Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Кафедра информатики

Н. А. Волорова, А. С. Летохо

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Методическое пособие
для студентов специальности 1-31 03 04 «Информатика»
дневной формы обучения

В 2-х частях

Часть 2

Минск БГУИР 2012

УДК 519.2(076)
ББК 22.17я7
В68

Р е ц е н з е н т:
доцент кафедры высшей математики Белорусского государственного университета информатики и радиоэлектроники,
кандидат физико-математических наук О. А. Феденя

Волорова, Н. А.

В68 Теория вероятностей и математическая статистика: метод. пособие для студ. спец. 1-31 03 04 «Информатика» днев. формы обуч. : В 2 ч. Ч. 2 / Н. А. Волорова, А. С. Летохо. – Минск : БГУИР, 2012. – 64 с.

ISBN 978-985-488-732-6 (ч. 2).

В пособии приведены краткие теоретические сведения по теории вероятностей и математической статистике, примеры решения типовых задач, а также условия задач, рекомендуемых для проведения контрольных работ, при приеме зачетов и экзаменов, на практических занятиях и при самостоятельной работе студентов.

УДК 519.2(076)
ББК 22.17я7

Часть первая издана в БГУИР в 2006 г. : Теория вероятностей и математическая статистика: учеб.-метод. пособие для студ. спец. «Информатика» дневн. формы обуч. : В 2 ч. Ч. 1 / Н. А. Волорова, А. С. Летохо. – Минск : БГУИР, 2006. – 75 с.

ISBN 978-985-488-732-6 (ч. 2)
ISBN 978-985-488-013-6
ISBN 985-488-013-3

Содержание

ВВЕДЕНИЕ

Математической статистикой называется наука, занимающаяся методами обработки экспериментальных данных (ЭД), полученных в результате наблюдений над случайными явлениями.

Перед любой наукой ставятся следующие задачи:

- описание явлений;
- анализ и прогноз;
- выборка оптимальных решений.

Применительно к математической статистике приведём пример задачи первого типа: пусть имеется статистический материал, представляющий собой случайные числа. Требуется его упростить, представить в виде таблиц и графиков, обеспечивающих наглядность и информативность статистического материала.

Пример задачи второго типа: оценка (хотя бы приблизительная) характеристик случайных величин, например, математического ожидания, дисперсии и т.д. Какова точность полученных оценок?

Одной из характерных задач третьего типа является задача проверки правдоподобия гипотез, которая формулируется следующим образом: можно ли предполагать, что имеющаяся совокупность случайных чисел не противоречит некоторой гипотезе (например, о виде распределения, наличии корреляционной зависимости и т.д.).

Математическая статистика рассматривает задачи всех трех типов: способы описания результатов опыта, способы обработки опытных данных и оценки по ним характеристик случайного явления, способы выбора разумных решений.

1. БАЗОВЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Эмпирическая функция распределения

Под *генеральной совокупностью* понимают все возможные значения параметра, которые могут быть зарегистрированы в ходе неограниченного по времени наблюдения за объектом. Такая совокупность состоит из бесконечного множества элементов. В результате наблюдения за объектом формируется ограниченная по объему совокупность значений параметра x_1, x_2, ..., x_n. Такие данные представляют собой *выборку* из генеральной совокупности. Наблюдаемые значения x_i называют *вариантами*, а их количество – объемом выборки n. Для того чтобы по результатам наблюдения можно было делать какие-либо выводы, выборка должна быть *репрезентативной* (представительной), т. е. правильно представлять пропорции генеральной совокупности. Это требование выполняется, если объем выборки достаточно велик, а каждый элемент генеральной совокупности имеет одинаковую вероятность попасть в выборку.

Пусть в полученной выборке значение x_1 параметра наблюдалось n_1 раз, значение x_2 – n_2 раз, значение x_k – n_k раз, $n_1 + n_2 + ... + n_k = n$. Совокупность значений, записанных в порядке их возрастания, называют *вариационным рядом*, величины n_i – частотами, а их отношения к объему выборки $n_i = n_i / n$ – *относительными частотами* (ча́стостями). Очевидно, что сумма относительных частот равна единице. Другой формой вариационного ряда является ряд накопленных частот, называемый *кумулятивным рядом*.

Под распределением понимают соответствие между наблюдаемыми вариантами и их частотами или частостями. Пусть n_x – количество наблюдений, при которых случайные значения параметра X меньше x. Частость события $X < x$ равна n_x/n. Это отношение является функцией от x и от объема выборки: $F_n(x) = n_x/n$. Величина $F_n(x)$ обладает всеми свойствами функции распределения:

- $F_n(x)$ – неубывающая функция, ее значения принадлежат отрезку $[0 – 1]$;
- если x_1 – наименьшее значение параметра, а x_k – наибольшее, то $F_n(x) = 0$, когда $x \le x_1$, и $F_n(x) = 1$, когда $x > x_k$.

Функция $F^*_n(x)$ определяется по ЭД, поэтому ее называют эмпирической функцией распределения. В отличие от эмпирической функции $F^*_n(x)$ функцию распределения $F(x)$ генеральной совокупности называют теоретической функцией распределения, она характеризует не частость, а вероятность события $X < x$. При большом объеме наблюдений теоретическую функцию распределения $F(x)$ можно заменить эмпирической функцией $F^*_n(x)$.

Основные свойства функции $F^*_n(x)$:

1. $0 \le F^*_n(x) \le 1$.
2. $F^*_n(x)$ – неубывающая ступенчатая функция.

3. $F_n(x) = 0,\ x \le x_1$.
4. $F_n(x) = 1,\ x > x_n$.

Пример 1.1 . Задана выборка случайной величины

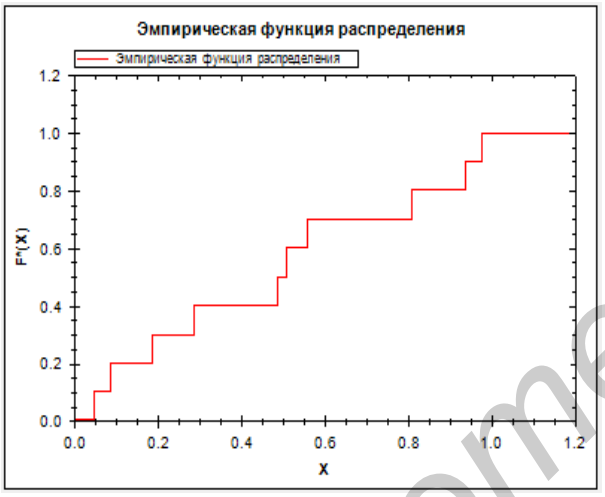
X = { 0,56; 0,79; 0,29; 0,56; 0,14; 1,00; 0,98; 1,00; 0,19; 0,08}.

Построить график эмпирической функции распределения $F_n(x)$.

Решение. Вариационный ряд случайной величины имеет вид

X_i = { 0,08; 0,14; 0,19; 0,29; 0,56; 0,56; 0,79; 0,98; 1,00; 1,00; }.

Выделяем полуинтервалы $(-\infty; 0,08]$, $(0,08; 0,14]$, $(0,14; 0,19]$, …,$(1,0; +\infty]$. На полуинтервале $(-\infty; 0,08]$ $F_n(x) = 0/10 = 0$. При $0,08 < x \le 0,19$ $F_n(x) = 1/10 = 0,1$. Аналогично определяем значения $F_n(x)$ на остальных полуинтервалах:



$$F(x) = \begin{cases} 0 \quad при \quad -\infty < x \le 0,08, \\ 0,1 \ при \quad 0,08 < x \le 0,14, \\ 0,2 \ при \quad 0,14 < x \le 0,19, \\ 0,3 \ при \quad 0,19 < x \le 0,29, \\ 0,4 \ при \quad 0,29 < x \le 0,56, \\ 0,6 \ при \quad 0,56 < x \le 0,79, \\ 0,7 \ при \quad 0,79 < x \le 0,98, \\ 0,8 \ при \quad 0,98 < x \le 1,0, \\ 1,0 \ при \quad 0,1 < x < +\infty. \end{cases}$$

Рис. 1.1. График функции $F_n(x)$

Замечание. В каждой точке оси x, соответствующей значениям x_i, функция $F_n(x)$ имеет скачок. В точке разрыва $F_n(x)$ непрерывна слева.

1.2. Гистограмма

При большом объеме выборки (понятие «большой объем» зависит от целей и методов обработки, в данном случае будем считать n большим, если $n > 40$) в целях удобства обработки и хранения сведений прибегают к группированию ЭД в интервалы. Количество интервалов следует выбрать так, чтобы в необходимой мере отразилось разнообразие значений параметра в совокупности и в то же время закономерность распределения не искажалась случайными колебаниями частот по отдельным разрядам. Существуют нестрогие рекомендации по выбору количества M и размера таких интервалов,

в частности, параметр M рекомендуется выбирать с помощью следующих соотношений:

$$\begin{aligned} M &= \text{int}\,(\sqrt{n}), \quad n \le 100 \\ M &= \text{int}\,((2...4)\cdot \lg(n)), \quad n > 100\,, \end{aligned} \tag{1.1}$$

где int(x) – целая часть числа x. Желательно, чтобы n без остатка делилось на M.

Графически статистический ряд отображают в виде гистограммы, полигона и ступенчатой линии. Часто гистограмму представляют как фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиною Δ, а высоты равны $m_i/(n\Delta)$. Такую гистограмму можно интерпретировать как графическое представление эмпирической функции плотности распределения $f_n(x)$, в ней суммарная площадь всех прямоугольников составит единицу. Гистограмма помогает подобрать вид теоретической функции распределения для аппроксимации ЭД.

Полигоном называют ломаную линию, отрезки которой соединяют точки с координатами по оси абсцисс, равными серединам интервалов, а по оси ординат – соответствующим частостям.

Порядок построения гистограммы следующий.

1. Построить вариационный ряд, т.е. расположить выборочные значения в порядке возрастания: $\hat{x}_1 \le \hat{x}_2 \le ... \le \hat{x}_n$.

2. Вся область возможных значений $\left[\hat{x}_1, \hat{x}_n\right]$ разбивается на M непересекающихся и примыкающих друг к другу интервалов.

A_i, B_i – соответственно левая и правая границы i-го интервала ($A_{i+1} = B_i$);

$\Delta_i = B_i - A_i$ – длина i-го интервала;

m_i – количество чисел в выборке, попадающих в i-й интервал.

При использовании *равноинтервального* метода построения гистограммы параметры A_i, B_i, Δ_i вычисляются следующим образом:

$$\Delta_i = \Delta = (\hat{x}_n - \hat{x}_1)/M; \quad A_i = \hat{x}_1 + (i-1)\Delta; \quad B_i = A_{i+1}; \quad i = 1,2,...,M.$$

Если при подсчете значений какое-то число в выборке точно совпадает с границей между интервалами, то необходимо в счетчик обоих интервалов прибавить по 0,5.

В случае применения *равновероятностного* метода границы A_i, B_i выбираются таким образом, чтобы в каждый интервал попадало одинаковое количество выборочных значений:

$$m_i = m = n / M.$$

В этом случае

$$A_1 = \hat{x}_1;\ B_1 = (\hat{x}_m + \hat{x}_{m+1})/2;\ A_2 = B_1;\ A_i = (\hat{x}_{(i-1)m} + \hat{x}_{(i-1)m+1})/2;\ i = 2,3,...,M.$$

3. Вычисляется средняя плотность вероятности для каждого интервала по формуле

$$f_i^* = m_i / (n \cdot \Delta_i).$$

4. На графике провести две оси: x и $f^*(x)$.

5. На оси x отмечаются границы всех интервалов.

6. На каждом интервале строится прямоугольник с основанием Δ_i и высотой $f_i^* = m_i / (n \cdot \Delta_i)$. Полученная при этом ступенчатая линия называется гистограммой, график которой приблизительно выглядит так, как показано на рис. 1.2.

Замечания.

1. Суммарная площадь всех прямоугольников равна единице.

2. В равновероятностной гистограмме площади всех прямоугольников одинаковы. По виду гистограммы можно судить о законе распределения случайной величины.

Достоинства использования гистограммы: простота применения, наглядность.

Рассмотренные представления ЭД являются исходными для последующей обработки и вычисления различных параметров.

Пример 1.2. Дан вариационный ряд выборки случайной величины X (n = 100). Построить гистограммы равноинтервальным методом.

$$Y_i = \{0,01; 0,02; 0,03; 0,05; 0,06; 0,08; 0,08; 0,09; 0,11; 0,12; 0,18; 0,19;$$
$$0,21; 0,23; 0,24; 0,24; 0,25; 0,26; 0,29; 0,30; 0,30; 0,30; 0,31; 0,32; 0,34;$$
$$0,35; 0,35; 0,36; 0,37; 0,41; 0,42; 0,44; 0,50; 0,53; 0,54; 0,54; 0,57; 0,58;$$
$$0,58; 0,59; 0,59; 0,59; 0,61; 0,61; 0,62; 0,62; 0,64; 0,65; 0,65; 0,66; 0,66;$$
$$0,66; 0,70; 0,72; 0,72; 0,73; 0,73; 0,73; 0,73; 0,74; 0,76; 0,77; 0,78; 0,78;$$
$$0,80; 0,81; 0,83; 0,84; 0,85; 0,86; 0,86; 0,87; 0,89; 0,90; 0,90; 0,90; 0,92;$$
$$0,93; 0,94; 0,94; 0,94; 0,95; 0,96; 0,97; 0,97; 0,97; 0,98; 0,99; 0,99; 0,99;$$
$$0,99; 0,99; 0,99; 0,99; 1,00; 1,00; 1,00; 1,00; 1,00; 1,00\}.$$

Разобьем область возможных значений $[X_1, X_n] = [0, 1]$ на M непересекающихся интервалов, где M выбирается в соответствии с (1.1):

$$M = \sqrt{100} = 10.$$

При использовании *равноинтервального* метода построения гистограммы параметры A_i, B_i, h_i вычисляются следующим образом:

$$h = h_i = \frac{y_n - y_1}{M}; \quad A_i = y_i + (i-1)h; \quad B_i = A_{i+1}; \quad i = 1,2,...,M.$$

Откуда:

$$h_i = h = \frac{b-a}{M} = 0{,}1.$$

A_i = [0; 0,1; 0,2; 0,3; 0,4; 0,6; 0,6; 0,7; 0,8; 0,9].

B_i = [0,1; 0,2; 0,3; 0,4; 0,6; 0,6; 0,7; 0,8; 0,9; 1,0].

Для каждого интервала посчитаем числа v_i – количество чисел в выборке, попадающих в *i*-й интервал, и вычислим среднюю плотность вероятности по формуле

$$f_i^* = \frac{v_i}{n \cdot h_i}.$$

v_i = [8; 4; 8; 9; 4; 9; 11; 11; 12; 24]

f_i^* = [0,8; 04; 0,8; 0,9; 04; 0,9; 1,1; 1,1; 1,2; 2,4].

Пример 1.3. Дан вариационный ряд выборки случайной величины X (n = 100). Построить гистограммы равноинтервальным методом.

$$\begin{aligned} Y_i = [&0{,}04;0{,}06;0{,}06;0{,}08;0{,}08;0{,}09;0{,}09;0{,}12;0{,}12;0{,}13;0{,}13;0{,}14;\\ &0{,}17;0{,}19;0{,}20;0{,}20;0{,}22;0{,}22;0{,}23;0{,}24;0{,}26;0{,}27;0{,}29;0{,}29;0{,}32;\\ &0{,}32;0{,}32;0{,}33;0{,}38;0{,}38;0{,}40;0{,}41;0{,}42;0{,}43;0{,}44;0{,}45;0{,}47;0{,}51;\\ &0{,}52;0{,}53;0{,}53;0{,}53;0{,}54;0{,}54;0{,}55;0{,}55;0{,}59;0{,}60;0{,}60;0{,}60;0{,}61;\\ &0{,}65;0{,}65;0{,}70;0{,}73;0{,}74;0{,}75;0{,}75;0{,}75;0{,}76;0{,}80;0{,}81;0{,}81;0{,}82;\\ &0{,}82;0{,}86;0{,}86;0{,}86;0{,}86;0{,}86;0{,}87;0{,}88;0{,}88;0{,}88;0{,}89;0{,}91;0{,}92;\\ &0{,}92;0{,}93;0{,}94;0{,}94;0{,}94;0{,}94;0{,}96;0{,}96;0{,}97;0{,}98;0{,}98;0{,}99;0{,}99;\\ &0{,}99;0{,}99;1{,}00;1{,}00;1{,}00;1{,}00;1{,}00;1{,}00;1{,}00;1{,}00\,]. \end{aligned}$$

Разобьем область возможных значений $[Y_1, Y_n] = [0, 1]$ на

$$M = \sqrt{100} = 10.$$

В случае применения *равновероятностного* метода количество попаданий в каждый интервал равно:

$$v_i = \nu = \frac{n}{M} = \frac{100}{10} = 10.$$

Границы отрезков A_i, B_i вычисляются следующим образом:

$$A_1 = y_1;\ \ B_1 = \frac{y_v + y_{v+1}}{2};\ \ A_2 = B_1;\ \ A_i = \frac{y_{(i-1)v} + y_{(i-1)v+1}}{2};\quad i = 2,3,\ldots,M.$$

$$A_i = [0;0{,}13;0{,}26;0{,}4;0{,}53;0{,}6;0{,}8;0{,}87;0{,}94;0{,}99].$$

$$B_i = [0{,}13;0{,}26;0{,}4;0{,}53;0{,}6;0{,}8;0{,}87;0{,}94;0{,}99;1{,}0].$$

Откуда:

$$h_i = B_i - A_i,$$

$h_i = [\,0{,}13;\ 0{,}13;\ 0{,}14;\ 0{,}13;\ 0{,}07;\ 0{,}2;\ 0{,}07;\ 0{,}07;\ 0{,}05;\ 0{,}007\,]$.

