МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Пензенский государственный университет
архитектуры и строительства»
(ПГУАС)

Г.А. Левова, О.В. Снежкина, С.Н. Ячинова

СТАТИСТИЧЕСКИЙ АНАЛИЗ СОВОКУПНОСТИ СЛУЧАЙНЫХ ВЕЛИЧИН

Рекомендовано Редсоветом университета
в качестве учебного пособия для студентов,
обучающихся по направлениям 08.03.01 «Строительство»,
21.03.02 «Землеустройство и кадастры»

Пенза 2014

УДК 511
ББК 22.11
Л34

Рецензенты: кандидат педагогических наук, доцент кафедры «Автоматизированные системы управления и программного обеспечения» О.В. Бочкарева (филиал Военного учебно-научного центра Сухопутных войск «Общевойсковая Академия ВС РФ», г. Пенза); кандидат технических наук, доцент кафедры «Механика» М.Б.Зайцев (ПГУАС)

Левова Г.А.

Л34

учеб. пособие / Г.А. Левова, О.В. Снежкина, С.Н. Ячинова. – Пенза: ПГУАС, 2014. – 88 с.

Учебное пособие представляет собой руководство к решению задач математической статистики. Излагаемые теоретические вопросы сопровождаются задачами, приводимыми с решениями. Содержит варианты заданий для самостоятельной работы.

Данное учебное пособие соответствует образовательным стандартам третьего поколения направления 08.03.01 «Строительство» и 21.03.02 «Землеустройство и кадастры» и рекомендуется при изучении дисциплины «Математика».

Пособие подготовлено на кафедре «Математика и математическое моделирование» и предназначено для студентов высших технических учебных заведений, может быть использовано преподавателями, инженерами и научными работниками, заинтересованными в освоении вероятностных методов для решения практических задач.

ПРЕДИСЛОВИЕ

Целью данного пособия является формирование у студентов навыков практического применения вероятностных методов для решения задач математической статистики.

Пособие состоит из трех разделов. Первый и второй разделы содержат основные вопросы теории вероятностей и математической статистики, которые сопровождаются решением примеров.

Третья глава – практическая, состоящая из заданий для самостоятельной работы с одним из вариантов их решения с подробными указаниями.

Настоящее учебное пособие подготовлено с учетом опыта преподавания теории вероятностей и математической статистики в высшем техническом учебном заведении по направлению « Строительство» и «Землеустройство и кадастры».

1. ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Предмет теории вероятностей и математической статистики

Вокруг нас происходит очень много событий, исходы которых предсказать заранее невозможно. Например, подбрасывая монету, мы не знаем, какой стороной она упадет. Стреляя однотипными снарядами без изменения наводки орудия в одну точку попасть невозможно. Производя повторные высокоточные измерения, например, скорости света или очень больших расстояний, обычно получают лишь приблизительно равные результаты, которые зависят от всевозможных случайностей.

Иначе обстоит дело, когда рассматриваются события, которые могут многократно наблюдаться при осуществлении одних и тех же условий, то есть если речь идет о массовых однородных случайных событиях. Достаточно большое число однородных случайных событий независимо от их природы подчиняется определенным закономерностям, установлением которых и занимается теория вероятностей.

Теория вероятностей – математическая наука, изучающая закономерности случайных явлений. Следовательно, предметом теории вероятностей является изучение закономерностей массовых однородных случайных событий.

Математическая статистика опирается на теорию вероятностей. Она оперирует непосредственно результатами наблюдений. Используя результаты, полученные согласно теории вероятностей, математическая статистика позволяет оценить значения искомых характеристик, выявить степень точности получаемых при обработке данных выводов.

Математическая статистика – раздел математики, изучающий методы сбора, систематизации и обработки результатов измерений с целью выявления статистических закономерностей.

Задачи математической статистики состоят в том, чтобы:

1. Указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.

2. Разработать методы анализа статистических данных в зависимости от целей исследования.

Первые работы, в которых зарождались основные понятия теории вероятностей, появились в XVI–XVII вв. Они представляли собой по-

пытки создания теории азартных игр с целью дать рекомендации игрокам (Д.Кардано, Б.Паскаль, П.Ферма, Х.Гюйгенс и др.).

Следующий этап развития теории вероятностей связан с именем Якоба Бернулли, который доказал теорему, получившую впоследствии название «закона больших чисел». Это было первое теоретическое обоснование накопленных ранее фактов.

Дальнейшее развитие теории вероятностей приходится на XVII–XIX вв. благодаря работам А. Муавра, П. Лапласа, К. Гаусса, С. Пуассона и др.

В XVII веке возникает и развивается параллельно с теорией вероятностей математическая статистика.

Наиболее плодотворный период развития «математики случайного» связан с именами русских математиков П.Л.Чебышева и его учеников А.А.Маркова и А.М.Ляпунова (XIX – начало XX в.). В этот период теория вероятностей становится отдельной математической наукой.

Большой вклад в развитие теории вероятностей и математической статистки внесли российские математики С.Н. Бернштейн, В.И. Романовский, А.Н. Колмогоров, А.Я. Хинчин, Б.В. Гнеденко, Н.В. Смирнов и др., а также ученые англо-американской школы Стьюдент, Р. Фишер, Э. Пирсон, Ю. Нейман, А. Вальд и др.

1.2. Основные понятия теории вероятностей

Одним из основных понятий теории вероятностей является понятие события.

Случайным событием (или просто *событием*) называется любой факт, который в результате испытания может произойти или не произойти. Событие – это возможный *исход*, результат испытания (опыта, эксперимента).

События могут быть совместными и несовместными. События называются *несовместными*, если наступление одного из них исключает наступление любого другого. Иначе события называются *совместными*.

Событие называется *достоверным*, если в результате испытания оно обязательно должно произойти.

Событие называется *невозможным*, если в результате испытания оно вообще не может произойти.

События называются *равновозможными*, если в результате испытания по условиям симметрии ни одно из этих событий не является объективно более возможным.

Несколько событий образуют *полную группу*, если они являются единственно возможными и несовместными исходами испытания. Это

означает, что в *результате испытания обязательно должно произойти одно и только одно из этих событий*.

Вероятность – одно из основных понятий теории вероятностей. Существует несколько определений этого понятия. Рассмотрим определение, которое называют классическим.

В практической деятельности важно уметь сравнивать события по степени возможности их наступления. Поэтому для сравнения событий нужна определенная мера.

Численная мера степени объективной возможности наступления события называется *вероятностью события*.

Это определение, качественно отражающее понятие вероятности события, не является математическим. Чтобы оно таким стало, необходимо определить его количественно.

Пусть исходы некоторого испытания образуют полную группу событий и равновозможны, т.е. единственно возможны, несовместны и равновозможны. Такие исходы называют *элементарными исходами*.

Случай называется *благоприятствующим* событию, если появление этого случая влечет за собой появление события.

Согласно классическому определению *вероятностью события* называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу. Вероятность события A определяется формулой

$$P(A)=\frac{m}{n},$$

где $P(A)$ – вероятность события A;

m – число элементарных исходов, благоприятствующих событию A;

n – число всех возможных элементарных исходов испытания.

Классическое определение вероятности следует рассматривать не как определение, а как метод вычисления вероятностей для испытаний, сводящихся к схеме случаев.

Отметим свойства вероятности события.

Свойство 1. Вероятность достоверного события равна единице.

Свойство 2. Вероятность невозможного события равна нулю.

Свойство 3. Вероятность случайного события есть положительное число, заключенное между нулем и единицей, т.е.

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

События, вероятности которых очень малы (близки к нулю) или очень велики (близки к единице), называются соответственно *практически невозможными* или *практически достоверными* событиями.

1.3. Понятие случайной величины. Виды случайных величин

Одним из важнейших понятий теории вероятностей является понятие случайной величины.

Случайной называют величину, которая в результате испытания примет одно и только одно возможное значение, наперед не известное и зависящее от случайных причин, которые заранее не могут быть учтены.

Возможным значением случайной величины называется конкретное значение, которое она может принимать.

Случайные величины обозначаются прописными буквами X, Y, Z, а их возможные значения строчными буквами x, y, z.

Случайные величины делятся на дискретные (прерывные) и непрерывные.

Дискретной (*прерывной*) называют случайную величину, которая принимает отдельные, изолированные возможные значения с определенными вероятностями.

Под *непрерывной* случайной величиной будем понимать величину, бесконечное множество значений которой есть некоторый интервал (конечный или бесконечный) числовой оси (строгое определение непрерывной случайной величины будет дано ниже).

Примерами дискретных случайных величин с конечным множеством значений могут служить число родившихся детей в течение суток в населенном пункте, количество бракованных изделий в данной партии, с бесконечным, но счетным множеством значений – число произведенных выстрелов до первого попадания. Дальность полета артиллерийского снаряда, расход электроэнергии на предприятии за месяц – примеры непрерывных случайных величин.

В теории вероятностей рассматриваются случайные величины, возможные значения которых определяются одним числом, – одномерные случайные величины. Кроме одномерных случайных величин изучаются величины, возможные значения которых определяются двумя, тремя, ..., n числами. Такие величины называются соответственно двумерными, трехмерными, ..., n – мерными.

Наиболее полным описанием случайной величины является ее закон распределения.

1.4. Функция распределения вероятностей и плотность распределения вероятностей случайной величины

Функцией распределения называют функцию $F(x)$, определяющую вероятность того, что случайная величина X примет значение, меньшее x, т.е.

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

Дадим более точное определение непрерывной случайной величины.

Случайную величину называют *непрерывной*, если ее функция распределения есть непрерывная, кусочно-дифференцируемая функция с непрерывной производной.

Функция распределения обладает следующими свойствами.

Свойство 1. Значения функции распределения принадлежат отрезку [0,1]:

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

Свойство 2. $F(x)$ – неубывающая функция, т.е.

$$F(x_2) \ge F(x_1), \text{ если } x_2 > x_1.$$

Следствие 1. Вероятность того, что случайная величина примет значение, заключенное в интервале (a, b), равна приращению функции распределения на этом интервале:

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

Следствие 2. Вероятность того, что непрерывная случайная величина X примет одно определенное значение, равна нулю.

Плотностью распределения вероятностей непрерывной случайной величины называют функцию $f(x)$ – первую производную от функции распределения $F(x)$:

$$f(x) = F'(x).$$

Свойства плотности распределения.

Свойство 1. Плотность распределения – неотрицательная функция:

$$f(x) \ge 0.$$

Свойство 2. Несобственный интеграл от плотности распределения в пределах от $-\infty$ до ∞ равен единице:

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

1.5. Числовые характеристики непрерывных случайных величин

Рассмотрим числовые характеристики непрерывных случайных величин.

Математическим ожиданием непрерывной случайной величины X, возможные значения которой принадлежат отрезку $[a, b]$, называют определенный интеграл

$$M(X) = \int_a^b x\, f(x)dx\,.$$

Если возможные значения принадлежат всей оси Ox, то

$$M(X) = \int_{-\infty}^{\infty} x\, f(x)dx\,.$$

Дисперсией непрерывной случайной величины называют математическое ожидание квадрата ее отклонения.

Если возможные значения X принадлежат отрезку $[a, b]$, то

$$D(X) = \int_a^b [x - M(X)]^2 f(x)d(x)\,.$$

Если возможные значения принадлежат всей оси Ox, то

$$D(X) = \int_{-\infty}^{\infty} [x - M(X)]^2 f(x)d(x)\,.$$

Средним квадратическим отклонением непрерывной случайной величины называют квадратный корень из дисперсии:

$$\sigma(X) = \sqrt{D(X)}\,.$$

1.6. Законы распределения случайных величин. Нормальный закон распределения. Кривая Гаусса, свойства, график

Наиболее полным описанием случайной величины является ее закон распределения.

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

Для дискретной случайной величины закон распределения можно задать таблично, аналитически (в виде формулы) и графически.

Основными законами распределения дискретных случайных величин являются: биноминальный закон распределения, распределение Пуассона, геометрическое распределение, гипергеометрическое распределение.

Основными законами распределения непрерывных случайных величин являются: равномерный закон распределения, показательный (экспоненциальный) закон распределения, нормальный закон распределения, логарифмически-нормальное распределение.

Биноминальным называют распределение вероятностей, определяемое формулой Бернулли:

$$P_n(k) = C_n^k p^k q^{n-k},$$

где $0 < p < 1$, $q = 1 - p$, $k = 0, 1, ..., n$.

Если число испытаний n велико, а вероятность появления события в каждом испытании очень мала ($p \le 0{,}1$), то используют приближенную формулу

$$P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

где k – число появлений события в n независимых испытаниях, $\lambda = np$ (среднее число появлений события в n испытаниях), и говорят, что случайная величина распределена по *закону Пуассона.*

Дискретная случайная величина имеет *геометрическое распределение*, если она принимает значения $k = 1, 2, ...$ с вероятностями

$$P(X = k) = q^{k-1} p.$$

Вероятности p_i образуют геометрическую прогрессию с первым членом p и знаменателем q (отсюда название «геометрическое распределение»).

Равномерным называют распределение вероятностей непрерывной случайной величины X, если на интервале (a, b), которому принадлежат все возможные значения X, плотность распределения сохраняет постоянное значение, а именно $f(x) = \frac{1}{b-a}$; вне этого интервала $f(x) = 0$.

Подробнее остановимся на нормальном законе распределения, который наиболее часто встречается на практике. Главная особенность, выделяющая его среди других законов, состоит в том, что он является

предельным законом, к которому приближаются другие законы распределения при весьма часто встречающихся типичных условиях.

Непрерывная случайная величина X имеет *нормальный закон распределения* (*закон Гаусса*) с параметрами a и σ, если её плотность вероятности имеет вид:

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

Чтобы задать нормальное распределение достаточно знать его параметры. Покажем, что вероятностный смысл этих параметров таков: a – математическое ожидание; σ – среднее квадратическое отклонение нормального распределения.

По определению математического ожидания непрерывной случайной величины:

$$M(X)=\int_{-\infty}^{\infty}xf(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}xe^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Введем новую переменную $z=\frac{x-a}{\sigma}$. Тогда $x=\sigma z+a, dx=\sigma dz$, пределы интегрирования не меняются и , следовательно,

$$M(X)=\frac{\sigma}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z+a)\,e^{-\frac{z^2}{2}}dz=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma z e^{-\frac{z^2}{2}}dz+\frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{z^2}{2}}dz.$$

Первый интеграл равен нулю как интеграл от нечетной функции по симметричному относительно начала координат промежутку. Второй интеграл – интеграл Эйлера – Пуассона.

$$\int_{-\infty}^{\infty}e^{-\frac{z^2}{2}}dz=\sqrt{2\pi}.$$

Итак, $M(X)=a$, т.е. математическое ожидание нормального распределения равно параметру a.

По определению дисперсии непрерывной случайной величины, учитывая, что $M(X)=a$, имеем

$$D(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-a)^2e^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Сделаем ту же замену переменной $z=\frac{x-a}{\sigma}$, как и при вычислении предыдущего интеграла. Тогда

$$D(X)=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z\cdot ze^{-\frac{z^2}{2}}dz.$$

Применяя метод интегрирования по частям, положив $u=z, dv=ze^{-\frac{z^2}{2}}dz$, получим

$$D(X)=\sigma^2.$$

Следовательно,

$$\sigma(X)=\sqrt{D(X)}=\sqrt{\sigma^2}=\sigma.$$

Итак, среднее квадратическое отклонение нормального распределения равно параметру σ.

Кривую нормального закона распределения называют *нормальной* или *гауссовой кривой.*

Исследуем функцию

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

1. Функция определена на всей оси x.

2. При всех значениях x функция принимает положительные значения $f(x)>0$, т.е. нормальная кривая расположена над осью Ox.

3. Предел функции при неограниченном возрастании x (по абсолютной величине) равен нулю: $\lim\limits_{|x|\to\infty} y=0$, т.е. ось Ox является горизонтальной асимптотой графика.

4. Исследуем функцию на экстремум. Найдем первую производную:

$$y'=-\frac{x-a}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Легко видеть, что $y'=0$ при $x=a$, $y'>0$ при $x<a$, $y'<0$ при $x>a$.

Следовательно, при $x=a$ функция имеет максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$.

5. График функции симметричен относительно прямой $x=a$, так как $x-a$ содержится в аналитическом выражении функции в квадрате.

6. Исследуем функцию на точки перегиба. Найдем вторую производную:

$$y'' = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \left[1 - \frac{(x-a)^2}{\sigma^2}\right].$$

При $x = a + \sigma$ и $x = a - \sigma$ вторая производная равна нулю, а при переходе через эти точки она меняет знак (в обеих этих точках значение функции равно $\frac{1}{\sigma\sqrt{2\pi e}}$). Таким образом, точки графика $\left(a - \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ и $\left(a + \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ являются точками перегиба.

На рис. 1 изображена нормальная кривая при $a = 1$ и $\sigma = 2$.

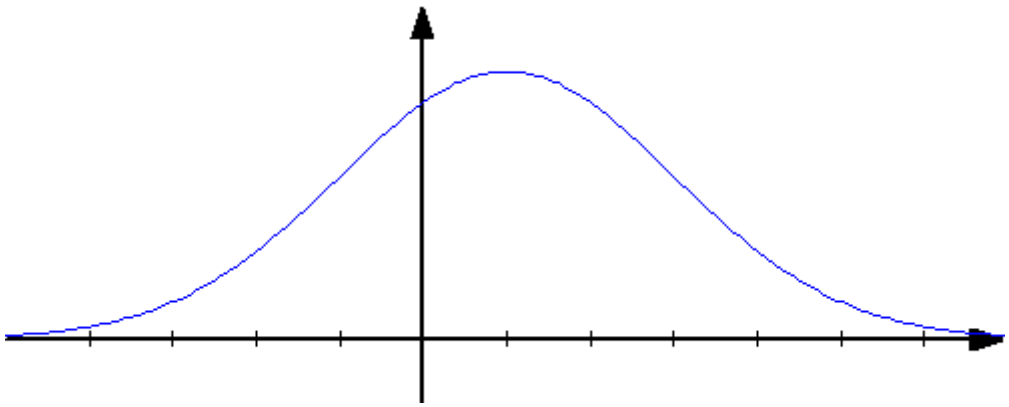

Рис.1

Выясним, как будет меняться нормальная кривая при изменении параметров a и σ.

Если $\sigma = \text{const}$ и меняется параметр a, т.е. центр симметрии распределения, то нормальная кривая будет смещаться вдоль оси абсцисс, не меняя формы (рис. 2).

σ=const, $a_1 < a_2 < a_3$.

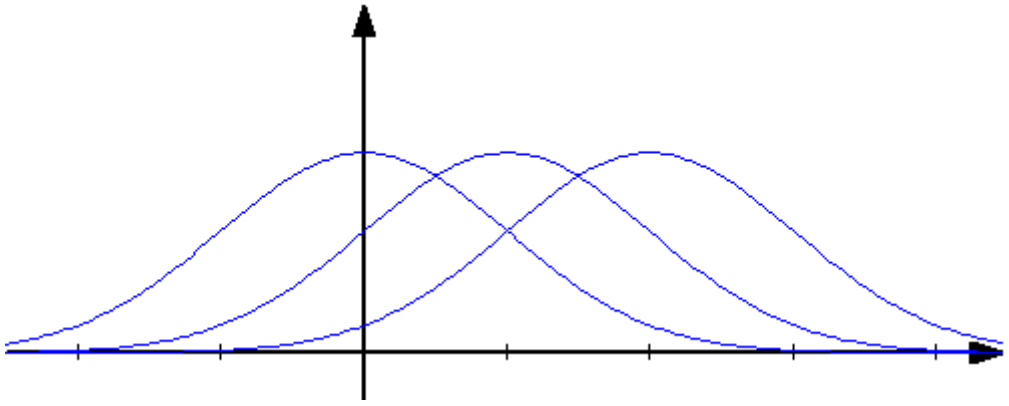

Рис.2

Если $a = \text{const}$ и меняется параметр σ, то меняется ордината максимума кривой $f_{\max}(a) = \dfrac{1}{\sigma\sqrt{2\pi}}$. При увеличении σ ордината максимума кривой уменьшается, но так как площадь под любой кривой распределения должна оставаться равной единице, то кривая становится более плоской, растягиваясь вдоль оси абсцисс; при уменьшении σ нормальная кривая вытягивается вверх, одновременно сжимаясь с боков (рис. 3).

