	1,291	1,290	1,288	1,286
30	1,351	1,343	1,337	1,331	1,326	1,321	1,317	1,313	1,310	1,306	1,303	1,301	1,298	1,296	1,293	1,291	1,289	1,287	1,286	1,284	1,282

Таблица П.6

Распределение Стьюдента
(к построению доверительных интервалов)

Значения t_β						
n-1	β					
	0,8	0,9	0,95	0,99	0,999	0,9999
10	1,3722	1,8125	2,2281	3,1693	4,5868	6,2119
11	1,3634	1,7959	2,2010	3,1058	4,4369	5,9232
12	1,3562	1,7823	2,1788	3,0545	4,3178	5,6950
13	1,3502	1,7709	2,1604	3,0123	4,2209	5,5134
14	1,3450	1,7613	2,1448	2,9768	4,1403	5,3644
15	1,3406	1,7531	2,1315	2,9467	4,0728	5,2387
16	1,3368	1,7459	2,1199	2,9208	4,0149	5,1339
17	1,3334	1,7396	2,1098	2,8982	3,9651	5,0431
18	1,3304	1,7341	2,1009	2,8784	3,9217	4,9663
19	1,3277	1,7291	2,0930	2,8609	3,8833	4,8988
20	1,3253	1,7247	2,0860	2,8453	3,8496	4,8382
21	1,3232	1,7207	2,0796	2,8314	3,8193	4,7847
22	1,3212	1,7171	2,0739	2,8188	3,7922	4,7358
23	1,3195	1,7139	2,0687	2,8073	3,7676	4,6939
34	1,3070	1,6909	2,0322	2,7284	3,6007	4,4052
25	1,3163	1,7081	2,0595	2,7874	3,7251	4,6194
26	1,3150	1,7056	2,0555	2,7787	3,7067	4,5868
27	1,3137	1,7033	2,0518	2,7707	3,6895	4,5565
28	1,3125	1,7011	2,0484	2,7633	3,6739	4,5309
29	1,3114	1,6991	2,0452	2,7564	3,6595	4,5053
30	1,3104	1,6973	2,0423	2,7500	3,6460	4,4820
35	1,3062	1,6896	2,0301	2,7238	3,5911	4,3889
40	1,3031	1,6839	2,0211	2,7045	3,5510	4,3213
45	1,3007	1,6794	2,0141	2,6896	3,5203	4,2689
50	1,2987	1,6759	2,0086	2,6778	3,4960	4,2282
55	1,2971	1,6730	2,0040	2,6682	3,4765	4,1956
60	1,2958	1,6706	2,0003	2,6603	3,4602	4,1688
70	1,2938	1,6669	1,9944	2,6479	3,4350	4,1269
80	1,2922	1,6641	1,9901	2,6387	3,4164	4,0955
90	1,2910	1,6620	1,9867	2,6316	3,4019	4,0722
100	1,2901	1,6602	1,9840	2,6259	3,3905	4,0536

Таблица П.7

Распределение Пирсона

(к построению доверительного интервала дисперсии)

Значения χ^2										
r	$\beta = 0{,}8$		$\beta = 0{,}9$		$\beta = 0{,}95$		$\beta = 0{,}99$		$\beta = 0{,}999$	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
	0,1000	**0,9000**	**0,0500**	**0,9500**	**0,0250**	**0,9750**	**0,0050**	**0,9950**	**0,0005**	**0,9995**
10	15,987	4,865	18,307	3,940	20,483	3,247	25,188	2,156	31,419	1,265
11	17,275	5,578	19,675	4,575	21,920	3,816	26,757	2,603	33,138	1,587
12	18,549	6,304	21,026	5,226	23,337	4,404	28,300	3,074	34,821	1,935
13	19,812	7,041	22,362	5,892	24,736	5,009	29,819	3,565	36,477	2,305
14	21,064	7,790	23,685	6,571	26,119	5,629	31,319	4,075	38,109	2,697
15	22,307	8,547	24,996	7,261	27,488	6,262	32,801	4,601	39,717	3,107
16	23,542	9,312	26,296	7,962	28,845	6,908	34,267	5,142	41,308	3,536
17	24,769	10,085	27,587	8,672	30,191	7,564	35,718	5,697	42,881	3,980
18	25,989	10,865	28,869	9,390	31,526	8,231	37,156	6,265	44,434	4,439
19	27,204	11,651	30,144	10,117	32,852	8,907	38,582	6,844	45,974	4,913
20	28,412	12,443	31,410	10,851	34,170	9,591	39,997	7,434	47,498	5,398
21	29,615	13,240	32,671	11,591	35,479	10,283	41,401	8,034	49,010	5,895
22	30,813	14,041	33,924	12,338	36,781	10,982	42,796	8,643	50,510	6,404
23	32,007	14,848	35,172	13,091	38,076	11,689	44,181	9,260	51,999	6,924
24	33,196	15,659	36,415	13,848	39,364	12,401	45,558	9,886	53,478	7,453
25	34,382	16,473	37,652	14,611	40,646	13,120	46,928	10,520	54,948	7,991
26	35,563	17,292	38,885	15,379	41,923	13,844	48,290	11,160	56,407	8,537
27	36,741	18,114	40,113	16,151	43,195	14,573	49,645	11,808	57,856	9,093
28	37,916	18,939	41,337	16,928	44,461	15,308	50,994	12,461	59,299	9,656
29	39,087	19,768	42,557	17,708	45,722	16,047	52,335	13,121	60,734	10,227
30	40,256	20,599	43,773	18,493	46,979	16,791	53,672	13,787	62,160	10,804
35	46,059	24,797	49,802	22,465	53,203	20,569	60,275	17,192	69,197	13,788
40	51,805	29,051	55,758	26,509	59,342	24,433	66,766	20,707	76,096	16,906
45	57,505	33,350	61,656	30,612	65,410	28,366	73,166	24,311	82,873	20,136
50	63,167	37,689	67,505	34,764	71,420	32,357	79,490	27,991	89,560	23,461
55	68,796	42,060	73,311	38,958	77,380	36,398	85,749	31,735	96,161	26,865
60	74,397	46,459	79,082	43,188	83,298	40,482	91,952	35,534	102,697	30,339
70	85,527	55,329	90,531	51,739	95,023	48,758	104,215	43,275	115,577	37,467
80	96,578	64,278	101,879	60,391	106,629	57,153	116,321	51,172	128,264	44,792
90	107,565	73,291	113,145	69,126	118,136	65,647	128,299	59,196	140,780	52,277
100	118,498	82,358	124,342	77,929	129,561	74,222	140,170	67,328	153,164	59,895