Для каждого интервала посчитаем числа v_i – количество чисел в выборке, попадающих в *i*-й интервал, и вычислим среднюю плотность вероятности по формуле

$$f_i^* = \nu_i / (n \cdot h_i).$$

$f_i^* = [\,0{,}76;\ 0{,}76;\ 0{,}8;\ 0{,}9;\ 0{,}4;\ 0{,}9;\ 1{,}1;\ 1{,}1;\ 1{,}2;\ 2{,}4\,]$.

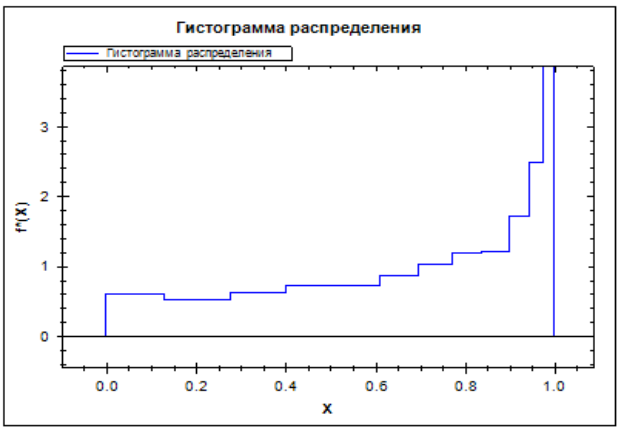


Рис. 1.2. Гистограмма распределения (равноинтервальный метод)

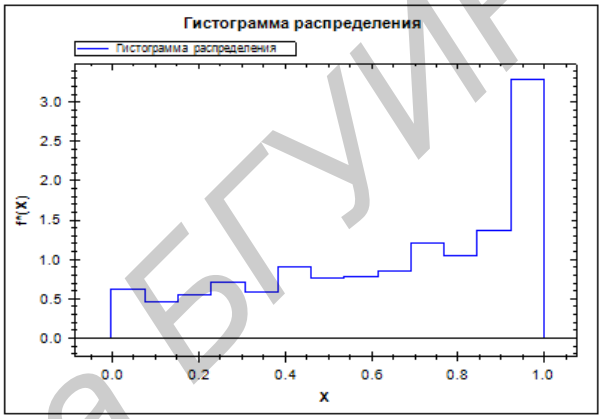


Рис.1.3. Гистограмма распределения (равновероятностный метод)

2. ТОЧЕЧНАЯ ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.1. Сущность задачи точечного оценивания параметров

Статистической оценкой θ* параметра θ распределения называется приближенное значение параметра, вычисленное по результатам эксперимента (по выборке).

1. Оценка θ* называется *несмещенной*, если M[θ*] = θ. Несмещенность – минимальное требование к оценкам.

2. Оценка θ* называется состоятельной, если при увеличении числа *n* она сходится по вероятности к значению параметра θ:

$$\lim_{n\to\infty}(P(|\theta^* - \theta| < \varepsilon)) = 1,$$

где ε − любое положительное число.

Несмещенная оценка является состоятельной, если

$$\lim_{n\to\infty} D[\theta^*] = 0.$$

3. Несмещенная оценка θ* является эффективной, если ее дисперсия минимальна по отношению к дисперсии любой другой оценки.

Точечная оценка предполагает нахождение единственной числовой величины, которая и принимается за значение параметра. Такую оценку целесообразно определять в тех случаях, когда объем ЭД достаточно велик.

Задача точечной оценки параметров в типовом варианте постановки состоит в следующем.

Имеется: выборка наблюдений (x_1, x_2, ..., x_n) за случайной величиной *X*. Объем выборки *n* фиксирован.

Известен вид закона распределения величины *X*, например, в форме плотности распределения $f(\theta,x)$, где θ – неизвестный (в общем случае векторный) параметр распределения. Параметр является неслучайной величиной.

Требуется найти оценку θ* параметра θ закона распределения.

Ограничения: выборка представительная.

Существует несколько методов решения задачи точечной оценки параметров, наиболее употребительными из них являются методы максимального (наибольшего) правдоподобия, моментов и квантилей.

2.2. Метод максимального правдоподобия

Метод предложен Р. Фишером в 1912 г. Метод основан на исследовании вероятности получения выборки наблюдений (x_1, x_2, ..., x_n). Эта вероятность равна

$$f(x_1, \theta)\, f(x_2, \theta) \ldots f(x_n, \theta)\, dx_1\, dx_2 \ldots dx_n.$$

Совместная плотность вероятности

$$L(x_1, x_2 \ldots, x_n ; \theta) = f(x_1, \theta)\, f(x_2, \theta) \ldots f(x_n, \theta),$$

рассматриваемая как функция параметра θ, называется *функцией правдоподобия*.

В качестве оценки θ* параметра θ следует взять то значение, которое обращает функцию правдоподобия в максимум. Для нахождения оценки необходимо заменить в функции правдоподобия θ на q и решить уравнение

$$dL / d\theta^* = 0.$$

Для упрощения вычислений переходят от функции правдоподобия к ее логарифму ln$L$. Если параметр распределения − векторная величина

$$\theta^* = (q_1, q_2, \ldots, q_n),$$

то оценки максимального правдоподобия находят из системы уравнений

$$\begin{cases} d \ln L(q_1, q_2, \ldots, q_n) / dq_1 = 0; \\ d \ln L(q_1, q_2, \ldots, q_n) / dq_2 = 0; \\ \quad . \; . \; . \; . \; . \; . \; . \; . \; . \\ d \ln L(q_1, q_2, \ldots, q_n) / dq_n = 0. \end{cases}$$

Для проверки того, что точка оптимума соответствует максимуму функции правдоподобия, необходимо найти вторую производную от этой функции. И если вторая производная в точке оптимума отрицательна, то найденные значения параметров максимизируют функцию.

Пример 2.1. Будем считать, что случайная величина X имеет нормальное распределение. Необходимо найти оценки максимального правдоподобия параметров m и S этого распределения.

Решение. Функция правдоподобия для выборки ЭД объемом n:

$$L(\alpha,\sigma) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \alpha)^2}{2\sigma^2} \right].$$

Логарифм функции правдоподобия

$$\ln L(\alpha,\sigma) = -n\ln\sqrt{2\pi} - n\ln\sigma - \left\{ -\sum_{i=1}^{n} \frac{(x_i - \alpha)^2}{2\sigma^2} \right\}.$$

Система уравнений для нахождения оценок параметров

$$\frac{\partial \ln L(\alpha,\sigma)}{\partial \alpha} = \left\{ \sum_{i=1}^{n} \frac{(x_i - \alpha)}{\sigma^2} \right\} = 0;$$

$$\frac{\partial \ln L(\alpha,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \alpha)^2}{\sigma^3} = 0.$$

Из первого уравнения следует

$$\sum_{i=1}^{n}(x_i - \alpha) = 0$$

или окончательно

$$\alpha = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Таким образом, среднее арифметическое является оценкой максимального правдоподобия для математического ожидания.

Из второго уравнения можно найти

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \alpha)^2}{n}.$$

Эмпирическая дисперсия является смещенной. После устранения смещения

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \alpha)^2}{n-1}.$$

2.3. Метод моментов

Метод предложен К. Пирсоном в 1894 г. Сущность метода:

- выбирается столько эмпирических моментов, сколько требуется оценить неизвестных параметров распределения. Желательно применять моменты младших порядков, так как погрешности вычисления оценок резко возрастают с увеличением порядка момента;
- вычисленные по ЭД оценки моментов приравниваются к теоретическим моментам;
- параметры распределения определяются через моменты и составляются уравнения, выражающие зависимость параметров от моментов, в результате получается система уравнений. Решение этой системы дает оценки параметров распределения генеральной совокупности.

Пример 2.2. Предположим, что случайная величина X имеет гамма-распределение. Необходимо найти оценки параметров этого распределения (можно отметить, что нормальное распределение является частным случаем гамма-распределения).

Решение. Функция плотности гамма-распределения имеет вид

$$f(\chi,\lambda) = \frac{\lambda^{\nu}}{\Gamma(\nu)}\chi^{\nu-1}\exp(-\lambda\chi), \qquad \chi \geq 0, \quad \lambda > 0, \quad \nu \geq 0.$$

Распределение характеризуется двумя параметрами ν и λ, поэтому следует выразить один параметр через оценку математического ожидания, а другой – через оценку дисперсии. Математическое ожидание и дисперсия этого распределения равны ν/λ и ν/λ^2 соответственно. Пусть их оценки определены и равны

$$\alpha_1 = 27{,}51, \quad \mu_2 = 0{,}91.$$

Составим систему уравнений для оцениваемых параметров:

$$\frac{\nu}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \alpha_1,$$

$$\frac{\nu}{\lambda^2} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \alpha_1)^2 = \mu_2.$$

Разделив оценку математического ожидания на оценку дисперсии, получим

$$\lambda = \alpha_1/\mu_2 = 30{,}12 .$$

Метод моментов позволяет получить состоятельные, достаточные оценки, они при довольно общих условиях распределены асимптотически нормально. Смещение удается устранить введением поправок. Эффективность оценок невысокая, т.е. даже при больших объемах выборок дисперсия оценок относительно велика (за исключением нормального распределения, для которого метод моментов дает эффективные оценки). В реализации метод моментов проще метода максимального правдоподобия. Напомним, что метод целесообразно применять для оценки не более чем четырех параметров, так как точность выборочных моментов резко падает с увеличением их порядка.

2.4. Метод квантилей

Сущность метода квантилей схожа с методом моментов: выбирается столько квантилей, сколько требуется оценить параметров; неизвестные теоретические квантили, выраженные через параметры распределения, приравниваются к эмпирическим квантилям. Решение полученной системы уравнений дает искомые оценки параметров.

Дисперсия $D(x_G)$ выборочной квантили обратно пропорциональна квадрату плотности распределения

$$D(x_G) = [G(1-G)] / [nf^2(x_G)]$$

в окрестностях точки x_G. Поэтому следует выбирать квантили вблизи тех значений x, в которых плотность вероятности максимальна.

Пример 2.3. Оценить методом квантилей параметры нормального распределения случайной величины.

Решение. Так как требуется определить два параметра распределения m и S, то выберем из вариационного ряда две эмпирические квантили. Например, можно взять

$G_1 = 5/44 = 0{,}114;$ $\quad x_{G1} = 26{,}13;$

$G_2 = 31/44 = 0{,}705;$ $\quad x_{G2} = 28{,}01.$

Используя стандартные функции математических пакетов, для выбранных значений G_1 и G_2 определим значения аргументов теоретической функции распределения для стандартизованной переменной:

$$U_{G1} = -1,207; \quad U_{G2} = 0{,}538.$$

Составим систему из двух уравнений:

$$U_{G1} = (x_{G1} - m) / S;$$

$$U_{G1} = (x_{G2} - m) / S.$$

Решение системы позволит найти искомые оценки параметров:

$$m = (U_{G2} \cdot x_{G1} - U_{G1} \cdot x_{G2}) / (U_{G1} - U_{G1}) = 27{,}42; \quad S = (x_{G1} - m) / U_{G1} = 1{,}07.$$

Метод квантилей позволяет получить асимптотически нормальные оценки, однако они несут в себе некоторый субъективизм, связанный с относительно произвольным выбором квантилей. Эффективность оценок не выше метода моментов. Определение оценок может приводить к необходимости численного решения достаточно сложных систем уравнений.

2.5. Точечные оценки числовых характеристик

1. Несмещенная состоятельная оценка *математического ожидания*, называемая выборочным средним, вычисляется по формуле

$$\overline{x} = \sum_{i=1}^{n} x_i / n.$$

2. Несмещенная состоятельная оценка *дисперсии*

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2.$$

3. Смещенная состоятельная оценка *дисперсии*

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2.$$

4. Несмещенная состоятельная оценка дисперсии

$$S_1^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - m_x\right)^2.$$

5. Состоятельная оценка *среднеквадратичного отклонения*

$$S_0 = \sqrt{S_0^2}.$$

6. Несмещенная состоятельная оценка *корреляционного момента*

$$\hat{K}_{XY} = \frac{1}{n-1} \cdot \sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y}),$$

где x_k, y_k − значения, которые приняли случайные величины X, Y в k-м опыте;

$\overline{x}$, $\overline{y}$ − средние значения случайных величин X и Y соответственно.

7. Состоятельная оценка *коэффициента корреляции*

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

8. Выборочный *начальный момент* k-го порядка определяется по формуле

$$\hat{\alpha}_k = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i)^k.$$

9. Выборочный *центральный момент* k-го порядка равен

$$\hat{\mu}_k = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^k.$$

10. В случае *неравноточных* измерений несмещенная состоятельная оценка математического ожидания равна

$$\tilde{x} = \sum_{i=1}^{n} \left(x_i / D[\xi_i] \right) \Big/ \sum_{i=1}^{n} \left(1 / D[\xi_i] \right),$$

где $D[\xi_i]$ − дисперсия случайной величины в i-м опыте.

11. Несмещенная состоятельная и эффективная оценка вероятности в схеме независимых опытов Бернулли:

$$p^* = \omega = m/n,$$

где m − число успешных опытов.

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Сущность задачи проверки статистических гипотез

Статистическая гипотеза представляет собой некоторое предположение о законе распределения случайной величины или о параметрах этого закона, формулируемое на основе выборки. Примерами статистических гипотез являются предположения: генеральная совокупность распределена по экспоненциальному закону; математические ожидания двух экспоненциально распределенных выборок равны друг другу. В первой из них высказано предположение о виде закона распределения, а во второй – о параметрах двух распределений. Гипотезы, в основе которых нет никаких допущений о конкретном виде закона распределения, называют непараметрическими, в противном случае – параметрическими.

Гипотезу, утверждающую, что различие между сравниваемыми характеристиками отсутствует, а наблюдаемые отклонения объясняются лишь случайными колебаниями в выборках, на основании которых производится сравнение, называют нулевой (основной) гипотезой и обозначают H_0. Наряду с основной гипотезой рассматривают и альтернативную (конкурирующую, противоречащую) ей гипотезу H_1. И если нулевая гипотеза будет отвергнута, то будет иметь место альтернативная гипотеза.

Различают простые и сложные гипотезы. Гипотезу называют *простой*, если она однозначно характеризует параметр распределения случайной величины. Например, если λ является параметром экспоненциального распределения, то гипотеза H_0 о равенстве $\lambda = 10$ – простая гипотеза. *Сложной* называют гипотезу, которая состоит из конечного или бесконечного множества простых гипотез. Сложная гипотеза H_0 о неравенстве $\lambda > 10$ состоит из бесконечного множества простых гипотез H_0 о равенстве $\lambda = b_i$, где b_i – любое число, большее 10.

Проверка гипотезы основывается на вычислении некоторой случайной величины – критерия, точное или приближенное распределение которого известно. Обозначим эту величину через z, ее значение является функцией от элементов выборки

$$z = z(x_1, x_2, \ldots, x_n).$$

Процедура проверки гипотезы предписывает каждому значению критерия одно из двух решений – принять или отвергнуть гипотезу. Тем самым все выборочное пространство и соответственно множество значений критерия делятся на два непересекающихся подмножества S_0 и S_1. Если значение критерия z попадает в область S_0, то гипотеза принимается, а если в область S_1, гипотеза отклоняется. Множество S_0 называется *областью принятия гипотезы,* или *областью допустимых значений*, а множество S_1 – *областью отклонения гипотезы,* или *критической областью*. Выбор одной области однозначно определяет и другую область.

Принятие или отклонение гипотезы H_0 по случайной выборке соответствует истине с некоторой вероятностью и, соответственно, возможны два рода ошибок.

Ошибка первого рода возникает с вероятностью α тогда, когда ***отвергается верная гипотеза*** H_0 и принимается конкурирующая гипотеза H_1.

Ошибка второго рода возникает с вероятностью β в том случае, когда ***принимается неверная гипотеза*** $\boldsymbol{H_0}$, в то время как справедлива конкурирующая гипотеза H_1.