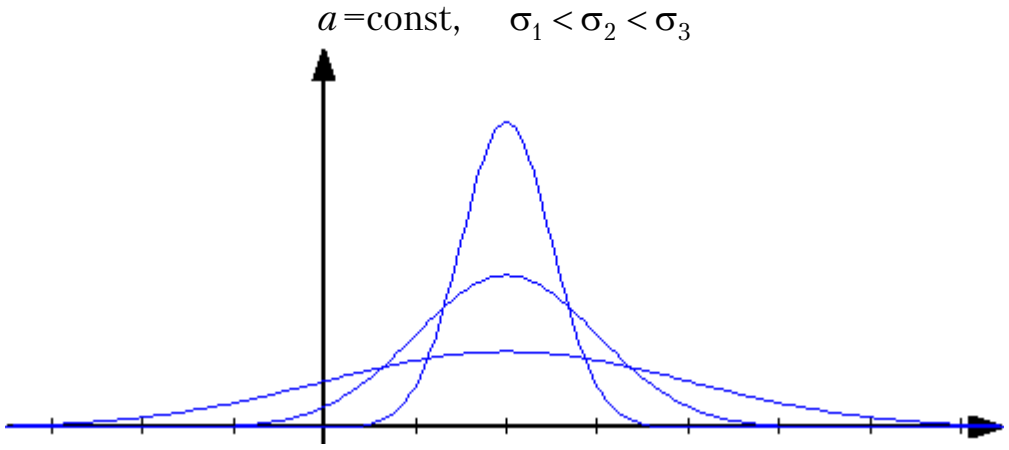

Рис.3

Таким образом, параметр a (математическое ожидание) характеризует положение, а параметр σ – форму нормальной кривой.

1.7. Вероятность попадания в заданный интервал нормальной случайной величины

Известно, что если случайная величина X задана плотностью распределения $f(x)$, то вероятность того, что X примет значение, принадлежащее интервалу (α, β), такова:

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

Пусть случайная величина X распределена по нормальному закону. Тогда вероятность того, что X примет значение, принадлежащее интервалу (α, β), равна

$$P(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{(x-a)^2}{\sigma^2}} dx .$$

Преобразуем эту формулу таким образом, чтобы можно было пользоваться готовыми таблицами. Введем новую переменную $z=\frac{x-a}{\sigma}$. Отсюда $x=\sigma z+a, dx=\sigma dz$. Найдем новые пределы интегрирования. Если $x=a$, то $z=\frac{\beta-\alpha}{\sigma}$.

Таким образом, имеем

$$P(\alpha<X<\beta)=\frac{1}{\sigma\sqrt{2\pi}}\int_{(\alpha-a)/\sigma}^{(\beta-a)/\sigma}e^{-\frac{z^2}{2}}(\sigma dz)=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{(\alpha-a)/\sigma}^{0}e^{-\frac{z^2}{2}}dz+\frac{1}{\sqrt{2\pi}}\int_{0}^{(\beta-a)/\sigma}e^{-\frac{z^2}{2}}dz=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{0}^{(\beta-a)/\sigma}e^{-\frac{z^2}{2}}dz-\frac{1}{\sqrt{2\pi}}\int_{0}^{(\alpha-a)/\sigma}e^{-\frac{z^2}{2}}dz.$$

Пользуясь функцией Лапласа

$$\Phi(x)=\frac{1}{2\pi}\int_{0}^{x}e^{-\frac{z^2}{2}}dz,$$

окончательно получим, что вероятность попадания случайной величины X, распределенной по нормальному закону, в интервал (α,β) равна

$$P(\alpha<X<\beta)=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right).$$

1.8. Вероятность отклонения случайной величины от математического ожидания

Часто требуется вычислить вероятность того, что отклонение нормально распределенной случайной величины X по абсолютной величине меньше заданного положительного числа δ, т.е. требуется найти вероятность осуществления неравенства $|X-a|<\delta$.

Заменим это неравенство равносильным ему двойным неравенством

$$-\delta<X-a<\delta, \text{ или } a-\delta<X<a+\delta.$$

Пользуясь формулой

$$P(\alpha<X<\beta)=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right),$$

получим

$$P(|X-a|<\delta)=P(a-\delta<X<a+\delta)=$$
$$=\Phi\left[\frac{(a+\delta)-a}{\sigma}\right]-\Phi\left[\frac{(a-\delta)-a}{\sigma}\right]=\Phi\left(\frac{\delta}{\sigma}\right)-\Phi\left(-\frac{\delta}{\sigma}\right).$$

Принимая во внимание равенство

$$\Phi\left(-\frac{\delta}{\sigma}\right)=-\Phi\left(\frac{\delta}{\sigma}\right)$$

(функция Лапласа – нечетная), окончательно имеем

$$P(|X-a|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

В частности, при $a=0$

$$P(|X|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

Если две случайные величины нормально распределены и $a=0$, то вероятность принять значение, принадлежащее интервалу $(-\delta,\delta)$, больше у той величины, которая имеет меньшее значение σ. Это полностью соответствует вероятностному смыслу параметра σ (σ – среднее квадратическое отклонение; оно характеризует рассеяние случайной величины вокруг ее математического ожидания).

1.9. Генеральная, выборочная совокупность. Повторная и бесповторная выборка. Вариационный ряд

Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных – сведений о том, какие значения принял в результате наблюдений интересующий признак.

Различные значения x_i признака X называются *вариантами*, а последовательность вариант, записанных в порядке возрастания или убывания – *вариационным рядом*. Число наблюдений называют *частотой*, отношение числа наблюдений к объему выборки – *относительной частотой*.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот.

Совокупность объектов, или, точнее, совокупность значений какого-то признака объекта, называется *генеральной совокупностью*. Полное обследование генеральной совокупности практически невозможно или неэкономично. Поэтому из генеральной совокупности делают выборку, т.е. исследуют только некоторые её объекты. *Выборочной совокупностью* или просто *выборкой* называют совокупность случайно отобранных объектов из генеральной совокупности.

Выборку можно рассматривать как некий эмпирический аналог генеральной совокупности. Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности (по выборке) выносить суждение о ее свойствах в целом.

С помощью выборки оценивают генеральную совокупность по вероятностным свойствам. Чтобы оценка была достоверной, выборка должна быть *репрезентативной* (*представительной)*, т.е. ее вероятностные свойства должны совпадать или быть близкими к свойствам генеральной совокупности.

Представительную выборку можно получить, если выбирать объекты для исследований случайно, т.е. гарантировать всем объектам генеральной совокупности одинаковую вероятность исследования.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности.

Выборки подразделяют на повторные и бесповторные.

Повторной называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

Бесповторная выборка более независимая и представительная.

Важнейшей задачей выборочного метода является оценка параметров (характеристик) генеральной совокупности по данным выборки.

1.10. Полигоны и гистограммы

Для графического изображения вариационных рядов наиболее часто используются полигон и гистограмма.

Полигон, как правило, служит для изображения дискретного вариационного ряда и представляет собой ломаную.

Полигоном частот называется ломаная, отрезки которой соединяют точки $(x_1,\ n_1)$, $(x_2,\ n_2)$, ..., $(x_k,\ n_k)$, где x_i – варианты выборки и n_i – соответствующие им частоты.

Полигоном относительных частот называется ломаная, отрезки которой соединяют точки (x_1, ω_1), (x_2, ω_2), ..., (x_k, ω_k), где x_i – варианты выборки и ω_i – соответствующие им относительные частоты (рис.4).

Рис. 4

Гистограмма служит только для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частоты).

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{n_i}{h}$ (рис.5).

Площадь i-го частичного прямоугольника равна $\frac{h n_i}{h} = n_i$ – сумме частот вариант i-го интервала, следовательно, *площадь гистограммы частот равна сумме всех частот*, т.е. *объему выборки*.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат

частичные интервалы длиною h, а высоты равны отношению $\frac{\omega_i}{h}$ (плотность относительной частоты).

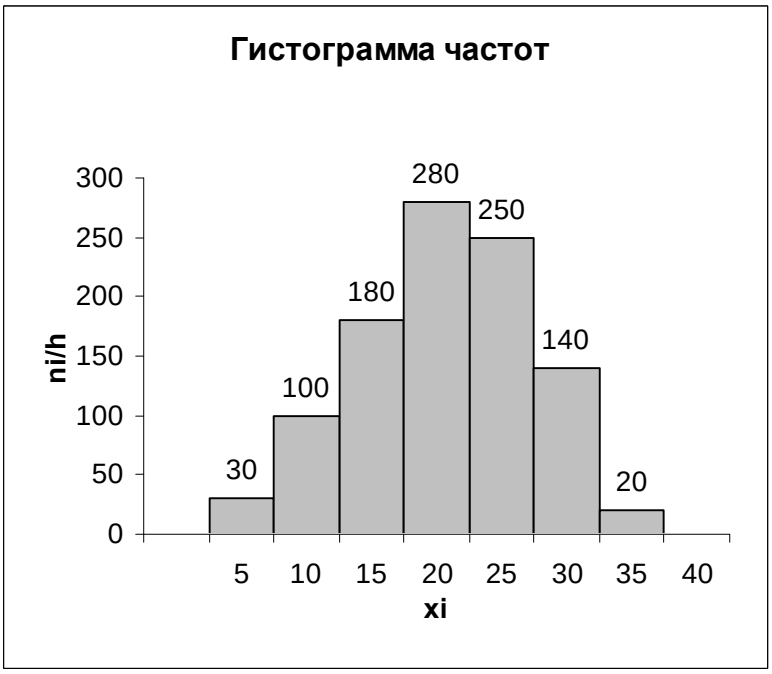

Рис. 5

Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{\omega_i}{h}$. Площадь i-го частичного прямоугольника равна $\frac{h\omega_i}{h} = \omega_i$ – относительной частоте вариант, попавших в i-ый интервал. Следовательно, *площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.*

1.11. Эмпирическая функция распределения

Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

Эмпирической функцией распределения $F^*(x)$ называется относительная частота (частость) того, что признак (случайная величина X) примет значение, меньше заданного x.

По определению,

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

где n_x – число вариант, меньших x;

n – объем выборки.

Эмпирическая функция обладает следующими свойствами.

Свойство 1. Значения эмпирический функции принадлежат отрезку [0;1].

Свойство 2. $F^*(x)$ – неубывающая функция.

Свойство 3. Если x_1 – наименьшая варианта, то $F^*(x)=0$ при $x\le x_1$; если x_k – наибольшая варианта, то $F^*(x)=1$ при $x>x_k$.

На рис. 6 изображен график эмпирической функции

$$F^*(x)=\begin{cases}0 & \text{при} \quad x\le 3,\\ 0{,}2 & \text{при} \quad 3<x\le 4,\\ 0{,}3 & \text{при} \quad 4<x\le 7,\\ 0{,}7 & \text{при} \quad 7<x\le 10,\\ 1 & \text{при} \quad x>10.\end{cases}$$

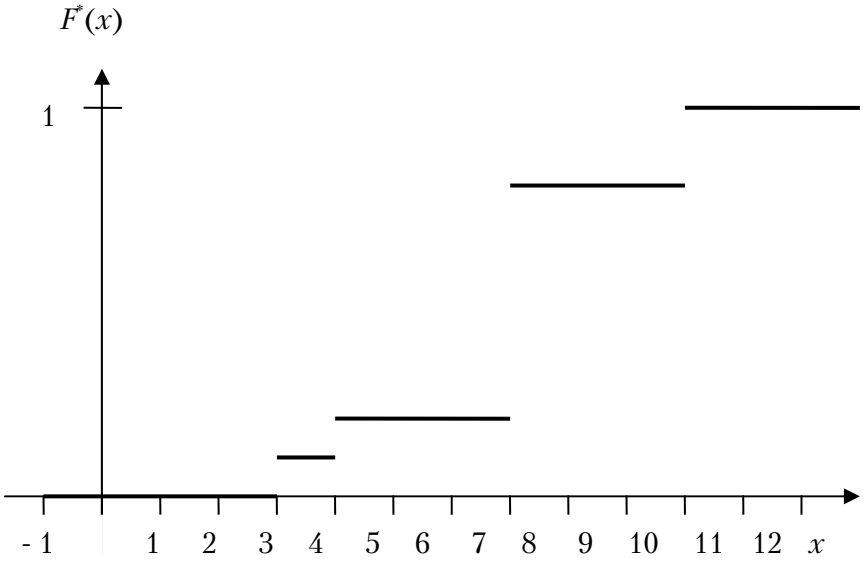

Рис.6

1.12. Числовые характеристики выборки

Числовыми характеристиками выборки являются: выборочная средняя, выборочная дисперсия, выборочное среднее квадратическое отклонение. Дадим определение каждой числовой характеристике выборки.

Выборочной средней $\overline{x_B}$ называют среднее арифметическое значение признака выборочной совокупности.

Если все значения $x_1, x_2, ..., x_n$ признака выборки объема n различны, то

$$\overline{x_B} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Если же все значения признака $x_1, x_2, ..., x_k$ имеют соответственно частоты $n_1, n_2, ..., n_k$, причем $n_1 + n_2 + ... + n_k = n$, то

$$\overline{x_B} = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n} \text{ или } \overline{x_B} = \frac{\sum_{i=1}^{k} n_i x_i}{n}.$$

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\overline{x_B}$.

Если все значения $x_1, x_2, ..., x_n$ признака выборки объема n различны, то

$$D_B = \frac{\sum_{i=1}^{n} (x_i - \overline{x_B})^2}{n}.$$

Если же все значения признака $x_1, x_2, ..., x_k$ имеют соответственно частоты $n_1, n_2, ..., n_k$, причем $n_1 + n_2 + ... + n_k = n$, то

$$D_B = \frac{\sum_{i=1}^{n} n_i (x_i - \overline{x_B})^2}{n}.$$

Выборочным средним квадратическим отклонением называют квадратный корень из выборочной дисперсии:

$$\sigma_B = \sqrt{D_B}.$$

В качестве оценки генеральной дисперсии принимают исправленную дисперсию

$$s^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x_B})^2}{n-1}.$$

Для оценки среднего квадратического отклонения генеральной совокупности используют «исправленное» среднее квадратическое отклонение, которое равно квадратному корню из исправленной дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - x_B)^2}{n-1}}.$$

Кроме выборочной средней и выборочной дисперсии применяются и другие характеристики вариационного ряда. Рассмотрим главные из них.

Модой M_0 называют варианту, которая имеет наибольшую частоту.

Медианой m_e называют варианту, которая делит вариационный ряд на две части, равные по числу вариант.

Размахом варьирования R называют разность между наибольшей и наименьшей вариантами:

$$R = x_{\max} - x_{\min}.$$

Размах является простейшей характеристикой рассеяния вариационного ряда.

Коэффициентом вариации V называют выраженное в процентах отношение выборочного среднего квадратического отклонения к выборочной средней:

$$V = \frac{\sigma_B}{x_B} \cdot 100\%.$$

Коэффициент вариации служит для сравнения величин рассеяния по отношению к выборочной средней двух вариационных рядов: тот из рядов имеет большее рассеяние по отношению к выборочной средней, у которого коэффициент вариации больше. Коэффициент вариации – безразмерная величина, поэтому он пригоден для сравнения рассеяний вариационных рядов, варианты которых имеют различную размерность.

Для вычисления сводных характеристик выборки удобно пользоваться эмпирическими моментами.

Обычным эмпирическим моментом порядка k называют среднее значение *k*-х степеней разности $x_i - C$:

$$M'_k = \frac{\sum n_i (x_i - C)^k}{n},$$

где x_i – наблюдаемая варианта;

n_i – частота варианты;

$n = \sum n_i$ – объем выборки;

C – произвольное постоянное число (ложный нуль).

Начальным эмпирическим моментом порядка k называют обычный момент порядка *k* при *C*=0

$$M_k = \frac{\sum n_i x_i^k}{n}.$$

Например, начальный эмпирический момент первого порядка равен выборочной средней:

$$M_1 = \frac{\sum n_i x_i}{n} = \overline{x_B}.$$

Центральным эмпирическим моментом порядка k называют обычный момент порядка *k* при $C = \overline{x_B}$

$$m_k = \frac{\sum n_i (x_i - \overline{x_B})^k}{n}.$$

Начальный эмпирический момент второго порядка равен выборочной дисперсии

$$m_2 = \frac{\sum n_i (x_i - \overline{x_B})^2}{n} = D_B.$$

Вычисление центральных моментов требует громоздких вычислений. Чтобы упростить расчеты, заменяют первоначальные варианты условными.

Зная обычные моменты, можно через них выразить центральные моменты:

$$m_2 = M'_2 - \left(M'_1\right)^2,$$

$$m_3 = M'_3 - 3M'_2 M'_1 + 2\left(M'_1\right)^3,$$

$$m_4 = M'_4 - 4M'_3 M'_1 + 6M'_2 \left(M'_1\right)^2 - 3\left(M'_1\right)^4.$$

Условным эмпирическим моментом порядка k называют начальный момент порядка *k*, вычисленный для условных вариант:

$$M_k^* = \frac{\sum n_i u_i^k}{n} = \frac{\sum n_i \left(\frac{x_i - C}{h} \right)^k}{n}.$$

В частности,

$$M_1^* = \frac{\sum n_i \left(\frac{x_i - C}{h} \right)}{n} = \frac{1}{h}\left[\frac{\sum n_i x_i}{n} - C \frac{\sum n_i}{n} \right] = \frac{1}{h}\left(\overline{x_B} - C \right).$$

Отсюда

$$\overline{x_B} = M_1^* h + C.$$

Таким образом, для нахождения выборочной средней, достаточно вычислить условный момент первого порядка, умножить его на h и к результату произведения прибавить ложный нуль C.

Выразим обычные моменты через условные:

$$M_k^* = \frac{1}{h^k} \frac{\sum n_i (x_i - C)^k}{n} = \frac{M_k'}{h^k}.$$

Следовательно, $M_k' = M_k^* h^k$.

Выразим центральные моменты через условные:

$$m_2 = [M_2^* - \left(M_1^* \right)^2] h^2,$$

$$m_3 = [M_3^* - 3M_2^* M_1^* + 2\left(M_1^* \right)^3] h^3,$$

$$m_4 = [M_4^* - 4M_3^* M_1^* + 6M_2^* \left(M_1^* \right)^2 - 3\left(M_1^* \right)^4] h^4.$$

Выборочную дисперсию можно вычислить по условным моментам первого и второго порядков:

$$D_B = [M_2^* - \left(M_1^* \right)^2] h^2.$$

Для оценки отклонения эмпирического распределения от нормального используют такие характеристики, как асимметрия и эксцесс.

Асимметрия эмпирического распределения определяется равенством

$$a_s = \frac{m_3}{\sigma_B^3},$$

где m_3 – центральный эмпирический момент третьего порядка.

Эксцесс эмпирического распределения определяется равенством

$$e_k = \frac{m_4}{\sigma_B^4} - 3,$$

где m_4 – центральный эмпирический момент четвертого порядка.

Эксцесс является показателем «крутости» вариационного ряда по сравнению с нормальным распределением, т.е. большего или меньшего подъема кривой теоретического распределения.

1.13. Метод произведений

Метод произведений дает удобный способ вычисления условных моментов различных порядков вариационного ряда с равноотстоящими вариантами. Зная же условные моменты, нетрудно найти начальные и центральные эмпирические моменты.

Методом произведений удобно вычислять выборочную среднюю и выборочную дисперсию. Для их вычисления удобно пользоваться расчетной таблицей 1, которая составляется следующим образом:

1) в первый столбец таблицы записывают выборочные (первоначальные) варианты, располагая их в возрастающем порядке;

2) во второй столбец записывают частоты вариант; складывают все частоты и их сумму (объем выборки n) помещают в нижнюю клетку столбца;

3) в третий столбец записывают условные варианты $u_i = \frac{x_i - C}{h}$, причем в качестве ложного нуля C выбирают варианту, которая расположена примерно в середине вариационного ряда и имеет наибольшую частоту, полагают h равным разности между любыми двумя соседними вариантами; практически же третий столбец заполняется так: в клетке строки, содержащей выбранный ложный нуль, пишут 0; в клетках над нулем пишут последовательно –1, –2, –3 и т.д., а под нулем – 1, 2, 3 и т.д.;

4) умножают частоты на условные варианты и записывают их произведения $n_i u_i$ в четвертый столбец; сложив все полученные числа, их сумму $\sum n_i u_i$ помещают в нижнюю клетку столбца;

5) умножают частоты на квадраты условных вариант и записывают их произведения $n_i u_i^2$ в пятый столбец; сложив все полученные числа, их сумму $\sum n_i u_i^2$ помещают в нижнюю клетку столбца;

6) умножают частоты на квадрат условных вариант, увеличенных каждая на единицу, и записывают произведения $n_i(u_i+1)^2$ в шестой контрольный столбец, сложив все полученные числа, их сумму $\sum n_i(u_i+1)^2$ помещают в нижнюю клетку столбца.