П. 8. Нормально-вероятностная бумага

Список использованной литературы

1. Митропольский А.К. Техника статистических вычислений.
М.: Наука, 1971. 576 с.
2. Гмурман В.Е. Теория вероятностей и математическая статистика.
М.: Высшая школа, 2006. 479 с.
3. Вентцель Е.С. Теория вероятностей. М.: КНОРУС, 2010. 664 с.
4. Дунин-Барковский И.В., Смирнов Н.В. Теория вероятностей и
математическая статистика в технике. М.: Гостехиздат, 1955. 556 с.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ ... 3

1. ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ ... 4

1.1. Неравенство Чебышева ... 5
1.2. Закон больших чисел (теорема П.Л.Чебышева) ... 7
1.3. Обобщённая теорема Чебышева ... 10
1.4. Теорема Маркова ... 11
1.5. Теорема Я. Бернулли ... 12
1.6. Теорема Пуассона ... 14
2. ВЫБОРОЧНЫЕ ОЦЕНКИ И ИХ СВОЙСТВА ... 16

2.1. Требования к выборочным оценкам ... 16
2.2. Свойства выборочных оценок математического ожидания ... 17
2.3. Свойства выборочных оценок дисперсии ... 18
2.4. Свойства выборочных оценок вероятности случайного события ... 21
3. ОБРАБОТКА ОПЫТОВ ... 23

3.1. Простая статистическая совокупность. Статистический ряд. Гистограмма ... 23
3.2. Числовые характеристики статистического распределения ... 26
3.3. Выравнивание статистических рядов ... 28
4. ПРЕДВАРИТЕЛЬНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ... 34

4.1. Оценка математических ожиданий и средних квадратических отклонений ... 34
4.1.1. Оценка математических ожиданий и средних квадратических отклонений для средних выборок ... 35
4.1.2. Оценка математических ожиданий и средних квадратических отклонений для представительных выборок ... 36
4.2. Построение статистических функций распределения ... 39
на нормально–вероятностной бумаге ... 39
4.2.1. Средняя выборка ... 44
4.2.2. Представительная выборка ... 46
5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ ... 49

5.1. Проверка гипотезы нормальности статистической функции распределения ... 49
5.1.1. Проверка гипотезы нормальности статистической функции распределения для средних выборок ... 52
5.1.2. Проверка гипотезы нормальности статистической функции распределения для представительных выборок ... 55
5.2. Оценка значимости расхождений статистических оценок ... 60
5.2.1. Оценка расхождений средних значений ... 61
5.2.2. Оценка расхождений дисперсий ... 65
6. ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ... 67

6.1 Доверительный интервал математического ожидания ... 68
6.2 Доверительный интервал дисперсии ... 70
6.3 Доверительный интервал вероятности наблюдаемых событий ... 71
6.4 Доверительный интервал вероятности редких событий ... 75
7. РЕГРЕССИОННЫЙ АНАЛИЗ ... 77

7.1. Линейный регрессионный анализ ... 77
7.2. Значимость выборочной корреляции ... 79
7.3. Оценка адекватности линейной регрессии ... 80
7.3.1. Критерий Фишера. Надёжности регрессии ... 80
7.3.2. Коэффициент детерминации ... 81
8. ПЛАНИРОВАНИЕ ОБЪЁМА ВЫБОРОК ... 86

8.1. Планирование оценивания математического ожидания ... 87
8.2. Планирование оценивания дисперсии ... 89
8.3. Планирование оценивания вероятности наблюдаемых событий ... 92
8.4. Планирование оценивания вероятности редких событий ... 94

ПРИЛОЖЕНИЕ ... 95

Дуплякин
Вячеслав Митрофанович

Заслуженный деятель
науки и техники
Российской Федерации

Доктор технических наук

Профессор кафедры экономики
Самарского государственного
Аэрокосмического университета

Учебное издание

Дуплякин Вячеслав Митрофанович

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВЫБОРОЧНЫХ ДАННЫХ

Учебное пособие

Подписано в печать 06.05.2010. Формат 60×84 1/8
Электронное издание
Печ. л. 13,5
Арт.С – Э3/2010

Самарский государственный
Аэрокосмический университет
443086, Самара, Московское шоссе, 34

Изд-во Самарского государственного
Аэрокосмического университета.
443086, Самара, Московское шоссе, 34