Доверительная вероятность – это вероятность не совершить ошибку первого рода и принять верную гипотезу H_0.

Вероятность отвергнуть ложную гипотезу H_0 называется ***мощностью критерия***. Следовательно, при проверке гипотезы возможны четыре варианта исходов (табл. 3.1).

Таблица 3.1

Гипотеза H_0	Решение	Вероятность	Примечание
Верна	Принимается	$1 - \alpha$	Доверительная вероятность
	Отвергается	α	Вероятность ошибки первого рода
Неверна	Принимается	β	Вероятность ошибки второго рода
	Отвергается	$1 - \beta$	Мощность критерия

Целесообразно полагать одинаковыми значения вероятности выхода параметра θ* за нижний и верхний пределы интервала. Суммарная вероятность того, что параметр θ* выйдет за пределы интервала с границами $\theta^*_{1-\alpha/2}$ и $\theta^*_{\alpha/2}$, составляет величину α. Эту величину следует выбрать настолько малой, чтобы выход за пределы интервала был маловероятен. Если оценка параметра попала в заданный интервал, то в таком случае нет оснований подвергать сомнению проверяемую гипотезу, следовательно, гипотезу равенства $\theta^* = \theta$ можно принять. Но если после получения выборки окажется, что оценка выходит за установленные пределы, то в этом случае есть серьезные основания отвергнуть гипотезу H_0. Отсюда следует, что вероятность допустить ошибку первого рода равна α (равна уровню значимости критерия).

При заданном объеме выборки вероятность совершения ошибки первого рода можно уменьшить, снижая уровень значимости α. Однако при этом увеличивается вероятность ошибки второго рода β (снижается мощность критерия).

Единственный способ уменьшить обе вероятности состоит в увеличении объема выборки (плотность распределения оценки параметра при этом становится более «узкой»). При выборе критической области руководствуются правилом Неймана – Пирсона: следует так выбирать критическую область, чтобы вероятность α была мала, если гипотеза верна, и велика − в противном случае. Однако выбор конкретного значения α относительно произволен.

Употребительные значения лежат в пределах от 0,001 до 0,2. В целях упрощения ручных расчетов составлены таблицы интервалов с границами $\theta^*_{1-\alpha/2}$ и $\theta^*_{\alpha/2}$ для типовых значений α и различных способов построения критерия.

В зависимости от сущности проверяемой гипотезы и используемых мер расхождения оценки характеристики от ее теоретического значения применяют различные критерии. К числу наиболее часто применяемых критериев для проверки гипотез о законах распределения относят критерии хи-квадрат Пирсона, Колмогорова, Мизеса, Вилкоксона, о значениях параметров – критерии Фишера, Стьюдента.

3.2. Проверка гипотез о законе распределения

Обычно сущность проверки гипотезы о законе распределения ЭД заключается в следующем. Имеется выборка ЭД фиксированного объема, выбран или известен вид закона распределения генеральной совокупности. Необходимо оценить по этой выборке параметры закона, определить степень согласованности ЭД и выбранного закона распределения, в котором параметры заменены их оценками. Пока не будем касаться способов нахождения оценок параметров распределения, а рассмотрим только вопрос проверки согласованности распределений с использованием наиболее употребительных критериев.

3.2.1. Критерий хи-квадрат К. Пирсона

Использование этого критерия основано на применении такой меры (статистики) расхождения между теоретическим $F(x)$ и эмпирическим распределением $F^*_n(x)$, которая приближенно подчиняется закону распределения χ^2. Гипотеза H_0 о согласованности распределений проверяется путем анализа распределения этой статистики. Применение критерия требует построения статистического ряда.

Пусть выборка представлена статистическим рядом с количеством разрядов M. Наблюдаемая частота попаданий в i-й разряд n_i. В соответствии с теоретическим законом распределения ожидаемая частота попаданий в i-й разряд составляет F_i. Разность между наблюдаемой и ожидаемой частотами составит величину $(n_i - F_i)$. Для нахождения общей степени расхождения между $F(x)$ и $F^*_n(x)$ необходимо подсчитать взвешенную сумму квадратов разностей по всем разрядам статистического ряда:

$$\chi^2 = \sum_{i=1}^{M} \frac{(n_i - F_i)^2}{F_i}. \qquad (3.1)$$

Величина χ^2 при неограниченном увеличении n имеет χ^2-распределение (асимптотически распределена как χ^2). Это распределение зависит от числа степеней свободы k, т.е. количества независимых значений, слагаемых в выражении (3.1). Число степеней свободы равно числу y минус число линейных связей, наложенных на выборку. Одна связь существует в силу того, что любая частота может быть вычислена по совокупности частот в оставшихся $M - 1$ разрядах.

Кроме того, если параметры распределения не известны заранее, то имеется еще одно ограничение, обусловленное подгонкой распределения к выборке. Если по выборке определяются S параметров распределения, то число степеней свободы составит $k = M - S - 1$.

Область принятия гипотезы H_0 определяется условием $\chi^2 < \chi^2(k;\ \alpha)$, где $\chi^2(k;\ \alpha)$ – критическая точка χ^2 – распределения с уровнем значимости α. Вероятность ошибки первого рода равна α, вероятность ошибки второго рода четко определить нельзя, потому что существует бесконечно большое множество различных способов несовпадения распределений. Критерий рекомендуется применять при $n > 100$, допускается применение при $n > 40$, именно при таких условиях критерий состоятелен (как правило, отвергает неверную нулевую гипотезу).

Алгоритм проверки по критерию

1. Построить гистограмму равновероятностным способом.
2. По виду гистограммы выдвинуть гипотезу

$$H_0 : f(x) = f_0(x),$$

$$H_1 : f(x) \neq f_0(x),$$

где $f_0(x)$ − плотность вероятности гипотетического закона распределения.

3. Вычислить значение критерия по формуле

$$\chi^2 = n \sum_{i=1}^{M} \frac{\left(p_i - p_i^*\right)^2}{p_i} = \sum_{i=1}^{M} \frac{\left(\nu_i - np_i\right)^2}{np_i},$$

где $p_i^* = \nu_i / n$ – частота попадания в i-й интервал;

p_i − теоретическая вероятность попадания случайной величины в i-й интервал при условии, что гипотеза H_0 верна.

Замечания. После вычисления всех вероятностей p_i проверить, выполняется ли контрольное соотношение

$$\left| 1 - \sum_{i=1}^{M} p_i \right| \leq 0{,}01.$$

4. Из таблицы «Хи-квадрат» приложения выбирается значение $\chi^2_{\alpha,k}$, где α − заданный уровень значимости, а k − число степеней свободы, определяемое по формуле

$$k = M - 1 - S.$$

Здесь S − число параметров теоретического распределения, значения которых были получены из экспериментальных данных.

5. Если $\chi^2 > \chi^2_{\alpha,k}$, то гипотеза H_0 отклоняется. В противном случае нет оснований ее отклонить: с вероятностью 1 − β она верна, а с вероятностью − β неверна, но величина β неизвестна.

Пример 3.1. По выборке СВ объемом n = 200 построена гистограмма равновероятностным способом. Проверить гипотезу о соответствии выборки закону, имеющему следующую функцию распределения:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{2}{\pi} * \arcsin(x), & x \in [0,1]; \\ 1, & x > 1. \end{cases}$$

Доверительная вероятность равна 0,01.

i	A_i	B_i	h_i	v_i	f_i^*
0	0	0,15	0,15	20	0,64
1	0,15	0,33	0,18	20	0,54
2	0,33	0,48	0,15	20	0,66
3	0,48	0,60	0,11	20	0,90
4	0,60	0,69	0,09	20	1,08
5	0,69	0,76	0,07	20	1,36
6	0,76	0,86	0,09	20	1,05
7	0,86	0,92	0,06	20	1,54
8	0,92	0,98	0,05	20	1,80
9	0,98	1	0,02	20	5,11

Гистограмма и кривая теоретического распределения приведены на рис. 3.1.

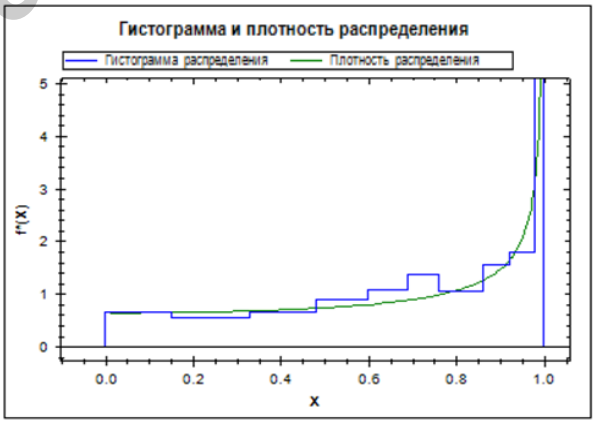


Рис. 3.1. Гистограмма и кривая теоретического распределения

Рассчитаем теоретическую вероятность попадания случайной величины в i-й интервал при условии, что гипотеза верна.

$$p_i = F(B_i) - F(A_i) = \frac{2}{\pi} \cdot (\arcsin(B_i) - \arcsin(A_i)).$$

Вычислим значение критерия по формуле

$$\chi^2 = n \sum_{i=1}^{M} \frac{\left(p_i - p_i^*\right)^2}{p_i} = \sum_{i=1}^{M} \frac{\left(\nu_i - np_i\right)^2}{np_i} .$$

Результаты промежуточных вычислений представлены в таблице.

Таблица 3.2

	$F(A_i)$	$F(B_i)$	p_i	p_i^*	$n * \frac{(p_i - p_i^*)^2}{p_i}$
1	0	0,095903	0,095903	0,1	0,034999
2	0,095903	0,214206	0,118303	0,1	0,566327
3	0,214206	0,318888	0,104682	0,1	0,041886
4	0,318888	0,409873	0,090985	0,1	0,178647
5	0,409873	0,485025	0,075152	0,1	1,6432
6	0,485025	0,549881	0,064856	0,1	3,808723
7	0,549881	0,659407	0,109526	0,1	0,16572
8	0,659407	0,744	0,084593	0,1	0,561228
9	0,744	0,872905	0,128905	0,1	1,29631
10	0,872905	1,000507	0,127602	0,1	1,194121

Полученное значение $\chi^2 = 5{,}36$.

Количество степеней свободы в нашем примере равно

$$k = M - 1 - S = 10 - 1 - 0 = 9.$$

Значение $S = 0$, так как закон распределения не зависит ни от каких параметров. Из таблицы «Хи-квадрат» приложения выберем значение $\chi^2_{\alpha,k}$, где $\alpha = 0{,}01$:

$$\chi^2_{\alpha,k} = 21{,}07 .$$

$\chi^2 < \chi^2_{\alpha,k}$, следовательно, нет оснований отклонять выдвинутую гипотезу.

3.2.2. Критерий А. Н. Колмогорова

Для применения критерия А. Н. Колмогорова ЭД требуется представить в виде вариационного ряда (ЭД недопустимо объединять в разряды). В качестве меры расхождения между теоретической $F(x)$ и эмпирической $F^*_n(x)$ функциями распределения непрерывной случайной величины X используется модуль максимальной разности:

$$d_n = \max \left| F(x) - F_n(x) \right|.$$

А. Н. Колмогоров доказал, что какова бы ни была функция распределения *F(x)* величины *X* при неограниченном увеличении количества наблюдений *n*, функция распределения случайной величины d_n асимптотически приближается к функции распределения:

$$K(\lambda) = P(d\sqrt{n} > \lambda) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2x^2\lambda^2).$$

Иначе говоря, критерий А. Н. Колмогорова характеризует вероятность того, что величина d_n не будет превосходить параметр *l* для любой теоретической функции распределения. Уровень значимости α выбирается из условия

$$P(d\sqrt{n} > \lambda) = \alpha = 1 - K(\lambda)$$

в силу предположения, что почти невозможно получить это равенство, когда существует соответствие между функциями *F(x)* и $F^*_n(x)$. Критерий А. Н. Колмогорова позволяет проверить согласованность распределений по малым выборкам, он проще критерия хи-квадрат, поэтому его часто применяют на практике. Но требуется учитывать два обстоятельства.

1. В соответствии с условиями его применения необходимо пользоваться следующим соотношением:

$$d = \max(d_n^+, d_n^-),$$

где

$$d_n^+ = \max\left|\frac{i}{n} - F(x_i)\right|; \quad d_n^- = \max\left|F(x_i) - \frac{i-1}{n}\right|.$$

2. Условия применения критерия предусматривают, что теоретическая функция распределения известна полностью – известны вид функции и значения ее параметров.

Последовательность действий при проверке гипотезы следующая.

1. Построить вариационный ряд.
2. Построить график эмпирической функции распределения $F^*(x)$.
3. Выдвинуть гипотезу:

$$H_0 : F(x) = F_0(x),$$
$$H_1 : F(x) \neq F_0(x),$$

где $F_0(x)$ − теоретическая функция распределения.

4. Построить график функции $F_0(x)$ в одной системе координат с функцией $F^*_n(x)$.
5. Определить максимальное по модулю отклонение между функциями $F^*_n(x)$ и $F_0(x)$.
6. Вычислить значение критерия

$$\lambda = \sqrt{n} \cdot \max\left|F^*(x) - F_0(x)\right|.$$

7. Принимают тот или иной уровень значимости (чаще всего 0,05 или 0,01). Тогда доверительная вероятность $\gamma = 1 - \alpha$.

8. Из таблицы вероятностей Колмогорова выбрать критическое значение λ_γ.

9. Если $\lambda > \lambda_\gamma$, то нулевая гипотеза H_0 отклоняется, в противном случае – принимается, хотя она может быть неверна.

Достоинства критерия Колмогорова по сравнению с критерием χ^2: возможность применения при очень маленьких объемах выборки ($n < 20$).

Недостаток: критерий можно использовать в том случае, если параметры θ_1, ..., θ_k распределения заранее известны, а эмпирическая функция распределения $F^*(x)$ должна быть построена по *несгруппированным* выборочным данным.

Пример 3.2. Дана выборка СВ объемом n = 30. Проверить гипотезу о соответствии выборки закону, имеющему следующую функцию распределения:

$$F(x)=\begin{cases} 0, & x<0; \\ \dfrac{2}{\pi}\cdot\arcsin(x), & x\in[0,1]; \\ 1, & x>1. \end{cases}$$

Построим вариационный ряд

$$X_i=[0,04;0,05;0,06;0,08;0,09;0,11;0,18;0,22;0,26;0,28;0,34;0,42;$$
$$0,49;0,51;0,56;0,58;0,62;0,64;0,65;0,66;0,68;0,73;0,74;0,75;0,80;$$
$$0,85;0,87;0,87;1,00;1,00]$$

и эмпирическую функцию распределения

$$F_0(X_i)=[0,03;0,07;0,1;0,13;0,17;0,2;0,23;0,27;0,3;0,33;0,37;0,4;$$
$$0,43;0,47;0,5;0,53;0,57;0,6;0,63;0,67;0,7;0,73;0,77;0,8;0,83;0,87;0,9;$$
$$0,93;0,97;0,1;].$$

Графики эмпирической и теоретической функций распределения приведены на рис. 3.2.

Рис. 3.2. Графики эмпирической и теоретической функций распределения

Максимальная разность по модулю между функцией $F_0(X_i)$ и $F(X)$ равна 0,14 при $y = 0,87$.

Вычислим значение статистики λ:

$$\lambda = \sqrt{n} \cdot \max\left|F(x) - F_0(x)\right| = 0{,}8.$$

Из таблицы функции Колмогорова выберем критическое значение:

$$\lambda_x = 1{,}63.$$

Поскольку $\lambda < \lambda_x$, следовательно, у нас нет оснований отклонять выдвинутую гипотезу.

3.2.3. Критерий Р. Мизеса

В качестве меры различия теоретической функции распределения $F(x)$ и эмпирической $F^*_n(x)$ по критерию Мизеса (критерию ω^2) выступает средний квадрат отклонений по всем значениям аргумента x:

$$\varpi_n^2 = \int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 dF(x).$$

Статистика критерия

$$n\,\omega_n^2 = \frac{1}{12n} + \sum_{j=1}^{n}\left(F(x_i) - \frac{i - 0{,}5}{n}\right)^2.$$

При неограниченном увеличении n существует предельное распределение статистики $n\omega_n^2$. Задав значение вероятности α, можно определить критические значения $n\omega_n^2(\alpha)$. Проверка гипотезы о законе распределения осуществляется обычным образом: если фактическое значение $n\omega_n^2$ окажется больше критического или равно ему, то согласно критерию Мизеса с уровнем значимости α гипотеза H_0 о том, что закон распределения генеральной совокупности соответствует $F(x)$, должна быть отвергнута.