Замечание 1. Целесообразно отдельно складывать отрицательные числа четвертого столбца (их сумму A_1 записывают в клетку строки, содержащей ложный нуль) и отдельно положительные числа (их сумму A_2 записывают в предпоследнюю клетку столбца); тогда $\sum n_i u_i = A_1 + A$.

Таблица 1

1	2	3	4	5	6	7	8
x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	$n_i(u_i+1)^4$

Замечание 2. При вычислении произведений $n_i u_i^2$ пятого столбца целесообразно числа $n_i u_i$ четвертого столбца умножать на u_i; аналогично заполняются шестой и седьмой столбцы.

Замечание 3. Восьмой столбец служит для контроля вычислений: если сумма $\sum n_i(u_i+1)^4$ окажется равной сумме $\sum n_i u_i^4 + 4\sum n_i u_i^3 + 6\sum n_i u_i^2 + 4\sum n_i u_i + \sum n_i$, то вычисления проведены правильно.

После того как расчетная таблица заполнена и проверена правильность вычислений, вычисляют условные моменты:

$$M_1^* = \frac{\sum n_i u_i}{n}, \quad M_2^* = \frac{\sum n_i u_i^2}{n},$$

$$M_3^* = \frac{\sum n_i u_i^3}{n}, \quad M_4^* = \frac{\sum n_i u_i^4}{n}.$$

Затем вычисляют выборочные среднюю и дисперсию по формулам:

$$\overline{x_в} = M_1^* h + C, \quad D_в = [M_2^* - (M_1^*)^2]h^2.$$

Пример 1. Определить числовые характеристики выборки методом моментов.

x_i	8,33	8,43	8,53	8,63	8,73	8,83	8,93
n_i	3	10	18	28	25	14	2

Решение. Выборка задана в виде распределения равноотстоящих вариант и соответствующих им частот, поэтому выборочные среднюю и дисперсию удобно найти методом произведений по формулам:

$$\overline{x_в} = M_1^* h + C, \quad D_в = [\, M_2^* - (M_1^*)^2]\, h^2,$$

где h – шаг (разность между двумя соседними вариантами);

C – ложный нуль;

$u_i = \dfrac{x_i - C}{h}$ – условная варианта;

$M_1^* = \dfrac{\sum m_i u_i}{n}$ – условный момент первого порядка;

$M_2^* = \dfrac{\sum m_i u_i^2}{n}$ – условный момент второго порядка.

Составим расчетную табл. 2. Для этого:

1) запишем варианты $\boldsymbol{x_i}$ в первый столбец;

2) запишем частоты $\boldsymbol{n_i}$ во второй столбец; сумму частот 100 запишем в нижнюю клетку столбца;

3) в качестве ложного нуля выберем варианту C = 8,63, которая имеет наибольшую частоту; в клетке четвертого столбца, которая принадлежит строке, содержащей ложный нуль, пишем 0; над нулем последовательно –1, –2, –3, а под нулем 1, 2, 3;

4) произведения частот $\boldsymbol{n_i}$ на условные варианты $\boldsymbol{u_i}$ запишем в пятый столбец; отдельно находим сумму $[-9+(-20)+(-18)=-47]$ отрицательных чисел и отдельно сумму $[25+28+6=59]$ положительных чисел; сложив эти числа, их сумму $[-47+59=12]$ записываем в нижнюю клетку пятого столбца;

5) произведения частот на квадраты их условных вариант, т.е. $\boldsymbol{n_i u_i^2}$, запишем в шестой столбец (удобнее перемножить числа каждой строки четвертого и пятого столбцов: $u_i \cdot n_i u_i = n_i u_i^2$); сумму чисел столбца $(27+40+18+25+56+18=184)$ записываем в нижнюю клетку шестого столбца;

6) произведения частот на кубы их условных вариант, т.е. $\boldsymbol{n_i u_i^3}$, запишем в седьмой столбец (удобнее перемножить числа каждой строки четвертого и шестого столбцов: $u_i \cdot n_i u_i^2 = n_i u_i^3$); сумму чисел столбца $(-81+(-80)+(-18)+25+112+54=12)$ записываем в нижнюю клетку седьмого столбца;

7) восьмой столбец заполняем аналогично тому, как заполняли седьмой столбец, перемножаем числа каждой строки четвертого и седьмого столбцов:

$u_i \cdot n_i u_i^3 = n_i u_i^4$; сумму чисел столбца $(243+160+18+25+224+162=832)$ записываем в нижнюю клетку восьмого столбца;

8) произведения частот на квадраты условных вариант, увеличенных на единицу, т.е. $\boldsymbol{n_i(u_i+1)^4}$, запишем в девятый контрольный столбец; сумму чисел столбца $(48+10+28+400+1134+512=2132)$ записываем в нижнюю клетку девятого столбца.

Шаг (разность между двумя соседними вариантами) h=0,1; ложный нуль (варианта, которая имеет наибольшую частоту) C= 8,63.

В результате получим расчетную табл. 2.

Таблица 2

i	x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	$n_i(u_i+1)^4$
1	8,33	3	-3	-9	27	-81	243	48
2	8,43	10	-2	-20	40	-80	160	10
3	8,53	18	-1	-18	18	-18	18	0
4	8,63	28	0	0	0	0	0	28
5	8,73	25	1	25	25	25	25	400
6	8,83	14	2	28	56	112	224	1134
7	8,93	2	3	6	18	54	162	512
$\sum$		100		12	184	12	832	2132

Для контроля вычислений пользуются тождеством:

$$\sum n_i(u_i+1)^4 = \sum n_i u_i^4 + 4\sum n_i u_i^3 + 6\sum n_i u_i^2 + 4\sum n_i u_i + \sum n_i .$$

Контроль:

$$\sum n_i(u_i+1)^4 = 2132,$$

$$\sum n_i u_i^4 + 4\sum n_i u_i^3 + 6\sum n_i u_i^2 + 4\sum n_i u_i + \sum n_i =$$
$$= 832 + 4\cdot 12 + 6\cdot 184 + 4\cdot 12 + 100 = 832 + 48 + 1104 + 14 + 100 = 2132.$$

Совпадение контрольных сумм свидетельствует о правильности вычислений.

Вычислим условные моменты первого, второго, третьего и четвертого порядков:

$$M_1^* = \frac{\sum n_i u_i}{n} = \frac{12}{100} = 0{,}12;\ M_2^* = \frac{\sum n_i u_i^2}{n} = \frac{184}{100} = 1{,}84;$$

$$M_3^* = \frac{\sum n_i u_i^3}{n} = \frac{12}{100} = 0{,}12;\ M_4^* = \frac{\sum n_i u_i^4}{n} = \frac{832}{100} = 8{,}32.$$

Вычислим выборочную среднюю:

$$\bar{x}_в = M_1^* h + C = 0{,}12\cdot 0{,}1 + 8{,}63 = 0{,}012 + 8{,}63 = 8{,}643.$$

Найдем выборочную дисперсию, выборочное среднее квадратическое отклонение:

$$D_{\text{в}} = [\, M_2^* - (M_1^*)^2]\, h^2 = [1{,}84 - 0{,}12^2] \cdot 0{,}1^2 = 1{,}8256 \cdot 0{,}01 = 0{,}018256,$$

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{0{,}018256} = 0{,}13511.$$

«Исправленная» выборочная дисперсия вычисляется следующим образом:

$$s = \sqrt{D_{\text{в}} \frac{n}{n-1}} = \sqrt{0{,}018256 \cdot \frac{100}{99}} = 0{,}1357954.$$

Найдем коэффициент вариации, который показывает относительный разброс вокруг выборочной средней $\overline{x_в}$:

$$V = \frac{\sigma_{\text{в}}}{x_{\text{в}}} \cdot 100\% = \frac{0{,}13511}{8{,}643} \cdot 100\% = 0{,}01563 \cdot 100\% = 1{,}563\%.$$

Вычислим размах варьирования, характеризующий рассеяние вариационного ряда:

$$R = x_{\max} - x_{\min} = 8{,}98 - 8{,}28 = 0{,}7.$$

Найдем центральные эмпирические моменты второго, третьего и четвертого порядков:

$$m_2 = [M_2^* - (M_1^*)^2] \cdot h^2 = [1{,}84 - (0{,}12)^2] \cdot (0{,}1)^2 = 0{,}018256,$$

$$m_3 = [M_3^* - 3 M_1^* M_2^* + 2 (M_1^*)^3] \cdot h^3 =$$
$$= [0{,}12 - 3 \cdot 0{,}12 \cdot 1{,}84 + 2\,(0{,}12)^3] \cdot (0{,}1)^3 = -0{,}000538944,$$

$$m_4 = [M_4^* - 4 M_1^* M_3^* + 6 (M_1^*)^2 M_2^* - 3 (M_1^*)^4] \cdot h^4 =$$
$$= [8{,}32 - 4 \cdot 0{,}12 \cdot 0{,}12 + 6 \cdot (0{,}12)^2 \cdot 1{,}84 - 3 \cdot (0{,}12)^4] \cdot (0{,}1)^4 = 0{,}000842075.$$

Найдем асимметрию и эксцесс, учитывая, что выборочное среднее квадратическое отклонение и центральные эмпирические моменты третьего и четвертого порядков были найдены ранее:

$$a_s = \frac{m_3}{s^3} = \frac{-0{,}000538944}{(0{,}1357)^3} = \frac{-0{,}000538944}{0{,}002498846} = -0{,}21658,$$

$$e_k = \frac{m_4}{s^4} - 3 = \frac{0{,}000842075}{(0{,}1357)^4} - 3 = \frac{0{,}000842075}{0{,}000339093} - 3 = 2{,}48332 - 3 = -0{,}51668.$$

1.14. Интервальные оценки. Доверительные интервалы для оценки математического ожидания нормального распределения при известном среднем квадратическом отклонении

Интервальной называют оценку, которая определяется двумя числами – концами интервала, покрывающего оцениваемый параметр. Интервальные оценки позволяют установить точность и надежность оценок.

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надежностью γ.

Пусть количественный признак X генеральной совокупности распределен нормально. Известно среднее квадратическое отклонение σ этого распределения. Необходимо оценить неизвестное математическое ожидание a по выборочной средней $\overline{x}$. Найдем доверительные интервалы, покрывающие параметр a с надежностью γ.

Для этого будем рассматривать выборочную $\overline{x}$ среднюю как случайную величину $\overline{X}$ и выборочные значения признака $x_1, x_2, ..., x_n$ – как одинаково распределенные независимые случайные величины $X_1, X_2, ..., X_n$. Математическое ожидание каждой из этих величин равно a и среднее квадратическое отклонение – σ.

Если случайная величина X распределена нормально, то выборочная средняя $\overline{X}$, найденная по независимым наблюдениям, также распределена нормально. Известно, что

$$M(\overline{X}) = a, \quad \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Требуется выполнение соотношения

$$P\left(\left|X - a\right| < \delta\right) = \gamma,$$

где γ – заданная надежность.

Пользуясь формулой

$$P\left(\left|X - a\right| < \delta\right) = 2\Phi\left(\frac{\delta}{\sigma}\right),$$

заменив X на $\overline{X}$ и σ на $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, получим

$$P\left(\left|\overline{X} - a\right| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi\,(t),$$

где $t = \frac{\delta\sqrt{n}}{\sigma}$.

Из последнего равенства $\delta = \frac{t\sigma}{\sqrt{n}}$. Можем записать

$$P\left(\left|\overline{X} - a\right| < \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t).$$

Принимая, что вероятность P задана и равна γ, имеем:

$$P\left(\overline{x} - \frac{t\sigma}{\sqrt{n}} < a < \overline{x} + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

Таким образом, можно с надежностью γ утверждать, что доверительный интервал $\left(\overline{x} - \frac{t\sigma}{\sqrt{n}}, \overline{x} + \frac{t\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a, точность оценки равна $\delta = \frac{t\sigma}{\sqrt{n}}$.

Число t определяется из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \frac{\gamma}{2}$. По таблице функции Лапласа (см. прил. 4) находят аргумент t, которому соответствует значение функции Лапласа, равное $\frac{\gamma}{2}$.

Оценку $\left|\overline{x} - a\right| < \frac{t\sigma}{\sqrt{n}}$ называют классической. Из формулы $\delta = \frac{t\sigma}{\sqrt{n}}$, определяющей точность классической оценки, можно сделать следующие выводы:

1) при возрастании объема выборки n число δ убывает и, следовательно, точность оценки увеличивается;

2) увеличение надежности оценки $\gamma = 2\Phi(t)$ приводит к увеличению t ($\Phi(t)$ – возрастающая функция), следовательно, и к возрастанию δ; другими словами, увеличение надежности классической оценки влечет за собой уменьшение ее точности.

1.15. Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном среднем квадратическом отклонении

Пусть количественный признак X генеральной совокупности распределен нормально. Среднее квадратическое отклонение σ неизвестно. Необходимо оценить неизвестное математическое ожидание a с помощью доверительных интервалов.

По данным выборки построим случайную величину, которая имеет распределение Стьюдента с $k = n - 1$ степенями свободы:

$$T = \frac{\overline{X} - a}{S} \cdot \sqrt{n},$$

где $\overline{X}$ – выборочная средняя;

S – «исправленное» среднее квадратическое отклонение;

n – объем выборки.

Распределение Стьюдента определяется параметром n – объемом выборки и не зависит от неизвестных параметров a и σ.

Плотность распределения Стьюдента

$$S(t,n) = B_n \left[1 + \frac{t^2}{n-1} \right]^{-\frac{n}{2}},$$

где $B_n = \dfrac{\Gamma\left(\dfrac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\dfrac{n-1}{2}\right)}$.

Так как $S(t,n)$ четная функция, зависящая от t, то вероятность выполнения неравенства $\left| \dfrac{\overline{X} - a}{S} \cdot \sqrt{n} \right| < t_\gamma$ определяется следующим образом:

$$P\left(\left| \frac{\overline{X} - a}{S} \cdot \sqrt{n} \right| < t_\gamma \right) = 2 \int_0^{t_\gamma} S(t,n)dt = \gamma.$$

Заменяя неравенство в круглых скобках равносильным ему двойным неравенством, получим

$$P\left(\overline{X} - \frac{t_\gamma S}{\sqrt{n}} < a < \overline{X} + \frac{t_\gamma S}{\sqrt{n}} \right) = \gamma.$$

Таким образом, пользуясь распределением Стьюдента, нашли доверительный интервал $\left(\overline{x} - \dfrac{t_\gamma S}{\sqrt{n}} < a < \overline{x} + \dfrac{t_\gamma S}{\sqrt{n}} \right)$, покрывающий неизвестный параметр a с надежностью γ. Случайные величины $\overline{X}$ и S заменены величинами $\overline{x}$ и s, найденными по выборке. По таблице прил. 3 и по данным n и γ можно найти t_γ.

Заметим, что для малых выборок ($n < 30$) замена распределения нормальным приводит к грубым ошибкам. Сужается доверительный интервал и тем самым повышается точность оценки. Это объясняется тем, что малая выборка содержит малую информацию об интересующем признаке.

Пример 2. Предполагая, что случайная величина распределена по нормальному закону и, зная ее числовые характеристики (пример 1), найти:

1. Доверительный интервал, покрывающий истинные размеры измеряемой величины с надежностью $\gamma = 0{,}95$, считая, что среднее квадратическое отклонение σ известно.

2. Доверительный интервал, покрывающий истинные размеры измеряемой величины с надежностью $\gamma = 0{,}95$, считая, что среднее квадратическое отклонение σ неизвестно.

3. Погрешность, с которой выборочная средняя $\overline{x}_{\text{в}}$ оценивает истинный размер, приняв надежность $\gamma = 0{,}99$.

4. Минимальное число измерений, которое нужно произвести, чтобы с надежностью $\gamma = 0{,}98$ можно было бы утверждать, что, принимая $\overline{x}_{\text{в}}$ за истинный размер измеряемой величины, совершается ошибка, не превышающая $\delta = 0{,}05$.

Решение. **1.** Известно, что $n = 7$, $\overline{x}_{\text{в}} = 8{,}643$, $\sigma_{\text{в}} = 0{,}13511$.

Требуется найти доверительный интервал

$$\overline{x}_{\text{в}} - t\frac{\sigma}{\sqrt{n}} < a < \overline{x}_{\text{в}} + t\frac{\sigma}{\sqrt{n}},$$

где $t\dfrac{\sigma}{\sqrt{n}} = \delta$ – точность оценки;

n – количество интервалов;

t – значение аргумента функции Лапласа $\Phi(t)$ (см. прил. 1), при котором $\Phi(t) = \dfrac{\gamma}{2}$.

Все величины, кроме t, известны. Найдем t из соотношения

$$\Phi(t) = \frac{0{,}95}{2} = 0{,}475.$$

По таблице прил. 4 находим $t = 1{,}96$. Тогда точность оценки:

$$\delta = 1{,}96 \cdot \frac{0{,}13511}{\sqrt{7}} = \frac{0{,}2648156}{2{,}6457513} = 0{,}1000909.$$

Доверительные границы:

$$\overline{x}_{\text{в}} - \delta = 8{,}643 - 0{,}100 = 8{,}543,\ \overline{x}_{\text{в}} + \delta = 8{,}643 + 0{,}100 = 8{,}743.$$

Окончательно получаем искомый доверительный интервал

$$8{,}543 < a < 8{,}743.$$

2. При неизвестном σ и объеме выборки $n < 30$ для нахождения доверительного интервала, покрывающего истинный размер, пользуются формулой

$$\bar{x}_{\text{в}} - t_{\gamma}\frac{s}{\sqrt{n}} < a < \bar{x}_{\text{в}} + t_{\gamma}\frac{s}{\sqrt{n}},$$

где s – «исправленное» выборочное среднее квадратическое отклонение;

t_{γ} находят по таблице прил. 2 и по заданным n и γ.

Известно, что $n = 7$, $s = 0{,}1357954$.

По таблице находим $t_{\gamma} = t(\gamma, n) = t(0{,}95, 7) = 2{,}45$. Точность оценки:

$$\delta = t_{\gamma}\frac{s}{\sqrt{n}} = 2{,}45 \cdot \frac{0{,}1357954}{2{,}6457513} = 0{,}12575.$$

Найдем доверительные границы:

$$\bar{x}_{\text{в}} - \delta = 8{,}643 - 0{,}126 = 8{,}517, \quad \bar{x}_{\text{в}} + \delta = 8{,}643 + 0{,}126 = 8{,}769.$$

Искомый доверительный интервал

$$8{,}517 < a < 8{,}769.$$

3. Пользуясь таблицей прил. 2, по $\gamma = 0{,}99$ и $n = 7$ находим $t_{\gamma} = t(\gamma, n) = t(0{,}99, 7) = 3{,}71$.

Найдем точность оценки:

$$\delta = t_{\gamma}\frac{\sigma_{\text{в}}}{\sqrt{n}} = 3{,}71 \cdot \frac{0{,}13511}{2{,}6457513} = 0{,}1894578.$$

С надежностью 0,99 погрешность, с которой выборочная средняя оценивает истинный размер, равна 0,1895.

4. Воспользуемся формулой, определяющей точность оценки математического ожидания генеральной совокупности по выборочной средней $\delta = \frac{t\,s}{\sqrt{n}}$. Отсюда $n = \frac{t^2 s^2}{\delta^2}$.

По условию, $\gamma = 0{,}98$. Следовательно, $\Phi(t) = \frac{0{,}98}{2} = 0{,}49$. По таблице прил. 1 найдем $t = 2{,}34$. Подставив $t = 2{,}34$, $s = 0{,}13579$ и соответствующую точность оценки δ.

При $\delta = 0{,}05$ искомый объем выборки

$$n = \left(\frac{2{,}34 \cdot 0{,}13579}{0{,}05} \right)^2 = 40 .$$

1.16. Статистическая гипотеза. Нулевая и конкурирующая, простая и сложная. Ошибки первого и второго рода

Часто необходимо знать закон распределения генеральной совокупности. Если закон распределения неизвестен, но есть основания предположить, что он имеет определенный вид, в этом случае выдвигают гипотезу: генеральная совокупность распределена по закону A.

Возможны и другие гипотезы: о равенстве параметров двух или нескольких распределений, о независимости выборок и многие другие.

Статистической гипотезой называется любое предложение о виде или параметрах неизвестного закона распределения.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. Поэтому эти гипотезы целесообразно различать.

Проверяемую гипотезу называют *нулевой (основной)* и обозначают H_0. Наряду с нулевой гипотезой H_0 рассматривают *альтернативную*, или *конкурирующую*, гипотезу H_1, которая является логическим отрицанием нулевой.

Различают гипотезы, которые содержат только одно и более одного предположений.