Пример 3.3. Дана выборка СВ объемом n = 30. Проверить гипотезу о соответствии выборки закону, имеющему следующую функцию распределения:

$$F(x) = \begin{cases} 0, & x < 0; \\ \dfrac{2}{\pi} \cdot \arcsin(x), & x \in [0,1]; \\ 1, & x > 1. \end{cases}$$

Исходные данные и результаты вычислений представлены в табл. 3.3, где приняты следующие обозначения:

$F_n(x_i) = (i - 0{,}5)/30$ – значение эмпирической функции распределения;

$F(x_i)$ – значение теоретической функции распределения, соответствует значению функции нормального распределения в точке x_i:

$$D_i = [F_n(x_i) - F(x_i)]^2.$$

Таблица 3.3

i	x_i	$F_n(x_i)$	$F(x_i)$	D_i
1	2	3	4	5
1	0,023	0,017	0,015	0,001
2	0,092	0,05	0,059	0,002
3	0,218	0,083	0,140	0,008
4	0,358	0,117	0,233	0,022
5	0,413	0,15	0,271	0,024
6	0,458	0,183	0,303	0,023
7	0,481	0,217	0,320	0,019
8	0,507	0,25	0,339	0,015
9	0,518	0,283	0,347	0,009
10	0,526	0,317	0,353	0,005
11	0,569	0,35	0,385	0,033
12	0,576	0,383	0,391	0,002
13	0,611	0,417	0,419	0,001
14	0,636	0,45	0,439	0,0
15	0,673	0,483	0,470	0,0
16	0,751	0,517	0,541	0,003
17	0,754	0,55	0,744	0,001
18	0,804	0,583	0,595	0,002
19	0,853	0,617	0,650	0,004
20	0,863	0,65	0,662	0,002
21	0,864	0,683	0,664	0,0
22	0,898	0,717	0,710	0,001
23	0,901	0,75	0,714	0,0
24	0,937	0,783	0,773	0,001
25	0,948	0,817	0,794	0,0
26	0,973	0,85	0,852	0,001
27	0,994	0,883	0,930	0,006
28	0,997	0,917	0,954	0,005
29	1,0	0,95	0,987	0,005
30	1,0	0,983	0,987	0,002

Критическое значение статистики критерия Мизеса при заданном уровне значимости равно 0,744. Фактическое значение статистики

$$n \cdot \varpi_n^2 = \frac{1}{12 \cdot 30} + \sum_{i=1}^{10} \Delta_i = 0,083 ,$$

что меньше критического значения. Следовательно, гипотеза H_0 не противоречит имеющимся данным.

4. ИНТЕРВАЛЬНАЯ ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

4.1. Сущность задачи интервального оценивания параметров

Интервальный метод оценивания параметров распределения случайных величин заключается в определении интервала (а не единичного значения), в котором с заданной степенью достоверности будет заключено значение оцениваемого параметра. *Интервальная оценка* характеризуется двумя числами – концами интервала, внутри которого предположительно находится истинное значение параметра. Иначе говоря, вместо отдельной точки для оцениваемого параметра можно установить интервал значений, одна из точек которого является своего рода «лучшей» оценкой.

Постановка задачи интервальной оценки параметров заключается в следующем.

Имеется: выборка наблюдений (x_1, x_2, …, x_n) за случайной величиной X. Объем выборки n фиксирован.

Необходимо с доверительной вероятностью $g = 1 - \alpha$ определить интервал

$$\theta_0 - \theta_1 \quad (\theta_0 < \theta_1),$$

который накрывает истинное значение неизвестного скалярного параметра θ (здесь, как и ранее, величина θ является постоянной, поэтому некорректно говорить, что значение θ попадает в заданный интервал).

Ограничения: выборка представительная, ее объем достаточен для оценки границ интервала.

Эта задача решается путем построения доверительного утверждения, которое состоит в том, что интервал от θ_0 до θ_1 накрывает истинное значение параметра θ с доверительной вероятностью не менее g. Величины θ_0 и θ_1 называются нижней и верхней доверительными границами (НДГ и ВДГ соответственно). Доверительные границы интервала выбирают так, чтобы выполнялось условие

$$P(\theta_0 \le \theta^* \le \theta_1) = g.$$

В инженерных задачах доверительную вероятность g назначают в пределах от 0,95 до 0,99. В доверительном утверждении считается, что статистики θ_0 и θ_1 являются случайными величинами и изменяются от выборки к выборке. Это означает, что доверительные границы определяются неоднозначно, существует бесконечное количество вариантов их установления.

На практике применяют два варианта задания доверительных границ:

- устанавливают симметрично относительно оценки параметра, т.е.

$$\theta_0 = \theta^* - \varepsilon_g, \qquad \theta_1 = \theta^* + \varepsilon_g,$$

где ε_g выбирают так, чтобы выполнялось доверительное утверждение. Следовательно, величина абсолютной погрешности оценивания ε_g равна половине доверительного интервала;

- устанавливают из условия равенства вероятностей выхода за верхнюю и нижнюю границу:

$$P(\theta > \theta^* + \varepsilon_{1,g}) = P(\theta < \theta^* - \varepsilon_{2,g}) = \alpha / 2.$$

В общем случае величина $\varepsilon_{1,g}$ не равна $\varepsilon_{2,g}$. Для симметричных распределений случайного параметра θ^* в целях минимизации величины интервала значения $\varepsilon_{1,g}$ и $\varepsilon_{2,g}$ выбирают одинаковыми, следовательно, в таких случаях оба варианта эквивалентны.

Нахождение доверительных интервалов требует знания вида и параметров закона распределения случайной величины θ^*. Для ряда практически важных случаев этот закон можно определить из теоретических соображений.

4.2. Доверительный интервал для математического ожидания

Пусть по выборке достаточно большого объема, $n > 30$, и при заданной доверительной вероятности $1 - \alpha$ необходимо определить доверительный интервал для математического ожидания m_x, в качестве оценки которого используется среднее арифметическое:

$$m_x^* = \alpha_1^* = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Закон распределения оценки математического ожидания близок к нормальному (распределение суммы независимых случайных величин с конечной дисперсией асимптотически нормально).

Для симметричных функций минимальный интервал тоже будет симметричным относительно оценки m_x. В этом случае выражение для доверительной вероятности имеет вид

$$P(|\, m^*_x - m_x \,| < \varepsilon) = 1 - \alpha,$$

где ε – абсолютная погрешность оценивания.

Определим оценку дисперсии случайного параметра m^*_x, учитывая, что этот параметр равен среднему арифметическому одинаково распределенных случайных величин x_i (следовательно, их дисперсии $D(x_i)$ одинаковы и равны S^2):

$$D[\alpha_1^*] = D\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} D[X_i]\right] = \frac{nS^2}{n^2} = \frac{S^2}{n}.$$

Итак, случайная величина m^*_x распределена по нормальному закону с параметрами m^*_x и S^2/n. Для установления необходимых соотношений целесообразно перейти к центрированным и нормированным величинам. Выражение $m^*_x - m_x$ можно трактовать как центрирование случайной величины m^*_x. Нормирование осуществляется делением на величину среднеквадратичного отклонения оценки m^*_x:

$$P\left(\frac{\left|m_x^* - m_x\right|}{\sqrt{S^2/n}} \le \frac{\varepsilon}{\sqrt{S^2/n}}\right) = \gamma.$$

Для стандартизованной величины вероятность соблюдения неравенства определяется по функции нормального распределения:

$$P(|z| \le \beta) = \frac{1}{\sqrt{2\pi}} \int_{-\beta}^{\beta} \exp(-t^2/2)dt = \Phi(\beta) - \Phi(-\beta) = -1 + 2\Phi(\beta) = 1 - \alpha,$$

где $\beta = \frac{\varepsilon}{\sqrt{S^2/n}}$. Значение β равно квантили $u_{1-\alpha/2}$ стандартного нормального распределения уровня $1-\alpha/2$.

Окончательно можно записать

$$u_{1-\alpha/2}^2 = \frac{n\varepsilon}{S^2} \qquad (4.1)$$

При фиксированном объеме выборки из (4.1) следует, что чем больше доверительная вероятность $1 - \alpha$, тем шире границы доверительного интервала (тем больше ошибка в оценке математического ожидания). Это равенство позволяет определить необходимый объем выборки для получения оценки математического ожидания с заданной надежностью и требуемой точностью (погрешностью):

$$N = S^2 u^2_{1-\alpha/2} / \varepsilon^2.$$

Если перейти к относительной погрешности $\varepsilon_0 = \varepsilon / m^*_x$, то

$$n = S^2 u^2_{1-\alpha/2} / (\varepsilon_0^2 m^{*2}_x). \qquad (4.2)$$

Таким образом, чтобы снизить относительную погрешность на порядок, необходимо увеличить объем выборки на два порядка. Приведенная формула часто используется в статистическом моделировании для определения необходимого количества испытаний модели.

Во многих случаях предположение о нормальном распределении случайной величины m^*_x становится приемлемым при $n > 4$ и вполне хорошо оправдывается при $n > 10$. Оценка m^*_x вполне пригодна для применения вместо m_x. Но не так обстоит дело с дисперсией, правомочность ее замены на S^2 не обоснована даже в указанных случаях. При небольшом объеме выборки, $n < 30$, закон распределения оценки дисперсии S^2 принимать за нормальный неоправданно. Ее распределение следует аппроксимировать распределением хи-квадрат как суммы квадратов центрированных величин (хи-квадрат распределение сходится к нормальному при количестве слагаемых, превышающем 30). Но это утверждение обосновано только для случая, когда случайная величина X распределена нормально.

С учетом сделанных допущений величина z будет подчиняться закону распределения Стьюдента с $n - 1$ степенями свободы (одна степень свободы использована для определения оценки дисперсии). Распределение Стьюдента симметричное, поэтому полученное соотношение между точностью, надежностью оценки и объемом выборки сохраняется, меняются только значения квантилей. Вместо квантили нормального распределения $u_{1-\alpha/2}$ следует взять квантиль $t_{1-\alpha/2(n-1)}$ распределения Стьюдента с $(n - 1)$ степенями свободы. Значения критических точек распределения Стьюдента для некоторых вероятностей и различных степеней свободы представлены в таблицах. Сравнение таблиц показывает, что квантили распределения Стьюдента больше квантилей нормального распределения того же уровня надежности при малом n. Иначе говоря, применение нормального распределения при небольшом объеме выборки ЭД приводит к неоправданному завышению точности оценки.

Пример 4.1. По выборке СВ объемом $n = 20$ найти точечную оценку математического ожидания случайной величины и доверительный интервал для оценки математического ожидания случайной величины для уровня значимости 0,99.

Решение. Исходная выборка имеет вид

$$X = [0{,}14; 0{,}21; 0{,}30; 0{,}32; 0{,}37; 0{,}42; 0{,}44; 0{,}48; 0{,}61; 0{,}66; 0{,}68; 0{,}70;$$
$$0{,}70; 0{,}76; 0{,}76; 0{,}86; 0{,}89; 0{,}91; 0{,}96; 0{,}98].$$

Найдем точечную оценку математического ожидания по формуле

$$\overline{x} = \sum_{i=1}^{n} x_i / n = 0{,}64.$$

Найдем точечную несмещенную оценку дисперсии:

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = 0{,}14.$$

Построим доверительный интервал для математического ожидания:

$$\overline{X} - \frac{s \cdot t_{\gamma,n-1}}{\sqrt{n-1}} \le m_x < \overline{X} + \frac{s \cdot t_{\gamma,n-1}}{\sqrt{n-1}},$$

где t – значение, взятое из таблицы Стьюдента для заданного уровня значимости $\gamma = 0{,}99$ и объема выборки $n = 20$, $n = 20$; откуда

$$\frac{s \cdot t_{\gamma,n-1}}{\sqrt{n-1}} = 0{,}14.$$

Следовательно,

$$0{,}5 \le m_x \le 0{,}78.$$

4.3. Доверительный интервал для дисперсии

По выборке при заданной надежности $1 - \alpha$ необходимо определить доверительный интервал для дисперсии μ_2, оценка которой

$$\mu_2^* = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x^*)^2.$$

Если стандартизовать оценку дисперсии, то величина $(n-1)S^2/\mu_2$ имеет распределение хи-квадрат с $(n-1)$ степенями свободы. Из этого вытекает вероятностное утверждение относительно выборочной дисперсии:

$$P[(n-1)S^2/\mu_2 > c_\alpha^2(n-1)] = \alpha. \tag{4.3}$$

Функция хи-квадрат несимметричная, поэтому границы интервала $c_1^2(n-1)$ и $c_2^2(n-1)$ выбирают из условия равной вероятности выхода за их пределы:

$$P\left[(n-1)S^2/\mu_2 < c_1^2(n-1)\right] = P\left[(n-1)S^2/\mu_2 > c_2^2(n-1)\right] = \alpha/2$$

или (4.4)

$$P[(n-1)S^2/c_1^2(n-1) < \mu_2] = P[(n-1)S^2/c_2^2(n-1) > \mu_2] = \alpha/2.$$

Значения границ соответствуют квантилям распределения хи-квадрат со значениями уровней $\alpha/2$ и $1-\alpha/2$, количество степеней свободы равно $n-1$.

Нижняя граница $c_1^2(n-1)$ равна квантили $c_{\alpha/2}^2(n-1)$, а верхняя – квантили $c_{1-\alpha/2}^2(n-1)$. Если воспользоваться критическими точками распределения, то следует записать

$$c_1^2(n-1) = c^2(1-\alpha/2;\quad n-1),$$
$$c_2^2(n-1) = c^2(\alpha/2;\quad n-1).$$

Пример 4.2. Для выборки, приведенной в примере 4.1, найти доверительный интервал для дисперсии.

Решение.

Доверительный интервал для дисперсии определяется так:

$$\frac{ns^2}{\chi^2_{\frac{1-\gamma}{2},n-1}} \le D_x < \frac{ns^2}{\chi^2_{\frac{1+\gamma}{2},n-1}},$$

где $\chi^2_{\frac{1-\gamma}{2},n-1}$, $\chi^2_{\frac{1+\gamma}{2},n-1}$ – значения, взятые из таблицы «Хи-квадрат» для заданного уровня значимости и объема выборки. Для рассматриваемого случая

$$\chi^2_{\frac{1-\gamma}{2},n-1} = 37,6;$$

$$\chi^2_{\frac{1+\gamma}{2},n-1} = 8,26;$$

$$0,074 \le D_x < 0,339.$$

4.4. Доверительный интервал для вероятности

Пусть случайная величина X имеет только два возможных значения: 0 и 1. В результате проведения достаточно большого количества наблюдений эта случайная величина приняла единичное значение n раз. Необходимо при заданной надежности $1-\alpha$ определить доверительный интервал для вероятности p, оценка которой соответствует частоте $\omega^* = m^*/n$.

Оценка ω^* вероятности p является состоятельной, эффективной и несмещенной. Если оцениваемая вероятность не слишком мала и не слишком велика $(0,05 < p < 0,95)$, то можно считать, что распределение случайной величины ω^* близко к нормальному. Этим допущением можно пользоваться, если np и $n(1-p)$ больше четырех. Параметры нормального распределения частоты $m_x^* = p$, $S^2 = p(1-p)/n$ (дисперсия $S^2(m)$ количества успехов m составляет величину $np(1-p)$, а дисперсия частоты $S^2(m)/n^2$. Тогда по аналогии с определением доверительного интервала для математического ожидания нормально распределенной величины ω^* можно записать

$$\varepsilon = \left|\omega^* - p\right| = u_{1-\alpha/2} S = u_{1-\alpha/2}\left(p(1-p)/n\right)^{0,5}, \qquad (4.5)$$

где $u_{1-\alpha/2}$ – квантиль стандартизованного нормального распределения. Расчетные формулы для границ доверительного интервала имеют вид:

$$p_1 = \omega^* - u_{1-\alpha/2}[\omega^*(1-\omega^*)/n]^{0,5},$$

$$p_2 = \omega^* + u_{1-\alpha/2}[\omega^*(1-\omega^*)/n]^{0,5}.$$

Более общие результаты получены с учетом того, что случайная величина w^* распределена по биномиальному закону:

$$F(\omega^*) = \sum_{k=0}^{m} C_n^k p^k (1-p)^{n-k},$$

где C_k^n – число сочетаний из n по k.

Исходя из этого положения, для практического применения получены значения нижней p_1 и верхней p_2 доверительных границ

$$p_1 = \frac{\chi^2_{\alpha/2}(2m)}{2n - m + 1 + 0,5\chi^2_{\alpha/2}(2m)}; \qquad (4.6)$$

$$p_2 = \frac{\chi^2_{\alpha/2}(2(m+1))}{2n - m + 0{,}5 \cdot \chi^2_{1-\alpha/2}(2(m+1))}, \qquad (4.7)$$

где $\chi^2_{\xi}(k)$ – квантиль распределения хи-квадрат уровня ξ с числом степеней свободы k.

Формулы (4.6) и (4.7) можно применять и в тех случаях, когда частость ω^* события близка (равна) нулю или близка (равна) количеству экспериментов n соответственно. В первом случае НДГ p_1 принимается равной нулю и рассчитывается только ВДГ p_2. Во втором случае рассчитывается НДГ p_1, а верхняя граница $p_2 = 1$.

Пример 4.3. В результате наблюдения за 58 изделиями не было зафиксировано ни одного отказа. Определить доверительный интервал для вероятности отказа с надежностью 0,9.

Решение. Нижнюю доверительную границу p_1 следует принять равной нулю, ВДГ

$$p_2 = \frac{\chi^2_{0{,}95}(2)}{116 - 0 + 0{,}5 \cdot \chi^2_{0{,}95}(2)} = \frac{6{,}0}{119} = 0{,}05.$$

Таким образом, доверительный интервал с нижней границей 0 и верхней границей 0,05 с вероятностью 0,9 накрывает истинное значение вероятности отказа изделий.

Пример 4.4. Среди стандартных изделий одной фабрики в среднем 15 % относится ко второму сорту. С какой вероятностью можно утверждать, что процент p изделий второго сорта среди 1000 стандартных изделий данной фабрики отличается от 15 % не более чем на 2 %?

Решение. Здесь $n = 1000$, $w = 0{,}15$, $\Delta = 0{,}02$. Из (4.5) получим

$$t = \Delta\sqrt{\frac{h}{w(1-w)}} = 0{,}02\sqrt{\frac{1000}{0{,}15 \cdot 0{,}85}} = 1{,}77.$$

Тогда

$$P\left\{\left|p - 0{,}15\right| \le 0{,}02\right\} = \Phi(1{,}77) = 0{,}9233.$$

5. ОБРАБОТКА ОДНОТИПНЫХ ВЫБОРОК ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

5.1. Однотипные выборки экспериментальных данных и задачи их обработки

По результатам экспериментов может накапливаться целый ряд выборок по однотипным средствам и комплексам. Однотипность не означает равноценности объектов по их показателям. Неоднородность означает, что выборки принадлежат различным законам распределения, которые различаются или только параметрами при одном и том же виде, или видом и параметрами распределения.

Задачи обработки однотипных выборок подразделяются на две группы. К первой группе относятся задачи объединения выборок. Простое слияние однотипных, но неоднородных выборок для последующей оценки показателей по объединенной выборке приводит к снижению качества оценок или даже к их полной непригодности. Необходимо применение специальных приемов объединения разнородных сведений в интересах использования всей содержащейся в выборках информации. Таким образом, при объединении выборок необходимо сначала проверить их однородность. Однородные выборки сливаются в одну общую выборку, которая обрабатывается с помощью обычных методов. Неоднородные выборки обрабатываются раздельно или объединяются с помощью специальных приемов.

Вторая группа задач связана с сопоставлением параметров распределения выборок, т.е. с определением существенных различий в значениях параметров однотипных выборок. Наиболее широкое распространение получил один из видов подобного рода задач, так называемый дисперсионный анализ. В дисперсионном анализе исследуются методы проверки гипотезы о равенстве математических ожиданий случайных величин, представленных выборками ограниченного объема. Непосредственное сравнение оценок математических ожиданий совокупности выборок оказывается менее эффективным, чем сопоставление оценок дисперсий; это обстоятельство и дало наименование методу.

Итак, пусть имеются m $(m \geq 2)$ однородных выборок, каждая выборка имеет свой объем n_i. Априорных сведений об однородности или неоднородности различных выборок нет.

$$\begin{matrix} x_{11} & x_{12} & \ldots & x_{1n} \\ y_{21} & y_{22} & \ldots & y_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ w_{m1} & w_{m2} & \ldots & w_{mn} \end{matrix} . \qquad (5.1)$$

Эта совокупность состоит из m слоев (строк). Каждая *i*-я строка $(i = 1, \ldots, m)$ представляет собой однородную случайную выборку результатов

наблюдений за значениями случайной величины $X, Y, \ldots, W$ соответственно. Слой характеризуется своим, в общем случае векторным, параметром Θ_i распределения и может иметь свои статистики, т.е. свои функции от выборочных значений.

5.2. Объединение выборок

По возможности объединения информации из совокупности однотипных выборок можно выделить три типовые ситуации:

- различные слои представляют собой однородные выборки. В такой идеализированной ситуации выборки можно объединить и определять искомые параметры, используя традиционный аппарат математической статистики;
- совокупность слоев частично неоднородна. Тогда однородные слои, если таковые обнаружатся, целесообразно объединить, а оставшиеся неоднородные группы выборок обрабатывать раздельно;
- слои полностью или частично неоднородны. Но в дополнение к результатам наблюдений имеется априорная информация о взаимосвязи параметров Θ_i различных выборок. Чем выше уровень априорной информированности о взаимосвязях параметров, тем потенциально более высокой эффективности оценок показателей можно достичь.

Следовательно, объединение слоев всегда следует начинать с проверки однородности выборок.

5.2.1. Объединение однородных выборок

Постановка задачи проверки однородности выборок формулируется следующим образом.

Имеются результаты наблюдений в виде совокупности выборок типа (5.1), задан уровень значимости *a* для проверки статистической гипотезы об однородности выборок.

Необходимо проверить однородность слоев.

Допущение: законы распределения случайных величин для различных слоев не известны.

На практике используется последовательная процедура проверки и попарного объединения выборок. В качестве исходной выборки можно взять любую, например, наибольшую по количеству элементов. В качестве второй выбирается любая из оставшихся выборок. Эти две выборки проверяются на однородность. При ее наличии выборки объединяются в одну, а при ее отсутствии вторая выборка остается самостоятельной. Указанную проверку и объединение повторяют для всех слоев исходной выборки.

Определение однородности двух выборок проводится на основе проверки статистической гипотезы H_0 о том, что выборки принадлежат одному, пусть и неизвестному, закону распределения. При этом может быть применен критерий Вилкоксона (Вилкоксона – Мана – Уитни).

Проверка однородности выборок по критерию Вилкоксона состоит в следующем. Пусть для случайной величины X имеется выборка объемом n_x и для случайной величины Y выборка объемом n_y. По этим выборкам необходимо с уровнем значимости α проверить гипотезу H_0 о том, что функция распределения $F(x)$ случайной величин X равна функции распределения $F(y)$ случайной величины Y. Конкурирующая гипотеза – функции распределения случайных величин различны: $F(x) < F(y)$ или $F(x) > F(y)$, т.е. критическая область двусторонняя.

Сущность проверки основана на простой идее: если верна гипотеза H_0, то нельзя ожидать преобладания наблюдений одной из выборок на любом из концов вариационного ряда, иначе говоря, результаты наблюдений из каждого слоя должны быть рассеяны по всему вариационному ряду. Такая проверка осуществляется только по порядковым соотношениям $x > y$ и $x < y$ между элементами выборок.

Пусть $n_x > 3$, $n_y > 3$ и суммарный объем обеих выборок не превосходит 25. Проверка гипотезы осуществляется поэтапно:

- из выборок исключаются одинаковые элементы (вероятность совпадения элементов весьма невелика, поэтому число исключаемых членов выборок не будет большим);
- на основе элементов обеих выборок строится общий вариационный ряд, индексы и конкретные значения элементов можно опустить. В результате получится просто последовательность букв *y* и *x*, например *xxxyxyyxxxyyy*;
- подсчитывается сумма порядковых номеров *u* вариант первой (меньшей по объему) выборки. В приведенном примере $n_x > n_y$ ($n_x = 7$ и $n_y = 6$), поэтому первой будем считать выборку для величины *Y*. Буква *y* встречается на четвертом, шестом, седьмом, одиннадцатом, двенадцатом и тринадцатом местах, следовательно,

$$u = 4 + 6 + 7 + 11 + 12 = 53.$$

Случайная величина *u* имеет распределение Вилкоксона. Для нее построена специальная таблица нижних критических точек распределения (см. приложение);

- по таблице критических точек для $n_x = 7, n_y = 6$, заданного уровня значимости, например $\alpha = 0,05$ (критическая область двусторонняя, следовательно, каждая сторона критической области соответствует уровню значимости $\alpha / 2 = 0,025$), определяется нижняя критическая точка $u_{\text{н}}$. В данном случае $u_{\text{н}} = 27$;
- вычисляется верхняя критическая точка $u_{\text{в}} = (n_y + n_x + 1) \cdot n_y - u_{\text{н}}$. Для рассматриваемого примера

$$u_{\text{в}} = (6 + 7 + 1) \cdot 6 - 27 = 57;$$

- если $u < u_{\text{н}}$ или $u > u_{\text{в}}$, то нулевую гипотезу отвергают. В противном случае нет оснований для отклонения нулевой гипотезы. В приведенном примере нулевая гипотеза об однородности выборок принимается.

Сумма порядковых номеров вариант первой выборки с увеличением общего объема выборок стремится к нормальному распределению. *Нормальное распределение можно применять, если* $n_x > 3$, $n_y > 3$ *и объем хотя бы одной из выборок превосходит* 25. В таком случае значение нижней критической точки величины *и* при $n_x \cdot n_y$ равно

$$u_n = \frac{(n_x + n_y + 1)\, n_y - 1}{2} - z_{(1-\alpha/2)} \sqrt{\frac{(n_x + n_y + 1)\, n_x\, n_y}{12}}, \quad (5.2)$$

где $z_{(1-\alpha/2)}$ – квантиль уровня $1 - \alpha / 2$ стандартизованной нормальной случайной величины.

Остальные этапы проверки ничем не отличаются от рассмотренных выше, применительно к малому объему слоев. В результате выполнения рассмотренных процедур однородные выборки будут объединены.

5.2.2. Объединение неоднородных выборок

Одним из простых и рациональных способов слияния является линейное объединение оценок показателей независимо от степени однородности имеющейся информации. При таком способе объединения неоднородной информации общая выборка рассматривается как смесь из *m* выборок однотипных наблюдений, каждая из которых имеет свои значения показателей. Подобное объединение возможно для несмещенных выборочных средних оценок (типа центральных моментов распределения, вероятностей свершения событий).

Пусть имеются выборочные средние оценки q_i отдельных слоев. Задача состоит в нахождении функции $\Theta = z(\Theta_1, \ldots, \Theta_m)$, которая была бы лучшей, в смысле принятого критерия, объединенной оценкой Θ^* параметра Θ. Типичным критерием оптимальности оценки является минимум дисперсии оценки. В качестве оценочной функции можно взять любую, но использование сложных функций вызывает труднопреодолимые препятствия по нахождению несмещенных и эффективных оценок. Лучше взять простую линейную комбинацию $\Theta = \sum_{i=1}^{m} v_i \Theta_i$. Коэффициенты выбирают из условия $\sum_{i=1}^{m} v_i = 1$, что обеспечивает получение несмещенной объединенной оценки. Значения коэффициентов v_i, обеспечивающие минимум дисперсии искомой оценки $\mu_2(\Theta) = \sum_{i=1}^{n} v_i^2 \mu_2(\Theta_i)$, равны

$$v_i = \mu_2^{-1}(\Theta) / (\sum_{i=1}^{n} \frac{1}{\mu_2(\Theta)}).$$

Применение рассмотренного подхода предполагает знание дисперсий оценок, которые, как правило, не известны. Замена дисперсии ее выборочной оценкой приводит к трудно оцениваемому смещению величины Θ^*. Преодоление данного недостатка возможно на основе объединения выборок с учетом доли каждой выборки в общем объеме имеющихся сведений, т.е. коэффициенты v_i характеризуют относительный вклад каждого слоя в общую оценку. Значение коэффициента v_i можно определить как отношение объема данной выборки к общему объему всех. Линейное объединение оценок приводит к их усреднению по всем выборкам. Иначе говоря, значение некоторого показателя в данном случае следует рассматривать как среднее значение случайной величины, принимающей значение Θ_i с вероятностью v_i.

Пример 5.1. По результатам наблюдения за пропускной способностью канала в различные дни испытаний сформированы упорядоченные выборки (табл. 5.1). При уровне значимости $\alpha = 0{,}05$ необходимо проверить однородность выборок.

Решение. Возьмем в качестве исходной выборку *X*, соответствующую первому дню испытаний, и проверим ее на однородность с выборкой *Y*, составленной из результатов второго дня испытаний. Перечислим последовательность элементов в общем вариационном ряду, составленном из элементов первой и второй выборок: *yyxyxyxxyxyx*.

Таблица 5.1

День испытаний	Пропускная способность						
	1	2	3	4	5	6	7
1	259,14	260,06	260,97	262,43	267,83	273,14	—
2	253,68	258,14	259,49	260,18	263,65	271,39	274,12
3	256,36	259,36	262,84	265,94	270,33	270,44	271,63

Сумма порядковых номеров вариант первого дня испытаний $(n_1 < n_2)$ составит

$$u = 3 + 5 + 7 + 8 + 10 + 12 = 45.$$

Количество элементов в обеих выборках меньше 25, поэтому следует воспользоваться распределением Вилкоксона для проверки гипотезы H_0 об однородности выборок. Значение нижней критической точки для двусторонней критической области при заданном уровне $\alpha / 2 = 0{,}025$, количестве наблюдений $n_1 = 6$, $n_2 = 7$ определим по табл. П.5. Оно составит

$$u_{\text{н}} = 27.$$

Значение верхней критической точки распределения равно

$$u_{\text{в}} = (n_1 + n_2 + 1)n_1 - u_{\text{н}} = (6 + 7 + 1)6 - 27 = 57.$$

Значение величины u превышает $u_{\text{н}}$ и меньше $u_{\text{в}}$, поэтому нет оснований отвергать нулевую гипотезу об однородности выборок. Обозначим объединенную выборку через *X*.

Проверим однородность объединенной выборки *X* и результатов третьего дня наблюдений *W*. Построим общий вариационный ряд из элементов выборки *X* и выборки *W*:

xwxxwxxxxxwxwxwwxwxx.

Сумма порядковых номеров вариант третьего дня испытаний (этих вариант меньше, чем в объединенном ряду *X*) составит

$$u = 2 + 5 + 11 + 13 + 15 + 16 + 18 = 80.$$

Воспользуемся распределением Вилкоксона и определим при уровне значимости $\alpha / 2 = 0,025$, $n_1 = 7$, $n_2 = 13$ нижнюю критическую точку $u_{\text{н}} = 48$ (табл. П.5). Верхняя критическая точка

$$u_{\text{в}} = (7 + 13 + 1)7 - 48 = 99.$$

В соответствии с выбранным критерием нет оснований отвергать нулевую гипотезу, следовательно, все три выборки однородны и их можно объединить в одну.

6. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

6.1. Задачи дисперсионного анализа

При исследовании однотипных величин возникают задачи их сравнения. Сравнение случайных величин производится путем сопоставления законов распределения или их моментов.

Законы распределения можно сопоставить на основе критерия Вилкоксона при нулевой гипотезе H_0 о равенстве законов распределения двух случайных величин $F_x = F_y$ и конкурирующей гипотезе H_1 в виде $F_x < F_y$ или $F_x > F_y$. В этих случаях критическая область является односторонней. Поэтому нижнюю критическую точку и квантиль распределения находят при уровне значимости α. Содержание остальных этапов проверки гипотез сохраняется. Следует отметить принятие гипотезы H_1 о том, что

$$F_x < F_y \text{ означает } X > Y.$$

Действительно, неравенство $F_x(x) < F_y(x)$ равносильно неравенству

$$P(X < x) < P(Y < x),$$

следовательно, $X > Y$.

Аналогично, если справедлива гипотеза $F_x > F_y$, то $X < Y$.

Вполне естественно сопоставление случайных величин на основе моментов проводить путем сравнения их математических ожиданий. *Однофакторный дисперсионный анализ* позволяет установить, оказывает ли существенное влияние некоторый фактор Ф, который имеет несколько уровней, на исследуемую случайную величину.

Задача сравнения выборок случайных величин формулируется следующим образом.

Имеются результаты наблюдений в виде совокупности слоев типа (5.1), задан уровень значимости α для проверки статистической гипотезы. В данном случае отдельные слои трактуются как выборки одной и той же случайной величины, полученные по результатам наблюдения за одним объектом при различных значениях фактора Ф (количество уровней фактора равно *m*).

Требуется проверить нулевую гипотезу H_0 о равенстве математических ожиданий случайных величин всех выборок.

Допущения: генеральные совокупности, соответствующие каждому слою, распределены *нормально*; *дисперсии* слоев *одинаковы*.

Основная идея дисперсионного анализа состоит не в сопоставлении математических ожиданий случайных величин, а в сравнении оценки «*факторной дисперсии*», порождаемой воздействием фактора, и оценки «*остаточной дисперсии*», обусловленной случайными причинами. Если различие между этими оценками значимо, то фактор оказывает существенное

влияние на случайную величину, в противном случае влияние фактора несущественно.