Простой называют гипотезу, содержащую только одно предположение.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Выдвинутую гипотезу проверяют статистическими методами и поэтому ее называют *статистической*. В результате статистической проверки гипотезы могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Правило, по которому гипотеза H_0 отвергается или принимается, называется *статистическим критерием*.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества. Одно подмножество содержит значения критерия, при которых нулевая гипотеза отвергается (критическая область), а другое – при которых она принимается (область допустимых значений).

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью допустимых значений (областью принятия) гипотезы называют совокупность значений критерия, при которых гипотезу принимают.

Вероятность допустить ошибку первого рода, т.е. отвергнуть гипотезу H_0, когда она верна, называется *уровнем значимости критерия*.

Вероятность не допустить ошибку второго рода, т.е. отвергнуть гипотезу H_0, когда она неверна, называется *мощностью критерия*.

Критическая область должна быть такой, чтобы при заданном уровне значимости мощность критерия была максимальной.

Основной принцип проверки статистических гипотез заключается в следующем: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

По своему прикладному содержанию статистические гипотезы можно подразделить на несколько основных типов:

– о равенстве числовых характеристик генеральных совокупностей;

– о числовых значениях параметров;

– о законе распределения;

– об однородности выборок (т.е. принадлежности их одной и той же генеральной совокупности).

1.17. Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

Одной из важнейших задач математической статистики является установление теоретического закона распределения случайной величины, характеризующей изучаемый признак по опытному (эмпирическому) распределению, представляющему вариационный ряд.

Если закон распределения генеральной совокупности неизвестен, но есть предположения о том, что он имеет определенный вид, то проверяют нулевую гипотезу: генеральная совокупность распределена по закону A.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится при помощи критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: χ^2 («хи квадрат») К. Пирсона, Колмогорова, Смирнова и др. Рассмотрим применение критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности. Будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

Критерий Пирсона устанавливает на принятом уровне значимости согласие или несогласие гипотезы с данными наблюдений.

Для того чтобы при заданном уровне значимости проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия:

$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n'_i)^2}{n'_i}.$$

Затем по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = s - 3$ найти критическую точку $\chi^2_{\text{кр}}(\alpha; k)$.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$ – нулевую гипотезу отвергают.

Замечание 1. Объем выборки должен быть достаточно велик, не менее 50. Каждая группа должна содержать не менее 5–8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Замечание 2. Для контроля вычислений формулу нахождения наблюдаемого значения критерия преобразуют к виду

$$\chi^2_{\text{набл}} = \frac{\sum n_i^2}{n'_2} - n.$$

Пример 3. При уровне значимости $\alpha = 0{,}05$ проверить гипотезу о нормальном распределении случайной величины с помощью критерия согласия.

Решение. Чтобы использовать χ^2 (хи-квадрат) – критерий Пирсона – возьмем из примера 1 и 2 следующие результаты: интервалы, эмпирические частоты, выборочную среднюю $\overline{x}_{\text{в}} = 8{,}643$, выборочное среднее квадратическое отклонение $\sigma_{\text{в}} = 0{,}1351$.

При уровне значимости $\alpha = 0{,}05$ (надежность $\gamma = 0{,}95$) проверим гипотезу о том, что случайная величина X распределена по нормаль-

ному закону. Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, то нет оснований отвергнуть гипотезу о нормальном распределении генеральной совокупности.

Чтобы найти $\chi^2_{\text{набл}}$, составим табл. 3.

Пронормируем X, т.е. перейдем к случайной величине Z и вычислим концы интервалов:

$$z_i = \frac{x_i - \overline{x}_{\text{в}}}{\sigma_{\text{в}}}, \qquad z_{i+1} = \frac{x_{i+1} - \overline{x}_{\text{в}}}{\sigma_{\text{в}}}.$$

Вычислим теоретические частоты:

$$n'_i = n \cdot P_i,$$

где m – объем выборки (сумма всех частот), n = 100;

$P_i = \Phi(z_{i+1}) - \Phi(z_i)$ – вероятность попадания X в интервал (x_i, x_{i+1}); $\Phi(Z)$ – функция Лапласа.

Таблица 3

i	Интервалы (x_i, x_{i+1})	Эмпирическая частота n_i	Вероятность P_i	Теоретические частоты n'_i	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
1	8,28-8,38	3	0,0219	2,19	0,6561	0,2996
2	8,38-8,48	10	0,0875	8,75	1,5625	0,1786
3	8,48-8,58	18	0,2061	20,61	6,8121	0,3305
4	8,58-8,68	28	0,2872	28,72	0,5184	0,0181
5	8,68-8,78	25	0,2374	23,74	1,5876	0,0669
6	8,78-8,88	14	0,1161	11,61	5,7121	0,4920
7	8,88-8,98	2	0,0335	3,35	1,8225	0,5440
$\sum$		100	0,9897	98,97		1,9297

$$P_1 = \Phi\left(\frac{8,38 - 8,643}{0,1351}\right) - \Phi\left(\frac{8,28 - 8,643}{0,1351}\right) = \Phi(-1,95) - \Phi(-2,68) =$$

$$= -0,4744 + 0,4963 = 0,0219,$$

$$P_2 = \Phi\left(\frac{8,48 - 8,643}{0,1351}\right) - \Phi\left(\frac{8,38 - 8,643}{0,1351}\right) = \Phi(-1,21) - \Phi(-1,95) =$$

$$= -0,3869 + 0,4744 = 0,0875,$$

$$P_3 = \Phi\left(\frac{8,58 - 8,643}{0,1351}\right) - \Phi\left(\frac{8,48 - 8,643}{0,1351}\right) = \Phi(-0,47) - \Phi(-1,21) =$$

$$= -0,1808 + 0,3869 = 0,2061,$$

$$P_4 = \Phi\left(\frac{8{,}68-8{,}643}{0{,}1351}\right) - \Phi\left(\frac{8{,}58-8{,}643}{0{,}1351}\right) = \Phi\left(0{,}27\right) - \Phi\left(-0{,}47\right) =$$

$$= 0{,}1064 + 0{,}1808 = 0{,}2872,$$

$$P_5 = \Phi\left(\frac{8{,}78-8{,}643}{0{,}1351}\right) - \Phi\left(\frac{8{,}68-8{,}643}{0{,}1351}\right) = \Phi(1{,}01) - \Phi(0{,}27) =$$

$$= 0{,}3438 - 0{,}1064 = 0{,}2374,$$

$$P_6 = \Phi\left(\frac{8{,}88-8{,}643}{0{,}1351}\right) - \Phi\left(\frac{8{,}78-8{,}643}{0{,}1351}\right) = \Phi\left(1{,}75\right) - \Phi\left(1{,}01\right) =$$

$$= 0{,}4599 - 0{,}3438 = 0{,}1161,$$

$$P_7 = \Phi\left(\frac{8{,}98-8{,}643}{0{,}1351}\right) - \Phi\left(\frac{8{,}88-8{,}643}{0{,}1351}\right) = \Phi\left(2{,}49\right) - \Phi\left(1{,}75\right) =$$

$$= 0{,}4934 - 0{,}4599 = 0{,}0335.$$

Вычислим наблюдаемое значение критерия Пирсона по формуле

$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n_i')^2}{n_i'}.$$

По таблице критических точек распределения χ^2 (см. прил. 3), по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $k = s - 3$, где s – число интервалов, находим критическую точку правосторонней критической области.

В нашем случае $s = 7$, следовательно, $k = s - 3 = 7 - 3 = 4$, а $\chi^2_{\text{кр}}(0{,}05;4) = 9{,}5$.

Так как $1{,}9297 < 9{,}5$, т.е. $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, то нет оснований отвергнуть гипотезу о нормальном распределении случайной величины. Другими словами, расхождение между эмпирическими и теоретическими частотами незначимо (случайно).

2. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

2.1. Функциональная и корреляционная зависимости

Изучение разнообразных явлений сопровождается выяснением закономерностей, которым подчиняются характерные для данных явлений количественные соотношения или связи. При этом оказывается, что только для тех явлений, происхождение которых связывается с четко учтенными факторами, количественные соотношения или связи имеют вполне точный и определенный характер. Для таких явлений, изучаемых, например, в физике, химии, механике, астрономии, действуют функциональные связи между количественными признаками. Характеристика этих связей в виде уравнения, графика или таблицы подчиняется принципу строго определенного соответствия, составляющего сущность функциональной зависимости.

Вместе с тем для самых разнообразных явлений массового характера нельзя установить закономерность в количественных соотношениях между рассматриваемыми показателями, которая удовлетворяла бы принципу строгого соответствия. Нарушение этого принципа связано с тем, что изменение одного показателя определяется не только изменением другого основного показателя, но и влиянием ряда сопутствующих второстепенных факторов.

Так, при установлении взаимосвязи между показателями использования основных средств и уровня производительности труда на заводе выясняется, что на уровень производительности труда, помимо объема затрачиваемых основных средств, влияют еще и другие факторы (рационализация производственного процесса, организация труда и др.). Взаимосвязи между количеством осадков и показателями урожайности, толщиной покрова снега и показателями стока воды, показателями содержания марганца в стали и степенью ее прокаливаемости, начальной прочностью бетона и прочностью его через 28 дней и т. д также не удовлетворяют условию определенного соответствия.

Во всех таких примерах мы сталкиваемся с невозможностью учета влияния всех факторов на интересующие нас количественные соотношения между двумя основными величинами (показателями). Поэтому характеристика каждой такой взаимосвязи по данным отдельных наблюдений носит случайный характер и может выявить некоторые закономерности лишь по данным большого числа наблюдений.

Характерная особенность взаимосвязей в массовых явлениях состоит в том, что каждому значению одной величины x соответствует распределение значений y (т. е. несколько значений y с различными ве-

роятностями каждого из них), меняющееся с изменением x. Такое же соответствие имеет место между каждым значением величины y и связанными с ним значениями x.

В отличие от функциональной зависимости связь такого характера между двумя величинами называется *статистической*. Степень рассеяния возможных значений y, соответствующих каждому значению x, характеризует большую или меньшую тесноту связи между этими величинами. Это значит, что если влияние неучтенных факторов на изучаемую связь между величинами x и y незначительно, то степень рассеяния значений y мала, а связь между x и y имеет большую тесноту. Если же влияние неучтенных факторов значительно, то степень рассеяния значений y велика, а теснота связи между x и y мала.

Для выяснения математической сущности связей такого вида на конкретном примере обратимся к данным табл. 4 распределения 100 растений житняка по общему весу x и по весу семян y каждого растения.

При составлении таблицы растения житняка сгруппированы в отдельные классы по общему весу и по весу семян, а затем определены середины классов, т. е. средние значения этих весов по каждому классу. Середины классов растений по общему весу обозначены переменной x, а середины классов по весу семян — переменной y. Так, $x_1 = 25$ обозначает середину класса растений с общим весом от 20 до 30 г, $x_2 = 35$ — середину класса от 30 до 40 г и т.д., $y_1 = 13$ — середину класса растений с весом семян от 10,5 до 15,5 г, $y_2 = 18$ — середину класса от 15,5 до 20,5 г и т.д.

Таблица 4

x \ y	13	18	23	28	33	38	43	48	53	58	63	68	n_x
25	3	2											5
35		6	4										10
45		1	13	5									19
55		1	2	4	8	1							16
65			1		4	4	2						11
75					2	6	6	2					16
85							1	5					6
95								1	4	1			6
105									2	4	1	1	8
115										1		1	2
125												1	1
n_y	3	10	20	9	14	11	9	8	6	6	1	3	100

Символом n_x обозначена численность класса (частота) растений с соответственным общим весом x, а символом n_y — частота растений с соответственным весом семян y. Так, число 19 в крайнем справа столбце (n_x) означает количество растений с общим весом (в среднем) 45 г (т.е. от 40 до 50 г), а число 20 в нижней строке (n_y) — количество растений с весом семян в (среднем) 23 г (т.е. от 20,5 до 25,5 г).

Числами во внутренних клетках обозначены частоты соответственных комбинаций растений с некоторым общим весом x и с некоторым весом семян y. Так, число 8 означает количество растений с общим весом 55 г (от 50 до 60 г) и с весом семян 33 г (от 30,5 до 35,5 г); число 4 в третьем (внутреннем) столбце — количество растений с общим весом (в среднем) 35 г и с весом семян (в среднем) 23 г, а число 4 в третьей снизу (внутренней) строке — количество растений с общим весом 105 г и с весом семян 58 г.

В обобщенных обозначениях для чисел во внутренних клетках применяется символ n с двойным индексом n_{xy}. Так, приведенное выше число 8 следовало бы обозначить символом $n_{4,5}$, ибо оно указывает на количество растений с общим весом $x_4 = 55$ и с весом семян $y_5 = 33$. По этим признакам число 4 из третьего столбца следовало бы обозначить символом $n_{2,3}$, а число 4 из третьей снизу строки — символом $n_{9,10}$.

Рассмотренная структура таблицы распределения растений житняка по общему весу и по весу семян отдельных растений раскрывает на этом частном примере общую структуру так называемой *корреляционной таблицы,* связывающей значения изучаемых показателей x и y.

Таблица 5

	y_1	y_2	y_3	...	...	...	y_l	n_x
x_1		$n_{1,2}$	$n_{1,3}$	...	...	...	$n_{1,l}$	n_{x_1}
x_2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	...	...	...	$n_{2,l}$	n_{x_2}
x_3	$n_{3,1}$		$n_{3,3}$	...	...	...	$n_{3,l}$	n_{x_3}
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
x_k	$n_{k,1}$	$n_{k,2}$	$n_{k,3}$	...	...	...	$n_{k,l}$	n_{x_k}
n_y	n_{y_1}	n_{y_2}	n_{y_3}	...	...	...	n_{y_l}	N

Суммы чисел n, расположенных во внутренних клетках, по строкам дают частоты соответственных значений переменной x. Так,

$$\sum n_{1,j} = n_{1,1} + n_{1,2} + \ldots + n_{1,l} = n_{x_1};$$
$$\sum n_{3,j} = n_{3,1} + n_{3,2} + \ldots + n_{3,l} = n_{x_3}.$$

Суммы чисел n_{xy} по столбцам дают частоты соответственных значений переменной y.

Так,
$$\sum n_{i,1} = n_{1,1} + n_{2,1} + \ldots + n_{k,1} = n_{y_1};$$
$$\sum n_{i,4} = n_{1,4} + n_{2,4} + \ldots + n_{k,4} = n_{y_4}.$$

Суммирование всех чисел n_{xy} можно представить в виде последовательного суммирования сначала по каждой строке, а затем по крайнему правому столбцу n_x или в виде суммирования сначала по каждому столбцу, а затем по нижней строке n_y.

Совпадение результатов суммирования, выполненного в одном или в другом порядке, подтверждает правильность составления корреляционной таблицы:

$$\sum\sum n_{xy} = \sum n_x = \sum n_y = N.$$

Если совпадение результатов нарушено, то ошибка может быть устранена проверкой результатов суммирования по каждой строке и по каждому столбцу.

В частных случаях числа n_{xy} располагаются рядами, заполняющими не все клетки строк и столбцов. При этом совокупность чисел в каждой строке — это *ряд распределения значений y*, соответствующих данному значению x, а совокупность чисел в столбце — *ряд распределения значений x*, соответствующих данному значению y. По корреляционной табл. 4, составленной для растений житняка, можно, например, отметить, что значению $x_5 = 65$ соответствует $n_{x_5} = 11$ значений y со следующим рядом распределения этих значений:

значения y	23	33	38	43
их частоты	1	4	4	2.

Распределение значений y, соответствующих значению $x_{11} = 125$, состоит из одного значения $y_{12} = 68$.

Значению $x_3 = 45$ соответствует 19 значений y со следующим распределением:

значения y	18	23	28
их частоты	1	13	5.

Так же элементарно можно охарактеризовать распределения значений x, соответствующие тем или другим значениям y.

Корреляционная таблица, составленная на основании результатов наблюдения за значениями переменных x и y, позволяет после некоторой математической обработки ее данных подойти к разрешению двух основных задач корреляционного анализа: установлению формы корреляционной связи между переменными x и y и определению тесноты этой связи.

Рассмотрение в корреляционной таблице рядов распределения значений y, соответствующих последовательным значениям x, может выявить некоторые закономерности в смещении этих рядов.

Простейшие случаи, характерные для формы таких смещений, позволяют убедиться в том, что с возрастанием значений x в среднем растут или в среднем убывают значения y, что с возрастанием значений x значения y в среднем сначала возрастают, а затем убывают, или наоборот. К этим характеристикам связей между значениями x и y приводит внешний вид расположения рядов распределения значений y, соответствующих последовательным значениям x.

Так, по данным корреляционной табл. 4 распределения растений житняка смещение рядов распределения значений y показывает, что с возрастанием x (общего веса растения) возрастает в среднем и y (вес семян растения). Но эта связь выразится более отчетливо, если каждому значению x будет поставлено в соответствия частное среднее значение y, которое обозначим символом $\overline{y}_x$.

Вычисляя эти частные средние по правилу определения средней взвешенной, будем иметь:

$$\overline{y}_{x=25} = \frac{13 \cdot 3 + 18 \cdot 2}{3 + 2} = 15;$$

$$\overline{y}_{x=35} = \frac{18 \cdot 6 + 23 \cdot 4}{6 + 4} = 25;$$

$$\overline{y}_{x=45} = \frac{18 \cdot 1 + 23 \cdot 13 + 28 \cdot 5}{1 + 13 + 5} \approx 24{,}1.$$

С помощью таких средних, вычисленных для всех значений x, исходная табл. 4 приводится к форме, отражающей *связь между значениями x и соответствующими частными средними* $\overline{y}_x$*:*

Таблица 6

$1\,x$	25	35	45	55	65	75	85	95	105	115	125
$\overline{y}_x$	15,0	25,0	24,1	29,9	35,7	40,5	47,2	53,0	58,4	63,0	68,0

Графическое отображение данных табл. 3 в виде точек, соответствующих парам значений x и $\overline{y}_x$, с последовательным соединением этих точек отрезками прямых приводит к ломаной, которая называется *эмпирической линией регрессии* y по x. По этой линии, или, вернее, по взаимному расположению точек (вершин ломаной), можно наметить форму линии, около которой группируются точки $(x, \overline{y}_x)$ с наименьшими отклонениями. Такую линию называют теоретической линией регрессии, или просто линией регрессии y по x. *Зависимость* $\overline{y}_x = f(x)$, *соответствующая линии регрессии, называется уравнением регрессии* y *по* x, *или корреляционной зависимостью между* y и x.

Отыскание уравнения этой линии дает разрешение первой основной задачи корреляционного анализа — *установления формы корреляционной связи между переменными* x и y.

Если точки $(x; \overline{y}_x)$ располагаются около некоторой прямой, то линия регрессии называется прямой регрессии y по x, и соответствующая операция «выравнивания» ломаной сводится к аналитическому определению параметров линейной функции $\overline{y}_x = ax + b$, т. е. к линейной корреляции.

К этому типу корреляционной зависимости между y и x приводит, в частности, рассматриваемый пример распределения растений житняка по общему весу и по весу семян (рис. 7).

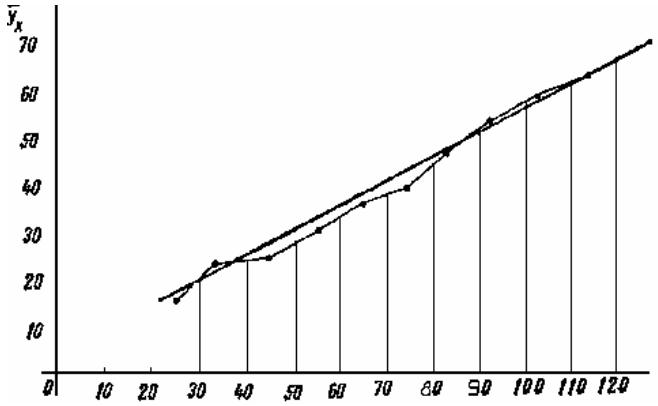

Рис. 7

Если же расположение точек, соответствующих парам значений x и $\overline{y}_x$, приводит к выравниванию ломаной с помощью какой-либо кривой (в простейших случаях — с помощью параболы или гиперболы), то соответствующее уравнение $\overline{y}_x = f(x)$ обозначает криволинейную корреляционную зависимость между y и x.

Здесь мы выяснили, таким образом, возможность установить наличие корреляционной связи между значениями x и соответствующими

частными средними значениями y. Но та же корреляционная таблица позволяет поставить вопрос о корреляционной связи между значениями y и соответствующими им частными средними значениями x.

Действуя аналогично, следует найти для каждого значения y соответствующую частную среднюю $\overline{x}_y$ и по парам значений $\left(\overline{x}_y; y\right)$ построить ломаную, являющуюся эмпирической линией регрессии x по y.