Дисперсионный анализ выполняется поэтапно. Такими этапами являются следующие:

- проверка выборок на принадлежность к нормальному закону распределения. Этап необходим, когда нет априорной информации о законах распределения слоев. Если принадлежность нормальному закону не подтвердится, то аппарат дисперсионного анализа, вообще говоря, применять нельзя. Некоторые исследователи допускают его применение при больших объемах выборок (объем каждой выборки должен быть не менее 30) независимо от вида закона распределения;
- проверка равенства оценок дисперсий во всех слоях выборки (проверка однородности дисперсий). Если однородность не подтвердится, то методы дисперсионного анализа не применимы;
- вычисление оценки факторной и остаточной дисперсии;
- сравнение средних значений величин методом дисперсионного анализа и формирование выводов по результатам сравнения.

6.2. Проверка однородности совокупности дисперсий

Для каждого слоя вычисляется несмещенная оценка дисперсии, обозначим эти оценки через $S_0^2(x), S_0^2(y), \ldots, S_0^2(w)$ соответственно. Числа степеней свободы этих оценок

$$k_1 = n_1 - 1, \quad k_2 = n_2 - 1, \ldots, \quad k_m = n_w - 1.$$

Гипотеза H_0 состоит в том, что выборки, по которым определены оценки дисперсии, получены из генеральных совокупностей, обладающих одинаковыми дисперсиями:

$$S_0^2(x) = S_0^2(y) = \ldots = S_0^2(w) = S_0^2,$$

при этом величина дисперсии S_0^2 остается неизвестной. Следует выяснить, являются ли величины $S_0^2(x), S_0^2(y), \ldots, S_0^2(w)$ оценками одной и той же генеральной дисперсии μ_2.

Рассмотрим сначала случай, когда объем выборок по слоям хотя бы частично различается. В такой ситуации применяется критерий однородности Бартлетта. Проверка однородности реализуется в несколько шагов.

Вычисляется усредненная оценка несмещенной дисперсии по всем слоям:

$$S_0^2 = \sum_{i=1}^{m} k_i \cdot S_0^2(i) / k, \qquad k = \sum_{i=1}^{m} k_i, \tag{6.1}$$

где $S_0^2(i)$ – несмещенная оценка дисперсии для слоя i.

Рассчитывается значение критерия

$$B = \frac{2{,}303\left[k \lg \mu_2 - \sum_{i=1}^{m} k_i \lg \mu_2(i)\right]}{1 + \frac{1}{3(m-1)}\left[\sum_{i=1}^{m} \frac{1}{k_i} - \frac{1}{k}\right]} \qquad (6.2)$$

Бартлетт установил, что случайная величина B при условии справедливости нулевой гипотезы распределена приближенно как хи-квадрат с $m-1$ степенями свободы, если все n_i больше трех. По заданному уровню значимости α, числу степеней свободы $m-1$ для правосторонней критической области определяется критическое значение $c^2_{\text{кр}}(m-1;\alpha)$. Если соблюдается условие

$$B < c^2_{\text{кр}}(m-1;\alpha),$$

то нет оснований отвергнуть нулевую гипотезу. Если $B > c^2_{\text{кр}}(m-1;\alpha)$, то нулевая гипотеза отвергается. Критерий Бартлетта чувствителен к отклонениям распределения от нормального, поэтому к результатам сравнения следует относится осторожно, а при одинаковом объеме всех слоев вместо критерия Бартлетта лучше применять критерий Кочрена (Кохрена).

Итак, если $k_1 = k_2 = \ldots = k_m$, то применяется *критерий Кочрена*:

$$G = S^2_{0\max} \Big/ \sum_{i=1}^{m} S^2_0(i), \qquad (6.3)$$

где S^2_0 – максимальная оценка дисперсии по всем слоям.

Критическая область для критерия Кочрена правосторонняя. Критическую точку $G_{\text{кр}}(k_1, m;\alpha)$ находят по таблице распределения Кочрена, (см. приложение). Критическая область определяется неравенством $G > G_{\text{кр}}(k_1, m;\alpha)$.

6.3. Сравнение факторной и остаточной дисперсий

Пусть все выборки (5.1) характеризуют одну случайную величину X при различных значениях фактора Ф, т.е. каждый слой соответствует одному количественному или качественному значению фактора. Сравнение дисперсий производится в следующем порядке:

- рассчитывается среднее значение (оценка математического ожидания) по всей совокупности наблюдений:

$$m^* = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij},$$

где $n = n_1 + n_2 + \ldots + n_m$, а x_{ij} – j-й элемент i-го слоя;

- вычисляются средние значения для всех слоев (групп):

$$m_{\text{гр}i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad i = 1, ..., m;$$

- определяется общая сумма квадратов отклонений наблюдаемых значений от оценки математического ожидания:

$$S_{\text{общ}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - m^*)^2; \tag{6.4}$$

- определяется факторная сумма квадратов отклонений средних по слоям от оценки математического ожидания (характеризует рассеяние между слоями):

$$S_{\text{факт}} = \sum_{i=1}^{m} n_i (m_{\text{гр}i} - m^*)^2; \tag{6.5}$$

- определяется остаточная сумма квадратов отклонений наблюдаемых значений внутри слоя от своей средней:

$$S_{\text{ост}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(m^*_{\text{гр}i} - x_{ij})^2. \tag{6.6}$$

Величина $S_{\text{факт}}$ характеризует влияние фактора Ф. Это положение можно пояснить следующим образом. Пусть фактор оказывает существенное влияние на величину X. Тогда результаты наблюдения для одного слоя, вообще говоря, отличаются от результатов, представленных в других слоях. Следовательно, различаются и средние значения по слоям, причем они тем больше отличаются от оценки математического ожидания по всей выборке, чем больше проявляется влияние фактора. Таким образом, сумма квадратов отклонений средних по слоям от общей средней и характеризует влияние фактора (возведение отклонений во вторую степень исключает взаимную компенсацию положительных и отрицательных отклонений).

Наблюдения внутри одного слоя различаются из-за воздействия случайных причин. Именно сумма квадратов отклонений наблюдаемых значений в каждом слое от среднего значения в слое и характеризует воздействие этих причин, т.е. величина $S_{\text{ост}}$ отражает суммарное влияние случайных причин на значение величины X.

Величина $S_{\text{общ}}$, как сумма квадратов отклонений конкретных значений от среднего значения, характеризует суммарное влияние фактора и случайных причин. Можно показать, что

$$S_{\text{общ}} = S_{\text{ост}} + S_{\text{факт}},$$

тогда для вычисления остаточной суммы квадратов можно воспользоваться более простым соотношением:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}.$$

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим оценки общей, факторной и остаточной дисперсий:

$$S^2_{0\ \text{общ}} = \frac{S_{\text{общ}}}{n-1}; \quad S^2_{0\ \text{факт}} = \frac{S_{\text{факт}}}{m-1}; \quad S^2_{0\ \text{ост}} = \frac{S_{\text{ост}}}{n-m}. \tag{6.7}$$

Если средние значения случайной величины, вычисленные по отдельным выборкам одинаковы, то оценки факторной и остаточной дисперсий являются несмещенными оценками генеральной дисперсии и различаются несущественно. Тогда сопоставление оценок этих дисперсий по критерию Р. Фишера

$$F = S^2_{0\ \text{факт}} / S^2_{0\ \text{ост}}$$

должно показать, что нулевую гипотезу о равенстве факторной и остаточной дисперсий отвергнуть нет оснований. Если $\mu_{2\,\text{факт}} < \mu_{2\,\text{ост}}$, то нет необходимости прибегать к вычислению критерия Р. Фишера – из неравенства сразу следует вывод о выполнении нулевой гипотезы. Итак, из справедливости гипотезы о равенстве средних величин по группам следует соблюдение гипотезы о равенстве факторной и остаточной дисперсий.

Если нулевая гипотеза о равенстве средних величин по слоям является ложной, то с увеличением расхождения между слоями возрастает оценка факторной дисперсии, а вместе с ней и величина критерия $F = \mu_{2\,\text{факт}} / \mu_{2\,\text{ост}}$. В результате значение F превысит критическое значение, и гипотеза о равенстве дисперсий будет отвергнута.

Рассуждая от противного, можно доказать справедливость утверждений: из справедливости (ложности) гипотезы о дисперсиях следует истинность (ложность) гипотезы о математических ожиданиях. Таким образом, вместо проверки нулевой гипотезы H_0 о равенстве средних значений для совокупности выборок следует проверить гипотезу о равенстве факторной и остаточной дисперсий.

Пример 6.1. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних значений по слоям применительно к результатам наблюдений (табл. 5.1). Предполагается, что выборки принадлежат нормальному распределению, а каждый слой соответствует некоторому значению фактора Ф.

Решение. Необходимо проверить однородность дисперсий, а затем непосредственно провести дисперсионный анализ. Проверим гипотезу об однородности дисперсий. Для этого вычислим:

- оценки математического ожидания по слоям (групповые средние)
$$m_{\text{гр}1} = 263,93; \quad m_{\text{гр}2} = 262,95; \quad m_{\text{гр}3} = 265,32;$$
- несмещенные оценки дисперсии по слоям
$$m_2^*(1) = 29,79; \quad m_2^*(2) = 54,20; \quad m_2^*(3) = 34,61;$$
- усредненную оценку несмещенной дисперсии по всем слоям
$$m_2^* = (29,79 \times 5 + 54,20 \times 6 + 34,61 \times 6) / 17 = 40,11.$$
- значение критерия Бартлетта

$$B = a / c = 0,56 / 1,08 = 0,52,$$

где $a = 2,303 \cdot (17 \lg 40,11 - (5 \cdot \lg 29,79 + 6 \cdot \lg 54,20 + 6 \cdot \lg 34,61)) = 0,56;$

$c = 1 + (1/5 + 1/6 + 1/6 - 1/17) / [3(3-1)] = 1,08.$

Критическое значение хи-квадрат для правосторонней области

$$c^2_{\text{кр}}(2; 0,05) = 6,0.$$

Поскольку величина B меньше $c^2_{\text{кр}}(2;0,05)$, отвергнуть нулевую гипотезу об однородности дисперсий нет оснований.

Дисперсионный анализ предусматривает вычисление:

- суммы квадратов

$$S_{\text{общ}} = 701,65; \quad S_{\text{факт}} = 19,81; \quad S_{\text{ост}} = 681,84;$$

- оценок дисперсий

$$\mu^*_{2\,\text{общ}} = 701,65 / 19 = 36,93; \quad \mu^*_{2\,\text{факт}} = 19,81 / 2 = 9,91; \quad \mu^*_{2\,\text{ост}} = 681,84 / 17 = 40,10.$$

Оценка факторной дисперсии меньше оценки остаточной дисперсии, поэтому можно сразу утверждать справедливость нулевой гипотезы о равенстве математических ожиданий по слоям выборки. Иначе говоря, в данном примере фактор Ф не оказывает существенного влияния на случайную величину.

7. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

7.1. Матрица данных

Многие объекты исследования характеризуются множеством параметров, и по результатам наблюдения за их функционированием формируются многомерные совокупности (матрицы) ЭД:

$$X = \begin{vmatrix} x_{11} & x_{12} & . & x_{1m} \\ x_{21} & x_{22} & . & x_{2m} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nm} \end{vmatrix}. \qquad (7.1)$$

Строки такой матрицы соответствуют результатам регистрации всех наблюдаемых параметров объекта в одном эксперименте, а столбцы содержат результаты наблюдений за одним параметром (фактором, вариантой) во всех экспериментах. Обозначим количество параметров через m $(m>1)$, а количество наблюдений – через n. В матрице элемент x_{ij} соответствует значению j-й варианты в i-м наблюдении.

Каждый столбец матрицы представляет собой случайную выборку значений одного параметра объекта.

Объектом исследования в многомерном анализе является многомерная случайная величина, представленная выборкой конечного объема. К такой выборке применимы все методы и оценки, рассмотренные при обработке одномерных ЭД. Параметры, характеризующие объект исследования, имеют разный физический смысл, и матрица данных существенно изменяется, если изменяются шкалы, в которых измеряются те или иные параметры. Матрицу данных целесообразно привести к стандартному виду, т.е. стандартизовать значения вариант (напомним, что среднее значение стандартизованной варианты равно нулю, дисперсия – единице). Стандартизованную матрицу будем обозначать через U. Переход от исходной к стандартизованной матрице осуществляется следующим образом.

1. По каждой варианте вычисляются оценки:
   - математического ожидания $m^*(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$;
   - дисперсии $\mu_2(x_j) = \sigma^2(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - m^*(x_i))^2$.
2. Вычисляются элементы стандартизованной матрицы

$$u_{ij} = (x_{ij} - m^*(x_j)) / \sigma(x_j), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m.$$

Элементы матрицы U являются безразмерными величинами. Именно матрица U будет являться объектом последующей обработки.

7.2. Корреляционный анализ

Величины, характеризующие различные свойства объектов, могут быть независимыми или взаимосвязанными. Различают два вида зависимостей между величинами (факторами): функциональную и статистическую.

При функциональной зависимости двух величин значению одной из них обязательно соответствует одно или несколько точно определенных значений другой величины.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями.

Более важным частным случаем статистической зависимости является корреляционная зависимость, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения.

При исследовании зависимости между одной величиной и такими характеристиками другой, как, например, моменты старших порядков (а не среднее значение), эта связь будет называться статистической, а не корреляционной.

Корреляционная связь описывает следующие виды зависимостей:

- причинную зависимость между значениями параметров. Примером такой зависимости является взаимосвязь пропускной способности канала передачи данных и соотношения сигнал/шум (на пропускную способность влияют и другие факторы – характер помех, амплитудно-частотные характеристики канала, способ кодирования сообщений и др.). Установить однозначную связь между конкретными значениями указанных параметров не удается;
- «зависимость» между следствиями общей причины. Подобная зависимость характерна, в частности, для скорости и безошибочности набора текста оператором (указанные факторы зависят от квалификации оператора).

Корреляционная зависимость определяется различными параметрами, среди которых наибольшее распространение получили показатели, характеризующие взаимосвязь двух случайных величин (парные показатели): корреляционный момент, коэффициент корреляции.

Оценка *корреляционного момента (коэффициента ковариации*) двух вариант x_j и x_k вычисляется по исходной матрице X:

$$K^*_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - m^*(x_j))\cdot(x_{ik} - m^*(x_k)). \qquad (7.2)$$

Коэффициент ковариации r_{jk} нормированных случайных величин называют коэффициентом корреляции, его оценка

$$r^*_{jk} = \frac{1}{n}\sum_{i=1}^{n} u_{ij}\cdot u_{ik} = \frac{\sum_{i=1}^{n}(x_{ij} - m^*(x_j))\cdot(x_{ik} - m^*(x_k))}{n\cdot S_j \cdot S_k}. \qquad (7.3)$$

Значение коэффициента корреляции лежит в пределах от –1 до +1. Если случайные величины X_j и X_k независимы, то коэффициент r_{jk} обязательно равен нулю, обратное утверждение неверно. Коэффициент r_{jk} характеризует значимость линейной связи между параметрами:

- при $r_{jk}=1$ значения x_{ij} и x_{ik} полностью совпадают, т.е. значения параметров принимают одинаковые значения. Иначе говоря, имеет место функциональная зависимость: зная значение одного параметра, можно однозначно указать значение другого параметра;
- при $r_{jk}=-1$ величины x_{ij} и x_{ik} принимают противоположные значения. И в этом случае имеет место функциональная зависимость;
- при $r_{jk}=0$ величины x_{ij} и x_{ik} практически не связаны друг с другом линейным соотношением. Это не означает отсутствия каких-то других (например нелинейных) связей между параметрами;
- при $\left|r_{jk}\right|>0$ и $\left|r_{jk}\right|<1$ однозначной линейной связи величин x_{ij} и x_{ik} нет. И чем меньше абсолютная величина коэффициента корреляции, тем в меньшей степени по значениям одного параметра можно предсказать значение другого.

Используя понятие коэффициента корреляции, матрице ЭД можно поставить в соответствие квадратную матрицу оценок коэффициентов корреляции (корреляционную матрицу)

$$r^*=\begin{vmatrix} r^*_{11} & r^*_{12} & \cdots & r^*_{1m} \\ r^*_{21} & r^*_{22} & \cdots & r^*_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r^*_{n1} & r^*_{n2} & \cdots & r^*_{nm} \end{vmatrix}. \qquad (7.4)$$

К числу характерных свойств корреляционной матрицы относят: симметричность относительно главной диагонали, $r^*_{jk}=r^*_{kj}$; единичные значения элементов главной диагонали, $r_{kk}=1$ (r_{kk} соответствует дисперсии стандартизованного параметра X_k).

Оценка коэффициента корреляции, вычисленная по ограниченной выборке, практически всегда отличается от нуля. Но из этого еще не следует, что коэффициент корреляции генеральной совокупности также отличен от нуля. Требуется оценить значимость выборочной величины коэффициента или в соответствии с постановкой задач проверки статистических гипотез проверить гипотезу о равенстве нулю коэффициента корреляции. Если гипотеза H_0 о равенстве нулю коэффициента корреляции будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением. Если гипотеза H_0 будет принята, то оценка коэффициента не значима, и величины линейно не связаны друг с другом.

Проверка гипотезы о значимости оценки коэффициента корреляции требует знания распределения этой случайной величины. Распределение величины r_{ik} изучено только для частного случая, когда случайные величины U_j и U_k распределены по нормальному закону.

В качестве критерия проверки нулевой гипотезы H_0 применяют случайную величину

$$t = \left| r_{ik}^* \right| \frac{\sqrt{n-2}}{\sqrt{1-r_{ik}^{*2}}}.$$

Если модуль коэффициента корреляции относительно далек от единицы, то величина t при справедливости нулевой гипотезы распределена по закону Стьюдента с $n-2$ степенями свободы. Конкурирующая гипотеза H_1 соответствует утверждению, что значение r_{ik} не равно нулю (больше или меньше нуля). Поэтому критическая область двусторонняя.

Проверка гипотезы H_0 о равенстве нулю генерального коэффициента парной корреляции двумерной нормально распределенной случайной величины осуществляется в следующей последовательности:

- вычисляется значение статистики t;
- при уровне значимости α для двусторонней области определяется критическая точка распределения Стьюдента $t_{\text{кр}}(n-2;\alpha)$;
- сравнивается значение статистики t с критическим значением $t_{\text{кр}}(n-2;\alpha)$. Если $t < t_{\text{кр}}(n-2;\alpha)$, то нет оснований отвергнуть нулевую гипотезу, иначе гипотеза H_0 отвергается (коэффициент корреляции значим).

Когда модуль величины r_{ik}^* близок к единице, распределение r_{ik}^* отличается от распределения Стьюдента, так как значение $\left| r_{ik}^* \right|$ ограничено справа единицей. В этом случае применяют преобразование

$$y_{ik} = 0,5 \cdot \ln[(1+\left| r_{ik} \right|) / (1-\left| r_{ik} \right|)].$$

Величина y_{ik} не имеет указанного ограничения, она при $n > 10$ распределена приблизительно нормально с центром

$$m_1^*(r_{ik}) = 0,5 \cdot \ln[(1+\left| r_{ik} \right|) / (1-\left| r_{ik} \right|)] + 0,5 \cdot \left| r_{ik} \right| / (n-1)$$

и дисперсией

$$S_0^2(r_{ik}) = 1 / (n-3).$$

Если значение центрированной и нормированной величины

$$(y_{ik} - m^*(r_{ik})) / S_0^2(r_{ik})$$

превышает значение квантили уровня $1-\alpha/2$ нормального распределения стандартизованной величины, то нулевая гипотеза отвергается.

Таким образом, *постановка задачи линейного корреляционного анализа* формулируется в следующем виде.

Имеется матрица наблюдений вида (7.1).

Необходимо определить оценки коэффициентов корреляции для всех или только для заданных пар параметров и оценить их значимость. Незначимые оценки приравниваются к нулю.

Допущения:

- выборка имеет достаточный объем. Минимально допустимым считается объем, когда количество наблюдений не менее чем в 5–6 раз превосходит количество факторов;
- выборки по каждому фактору являются однородными. Это допущение обеспечивает несмещенную оценку средних величин;
- матрица наблюдений не содержит пропусков.

Если необходима проверка значимости оценки коэффициента корреляции, то требуется соблюдение дополнительного условия – распределение вариант должно подчиняться нормальному закону.

Задача анализа решается в несколько этапов:

- проводится стандартизация исходной матрицы;
- вычисляются парные оценки коэффициентов корреляции;
- проверяется значимость оценок коэффициентов корреляции, незначимые оценки приравниваются к нулю. По результатам проверки делается вывод о наличии связей между вариантами (факторами).

Пример 7.1. Результаты наблюдений за характеристиками канала представлены в табл. 7.1.

Таблица 7.1

Номер варианта	Пропускная способность	Отношение сигнал/шум	Остаточное затухание		
			1020	1800	1400
	X_1	X_2	X_3	X_4	X_5
1	26,37	41,98	17,66	16,05	22,85
2	28,00	43,83	17,15	15,47	23,25
3	27,83	42,83	15,38	17,59	24,55
4	31,67	47,28	18,39	16,92	26,59
5	23,50	38,75	18,32	15,66	26,22
6	21,04	35,12	17,81	17,00	27,52
7	16,94	32,07	21,42	16,77	25,76
8	37,56	54,25	26,42	15,68	23,10
9	18,84	32,70	17,23	15,92	23,41
10	25,77	40,51	30,43	15,29	25,17
11	33,52	49,78	21,71	15,61	25,39
12	28,21	43,84	28,33	15,70	24,56
13	28,76	44,03	30,42	16,87	24,45
14	24,60	39,46	21,66	15,25	23,81
15	24,51	38,78	25,77	16,05	24,48

Необходимо определить наличие линейных корреляционных связей между пропускной способностью и остальными факторами. Предполагается, что выборки по всем вариантам подчиняются нормальному закону. Проверку гипотезы о значимости оценок коэффициентов корреляции произвести с уровнем значимости α, равным 0,1.

Решение. Стандартизация исходной матрицы начинается с вычисления выборочной средней m_1^*, несмещенной оценки дисперсии μ_2^* и среднеквадратичного отклонения S по каждой варианте (табл. 7.2).

Таблица 7.2

Оценка параметра распределения	Варианта				
	X_1	X_2	X_3	X_4	X_5
M_1	26,47	41,68	21,87	16,12	24,74
M_2	29,10	36,47	26,37	0,52	1,88
S	5,39	6,04	5,13	0,72	1,37

В результате перехода к величинам $u_{ij} = (x_{ij} - \mu_j)/\sigma_j$ формируется стандартизованная матрица исходных данных (табл. 7.3).

Таблица 7.3

Номер варианта	Пропускная способность	Отношение сигнал/шум	Остаточное затухание на частоте		
			1020	1800	1400
	U_1	U_2	U_3	U_4	U_5
1	-0,02	-0,05	-0,82	-0,10	-1,38
2	0,28	0,36	-0,92	-0,90	-1,09
3	0,25	0,19	-1,26	2,03	-0,14
4	0,96	0,93	-0,68	1,10	2,35
5	-0,55	-0,49	-0,69	-0,64	1,08
6	-1,01	-1,09	-0,79	1,21	2,03
7	-1,77	-1,59	-0,09	0,90	0,74
8	2,06	2,08	0,89	-0,61	-1,20
9	-1,42	-1,49	-0,90	-0,28	-0,97
10	-0,13	-0,19	1,67	-1,15	0,31
11	1,31	1,34	-0,03	-0,71	0,47
12	0,32	0,36	1,26	-0,58	-0,13
13	0,42	0,39	1,66	1,03	-0,21
14	-0,35	-0,37	-0,04	-1,21	-0,68
15	-0,36	-0,48	0,76	-0,19	-0,19

Оценки коэффициентов корреляции

$$r_{1k}^* = \frac{1}{15}\sum_{i=1}^{15} u_{1i}\, u_{ki}, \quad (k = 2, 3, 4)$$

представлены в табл. 7.4. В этой же таблице приведены значения статистик критерия Стьюдента $t = \left|r_{1k}^{*}\right| \sqrt{n-2} \Big/ \sqrt{1-r_{1k}^{*2}}$ для вычисленных оценок коэффициентов корреляции при $n = 15$.

Таблица 7.4

	X_2	X_3	X_4	X_5
R_{1j}	0,93	0,25	-0,13	–0,22
T	9,12	0,93	0,47	0,81

Критическое значение

$$t_{\text{кр}}(n-2;\alpha) = t_{\text{кр}}(13;0{,}1) = 1{,}77.$$

Статистика критерия больше критического значения только для r_{12}. Это означает, что только для указанного коэффициента оценка значима (коэффициент корреляции генеральной совокупности не равен нулю), а остальные коэффициенты следует признать равными нулю.

Корреляционная зависимость необязательно устанавливается только для двух величин, с ее помощью можно анализировать связи между несколькими вариантами (множественная корреляция). А кроме линейной существуют и другие виды корреляции.

ЗАДАЧИ ПО РАЗДЕЛАМ

1

1.1. Интервал движения поездов метро составляет 2 минуты. В табл. 1.1 приведены значения случайной величины X – времени ожидания поезда. Составить вариационный ряд, гистограммы распределения равноинтервальным и равновероятностным методами.

Таблица 1.1

0,000	0,002	0,007	0,025	0,089	0,312	1,068	1,604	0,014	0,045
1,747	1,677	0,341	0,952	0,645	1,297	1,981	0,214	1,452	0,787
1,654	0,838	0,143	1,317	0,618	1,853	1,555	0,653	1,922	1,653
0,617	0,828	1,413	1,030	1,459	1,483	1,769	1,265	1,669	0,635
0,787	1,004	0,941	0,612	1,200	1,692	1,356	0,908	1,245	1,295

1.2. В табл. 1.2 приведены численные значения промежутков времени Δt (в минутах) между появлениями клиентов в некотором банке. Построить гистограммы распределения равноинтервальным и равновероятностным методами.

Таблица 1.2

0,000	0,002	0,007	0,025	0,091	0,339	1,527	3,239	0,014	3,457
4,134	3,647	0,374	1,293	0,778	2,091	9,344	0,226	2,590	1,000
3,507	1,086	0,148	2,150	0,740	5,223	3,007	0,791	6,492	3,502
0,738	1,069	2,453	1,447	2,614	2,706	4,314	2,001	3,600	0,764
1,000	1,394	1,272	0,730	1,832	3,742	2,267	1,211	1,949	2,086

1.3. В табл. 1.3 приведены значения прибыли 50 фирм Q (1000 усл. ед.). Составить вариационный ряд, гистограммы распределения равноинтервальным и равновероятностным методами.

Таблица 1.3

4,744	9,127	7,201	8,650	11,536	9,013	10,255	10,390	9,268	7,354
0,232	15,103	11,902	10,216	11,470	10,954	6,739	12,697	13,084	6,088
14,593	8,671	14,227	15,190	9,202	11,047	9,124	7,351	9,832	12,271
7,126	10,744	9,715	5,536	8,917	9,823	8,383	9,766	10,687	10,582
11,245	5,854	10,387	2,917	6,739	6,748	10,954	11,101	7,024	11,587

1.4. В табл. 1.4 приведены значения промежутков времени t (в минутах) между вызовами такси. Составить вариационный ряд, гистограммы распределения равноинтервальным и равновероятностным методами.

Таблица 1.4

0,000	0,000	0,000	0,003	0,011	0,042	0,191	0,405	0,002	0,432
0,517	0,456	0,047	0,162	0,097	0,261	0,168	0,028	0,324	0,125
0,438	0,136	0,019	0,269	0,092	0,653	0,376	0,099	0,812	0,438
0,092	0,134	0,307	0,181	0,327	0,338	0,539	0,250	0,450	0,096
0,125	0,174	0,159	0,091	0,229	0,468	0,283	0,151	0,244	0,261

2

2.1. Для экспериментальных данных, приведенных в табл. 1.1, найти выборочное среднее $\overline{x}$, смещенную $\overline{S}$ и несмещенную $\overline{S_0}$ оценки дисперсии, начального момента четвертого порядка α_4и центрального момента третьего порядка μ_3.

2.2. Для экспериментальных данных, приведенных в табл. 1.2, найти выборочное среднее $\overline{x}$, смещенную $\overline{S}$ и несмещенную $\overline{S_0}$ оценки дисперсии, начального α_3и центрального μ_3 моментов.

2.3. Для экспериментальных данных, приведенных в табл. 1.3, найти выборочное среднее $\overline{x}$, смещенную $\overline{S}$ и несмещенную $\overline{S_0}$ оценки дисперсии, начального α_2и центрального μ_2 моментов.

2.4. Для экспериментальных данных, приведенных в табл. 1.4, найти выборочное среднее $\overline{x}$, смещенную $\overline{S}$ и несмещенную $\overline{S_0}$ оценки дисперсии, начального α_4и центрального μ_1 моментов.

3

3.1. Случайная величина *X* характеризуется данными, приведенными в таблице. Установить на уровне доверия $\gamma = 0{,}95$, что случайная величина *X* имеет логарифмически нормальное распределение. Использовать критерий χ^2 Пирсона.

Интервалы СВ *X*	90–100	100–110	110–120	120–130	130–140
m_i	10	160	100	60	20

3.2. Используя критерий Мизеса, показать, что случайная величина *X* распределена по показательному закону, приняв $\alpha = 0{,}1$.

0,001	0,001	0,003	0,012	0,046	0,169	0,763	1,620	0,007	1,728
2,067	1,824	0,187	0,646	0,389	1,046	4,672	0,113	1,295	0,500
1,754	0,543	0,074	1,075	0,370	2,612	1,504	0,396	3,245	1,751
0,369	0,534	1,227	0,724	1,307	1,353	2,157	1,000	1,800	0,382
0,500	0,697	0,636	0,365	0,916	1,871	1,134	0,606	0,975	1,043

3.3. Показать, что случайная величина *X*, приведенная в таблице, имеет равномерное распределение на отрезке [–1,1]. Принять $\alpha = 0{,}01$. Использовать критерий Колмогорова.

–1,000	–0,998	–0,993	–0,975	–0,911	–0,688	0,068	0,604	–0,986	0,645
0,747	–0,162	-0,857	0,317	–0,382	0,853	0,555	–0,347	0,922	0,653
–0,383	–0,172	0,414	0,030	0,459	0,483	0,769	0,265	0,669	–0,365

3.4. Проверить гипотезу о нормальности распределения случайной величины Q, выборочные значения которой приведены в таблице, с помощью критерия χ^2 Пирсона при $\gamma = 0{,}99$.

4,744	9,127	7,201	8,650	11,536	9,013	10,255	10,390	9,268	7,354
0,232	15,103	11,902	10,216	11,470	10,954	6,739	12,697	13,084	6,088
14,593	8,671	14,227	15,190	9,202	11,047	9,124	7,351	9,832	12,271
7,126	10,744	9,715	5,536	8,917	9,823	8,383	9,766	10,687	10,582
11,245	5,854	10,387	2,917	6,739	6,748	10,954	11,101	7,024	11,587

4

4.1. Сколько лиц в возрасте от 19 до 24 лет надо опросить, чтобы установить средний процент студентов с точностью до 0,5 %?

4.2. По данным 10 измерений некоторой величины d найдено среднее $\overline{d} = 20$ и выборочная дисперсия $\overline{S_d^2} = 25$. Найти границы, в которых с вероятностью 0,99 заключено истинное значение величины.

4.3. Найти с вероятностью 0,99 доверительный интервал для дисперсии генеральной совокупности в задаче 4.2.

4.4. При выборке в 250 человек оказалось, что 105 из них поддерживают определенного кандидата в парламент. Сколько человек следует опросить, чтобы с вероятностью 0,99 можно было бы утверждать, что доля избирателей, поддерживающих данного кандидата, отличается от истинной доли не более чем на 0,05?

5

5.1. При уровне значимости 0,01 проверить нулевую гипотезу H_0: $F_1(x) = F_2(x)$ об однородности двух выборок, объемы которых $n_1 = 6$, $n_2 = 7$ (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки):

x_i : {3, 4, 6, 10, 13, 17}

y_i : {1, 2, 5, 7, 16, 20, 22}

Принять в качестве конкурирующей гипотезу H_1: $F_1(x) \neq F_2(x)$.

5.2. При уровне значимости 0,05 проверить нулевую гипотезу об однородности двух выборок объемов: $n_1 = 40$ и $n_2 = 50$ при конкурирующей гипотезе H_1: $F_1(x) \neq F_2(x)$, если известно, что в общем вариационном ряду, составленном из вариант обеих выборок, сумма порядковых номеров вариант первой выборки $W_{\text{набл}} = 1800$.

5.3. Предложены два метода (А и В) увеличения выхода продукции. При уровне значимости 0,05 проверить нулевую гипотезу об их одинаковой эффективности по двум выборкам объемов $n_1 = 6$, $n_2 = 9$ (в первой строке приведены проценты прироста продукции в каждом опыте по методу А, во второй строке – по методу В):

x_i : {0,2 0,3 0,5 0,8 1,0 1,3}

y_i : {0,1 0,4 0,6 0,7 0,9 1,4 1,7 1,8 1,9}

Принять в качестве конкурирующей гипотезу: эффективность методов А и В различна.

5.4. При уровне значимости 0,01 проверить нулевую гипотезу об однородности двух выборок объемов: $n_1 = 40$ и $n_2 = 60$ при конкурирующей гипотезе H_1: $F_1(x) \neq F_2(x)$, если известно, что во общем вариационном ряду, составленном из вариант обеих выборок, сумма порядковых номеров вариант первой выборки $W_{\text{набл}} = 3020$.