Если взаимное расположение вершин этой ломаной, соответствующих парам значений y и $\overline{x}_y$, может дать приближенное представление о некоторой линии, то операция выравнивания приведет к уравнению $\overline{x}_y = \varphi\left(y\right)$, выражающему *корреляционную зависимость* между x и y.

Корреляционные уравнения $\overline{y}_x = f\left(x\right)$ и $\overline{x}_y = \varphi\left(y\right)$ называются также уравнениями регрессии. Первое уравнение называется уравнением регрессии y по x, а второе — уравнением регрессии x по y. Соответственно геометрические образы этих уравнений называются линиями регрессии y по x и x по y.

Термин «регрессия», принятый для обозначения корреляционного характера связи между изучаемыми показателями и для графического отображения этой связи в виде некоторой линии, отражает тенденцию смещения рядов распределения значений y с изменением соответственных значений x. Так, в табл. 4 с увеличением значений x соответствующие ряды распределений y смещаются в сторону больших значений y. Контуры фигуры, образуемой скоплением данных распределения в таблице, позволяют приближенно представить форму связи между изучаемыми показателями в виде линии регрессии, выравнивающей обнаруженное в таблице смещение.

При составлении эмпирической линии регрессии y по x может оказаться, что все точки $\left(x; \overline{y}_x\right)$ лежат на прямой или на кривой, уравнение которой выражается в виде

$$\overline{y}_x = f\left(x\right).$$

В таких случаях говорят, что между x и y существует точная корреляционная зависимость (линейная, если эта линия — прямая, и криволинейная — в общем случае). Эти результаты в отношении эмпирической линии регрессии могут иметь место и при малой, и при большой степени рассеяния значений y относительно линии регрессии. Такое различие в степени рассеяния характеризует тесноту изучаемой корреляционной зависимости — при малом рассеянии теснота считается большой, и наоборот.

То же может иметь место и при составлении эмпирической линии регрессии x по y.

Признаком наличия точной линейной корреляции является обращение в тождество соответствующего уравнения регрессии при подстановке в него любой пары значений $(x; \overline{y}_x)$ или $(\overline{x}_y; y)$.

Бывает и так, что значения $\overline{y}_x$ оказываются одинаковыми для всех значений x. Это свидетельствует об отсутствии корреляционной связи y по x.

То же имеет место и в отношении связи x по y.

Заключительной стадией отыскания формы связи является операция выравнивания. Она состоит в определении аналитическими методами параметров корреляционного уравнения, которому приближенно удовлетворяют значения x и $\overline{y}_x$ или y и $\overline{x}_y$, характеризующие количественные признаки изучаемых явлений. При этом возможно, что одно из рассматриваемых явлений непосредственно воздействует на другое (например, интенсивность орошения и урожайность культуры или рост выработки продукции и доля накладных расходов в общей сумме затрат предприятия), а также, то что оба явления, связь между которыми требуется установить, находятся под влиянием какого-либо третьего общего явления (например, показатели урожайности двух различных культур, находящихся в одних и тех же климатических условиях).

Соотношения между показателями, характерные для корреляционной зависимости, имеют не точный, а приближенный характер, ибо, как выше указывалось, при изучении этих соотношений остаются неучтенными различные дополнительные факторы, которые рассеивают воздействие одного из основных показателей на другой. Поэтому *второй задачей теории корреляции является измерение тесноты корреляционной связи*. Такая связь в виде корреляционного уравнения будет тем теснее (ближе к данным наблюдения над значениями изучаемых показателей), чем слабее рассеяние связи между этими показателями под влиянием дополнительных неучтенных факторов.

Практическое значение теории корреляции состоит в том, что она позволяет, используя опытные данные и известные сведения о значениях той или иной величины, определять границы, в которых должна заключаться другая величина, с ней связанная.

2.2. Линейная корреляция

Этот вид корреляционной зависимости весьма важен, так как очень многие корреляционные связи, характерные для количественных признаков наблюдаемых однородных фактов, близки к линейным. Данные наблюдения, представленные в виде корреляционной таблицы, и найденные из этой таблицы пары соответственных значений x и $\overline{y}_x$ или y и $\overline{x}_y$, используются для отыскания параметров уравнений прямых регрессии

$$\overline{y}_x = ax + b \quad \text{и} \quad \overline{x}_y = cy + d\,.$$

Эта операция, называемая *выравниванием,* обычно выполняется по способу наименьших квадратов, сущность которого состоит в таком подборе параметров линии регрессии, при котором достигается минимум $\sum_i \left(\overline{y}_{x_i\text{п}} - \overline{y}_{x_i}\right)^2$.

Разберем применение данного способа в общем виде для каждого из записанных уравнений регрессии. При этом для иллюстрации используем данные корреляционной табл. 4 распределения растений житняка по общему весу и по весу семян.

2.2.1. Уравнение прямой регрессии y по x

При отыскании по способу наименьших квадратов параметров линейной функции $y=ax+b$ на основании данных наблюдения о парах значений x и y, связанных однозначным соответствием, используется система нормальных уравнений

$$\begin{cases} a\sum x^2 + b\sum x = \sum xy, \\ a\sum x + bn = \sum y. \end{cases}$$

Здесь коэффициенты определяются простым суммированием слагаемых в соответствии с количеством пар значений x и y.

Если же требуется с помощью способа наименьших квадратов определить параметры уравнения, связывающего значения x *с* соответственными частными средними $\overline{y}_x$, по данным не простой, а корреляционной таблицы, то структура коэффициентов и свободных членов нормальных уравнений должна отразить все данные корреляционной таблицы.

а) Коэффициенты, соответствующие суммам $\sum x$ и $\sum x^2$, должны включать в операцию суммирования все значения x как повторяющиеся, так и неповторяющиеся. Количество значений $x = x_1$ определяется

числом n_{x_1}, поэтому сумма этих значений x равна $n_{x_1}x_1$. Аналогично сумма значений $x = x_2$ равна $n_{x_2}x_2$ и т. д. Отсюда сумма всех значений x выразится в виде

$$n_{x_1}x_1 + n_{x_2}x_2 + \ldots + n_{x_k}x_k = \sum n_x x.$$

Суммирование квадратов переменной x строится также и дает

$$n_{x_1}x_1^2 + n_{x_2}x_2^2 + \ldots + n_{x_k}x_k = \sum n_x x^2.$$

б) Свободный член, соответствующий сумме $\sum y$, должен представить сумму всех частных средних $\overline{y}_x$. При этом для каждого значения x количество соответственных частных средних $\overline{y}_x$ определяется количеством таких значений самого x. Поэтому значению $x = x_1$ соответствует n_{x_1} частных средних $\overline{y}_{x_1}$, значению $x = x_2$ соответствует n_{x_2} частных средних $\overline{y}_{x_2}$ и т.д. Сумма всех частных средних $\overline{y}_x$ имеет вид

$$n_{x_1}\overline{y}_{x_1} + n_{x_2}\overline{y}_{x_2} + \ldots + n_{x_k}\overline{y}_{x_k} = \sum n_x \overline{y}_x .$$

в) Свободный член, соответствующий сумме $\sum xy$, должен представить сумму всех возможных произведений значений x на соответствующие частные средние $\overline{y}_x$. Количество разных произведений здесь определяется количеством соответственных значений x. Поэтому сумма всех произведений вида $x\overline{y}_x$ имеет вид

$$n_{x_1}x_1\overline{y}_{x_1} + n_{x_2}x_2\overline{y}_{x_2} + \ldots + n_{x_k}x_k\overline{y}_{x_k} = \sum n_x x\overline{y}_x .$$

Удовлетворяющая указанным требованиям система нормальных уравнений для отыскания значений параметров уравнения прямой регрессии $\overline{y}_x = ax + b$ имеет следующий вид:

$$\begin{cases} a\sum n_x x^2 + b\sum n_x x = \sum n_x x\overline{y}_x \\ a\sum n_x x + b\sum n_x = \sum n_x \overline{y}_x . \end{cases}$$

Определение корней этой системы предварительно требует некоторого преобразования коэффициентов и свободных членов.

Коэффициенты системы преобразуются так:

$$\sum n_x = n_{x_1} + n_{x_2} + \ldots + n_{x_k} = N;$$

$$\sum n_x x = N\overline{x}, \text{ так как } \overline{x} = \frac{\sum n_x x}{\sum n_x};$$

$$\sum n_x x^2 = N\overline{x^2}.$$

Развернутая запись свободного члена $\sum n_x \overline{y}_x$ позволяет для каждого слагаемого воспользоваться переходом от частных средних $\overline{y}_x$ к соответственным частным значениям y.

В самом деле, если $x = x_1$, то

$$\overline{y}_{x_1} = \frac{n_{1,1}y_1 + n_{1,2}y_2 + \ldots + n_{1,l}y_l}{n_{x_1}}.$$

Поэтому

$$n_{x_1}\overline{y}_{x_1} = n_{1,1}y_1 + n_{1,2}y_2 + \ldots + n_{1,l}y_l$$

и аналогично

$$n_{x_2}\overline{y}_{x_2} = n_{2,1}y_1 + n_{2,2}y_2 + \ldots + n_{2,l}y_l,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$n_{x_k}\overline{y}_{x_k} = n_{k,1}y_1 + n_{k,2}y_2 + \ldots + n_{k,l}y_l.$$

Почленное сложение всех равенств дает в соответствии с принятой структурой корреляционной табл. 2

$$\sum n_x \overline{y}_x = n_{y_1}y_1 + n_{y_2}y_2 + \ldots + n_{y_l}y_l = \sum n_y y.$$

После приведения этого результата к выражению, содержащему среднее значение y, получится

$$\sum n_y y = N\overline{y}, \text{ так как } \overline{y} = \frac{\sum n_y y}{\sum n_y}.$$

Преобразование свободного члена $\sum n_x x\overline{y}_x$ выполняется аналогично. Здесь при $x = x_1$ слагаемое $n_{x_1}x_1y_{x_1}$ приводится к виду

$$n_{1,1}x_1y_1 + n_{1,2}x_1y_2 + \ldots + n_{1,l}x_1y_l.$$

Последующая запись всех остальных слагаемых такого же вида при $x = x_2$, $x = x_3, \ldots, x = x_k$ и суммирование соответствующих выражений дает результат $\sum n_x x\overline{y}_x = \sum\sum n_{xy}xy$.

Сохраняя эту запись для выполнения подсчетов, можно привести полученный результат к выражению со средним значением xy.

Двойной знак суммирования позволяет выполнять суммирование в любом порядке: сначала по горизонтали (меняя нумерацию частных значений y), а затем по вертикали (меняя нумерацию частных значений x), или, наоборот, сначала по вертикали, а затем по горизонтали.

По структуре корреляционной таблицы:

$$\sum\sum n_{xy} = \sum_x\sum_y n_{xy} = \sum_x n_x = N,$$

или $$\sum\sum n_{xy} = \sum_y \sum_x n_{xy} = \sum_y n_y = N.$$

Отсюда $$\sum\sum n_{xy} xy = N\overline{xy},$$

так как $$\overline{xy} = \frac{\sum\sum n_{xy} xy}{\sum\sum n_{xy}}.$$

В преобразованном виде система такова:

$$\begin{cases} aN\overline{x^2} + bN\overline{x} = N\overline{xy}, \\ aN\overline{x} + bN = N\overline{y}, \end{cases}$$

или $$\begin{cases} a\overline{x^2} + b\overline{x} = \overline{xy}, \\ a\overline{x} + b = \overline{y}. \end{cases}$$

Для определения параметра a достаточно после умножения членов второго уравнения на $\overline{x}$ почленно вычесть это уравнение из первого:

$$a\left(\overline{x^2} - \overline{x}^2\right) = \overline{xy} - \overline{x} \cdot \overline{y}, \text{ или } a = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}.$$

Параметр b определяется непосредственно из второго уравнения:

$$b = \overline{y} - a\overline{x}.$$

Подставляя полученное выражение в уравнение прямой регрессии y по x, т.е. $\overline{y}_x = ax + b$, получим

$$\overline{y}_x = ax + \overline{y} - a\overline{x},$$

или $$\overline{y}_x - \overline{y} = a\left(x - \overline{x}\right).$$

Коэффициент a в уравнении прямой регрессии называется *коэффициентом прямой регрессии y по x* и обозначается символом $\rho_{y/x}$.

Таким образом, $$\rho_{y/x} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}$$

и окончательная запись уравнения прямой регрессии y по x таково:

$$\overline{y}_x - \overline{y} = \rho_{y/x}\left(x - \overline{x}\right).$$

Составим такое уравнение с числовыми параметрами для распределения растений житняка по данным корреляционной таблицы 2 об общем весе (x) и весе семян (y) растений. Вычисление необходимых параметров можно проводить по нижеследующей системе подсчетов, соответствующей выполненному общему решению.

1) Составляем вспомогательную таблицу.

2) По данным табл. 6

$$\overline{x}=\frac{\sum n_x x}{100}=64,\ \overline{y}=\frac{\sum n_x \overline{y}_x}{100}=35{,}15,\ \overline{x}^2=4096,$$

$$\overline{x^2}=4665,\ \overline{xy}=2249{,}60 \text{ и } \overline{xy}=2563{,}75.$$

Таблица 7

n_x	$n_x x$	$n_x \overline{y}_x$	$n_x x^2$	$n_x x \overline{y}_x$
5	$5\cdot 25$	75	$5\cdot 625$	1875
10	$10\cdot 35$	200	$10\cdot 1225$	7000
19	$19\cdot 45$	457	$19\cdot 2025$	20565
16	$16\cdot 55$	478	$16\cdot 3025$	26290
11	$11\cdot 65$	393	$11\cdot 4225$	25545
16	$16\cdot 75$	648	$16\cdot 5625$	48600
6	$6\cdot 85$	283	$6\cdot 7225$	24055
6	$6\cdot 95$	318	$6\cdot 9025$	30210
8	$8\cdot 105$	469	$8\cdot 11025$	49245
2	$2\cdot 115$	126	$2\cdot 13225$	14490
1	$1\cdot 125$	68	$1\cdot 15625$	8500
$N=100$	6400	3515	466500	256375

При вычислении значений $n_x \overline{y}_x$ для точности и удобства подсчетов исходим из того, что $n_x \overline{y}_x = n_{i,1} y_1 + n_{i,2} y_2 + ... + n_{i,l} y_l$.

3) Определяем коэффициент регрессии y по x:

$$\rho_{x/y}=\frac{256375-224960}{466500-409600}=\frac{31405}{56900}=0{,}552.$$

4) Записываем уравнение прямой регрессии y по x:

$$\overline{y}_x - 35{,}15 = 0{,}552\left(x-64\right),$$

или окончательно $$\overline{y}_x = 0{,}552x - 0{,}178.$$

2.2.2. Уравнение прямой регрессии x по y

Система нормальных уравнений для отыскания параметров c и d уравнения прямой регрессии x по y, получаемая в результате применения способа наименьших квадратов, имеет вид

$$\begin{cases} c\sum n_y y^2 + d\sum n_y y = \sum n_y y \overline{x}_y, \\ c\sum n_y y + d\sum n_y = \sum n_y \overline{x}_y. \end{cases}$$

По аналогии с преобразованиями, проведенными для случая регрессии y по x, можно записать, что

$$\sum n_y = N; \sum n_y y = N\overline{y}; \sum n_y y^2 = N\overline{y^2}, \sum n_y \overline{x}_y = \sum n_x x = N\overline{x};$$

$$\sum n_y \overline{y}\, x_y = \sum\sum n_{yx} yx = N\overline{yx} = N\overline{xy}\,.$$

Нормальные уравнения можно переписать в упрощенном виде:

$$\begin{cases} cN\overline{y^2} + dN\overline{y} = N\overline{xy}, \\ cN\overline{y} + dN = N\overline{x}, \end{cases}$$

или

$$\begin{cases} c\overline{y^2} + d\overline{y} = \overline{xy}, \\ c\overline{y} + d = \overline{x}. \end{cases}$$

Для определения параметра c из членов первого уравнения вычитаются члены второго уравнения, умноженные на $\overline{y}$:

$$c\left(\overline{y^2} - \overline{y}^2\right) = \overline{xy} - \overline{x}\cdot\overline{y}\,,$$

или

$$c = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{y^2} - \overline{y}^2}.$$

Параметр d определяется непосредственно из второго уравнения:

$$d = \overline{x} - c\overline{y}\,.$$

Замена d этим выражением в уравнении прямой регрессии $\overline{x}_y = cy + d$ дает

$$\overline{x}_y = cy + \overline{x} - c\overline{y},$$

или

$$\overline{x}_y - \overline{x} = c\left(y - \overline{y}\right).$$

Коэффициент c в этом уравнении называют *коэффициентом прямой регрессии x по y* и обозначают символом $\rho_{x/y}$.

Таким образом,

$$\rho_{x/y} = \frac{\overline{xy} - \overline{xy}}{\overline{y^2} - \overline{y}^2},$$

и окончательная запись уравнения прямой регрессии x по y такова:

$$\overline{x}_y - \overline{x} = \rho_{x/y}\left(y - \overline{y}\right).$$

Заметим, что обе прямые регрессии, как видно из их уравнений, проходят через точку $\left(\overline{x}; \overline{y}\right)$.

На примере распределения растений житняка по данным корреляционной таблицы о весе семян (y) и общем весе (x) растений составим уравнение прямой регрессии x по y с числовыми параметрами. Все необходимые вычисления для подсчета параметров проводятся в таком же порядке, как это выполнено для уравнения прямой регрессии y по x.

1) Составляем вспомогательную таблицу.

2) По данным табл. 8

$$\bar{x} = 64;\ \bar{y} = 35{,}15;\ \bar{y}^2 = 1235{,}5225;\ \overline{y^2} = 1424{,}65;$$

$$\bar{x} \cdot \bar{y} = 2249{,}60;\ \overline{xy} = 2563{,}75.$$

3) Определяем коэффициент регрессии x по y:

$$\rho_{x/y} = \frac{256375 - 224960}{142465 - 123552{,}25} = \frac{31415}{18912{,}75} \approx 1{,}661.$$

Таблица 8

n_y	$n_y y$	$n_y \bar{x}_y$	$n_y y^2$	$n_y y \bar{x}_y$
3	3·3	75	3·169	975
10	10·18	360	10·324	6480
20	20·23	900	20·529	20700
9	9·28	445	9·784	12460
14	14·33	850	14·1089	28050
11	11·38	765	11·1444	29070
9	9·43	665	9·1849	28595
8	8·48	670	8·2304	32160
6	6·53	590	6·2809	31270
6	6·58	630	6·3364	36540
1	1·63	105	1·3969	6615
3	3·68	345	3·4624	23460
N=100	3515	6400	142465	256375

При вычислении значений $n_x \bar{x}_y$ для точности и удобства подсчетов исходим из того, что $n_{y_i} \bar{x}_{y_i} = n_{1,i} x_1 + n_{2,i} x_2 + \ldots + n_{k,i} x_k$.

4) Записываем уравнение прямой регрессии x по y:

$$\bar{x}_y - 64 = 1{,}661(y - 35{,}15),$$

или окончательно $\bar{x}_y = 1{,}661y + 5{,}616.$

Ниже будет показано, что оба уравнения прямых регрессии могут быть получены одним расчетом с помощью коэффициента корреляции.

2.3. Коэффициент корреляции

В рассмотренном примере корреляционной связи оба коэффициента регрессии $a = \rho_{y/x}$ и $c = \rho_{x/y}$ положительны. В таком случае корреляцию называют положительной, что имеет место при изменении изучаемых количественных признаков в одинаковом направлении (x и y одновременно возрастают или одновременно убывают).

Прямые при положительных коэффициентах регрессии образуют острые углы с соответствующими осями координат (рис.8) — у прямой регрессии y по x коэффициент регрессии $\rho_{y/x} = \operatorname{tg}\alpha$, где α — острый угол, образованный прямой I с осью Ox, а у прямой регрессии x по y коэффициент регрессии $\rho_{y/x} = \operatorname{tg}\beta$, где β — острый угол, образованный прямой II с осью Oy. При отрицательных коэффициентах регрессии прямые регрессии образуют с соответствующими осями тупые углы.

Для большей наглядности на рис. 8 показано положение прямых регрессии относительно новой системы координат с началом в точке $O_1(\overline{x}; \overline{y})$ пересечения этих прямых.

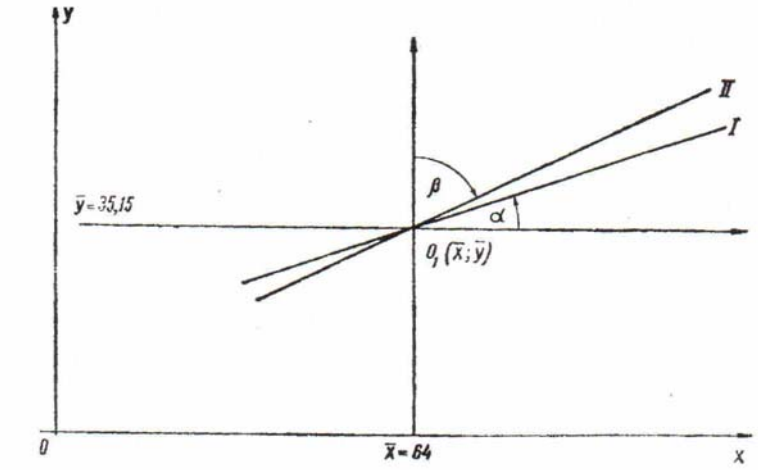

Рис. 8

Сами по себе значения коэффициентов регрессии не позволяют судить о тесноте связи между x и y. Это зависит от величины угла, образованного прямыми регрессии. Чем меньше этот угол, тем теснее корреляционная связь между x и y.

При слиянии этих двух прямых в одну имеет место линейная функциональная зависимость между x и y.

В качестве меры тесноты линейной корреляционной связи принимается коэффициент корреляции

$$r = \pm\sqrt{\rho_{y/x}\rho_{x/y}} = \pm\sqrt{\operatorname{tg}\alpha\operatorname{tg}\beta}$$

со знаком, совпадающим со знаками коэффициентов регрессии. При этом если прямые I и II совпадают, то $\beta = 90^\circ - \alpha$ и $\operatorname{tg}\beta = \operatorname{tg}\left(90^\circ - \alpha\right) =$

$= \operatorname{ctg}\alpha = \dfrac{1}{\operatorname{tg}\alpha}$. Но тогда $\operatorname{tg}\alpha\operatorname{tg}\beta = 1$ и, следовательно, $r = \pm 1$.

Обращение коэффициента корреляции в 1 или в -1 является, как это можно доказать, *необходимым и достаточным признаком линейной функциональной зависимости между* x и y.

Корреляционная таблица в таких случаях состоит из расположенных лишь на одной диагонали частот значений x и y.

Вместе с тем когда, по крайней мере, один из углов α или β равен нулю, то и $r = 0$, а значит, и между рассматриваемыми величинами не существует ни функциональной, ни корреляционной линейной зависимости. Однако в этом случае между x и y возможны нелинейные корреляционные и даже функциональные связи.

Корреляционная зависимость между x и y (для положительных коэффициентов регрессии) имеет место, когда коэффициент корреляции, как это можно доказать, выражается правильной дробью ($0 < r < 1$). При этом связь между переменными тем теснее, чем ближе коэффициент корреляции к единице.

Введенное определение коэффициента корреляции в виде $r = \pm\sqrt{\rho_{y/x}\rho_{x/y}}$ позволяет на основании выражений коэффициентов регрессии получить удобную формулу для непосредственного вычисления коэффициента корреляции.

Если обратиться к выражениям коэффициентов прямых регрессии

$$\rho_{y/x} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \quad \text{и} \quad \rho_{x/y} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{y^2} - \overline{y}^2},$$

то можно заметить, что знаменатели в обоих выражениях обозначают дисперсии соответствующих рядов распределений:

$$\overline{x^2} - \overline{x}^2 = \sigma_x^2 \quad \text{и} \quad \overline{y^2} - \overline{y}^2 = \sigma_y^2.$$

Отсюда можно получить для коэффициента корреляции формулу

$$r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_x \cdot \sigma_y},$$

которая сразу показывает, что между независимыми величинами корреляции не существует, так как для таких величин выполняется равенство $\overline{xy} = \overline{x} \cdot \overline{y}$.

Замечание. Последнее равенство является приближенным, а поэтому если коэффициент корреляции очень мал $(r < 0,4)$, считают, что линейной корреляции между x и y нет.

Записанная выше формула позволяет выразить каждый коэффициент регрессии через коэффициент корреляции.

Так, коэффициент регрессии y по x

$$\rho_{y/x} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} \cdot \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y} = r \frac{\sigma_y}{\sigma_x},$$

а коэффициент регрессии x по y

$$\rho_{x/y} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_y^2} = \frac{\sigma_x}{\sigma_y} \cdot \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y} = r \frac{\sigma_x}{\sigma_y}.$$

Такие выражения коэффициентов регрессии показывают, что составление уравнений прямых регрессии может быть облегчено, если будет найдено значение коэффициента корреляции. Для его вычисления следует использовать выражения числителя и знаменателя:

$$\overline{xy} = \overline{x} \cdot \overline{y} = \frac{\sum xy}{N} - \overline{xy}; \sigma_x = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{\frac{\sum x^2}{N} - \overline{x}^2}; \sigma_y = \sqrt{\frac{\sum y^2}{N} - \overline{y}^2}.$$

Тогда можно вычислить коэффициент корреляции по формуле

$$r = \frac{\frac{\sum xy}{N} - \overline{x} \cdot \overline{y}}{\sqrt{\frac{\sum x^2}{N} - x^{-2}} \cdot \sqrt{\frac{\sum y^2}{N} - \overline{y}^2}}.$$

Пример 4. В табл. 9 дана группировка 135 сахаропесочных заводов по размеру производственных основных средств в млн руб. (x) и по среднесуточной переработке свеклы в тыс. ц (y). Требуется определить коэффициент корреляции и составить уравнения регрессии.

Расположение рядов распределения значений y в табл. 9 позволяет наметить линейную корреляционную связь между x и y.

Для отыскания коэффициента корреляции составим вспомогательную табл. 10.

Таблица 9

x \ y	4	5	6	7	8	9	10	11	12	n_x
1,75	4	6	9	2						21
2,25	5	9	15	10	2	1				42
2,75		4	6	7	7		1			25
3,25	3	3	2	7	8	1				24
3,75			3	3	2	3	2			13
4,25				2	2	1	2	2	1	10
n_y	12	22	35	31	21	6	5	2	1	135

Таблица 10

n_x	$n_x x$	$n_x x^2$	$x n_x \overline{y}_x$	n_y	$n_y y$	$n_y y^2$	$y n_y \overline{x}_y$
21	21·1,75	21·3,0625	1,75·114	12	12·4	12·16	4·28,00
42	42·2,25	42·5,0625	2,25·250	22	22·5	22·25	4·51,50
25	25·2,75	25·7,5625	2,75·171	35	35·6	35·36	6·83,75
24	24·3,25	24·10,5625	3,25·161	31	31·7	31·49	7·87,75
13	13·3,75	13·14,0625	3,75·102	21	21·8	21·64	8·65,75
10	10·4,25	10·18,0625	4,25·93	6	6·9	6·81	9·21,00
				5	5·10	5·100	10·18,75
				2	2·11	2·121	11·8,50
				1	1·12	1·144	12·4,25
$N=135$	369,25	1082,9375	2533,25	135	891	6237	2533,25

Следует пояснить, что вторые множители в четвертом столбце $\left(n_x \overline{y}_x\right)$ получены из данных табл. 9 суммированием произведений каждого числа внутренней строки на соответствующее значение y (например, 114=4·4+6·5+9·6+2·7; 171= 4·5+6·6+7·7+8·7+1·10).

Так как суммирование этих вторых множителей дает сумму всех значений y, то сумма $\sum n_x \overline{y}_x = \sum n_y y$, и это равенство подтверждает правильность подсчета суммы всех значений y. Аналогична структура вторых множителей в последнем столбце.

Например, 28 = 4·1,75+5·2,25+3·3,25. Здесь суммирование вторых множителей дает сумму всех значений x, а потому *равенство* $\sum\limits_y n_x \overline{x}_y = \sum\limits_x n_x x$ служит для подтверждения правильности подсчета. Вместе с тем итоговые суммы по четвертому и последнему столбцам

являются в то же время суммами всех участвующих в таблице парных произведений xy. Отсюда

$$\sum_x x n_x \overline{y}_x = \sum_y y n_y \overline{x}_y = \sum xy .$$

По данным подсчетов имеем:

$$\overline{x} = \frac{\sum_x n_x x}{N} = \frac{369,25}{135} = 2,74; \overline{y} = \frac{891}{135} = 6,6; \overline{x}^2 = 7,5076; \overline{y}^2 = 43,56;$$

$$\overline{xy} = 18,084; \overline{xy} = \frac{2533,25}{135} = 18,764;$$

$$\overline{x^2} = \frac{1082,9375}{135} = 8,0218; \overline{y^2} = \frac{6237}{135} = 46,2;$$

$$\overline{x^2} - \overline{x}^2 = 8,0218 - 7,5076 = 0,5142; \overline{y^2} - \overline{y}^2 = 46,2 - 43,56 = 2,64.$$

Отсюда $\sigma_x = \sqrt{0,5142} = 0,717$ и $\sigma_y = \sqrt{2,64} = 1,625$ и коэффициент корреляции

$$r = \frac{18,764 - 18,084}{0,717 \cdot 1,625} = \frac{0,68}{1,1651} = 0,585.$$

Для составления уравнений прямых регрессии определяем коэффициенты регрессии:

$$\rho_{x/y} = 0,585 \cdot \frac{0,717}{1,625} = 0,585 \cdot 0,441 = 0,258;$$

$$\rho_{y/x} = \frac{0,585}{0,441} = 1,33.$$

Таким образом, уравнение прямой регрессии y по x

$$\overline{y}_x - 6,6 = 1,33(x - 2,74), \text{ или } \overline{y}_x = 1,33 + 2,96,$$

а уравнение прямой регрессии x по y

$$\overline{x}_y - 2,74 = 0,258(y - 6,6), \text{ или } \overline{x}_y = 0,258y + 1,05.$$

Сравнение коэффициента корреляции в этом примере $r = 0,585$, с коэффициентом корреляции в ранее рассмотренном примере распределения растений житняка

$$r = \sqrt{0,552 \cdot 1,661} = 0,957$$

показывает на большую тесноту связи между общим весом и весом семян. Это согласуется со структурой соответствующих корреляционных

таблиц. Табл. 4 распределения растений житняка характерна четким смещением рядов распределения значений y при малой степени рассеяния этих значений, а табл. 9 по сахаропесочным заводам дает малозаметное смещение рядов распределения значений y при значительной степени рассеяния этих значений.

2.4. Упрощенный способ вычисления коэффициента корреляции

Выше, в п.п. 2.2 и 2.3, при составлении уравнений прямых регрессии либо по данным корреляционной таблицы непосредственно вычислялись коэффициенты регрессии, либо по тем же данным предварительно вычислялся коэффициент корреляции. В обоих случаях вычисления были очень громоздкими (операции с многозначными числами).

Между тем при постоянных разностях для рассматриваемых в таблицах значений x и y (в табл. 4 $\Delta x = 10$ и $\Delta y = 5$, а в табл. 9 $\Delta x = 0{,}5$ и $\Delta y = 1$) можно заметно упростить вычисления, используя линейное преобразование переменных по формулам:

$$x = x_0 + u\Delta x \quad \text{и} \quad y = y_0 + v\Delta y,$$

где x_0 и y_0 — произвольно выбираемые значения из заданных значений переменных x и y;

u и v — новые переменные.

Так, для рассматриваемых значений x и y в табл. 4 можно провести преобразования

$$x = 25 + 10u \quad \text{и} \quad y = 13 + 5v,$$

при которых соответствие между значениями x и u, а также между y и v отражено в табл. 11а и 11б.

Если же применяются преобразования

$$x = 75 + 10u \quad \text{и} \quad y = 38 + 5v,$$

то получается другое соответствие (см. табл. 11в и 11г).

Преобразования второй серии обеспечивают большее упрощение вычислений, так как в этом случае все операции ведутся с меньшими по абсолютной величине числами.

Для обоснования этих линейных преобразований

$$x = x_0 + u\Delta x \quad \text{и} \quad y = y_0 + v\Delta y$$

можно показать, что операции над переменными x и y, связанные с вычислением коэффициента корреляции и коэффициентов регрессии,

сводятся при этих преобразованиях к аналогичным операциям над новыми переменными u и v.

Таблица 11

а

x	u
25	0
35	1
45	2
55	3
65	4
75	5
85	6
95	7
105	8
115	9
125	10

б

y	ν
13	0
18	1
23	2
28	3
33	4
38	5
43	6
48	7
53	8
58	9
63	10
68	11

в

x	u
25	-5
35	-4
45	-3
55	-2
65	-1
75	0
85	1
95	2
105	3
115	4
125	5

г

y	ν
13	-5
18	-4
23	-3
28	-2
33	-1
38	0
43	1
48	2
53	3
58	4
63	5
68	6

Прежде всего следует заметить, что средним значениям x и y соответствуют средние значения переменных u и v:

$$\overline{x} = \frac{\sum n_x x}{N} = \frac{\sum n_x \left(x_0 + u_x \Delta x\right)}{N} = \frac{N x_0 + \Delta x \sum n_x u_x}{N} = x_0 + \frac{N\overline{u}}{N}\Delta x = x_0 + \overline{u}\Delta x.$$

Отсюда, при зависимости $u = \dfrac{x - x_0}{\Delta x}$ будет и $\overline{u} = \dfrac{\overline{x} - x_0}{\Delta x}$.

Таким же образом можно установить, что

$$\overline{y} = y_0 + \overline{v}\Delta y, \text{ или } \overline{v} = \frac{\overline{y} - y_0}{\Delta y}.$$

Далее, разность $x - \overline{x} = \left(u - \overline{u}\right)\Delta x,$ а поэтому

$$\overline{x^2} - \overline{x}^2 = \frac{1}{N}\sum\left(x - \overline{x}\right)^2 = \frac{1}{N}\Delta x^2 \sum\left(u - \overline{u}\right)^2 =$$

$$= \frac{1}{N}\Delta x^2 \left[\sum u^2 - 2\overline{u}\sum u + N\overline{u}^2\right] = \Delta x^2\left(\overline{u^2} - \overline{u}^2\right).$$

Аналогично устанавливается, что $\overline{y^2} - \overline{y}^2 = \Delta y^2\left(\overline{v^2} - \overline{v}^2\right).$

Эти результаты показывают, что участвующие в вычислениях средние квадратические отклонения принимают вид $\sigma_x = \Delta x \sigma_u$ и $\sigma_y = \Delta y \sigma_v$.

Наконец, преобразование разности $\overline{xy} - \overline{x} \cdot \overline{y}$ дает

$$\overline{xy} - \overline{x}\,\overline{y} = \frac{\sum\sum n_{xy} xy}{N} - \overline{x}\,\overline{y} = \frac{\sum\sum n_{xy}\left(x_0 + u\Delta x\right)\left(y_0 + v\Delta y\right)}{N} -$$
$$-\left(x_0 + \overline{u}\Delta x\right)\left(y_0 + \overline{v}\Delta y\right) = \Delta x \Delta y\left(\overline{uv} - \overline{u}\,\overline{v}\right).$$

Таким образом, переход к новым переменным дает преобразованную форму коэффициента корреляции и коэффициентов регрессии:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y} = \frac{\overline{uv} - \overline{u} \cdot \overline{v}}{\sigma_u \sigma_v};$$

$$\rho_{y/x} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\Delta x \Delta y\left(\overline{uv} - \overline{u} \cdot \overline{v}\right)}{\Delta x^2\left(\overline{u^2} - \overline{u}^2\right)} = \frac{\Delta y}{\Delta x} \cdot \frac{\overline{uv} - \overline{u} \cdot \overline{v}}{\overline{u^2} - \overline{u}^2};$$

$$\rho_{x/y} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{y^2} - \overline{y}^2} = \frac{\Delta x}{\Delta y} \cdot \frac{\overline{uv} - \overline{u} \cdot \overline{v}}{\overline{v^2} - \overline{v}^2}.$$

Для составления уравнений регрессии с помощью новых переменных следует включать в корреляционную таблицу значения этих новых переменных, найденные по формулам:

$$u = \frac{x - x_0}{\Delta x} \quad \text{и} \quad v = \frac{y - y_0}{\Delta y}.$$

Удобней всего применять для этой цели исходную таблицу, помещая в ней значения u слева от соответственных значении x, а значения v — над соответственными значениями y. При этом вспомогательный характер значений u и v *в* таблице обычно оттеняется применением для них мелкого шрифта.

Для иллюстрации тех упрощений, которые достигаются введением новых переменных, используем этот способ на уже рассмотренном примере с распределением растений житняка. В виде значений x_0 и y_0 переменных x и y выгодней всего используются их средние или ближайшие к ним значения этих переменных. В примере с растениями житняка именно такую замену представляют данные второго преобразования. Поставленные во вторых столбцах табл. 11в и 11г числа получены таким образом: для значений переменной u преобразованием

$$\frac{x - 75}{10},$$

а для значений переменной v преобразованием

$$\frac{y - 38}{5}.$$

Вся операция по отысканию параметров уравнений регрессии проводится по отдельным этапам.

1) Корреляционная таблица 6 пополняется значениями u и v.

2) Для отыскания коэффициента корреляции составляется вспомогательная таблица (см. табл. 13) с вычислением ее итоговых данных.

Таблица 12

	v	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	
u	y / x	13	18	23	28	33	38	43	48	53	58	63	68	n_x
-5	25	3	2											5
-4	35		6	4										10
-3	45		1	12	5									19
-2	55		1	2	4	8	1							16
-1	65			1		4	4	2						11
0	75					2	6	6	2					16
1	85							1	5					6
2	95								1	4	1			6
3	105									2	4	1	1	8
4	115										1		1	2
5	125												1	1
	n_y	3	10	20	9	14	11	9	8	6	6	1	3	100

3) По данным подсчетов:

$$\overline{u} = -1{,}10;\ \overline{v} = -0{,}57;\ \overline{u^2} = 6{,}90;\ \overline{v^2} = 7{,}89;\ \overline{u}^2 = 1{,}21.$$

$$\overline{v}^2 = 0{,}3249;\ \overline{uv} = 6{,}91;\ \overline{u}\,\overline{v} = 0{,}627.$$

Следует заметить, что $\Delta x = 10;\ \Delta y = 5;\ x_0 = 75;\ y_0 = 38$, а также что формулы преобразования $x = x_0 + u\Delta x$ и $y = y_0 + v\Delta y$ позволяют: по найденным средним значениям новых переменных

$$\overline{u} = -1{,}1 \text{ и } \overline{v} = -0{,}57$$

сразу получить средние значения старых переменных:

$$\overline{x} = x_0 + \overline{u}\Delta x = 75 - 1{,}1 \cdot 10 = 75 - 11 = 64;$$

$$\overline{y} = y_0 + \overline{v}\Delta y = 38 - 0{,}57 \cdot 5 = 38 - 2{,}85 = 35{,}15.$$

Таблица 13

n_x	u	$n_x u$	$n_x u^2$	$u n_x \bar{\nu}_x$	n_y	ν	$n_y \nu$	$n_y \nu^2$
5	-5	-25	125	$-5\cdot(-23)=115$	3	-5	-15	75
10	-4	-40	160	$-4\cdot(-36)=144$	10	-4	-40	160
19	-3	-57	171	$-3\cdot(-53)=159$	20	-3	-60	180
16	-2	-32	64	$-2\cdot(-26)=52$	9	-2	-18	36
11	-1	-11	11	$-1\cdot(-5)=5$	14	0	-14	14
16	0	0	0	0	11	1	0	0
6	1	6	6	$1\cdot 11=11$	9	2	9	9
6	2	12	24	$2\cdot 18=36$	8	3	16	32
8	3	24	72	$3\cdot 33=99$	6	4	18	54
2	4	8	32	$4\cdot 10=40$	6	5	24	96
1	5	5	25	$5\cdot 6=30$	1	6	5	25
					3		18	108
$N=$ $=10$		$N\bar{u}=$ $=-110$	$N\overline{u^2}=$ $=690$	$N\overline{u\nu}=100$	$N=100$		$N\bar{\nu}=$ $=-57$	$N\overline{\nu^2}=$ $=789$

Совпадение с данными о значениях $\bar{x}$ и $\bar{y}$, найденных непосредственным вычислением, подтверждает правильность проведения упрощенных вычислений.

4) Определив значения трех разностей:

$$\overline{uv}-\overline{u}\,\overline{v}=6{,}91-0{,}627=6{,}283;$$

$$\overline{u^2}-\overline{u}^2=6{,}90-1{,}21=5{,}69;$$

$$\overline{v^2}-\overline{v}^2=7{,}89-0{,}3249=7{,}5651,$$

можно записать, что $\sigma_u=\sqrt{5{,}69}=2{,}385$ и $\sigma_v=\sqrt{7{,}5651}\approx 2{,}75$.

Отсюда определяется коэффициент корреляции

$$r=\frac{\overline{uv}-\overline{u}\cdot\overline{v}}{\sigma_u\cdot\sigma_v}=\frac{6{,}283}{2{,}385\cdot 2{,}75}=\frac{6{,}283}{6{,}559}\approx 0{,}958.$$

Коэффициент регрессии y по x

$$\rho_{y/x}=\frac{\sigma_y}{\sigma_x}r=\frac{\Delta y}{\Delta x}\cdot\frac{\sigma_v}{\sigma_u}r=\frac{5\cdot 2{,}75}{10\cdot 2{,}385}\cdot 0{,}958\approx 0{,}551.$$

Коэффициент регрессии x по y

$$\rho_{x/y}=\frac{\sigma_y}{\sigma_x}r=\frac{4{,}87}{2{,}74}\cdot 0{,}958\approx 1{,}663.$$

Расхождения полученных коэффициентов с результатами непосредственных вычислений относятся к третьим десятичным знакам, что связано с приближенным характером вычислений.

2.5. Простейшие случаи криволинейной корреляции

В некоторых случаях ломаная, соединяющая точки, соответствующие парам значений x и $\overline{y}_x$ располагается вблизи кривой. Ограничимся рассмотрением корреляционной связи $\overline{y}_x = f(x)$ для двух простейших кривых: параболы, соответствующей трехчлену $f(x) = ax^2 + bx + c$, и гиперболы, определяемой уравнением $f(x) = a + \frac{b}{x}$.

1) Отыскание параметров квадратного трехчлена по способу наименьших квадратов с использованием данных простой таблицы значений x и y подробно разобрано выше.

Если же значения x и y представлены данными корреляционной таблицы, то корреляционная связь отыскивается как уравнение регрессии $\overline{y}_x = ax^2 + bx + c$, причем параметры этого уравнения определяются из системы нормальных уравнений, отражающих в структуре своих коэффициентов и свободных членов все данные корреляционной таблицы:

$$\begin{cases} a\sum n_x x^4 + b\sum n_x x^3 + c\sum n_x x^2 = \sum x^2 n_x \overline{y}_x, \\ a\sum n_x x^3 + b\sum n_x x^2 + c\sum n_x x = \sum x n_x \overline{y}_x, \\ a\sum n_x x^2 + b\sum n_x x + cN = \sum n_x \overline{y}_x. \end{cases}$$

Заметим, что к выравниванию с помощью параболы второго порядка можно обращаться в тех случаях, когда использование линейной корреляции обнаруживает малую тесноту связи (значения коэффициента корреляции в границах 0,4–0,6).

В качестве примера применения способа наименьших квадратов для отыскания зависимости между y и x в форме уравнения параболы второго порядка используем уже рассмотренные выше данные табл. 8 группировки 135 сахаропесочных заводов по размеру основных производственных средств в млн руб. (x) и по среднесуточной переработке свеклы в тыс. ц (y). Использование этих данных для установления параболической корреляционной зависимости целесообразно в связи с отмеченной выше малой теснотой линейной связи между рассматриваемыми показателями.

Для составления системы нормальных уравнений необходимые данные получаются суммированием, выполненным по схеме вспомогательной таблицы.

По итоговым данным табл. 14 можно записать систему нормальных уравнений:

$$\begin{cases} 11204{,}17a + 3388{,}04b + 1082{,}94c = 7596{,}14, \\ 3388a + 1082b + 369{,}25c = 2533{,}25, \\ 1082{,}94a + 369{,}25b + 135c = 891. \end{cases}$$

Решение этой системы дает параметры:

$$a \approx -0{,}0216; \quad b \approx 0{,}677; \quad c \approx 5{,}07.$$

Таблица 14

x	n_x	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \overline{y}_x$	$n_x x \overline{y}_x$	$n_x x^2 \overline{y}_x$
1,75	21	36,75	64,31	112,55	196,97	114	199,50	349,14
2,25	42	94,50	212,63	478,38	1076,36	250	562,50	1139,06
2,75	25	68,75	189,06	520,00	1430,00	171	470,25	1293,19
3,25	24	78,00	253,50	823,92	2677,84	161	523,25	1700,56
3,75	13	48,75	182,81	685,49	2570,28	102	382,50	1434,38
4,25	10	42,50	180,63	767,70	3252,72	93	395,25	1679,81
	$N = 135$	369,25	1082,94	3388,04	11204,17	891	2533,25	7596,14

2) Рассмотрим корреляционную зависимость гиперболического типа, определяемую уравнением $\overline{y}_x = a + \dfrac{b}{x}$.

Пример 5. В табл. 15 дана группировка 44 предприятий по выпуску продукции в тыс. ед. (x) и средней себестоимости единицы в руб. (y). Составить корреляционное уравнение связи этими показателями.

Таблица 15

x	до 1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
y	16,50	13,75	13,31	12,50	13,52	12,75	12,30	12,83	12,28	12,34
Число предприятий	6	6	8	7	4	4	3	2	2	2

Ломаная, отображающая данные этой таблицы (рис. 9), позволяет обратиться к уравнению гиперболы.

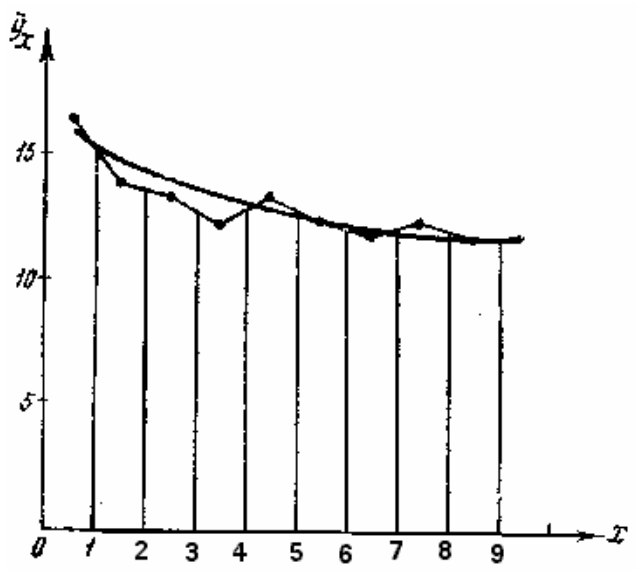

Рис. 9

Применим способ наименьших квадратов для определения пара метров искомого уравнения в виде $\bar{y}_x = a + \frac{b}{x}$.

Для функции $S = \sum\left(\bar{y}_x - a - \frac{b}{x}\right)^2$ необходимые условия минимума $\frac{\partial S}{\partial a} = 0$ и $\frac{\partial S}{\partial b} = 0$ приводят к системе

$$\begin{cases} \sum n_x \bar{y}_x = aN + b\sum n_x \cdot \frac{1}{x}, \\ \sum n_x \cdot \frac{\bar{y}_x}{x} = a\sum \frac{n_x}{x} + b\sum \frac{n_x}{x^2} \end{cases}$$

Суммирование выполняется на вспомогательной таблице (табл. 16).

Система нормальных уравнений

$$\begin{cases} 44a + 24b = 586,4, \\ 24a + 29b = 353,3 \end{cases}$$

определяет параметры:

$$a = \frac{8526,4}{700} = 12,8 \text{ и } b = \frac{1471,6}{700} = 2,1.$$

Отсюда искомое уравнение регрессии запишется так:

$$\bar{y}_x = 12,8 + \frac{2,1}{x}.$$

Соответствующая этому уравнению линия регрессии изображена вместе с ломаной на рис. 9.

Таблица 16

x	$\bar{y}_x$	n_x	$\frac{n_x}{x}$	$\frac{n_x}{x^2}$	$n_x\bar{y}_x$	$n_x\frac{\bar{y}_x}{x}$
0,5	16,50	6	12,000	24,000	99,00	198,000
1,5	13,75	6	4,000	2,664	82,50	55,000
2,5	13,31	8	3,200	1,280	100,48	40,192
3,5	12,50	7	2,000	0,574	87,50	24,100
4,5	13,52	4	0,888	0,196	54,08	12,178
5,5	12,75	4	0,728	0,132	51,00	9,273
6,5	12,30	3	0,452	0,072	36,90	5,677
7,5	12,83	2	0,266	0,036	25,66	3,421
8,5	12,28	2	0,235	0,028	24,56	2,889
9,5	12,34	2	0,210	0,022	25,68	2,591
		$N=44$	23,990	29,004	586,36	353,301

Для измерения тесноты связи при линейной корреляции введен *коэффициент корреляции*. Общим измерителем тесноты связи для всех случаев корреляции как линейной, так и криволинейной служат *корреляционные отношения* $\left(\eta_{y/x}$ и $\eta_{x/y}\right)$.

Рассмотрим корреляционное отношение для корреляционной зависимости, выражаемой уравнением $\bar{y}_x = f(x)$, которое устанавливает связь между частными средними $\bar{y}_x$, и соответственными значениями x.

В этом случае корреляционное отношение (его символ η) определяется формулой $\eta = \dfrac{\sigma(\bar{y}_x)}{\sigma_y}$, которая выражает *отношение среднего квадратического отклонения частных средних* $\bar{y}_x$ *от общей средней* y *к среднему квадратическому отклонению значений* y *от общей средней* $\bar{y}$.

Аналогично вводится понятие о корреляционном отношении $\eta_{x/y}$ с соответствующей формулой

$$\eta_{x/y} = \frac{\sigma(\bar{x}_y)}{\sigma_x}.$$

Здесь $\sigma(\bar{x}_y)$ означает среднее квадратическое отклонение частных средних $\bar{x}_y$ от общей средней $\bar{x}$, а σ_x — среднее квадратическое отклонение значений x от $\bar{x}$.

Это отношение в случаях линейной корреляции оказывается не меньше коэффициента корреляции, в чем можно убедиться на данных примера.

Ранее был найден коэффициент корреляции $r = 0{,}585$ по формуле

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y}.$$

При определении корреляционного отношения из имеющихся данных для вычисления r используется значение $\sigma_y = 1{,}625$. Следует определить еще $\sigma(\overline{y}_x) = \sqrt{\sigma^2(\overline{y}_x)}$. Но величина $\sigma^2(\overline{y}_x)$, выражающая дисперсию частных средних значений $\overline{y}_x$, определяется в виде:

$$\sigma^2(\overline{y}_x) = \frac{1}{N}\sum n_x \overline{y}_x^2 - \overline{y}^2.$$

Здесь

$$\sum n_x \overline{y}_x^2 = 21\left(\frac{114}{21}\right)^2 + 42\left(\frac{250}{42}\right)^2 + 25\left(\frac{171}{25}\right)^2 + 24\left(\frac{161}{24}\right)^2 + 13\left(\frac{102}{13}\right)^2 + 10\left(\frac{93}{10}\right)^2 \approx$$
$$\approx 6021{,}74$$

$$\overline{y_x^2} \approx \frac{6021{,}74}{135} \approx 44{,}61.$$

Отсюда $\sigma^2(\overline{y}_x) = 44{,}61 - 43{,}56 = 1{,}05$

и $\sigma(\overline{y}_x) = \sqrt{1{,}05} \approx 1{,}025.$

Таким образом, $\eta = \frac{1{,}025}{1{,}625} \approx 0{,}631.$

Ограничимся этим примером, опуская общий вывод о том, что в случаях линейной корреляции значение корреляционного отношения оказывается не меньше значения коэффициента корреляции, т. е. что $\eta \geq r$.

При этом знак равенства возможен только в случаях точной корреляционной связи:

$\eta_{y/x} = |r|$ при точной линейной корреляционной связи y по x;

$\eta_{x/y} = |r|$ при точной линейной корреляционной связи x по y;

$\eta_{y/x} = \eta_{x/y} = |r|$ при точной линейной корреляционной связи y по x, и x по y.

2.6. Понятие о множественной корреляции

Этот вид корреляционной зависимости возникает в тех случаях, когда рассматривается связь между тремя или большим числом признаков, характеризующих изучаемое явление.

Ограничиваясь линейной корреляционной связью между величиной z и аргументами x и y, общий вид которой

$$z \approx Ax + By\,,$$

заметим, что эту связь выгодней рассматривать в форме зависимости между отклонениями величин x, y и z от их средних x, y и z. Этим требуемая корреляционная зависимость приводится к виду

$$z - \overline{z} \approx A\left(x - \overline{x}\right) + B\left(y - \overline{y}\right).$$

Коэффициенты этого уравнения A и B выражают коэффициенты регрессии, которые определяются по формулам:

$$A = \frac{r_{xz} - r_{xy}r_{yz}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_x} \quad \text{и} \quad B = \frac{r_{yz} - r_{xy}r_{xz}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_y}.$$

Компонентами этих коэффициентов служат коэффициенты корреляции между x и y $\left(r_{xy}\right)$, x и z $\left(r_{xz}\right)$ и y и z $\left(r_{yz}\right)$, а также соответственные соотношения между средними квадратическими отклонениями величины z и каждого аргумента (x и y).

Такая структура коэффициентов регрессии A и B показывает, что для составления линейного корреляционного уравнения между тремя величинами требуется предварительное вычисление трех коэффициентов корреляции — между аргументами x и y, а также между каждым аргументом и величиной z. Эти же коэффициенты корреляции используются в выражении сводного коэффициента корреляции, определяющего тесноту корреляционной связи между тремя величинами x, y и z:

$$R = \sqrt{\frac{r_{xz}^2 - 2r_{xy}r_{xz}r_{yz} + r_{yz}^2}{1 - r_{xy}^2}}.$$

Этот коэффициент принимает значения $0 \le R \le 1$. При R=0 линейная связь между x, y, и z отсутствует, а при R=1 между ними имеется точная линейная связь $z = ax + by$.

3. СТАТИСТИЧЕСКИЙ АНАЛИЗ СЛУЧАЙНЫХ ВЕЛИЧИН

3.1. Формирование исходных данных к задачам

Для того чтобы получить свои личные числовые данные, необходимо взять свой номер по списку группы (A – предпоследняя цифра, B – последняя) и выбрать из таблицы 17 параметр m, а из таблицы 18 параметр n. Эти два числа нужно подставить в условия задач.

Таблица 17

A	0	1	2	3	4	5	6	7	8	9
m	4	3	5	1	3	2	4	2	1	5

Таблица 18

B	0	1	2	3	4	5	6	7	8	9
m	3	2	1	4	5	3	1	5	2	4

3.2. Численная обработка данных одномерной выборки

Выборка X объемом N=100 измерений задана табл. 19, где x_i – результаты измерений, m_{x_i} – частоты, с которыми встречаются значения x_i, $\sum_{i=1}^{7} m_{x_i} = 100$, $x_i = 0{,}2 \cdot m + (i-1) \cdot 0{,}3 \cdot n$.

Таблица 19

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
m_{x_i}	5	13	$30-(m+n)$	$30-(m+n)$	19	10	3

1. Построить полигон относительных частот $W_i = \dfrac{m_{x_i}}{N}$.

2. Вычислить среднее выборочное $\overline{X}$, выборочную дисперсию D_x и среднее квадратическое отклонение σ_x.

3. По критерию χ^2 проверить гипотезу о нормальном распределении генеральной совокупности при уровне значимости $\alpha = 0{,}05$.

Примечание. Для расчетов $\overline{X}$ и D_x рекомендуется перейти к условным значениям $u_i = \dfrac{x_i - C_x}{0{,}3 \cdot n}$ и, взяв за ложный нуль C_x значение с наибольшей частотой, использовать суммы $\sum_{i=1}^{7} m_{x_i} \cdot u_i$ и $\sum_{i=1}^{7} m_{x_i} \cdot u_i^2$.

3.3. Построение уравнения прямой регрессии

Двумерная выборка результатов совместных измерений признаков x и y объемом N=100 измерений задана корреляционной табл. 20, где $x_i = 0{,}2 \cdot m + (i-1) \cdot 0{,}3 \cdot n$, $y_j = 0{,}5 \cdot m + (j-1) \cdot 0{,}2 \cdot n$.

Таблица 20

	y_1	y_2	y_3	y_4	y_5	m_{x_i}
x_1	2	3	-	-	-	5
x_2	3	8	2	-	-	13
x_3	-	$12+n$	$12+n$	-	-	$30-(m+n)$
x_4	-	-	$16-m$		-	$30-(m+n)$
x_5	-	-	9	10	-	19
x_6	-	-	3	6	1	10
x_7	-	-	-	1	2	3
m_{y_j}	5	$19+m$	$42+n-m$	$31-n$	3	N=100

1. Найти $\overline{Y}$ и σ_y для выборки (см. табл. 21).

Таблица 21

y_j	y_1	y_2	y_3	y_4	y_5
m_{y_j}	5	$19+m$	$42+n-m$	$31-n$	3

(Расчеты $\overline{Y}$ и σ_y можно провести аналогично расчетам $\overline{X}$ и σ_x в задании 3.2.)

2. Построить уравнение прямой регрессии Y на X в виде $\overline{y}_x = ax + b$, $\overline{X}$ и σ_x следует взять из задачи 3.2.

3. На графике изобразить корреляционное поле, то есть нанести точки (x_i, y_j) и построить прямую $\overline{y}_x = ax + b$.

Примечание. Уравнение регрессии сначала рекомендуется найти в виде $\dfrac{\overline{y}_x - \overline{Y}}{\sigma_y} = r \cdot \dfrac{x - \overline{X}}{\sigma_x}$, где r – выборочный коэффициент корреляции, для расчета которого можно воспользоваться методом четырех полей.

3.4. Практические рекомендации по выполнению индивидуальных заданий

3.4.1. Численная обработка данных одномерной выборки

Выборка X объемом 100 измерений задана табл. 22, где x_i – результаты измерений, m_{x_i} – частоты с которыми встречаются значения x_i.

Таблица 22

i	1	2	3	4	5	6	7
x_i	0,2	1,4	2,6	3,8	5	6,2	7,4
m_{x_i}	5	13	25	25	19	10	3

1. Построить полигон относительных частот $W_i = \dfrac{m_{x_i}}{N}$.

Решение. Вычисляя относительные частоты $W_i = \dfrac{m_{x_i}}{N} = \dfrac{m_i}{100}$, получаем:

Таблица 23

i	1	2	3	4	5	6	7
x_i	0,2	1,4	2,6	3,8	5	6,2	7,4
m_{x_i}	5	13	25	25	19	10	3
W_i	0,05	0,13	0,25	0,25	0,19	0,1	0,03

Построим полигон относительных частот

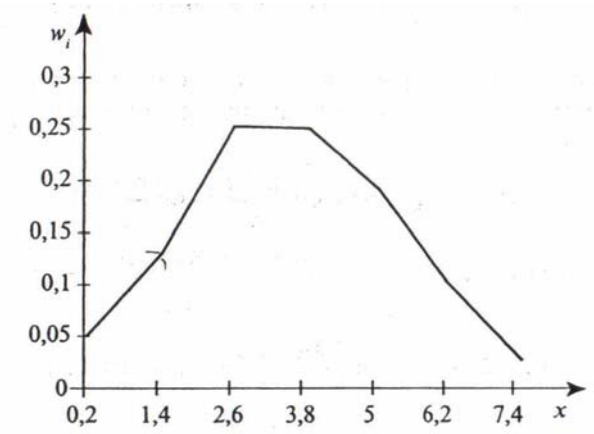

Рис. 10

2. Вычислить среднее выборочное $\overline{X}$, выборочную дисперсию D_x и среднее квадратическое отклонение σ_x.

Решение. Для вычисления $\overline{X}$, D_x и σ_x воспользуемся методом произведений. Введем условные варианты

$$u_i = \frac{x_i - C_x}{h_x},$$

где C_x – значение x_i, которому соответствует наибольшая частота, $C_x = 3{,}8$;

h_x – шаг выборки, $h_x = 1{,}2$.

Проверка:

$$\sum m_i (u_i + 1)^2 = \sum m_i u_i^2 + 2\sum m_i u_i + \sum m_i,$$

$$272 = 208 + 2\cdot(-18) + 100,\ 272{=}272.$$

Найдем теперь условные характеристики:

$$\overline{U} = \frac{\sum m_i u_i}{N} = \frac{-18}{100} = -0{,}18;$$

$$D_U = \frac{\sum m_i u_i^2}{N} - (\overline{U})^2 = \frac{208}{100} - (-0{,}18)^2 = 2{,}048;$$

$$\sigma_U = \sqrt{D_U} = \sqrt{2{,}048} = 1{,}43.$$

Таблица 24

i	u_i	m_i	$m_i u_i$	$m_i u_i^2$	$m_i (u_i + 1)^2$
1	-3	5	-15	45	20
2	-2	13	-26	52	13
3	-1	25	-25	25	0
4	0	25	0	0	25
5	1	19	19	19	76
6	2	10	20	40	90
7	3	3	9	27	48
Σ		100	-18	208	272

Возвращаясь к исходному вариационному ряду, с помощью равенств $x_i = h_x u_i + C_x$ получаем:

$$\overline{X} = h_x \overline{U} + C_x = 1{,}2\cdot(-0{,}18) + 3{,}8 = 3{,}58;$$

$$D_x = h_x^2 D_u = (1{,}2)^2 \cdot 2{,}048 = 2{,}949;$$

$$\sigma_x = h_x \sigma_u = 1{,}2 \cdot 1{,}43 = 1{,}72.$$

3. По критерию χ^2 проверить гипотезу о нормальном распределении генеральной совокупности при уровне значимости $\alpha = 0{,}05$.

Решение. Проверим гипотезу о нормальном распределении генеральной совокупности, используя критерий χ^2 (Пирсона) при $\alpha = 0{,}05$.

В основе критерия лежит сравнение частот m_i и теоретических частот m_i^T, вычисленных в предположении нормального распределения генеральной совокупности. Критерий Пирсона не подтверждает однозначно правильность или неправильность гипотезы, а только устанавливает её согласие или несогласие с данными выборки при данном уровне значимости. В качестве критерия выбирается величина

$$\chi^2 = \sum_{i=1}^{7} \frac{\left(m_i - m_i^T\right)^2}{m_i^T}.$$

Её значение сравнивают с критическим значением $\chi^2_{\text{кр}}$, определяемым по соответствующей таблице значений при заданном уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = p - 1 - r$, где p – число интервалов, r – число параметров нормального закона распределения. В данном случае $p = 7;\ r = 2;\ k = 4$.

По таблице распределения χ^2 с $k = 7 - 2 - 1 = 4$ степенями свободы при уровне значимости $\alpha = 0{,}05$ находим $\chi^2_{\text{кр}} = 9{,}49$.

Если в результате вычислений выполняется неравенство $\chi^2 < \chi^2_{\text{кр}}$, то гипотеза принимается при данном уровне значимости. Если же $\chi^2 > \chi^2_{\text{кр}}$, то гипотезу отвергают. Применим критерий Пирсона к данной выборке. Для этого составим расчетную таблицу, находя теоретические частоты m_i^T для нормального распределения по формуле

$$m_i^T = \frac{N h_x}{\sigma_x} \cdot \varphi\left(\frac{x_i - \overline{X}}{\sigma_x}\right),$$

где $\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$.

Складывая числа последнего столбца таблицы, получаем $\chi^2 = 1{,}65$.

Так как $\chi^2 < \chi^2_{\text{кр}}$ $(1{,}65 < 9{,}49)$, то гипотеза о нормальном распределении генеральной совокупности принимается.

Таким образом, с уровнем значимости $\alpha = 0{,}05$ можно считать, что генеральная совокупность распределена по нормальному закону с параметрами $a = \overline{X} = 3{,}58$, $\sigma_x = 1{,}72$.

Таблица 25

i	x_i	$z_i = \dfrac{x_i - \overline{X}}{\sigma_x}$	$\varphi(z_i)$	m_i^T	m_i	$m_i - m_i^T$	$\dfrac{\left(m_i - m_i^T\right)^2}{m_i^T}$
1	0,2	–1,97	0,06	4,42	5	0,58	0,076
2	1,4	–1,27	0,20	13,73	13	–5,58	0,038
3	2,6	–0,57	0,37	26,15	25	–1,15	0,051
4	3,8	0,13	0,44	30,58	25	–5,58	1,017
5	5	0,82	0,31	21,94	19	–2,94	0,393
6	6,2	1,52	0,14	9,66	10	0,34	0,012
7	7,4	2,22	0,04	2,61	3	0,39	0,059
Σ							1,65

Ответ. $\overline{X} = 3{,}58$, $D_x = 2{,}949$, $\sigma_x = 1{,}72$. Гипотеза о нормальном распределении генеральной совокупности принимается.

3.4.2. Построение уравнения прямой регрессии

Двумерная выборка результатов совместных измерений признаков x и y объемом N=100 измерений задана корреляционной табл. 26.

Таблица 26

i	j	1	2	3	4	5	m_{x_i}
	y_j / x_i	0,5	1,3	2,1	2,9	3,7	
1	0,2	2	3				5
2	1,4	3	8	2			13
3	2,6		9	16			25
4	3,8			15	10		25
5	5			9	10		19
6	6,2			3	6	1	10
7	7,4				1	2	3
m_{y_j}		5	20	45	27	3	N=100

1. Найти $\overline{Y}$ и σ_y для выборки (см. табл. 27).

Таблица 27

y_j	0,5	1,3	2,1	2,9	3,7
m_{y_j}	5	20	45	27	3

Решение. Для вычисления $\overline{Y}$, D_y и σ_y воспользуемся методом произведений. Введем условные варианты $\nu_j = \dfrac{y_j - C_y}{h_y}$, где C_y – значение y_j, которому соответствует наибольшая частота, $C_y = 2{,}1$, h_y – шаг выборки, $h_y = 0{,}8$.

Тогда, вычисляя ν_j, получим условный ряд:

Таблица 28

ν_j	–2	–1	0	1	2
m_j	5	20	45	27	3

Для этого ряда составим расчетную табл. 29.

Таблица 29

j	ν_j	m_j	$m_j\nu_j$	$m_j\nu_j^2$	$m_j(\nu_j+1)^2$
1	–2	5	–20	20	5
2	–1	20	–20	20	0
3	0	45	0	0	45
4	1	27	27	27	108
5	2	3	6	12	27
Σ		100	3	79	185

Проверка:

$$\sum m_j(\nu_j+1)^2 = \sum m_j\nu_j^2 + 2\sum m_j\nu_j + \sum m_j,$$

$$185 = 79 + 2\cdot 3 + 100,\ 185 = 185.$$

Условные характеристики:

$$\overline{V} = \frac{\sum m_j\nu_j}{N} = \frac{3}{100} = 0{,}03;$$

$$D_\nu = \frac{\sum m_j\nu_j^2}{N} - \left(\overline{V}\right)^2 = \frac{79}{100} - (0{,}03)^2 = 0{,}789;$$

$$\sigma_\nu = \sqrt{D_\nu} = \sqrt{0{,}789} = 0{,}89.$$

Возвращаясь к исходному вариационному ряду, с помощью равенств $y_j = h_y \nu_j + C_y$ получаем:

$$\overline{Y} = h_y \overline{V} + C_y = 0{,}8 \cdot 0{,}03 + 2{,}1 = 2{,}12;$$

$$D_y = h_y^2 D_\nu = (0{,}8)^2 \cdot 0{,}789 = 0{,}505;$$

$$\sigma_y = h_y \sigma_\nu = 0{,}8 \cdot 0{,}89 = 0{,}71.$$

2. Построить уравнение прямой регрессии Y на X в виде $\overline{y}_x = a x + b$. Уравнение прямой регрессии Y на X имеет вид:

$$\overline{y}_x - \overline{Y} = \frac{r \sigma_y}{\sigma_x} \cdot \left(x - \overline{X} \right).$$

Значения x_i и частоты их появления m_{x_i} совпадают с данными для задачи 4.1.

Следовательно,

$$\overline{X} = 3{,}584, \ \ \sigma_x = 1{,}72.$$

Значения $\overline{Y}$ и σ_y найдены в задаче: $\overline{Y} = 2{,}12$, $\sigma_y = 0{,}71$.

Коэффициент корреляции определяется по формуле

$$r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y},$$

где $\overline{XY} = \frac{1}{N} \sum_{i=1}^{7} \sum_{j=1}^{5} m_{ij} x_i y_j$.

Для нахождения $\overline{XY}$ воспользуемся корреляционной табл. 30.

Таблица 30

i	j	1	2	3	4	5	m_{x_i}	$\sum_{i=1}^{7} \sum_{j=1}^{5} m_{ij} x_i y_j$
	y_j / x_i	0,5	1,3	2,1	2,9	3,7		
1	0,2	2	3				5	0,98
2	1,4	3	8	2			13	22,54
3	2,6		9	16			25	117,78
4	3,8		15	10			25	229,9
5	5			9	10		19	239,5
6	6,2			3	6	1	10	169,88
7	7,4				1	2	3	46,22
m_{y_j}		5	20	45	27	3	N=100	$\sum$ = 856,8

Как следует из таблицы, $\overline{XY} = 8{,}568$

Таким образом,

$$r = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sigma_x \sigma_y} = \frac{8{,}568 - 3{,}584 \cdot 2{,}12}{1{,}72 \cdot 0{,}71} = 0{,}78.$$

Уравнение прямой регрессии Y на X имеет вид:

$$\overline{y}_x = \overline{Y} - \frac{r\sigma_y}{\sigma_x}\overline{X} + \frac{r\sigma_y}{\sigma_x}x.$$

Подставляя численные значения, получаем:

$$\overline{y}_x = 0{,}96 + 0{,}324\,x.$$

3. На графике изобразить корреляционное поле и построить прямую $\overline{y}_x = a\,x + b$.

Построим график прямой регрессии Y на X.

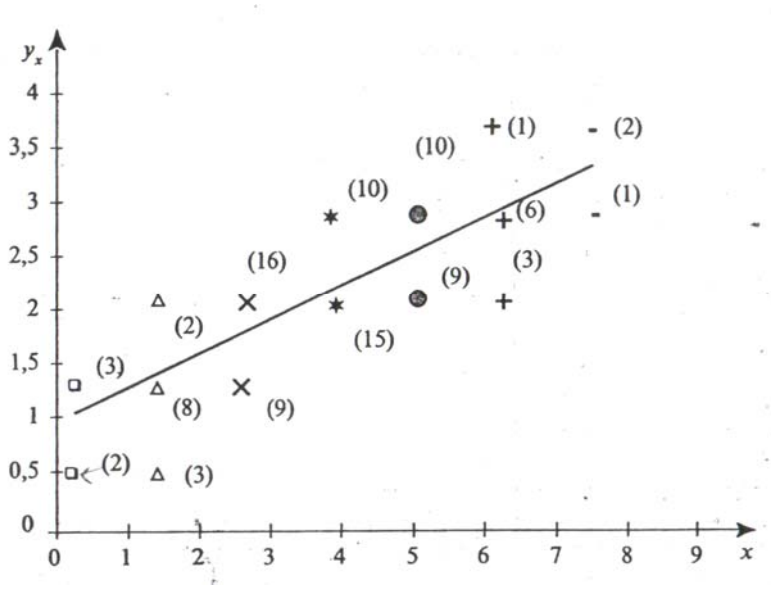

Рис.11

На графике рядом с точками указаны частоты их появления.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Афанасьев, В.В. Теория вероятностей [Текст] / В.В. Афанасьев. – М.: Владос, 2007. – 352 с.

2. Данилов, А.М. Теория вероятностей и математическая статистика [Текст]: учеб. пособие / А.М. Данилов, А.А. Данилов. – Пенза: ПГАСА, 1996. – 168 с.

3. Гнеденко, Б.В. Курс теории вероятностей [Текст] / Б.В. Гнеденко. – М.: Либроком, 2011. – 448 с.

4. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике [Текст]: учеб. пособие / В.Е. Гмурман. – М.: Юрайт, 2013. – 416 с.

5. Гмурман, В.Е. Теория вероятностей и математическая статистика [Текст]: учеб. пособие / В.Е. Гмурман. – М.: Юрайт, 2013. – 480 с.

6. Колде, Я.К. Практикум по теории вероятностей и математической статистике [Текст] / Я.К. Колде. – М.: Высш. шк., 1991. – 158 с.

7. Колемаев, В.А. Теория вероятностей и математическая статистика [Текст]: учеб. пособие / В.А. Колемаев, В.Н. Калинина. – М.: КноРус, 2012. – 376 с.

8. Кремер, Н.Ш. Теория вероятностей и математическая статистика [Текст] / Н.Ш. Кремер. – М.: ЮНИТИ-ДАНА, 2010. – 552 с.

9. Шапкин, А.С. Задачи с решениями по высшей математике, теории вероятностей, математической статистке, математическому программированию [Текст] / А.С. Шапкин, В.А. Шапкин. – М.: Дашков и Ко, 2013. – 432 с.

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-z^2}{2}} dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	**0,36**	0,1406	**0,72**	0,2642	**1,08**	0,3599
0,01	0,0040	**0,37**	0,1443	**0,73**	0,2673	**1,09**	0,3621
0,02	0,0080	**0,38**	0,1480	**0,74**	02703	**1,10**	0,3643
0,03	0,0120	**0,39**	0,1517	**0,75**	0,2734	**1,11**	0,3665
0,04	0,0160	**0,40**	0,1554	**0,76**	0,2764	**1,12**	0,3686
0,05	0,0199	**0,41**	0,1591	**0,77**	0,2794	**1,13**	0,3708
0,06	0,0239	**0,42**	0,1628	**0,78**	0,2823	**1,14**	0,3729
0,07	0,0279	**0,43**	0,1664	**0,79**	0,2852	**1,15**	0,3749
0,08	0,0319	**0,44**	0,1700	**0,80**	0,2881	**1,16**	0,3770
0,09	0,0359	**0,45**	0,1736	**0,81**	0,2910	**1,17**	0,3790
0,10	0,0398	**0,46**	0,1772	**0,82**	0,2939	**1,18**	0,3810
0,11	0,0438	**0,47**	0,1808	**0,83**	0,2967	**1,19**	0,3830
0,12	0,0478	**0,48**	0,1844	**0,84**	0,2995	**1,20**	0,3849
0,13	0,0517	**0,49**	0,1879	**0,85**	0,3023	**1,21**	0,3869
0,14	0,0557	**0,50**	0,1915	**0,86**	0,3051	**1,22**	0,3883
0,15	0,0596	**0,51**	0,1950	**0,87**	0,3078	**1,23**	0,3907
0,16	0,0636	**0,52**	0,1985	**0,88**	0,3106	**1,24**	0,3925
0,17	0,0675	**0,53**	0,2019	**0,89**	0,3133	**1,25**	0,3944
0,18	0,0714	**0,54**	0,2054	**0,90**	0,3159	**1,26**	0,3962
0,19	0,0753	**0,55**	0,2088	**0,91**	0,3186	**1,27**	0,3980
0,20	0,0793	**0,56**	0,2123	**0,92**	0,3212	**1,28**	0,3997
0,21	0,0832	**0,57**	0,2157	**0,93**	0,3238	**1,29**	0,4015
0,22	0,0871	**0,58**	0,2190	**0,94**	0,3264	**1,30**	0,4032
0,23	0,0910	**0,59**	0,2224	**0,95**	0,3289	**1,31**	0,4049
0,24	0,0948	**0,60**	0,2257	**0,96**	0,3315	**1,32**	0,4066
0,25	0,0987	**0,61**	0,2291	**0,97**	0,3340	**1,33**	0,4082
0,26	0,1026	**0,62**	0,2324	**0,98**	0,3365	**1,34**	0,4099
0,27	0,1064	**063**	0,2357	**0,99**	0,3389	**1,35**	0,4115
0,28	0,1103	**064**	0,2389	**1,00**	0,3413	**1,36**	0,4131
0,29	0,1141	**0,65**	0,2422	**1,01**	0,3438	**1,37**	0,4147
0,30	0,1179	**0,66**	0,2454	**1,02**	0,3461	**1,38**	0,4162
0,31	0,1217	**0,67**	0,2486	**1,03**	0,3485	**1,39**	0,4177
0,32	0,1255	**0,68**	0,2517	**1,04**	0,3508	**1,40**	0,4192
0,33	0,1293	**0,69**	0,2549	**1,05**	0,3531	**1,41**	0,4207
0,34	0,1331	**0,70**	0,2580	**1,06**	0,3554	**1,42**	0,4222
0,35	0,1368	**0,71**	0,2611	**1,07**	0,3577	**1,43**	0,4236

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,44	0,4251	**1,73**	0,4582	**2,04**	0,4793	**2,62**	0,4956
1,45	0,4265	**1,74**	0,4591	**2,06**	0,4803	**2,64**	0,4959
1,46	0,4279	**1,75**	0,4599	**2,08**	0,4812	**2,66**	0,4961
1,47	0,4292	**1,76**	0,4608	**2,10**	0,4821	**2,68**	0,4963
1,48	0,4306	**1,77**	0,4616	**2,12**	0,4830	**2,70**	0,4965
1,49	0,4319	**1,78**	0,4625	**2,14**	0,4838	**2,72**	0,4967
1,50	0,4332	**1,79**	0,4633	**2,16**	0,4846	**2,74**	0,4969
1,51	0,4345	**1,80**	0,4641	**2,18**	0,4854	**2,76**	0,4971
1,52	0,4357	**1,81**	0,4649	**2,20**	0,4861	**2,78**	0,4973
1,53	0,4370	**1,82**	0,4656	**2,22**	0,4868	**2,80**	0,4974
1,54	0,4382	**1,83**	0,4664	**2,24**	0,4875	**2,82**	0,4976
1,55	0,4394	**1,84**	0,4671	**2,26**	0,4881	**2,84**	0,4977
1,56	0,4406	**1,85**	0,4678	**2,28**	0,4887	**2,86**	0,4979
1,57	0,4418	**1,86**	0,4686	**2,30**	0,4893	**2,88**	0,4980
1,58	0,4429	**1,87**	0,4693	**2,32**	0,4898	**2,90**	0,4981
1,59	0,4441	**1,88**	0,4699	**2,34**	0,4904	**2,92**	0,4982
1,60	0,4452	**1,89**	0,4706	**2,36**	0,4909	**2,94**	0,4984
1,61	0,4463	**1,90**	0,4713	**2,38**	0,4913	**2,96**	0,4985
1,62	0,4474	**1,91**	0,4719	**2,40**	0,4918	**2,98**	0,4986
1,63	0,4484	**1,92**	0,4726	**2,42**	0,4922	**3,00**	0,49865
1,64	0,4495	**1,93**	0,4732	**2,44**	0,4927	**3,20**	0,49931
1,65	0,4505	**1,94**	0,4738	**2,46**	0,4931	**3,40**	0,49966
1,66	0,4515	**1,95**	0,4744	**2,48**	0,4934	**3,60**	0,499841
1,67	0,4525	**1,96**	0,4750	**2,50**	0,4938	**3,80**	0,499928
1,68	0,4535	**1,97**	0,4756	**2,52**	0,4941	**4,00**	0,499968
1,69	0,4545	**1,98**	0,4761	**2,54**	0,4945	**4,50**	0,499997
1,70	0,4554	**1,99**	0,4767	**2,56**	0,4948	**5,00**	0,499997
1,71	0,4564	**2,00**	0,4772	**2,58**	0,4951		
1,72	0,4573	**2,02**	0,4783	**2,60**	0,4953		

Приложение 2

Таблица значений $t_\gamma = t(\gamma, n)$

	γ				γ		
n	0,95	0,99	0,999	***n***	0,95	0,99	0,999
5	2,78	4,60	8,61	**20**	2,093	2,861	3,883
6	2,57	4,03	6,86	**25**	2,064	2,797	3,745
7	2,45	3,71	5,96	**30**	2,045	2,756	3,659
8	2,37	3,50	5,41	**35**	2,032	2,720	3,600
9	2,31	3,36	5,04	**40**	2,023	2,708	3,558
10	2,26	3,25	4,78	**45**	2,016	2,692	3,527
11	2,23	3,17	4,59	**50**	2,009	2,679	3,502
12	2,20	3,11	4,44	**60**	2,001	2,662	3,464
13	2,18	3,06	4,32	**70**	1,996	2,649	3,439
14	2,16	3,01	4,22	**80**	1,991	2,640	3,418
15	2,15	2,98	4,14	**90**	1,987	2,633	3,403
16	2,13	2,95	4,04	**100**	1,984	2,627	3,392
17	2,12	2,92	4,02	**120**	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Приложение 3

Критические точки распределения χ^2

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,89
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,4	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

О Г Л А В Л Е Н И Е

Учебное издание

Левова Галина Анатольевна
Снежкина Ольга Викторовна
Ячинова Светлана Николаевна

СТАТИСТИЧЕСКИЙ АНАЛИЗ СОВОКУПНОСТИ СЛУЧАЙНЫХ ВЕЛИЧИН
Учебное пособие

Р е д а к т о р В.С. Кулакова
В е р с т к а Н.А. Сазонова

Подписано в печать 22.01.14. Формат 60×84/16.
Бумага офисная «Снегурочка». Печать на ризографе.
Усл.печ.л. 5,1. Уч.-изд.л. 5,5. Тираж 80 экз.
Заказ № 10.

Издательство ПГУАС.
440028, г. Пенза, ул. Германа Титова, 28.