6

6.1. Произведено по четыре испытания на каждом из трех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

Номер испытания i	Уровни факторов		
	F_1	F_2	F_3
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34
$\overline{x}_{\text{гр j}}$	35	25	27

6.2. Произведено по пять испытаний на каждом из четырех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

Номер испытания i	Уровни факторов			
	F_1	F_2	F_3	F_4
1	36	56	52	39
2	47	61	57	57
3	50	64	59	63
4	58	66	58	61
5	67	66	79	65
$\overline{x}_{\text{гр } j}$	51,6	62,6	61,0	57,0

6.3. Произведено по семь испытаний на каждом из четырех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

Номер испытания i	Уровни факторов			
	F_1	F_2	F_3	F_4
1	51	52	56	54
2	59	58	56	58
3	53	66	58	62
4	59	69	58	64
5	63	70	70	66
6	69	72	74	67
7	72	74	78	69
$\overline{x}_{\text{гр } j}$	60,9	65,9	64,3	62,9

6.4. Произведено 13 испытаний, из них 4 – на первом уровне фактора, 4 – на втором, 3 – на третьем и 2 – на четвертом. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

Номер испытания i	Уровни факторов			
	F_1	F_2	F_3	F_4
1	1,38	1,41	1,32	1,31
2	1,38	1,42	1,33	1,33
3	1,42	1,44	1,34	-
4	1,42	1,45	-	-
$\overline{x}_{\text{гр}\,j}$	1,40	1,43	1,33	1,32

7

7.1. Для изучения надежности машины был собран статистический материал, приведенный в таблице, где Y – время непрерывной безотказной работы машины (в месяцах), X – количество предшествующих ремонтов. Установить тесноту связи между переменными X и Y и проверить значимость выборочного коэффициента корреляции при а = 0,05.

Y \ X	0	1	2	3
2–6	-	-	1	2
6–10	-	1	3	1
10–14	1	2	1	-
14–18	2	1	1	-
18–22	1	3	-	-

7.2. По выборке объема n = 100, извлеченной из двумерной нормальной генеральной совокупности (X,Y), найден выборочный коэффициент корреляции $r_{\text{В}}$ = 0,2. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_{\text{Г}} \neq 0$.

7.3. По выборке объема n = 100, извлеченной из двумерной нормальной генеральной совокупности (X,Y), составлена корреляционная таблица.

Y	X						
	2	7	12	17	22	27	n_y
100	2	4	-	-	-	-	6
120	-	6	2	-	-	-	8
130	-	-	3	50	2	-	55
140	-	-	1	10	6	-	17
150	-	-	-	4	7	3	14
n_x	2	10	6	64	15	3	N=100

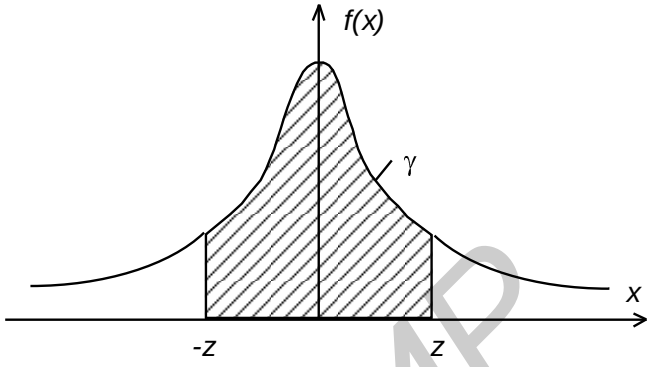
Требуется:
а) найти выборочный коэффициент корреляции;
б) при уровне значимости 0,01 проверить нулевую гипотезу о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе Н1: $r_{\text{Г}} \neq 0$.

7.4. По выборке объема n = 120, извлеченной из двумерной нормальной генеральной совокупности (X,Y), найден выборочный коэффициент корреляции $r_{\text{В}}$ = 0,4. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе Н1: $r_{\text{Г}} \neq 0$.

ПРИЛОЖЕНИЕ

1. Таблица функции Лапласа

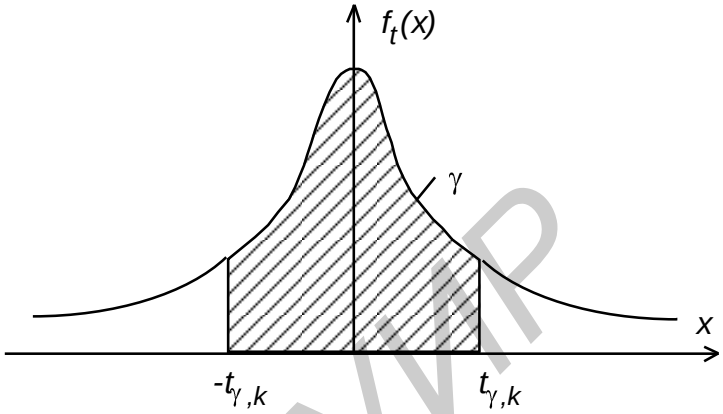
$$\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt = \gamma$$



z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
0,00	0,0000	0,66	0,4907	1,32	0,8132	1,98	0,9523
0,02	0,0160	0,68	0,5035	1,34	0,8198	2,00	0,9545
0,04	0,0319	0,70	0,5161	1,36	0,8262	2,05	0,9596
0,06	0,0478	0,72	0,5285	1,38	0,8324	2,10	0,9643
0,08	0,0638	0,74	0,5407	1,40	0,8385	2,15	0,9684
0,10	0,0797	0,76	0,5527	1,42	0,8444	2,20	0,9722
0,12	0,0955	0,78	0,5646	1,44	0,8501	2,25	0,9756
0,14	0,1113	0,80	0,5763	1,46	0,8557	2,30	0,9786
0,16	0,1271	0,82	0,5878	1,48	0,8611	2,35	0,9812
0,18	0,1428	0,84	0,5991	1,50	0,8664	2,40	0,9836
0,20	0,1585	0,86	0,6102	1,52	0,8715	2,45	0,9857
0,22	0,1741	0,88	0,6211	1,54	0,8764	2,50	0,9876
0,24	0,1897	0,90	0,6319	1,56	0,8812	2,55	0,9892
0,26	0,2051	0,92	0,6424	1,58	0,8859	2,60	0,9907
0,28	0,2205	0,94	0,6528	1,60	0,8904	2,66	0,9920
0,30	0,2358	0,96	0,6629	1,62	0,8948	2,70	0,9931
0,32	0,2510	0,98	0,6729	1,64	0,8990	2,75	0,9940
0,34	0,2661	1,00	0,6827	1,66	0,9031	2,80	0,9949
0,36	0,2812	1,02	0,6923	1,68	0,9070	2,85	0,9956
0,38	0,2961	1,04	0,7017	1,70	0,9109	2,90	0,9963
0,40	0,3108	1,06	0,7109	1,72	0,9146	2,95	0,9968
0,42	0,3255	1,08	0,7199	1,74	0,9181	3,00	0,9973
0,44	0,3401	1,10	0,7287	1,76	0,9216	3,10	0,9981
0,46	0,3545	1,12	0,7373	1,78	0,9249	3,20	0,9986
0,48	0,3688	1,14	0,7457	1,80	0,9281	3,30	0,9990
0,50	0,3859	1,16	0,7540	1,82	0,9312	3,40	0,9993
0,52	0,3969	1,18	0,7620	1,84	0,9342	3,50	0,9995
0,54	0,4108	1,20	0,7699	1,86	0,9371	3,60	0,9997
0,56	0,4245	1,22	0,7775	1,88	0,9399	3,70	0,9998
0,58	0,4381	1,24	0,7850	1,90	0,9426	3,80	0,9999
0,60	0,4515	1,26	0,7923	1,92	0,9451	3,90	0,9999
0,62	0,4647	1,28	0,7995	1,94	0,9476	4,00	0,9999
0,64	0,4778	1,30	0,8064	1,96	0,9500		

2. Таблица функции Стьюдента

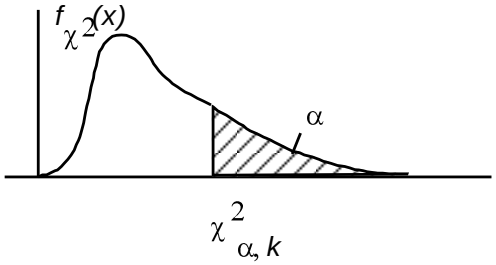
$$\gamma = \int_{-t_{\gamma,k}}^{t_{\gamma,k}} f_t(x)dx$$



k	γ			
	0,90	0,95	0,98	0,99
1	6,31	12,71	31,8	63,7
2	2,92	4,30	6,96	9,92
3	2,35	3,18	4,54	5,84
4	2,13	2,77	3,75	4,60
5	2,02	2,57	3,36	4,03
6	1,943	2,45	3,14	4,71
7	1,895	2,36	3,00	3,50
8	1,860	2,31	2,90	3,36
9	1,833	2,26	2,82	3,25
10	1,812	2,23	2,76	3,17
12	1,782	2,18	2,68	3,06
14	1,761	2,14	2,62	2,98
16	1,746	2,12	2,58	2,92
18	1,734	2,10	2,55	2,88
20	1,725	2,09	2,53	2,84
22	1,717	2,07	2,51	2,82
24	1,711	2,06	2,49	2,80
30	1,697	2,04	2,46	2,75
40	1,684	2,02	2,42	2,70

3. Таблица функции "Хи-квадрат"

$$P(\chi^2 > \chi^2_{\alpha,k}) = \alpha$$



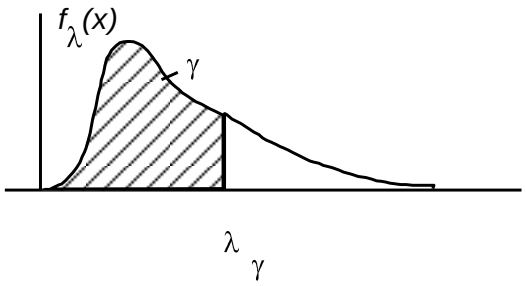
k	α					
	0,01	0,02	0,05	0,95	0,98	0,99
1	6,64	5,41	3,84	0,004	0,001	0,000
2	9,21	7,82	5,99	0,103	0,040	0,020
3	11,34	9,84	7,82	0,352	0,185	0,115
4	13,28	11,67	9,49	0,711	0,429	0,297
5	15,09	13,39	11,07	1,145	0,752	0,554
6	16,81	15,03	12,59	1,635	1,134	0,872
7	18,48	16,62	14,07	2,17	1,564	1,239
8	20,10	18,17	15,51	2,73	2,03	1,646
9	21,07	19,68	16,92	3,32	2,53	2,09
10	23,20	21,2	18,31	3,94	3,06	2,56
12	26,2	24,1	21,0	5,23	4,18	3,57
14	29,1	26,9	23,7	6,57	5,37	4,66
16	32,0	29,6	26,3	7,96	6,61	5,81
18	34,8	32,3	28,9	9,39	7,91	7,02
20	37,6	35,0	31,4	10,85	9,24	8,26
22	40,3	37,7	33,9	12,34	10,60	9,54
24	43,0	40,3	36,4	13,85	11,99	10,86
26	45,6	42,9	38,9	15,38	13,41	12,20
28	48,3	45,4	41,3	16,93	14,85	13,56
30	50,9	48,0	43,8	18,49	16,31	14,95

4. Таблица функции Р. Фишера (*F*-распределение)

Уровень значимости а = 0,10											
K_2	K_1										
	2	3	4	5	6	7	8	9	10	11	12
2	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41
3	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22
4	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90
5	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27
6	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90
7	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67
8	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50
9	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38
10	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28
11	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21
12	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15
13	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10
14	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05
Уровень значимости а = 0,05											
K_2	K_1										
	2	3	4	5	6	7	8	9	10	11	12
2	19,0	19,2	19,3	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4
3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57
8	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53
Уровень значимости а = 0,01											
K_2	K_1										
	2	3	4	5	6	7	8	9	10	11	12
2	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4
3	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1
4	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,6	14,5	14,4
5	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	10,0	9,9
6	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47
8	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67
9	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11
10	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71
11	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40
12	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16
13	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80

5. Таблица функции Колмогорова

$$P(0 \le \lambda < \lambda_\gamma) = \gamma$$



λ_γ	γ	λ_γ	γ	λ_γ	γ
0,50	0,0361	1,02	0,7500	1,54	0,9826
0,54	0,0675	1,06	0,7889	1,58	0,9864
0,58	0,1104	1,10	0,8223	1,62	0,9895
0,62	0,1632	1,14	0,8514	1,66	0,9918
0,66	0,2236	1,18	0,8765	1,70	0,9938
0,70	0,2888	1,22	0,8981	1,74	0,9953
0,74	0,3560	1,26	0,9164	1,78	0,9965
0,78	0,4230	1,30	0,9319	1,82	0,9973
0,82	0,4880	1,34	0,9449	1,86	0,9980
0,86	0,5497	1,38	0,9557	1,90	0,9985
0,90	0,6073	1,42	0,9646	1,94	0,9989
0,94	0,6601	1,46	0,9718	1,98	0,9992
0,98	0,7079	1,50	0,9778		

6. Таблица функции Мизеса

$P\{nw^2_n > nw^2_a\} = a$			
a	0,10	0,05	0,01
l – a	0,347	0,461	0,744

7. Таблица функции Вилкоксона

Объемы выборок		$a/2$			Объемы выборок		$a/2$		
n_1	n_2	0,05	0,025	0,005	n_1	n_2	0,05	0,025	0,005
6	6	28	26	23	7	7	39	36	32
	7	30	27	24		8	41	38	34
	8	31	29	25		9	43	40	35
	9	33	31	26		10	45	42	37
	10	35	32	27		11	47	44	38
	11	37	34	28		12	49	46	40
	12	38	35	30		13	52	48	41
8	8	51	49	43	9	9	66	62	56
	9	54	51	45		10	69	65	58
	10	56	53	47		11	72	68	61
	11	59	55	49		12	75	71	63
	12	62	58	51		13	78	73	65
	13	64	60	53		14	81	76	67
	14	67	62	54		15	84	79	69
10	10	82	78	71	11	11	100	96	87
	11	86	81	73		12	104	99	90
	12	89	84	76		13	108	103	93
	13	92	88	79		14	112	106	96
	14	96	91	81		15	116	110	99
	15	99	94	84		16	120	113	102
	16	103	97	86		17	123	117	105

8. Таблица функции Кочрена

Уровень значимости a = 0,05										
k	m									
	4	5	6	7	8	9	10	16	36	144
2	0,906	0,872	0,853	0,833	0,816	0,801	0,788	0,734	0,660	0,581
3	0,746	0,707	0,677	0,653	0,633	0,617	0,603	0,547	0,475	0,403
4	0,629	0,590	0,560	0,536	0,518	0,502	0,488	0,437	0,372	0,309
5	0,544	0,507	0,478	0,456	0,439	0,424	0,411	0,365	0,307	0,251
6	0,480	0,445	0,418	0,398	0,382	0,368	0,357	0,314	0,261	0,212
7	0,431	0,397	0,373	0,354	0,338	0,326	0,315	0,276	0,228	0,183
8	0,391	0,360	0,336	0,319	0,304	0,293	0,283	0,246	0,202	0,162
9	0,358	0,329	0,307	0,290	0,277	0,266	0,257	0,223	0,182	0,145
10	0,331	0,303	0,282	0,267	0,254	0,244	0,235	0,203	0,166	0,131
12	0,288	0,262	0,244	0,230	0,219	0,210	0,202	0,174	0,140	0,110
15	0,242	0,220	0,203	0,191	0,182	0,174	0,167	0,143	0,144	0,089
20	0,192	0,174	0,160	0,150	0,142	0,136	0,130	0,111	0,088	0,068
40	0,108	0,097	0,089	0,082	0,078	0,075	0,071	0,060	0,046	0,035
60	0,077	0,068	0,062	0,058	0,055	0,052	0,050	0,041	0,031	0,023
Уровень значимости a = 0,01										
k	m									
	4	5	6	7	8	9	10	16	36	144
2	0,959	0,937	0,917	0,899	0,882	0,867	0,854	0,795	0,707	0,606
3	0,834	0,793	0,761	0,734	0,711	0,691	0,674	0,606	0,515	0,423
4	0,721	0,676	0,641	0,613	0,590	0,570	0,554	0,488	0,406	0,325
5	0,633	0,588	0,553	0,526	0,504	0,485	0,470	0,409	0,335	0,264
6	0,564	0,520	0,487	0,461	0,440	0,433	0,408	0,353	0,286	0,223
7	0,508	0,466	0,435	0,411	0,391	0,375	0,362	0,311	0,249	0,193
8	0,463	0,423	0,393	0,370	0,352	0,337	0,323	0,278	0,221	0,170
9	0,425	0,387	0,359	0,338	0,321	0,307	0,295	0,251	0,199	0,152
10	0,393	0,357	0,331	0,311	0,295	0,281	0,270	0,230	0,181	0,138
12	0,343	0,310	0,286	0,268	0,254	0,242	0,232	0,196	0,154	0,116
15	0,288	0,260	0,239	0,223	0,210	0,200	0,192	0,161	0,125	0,093
20	0,229	0,205	0,188	0,175	0,165	0,157	0,150	0,125	0,096	0,071
40	0,128	0,114	0,103	0,096	0,090	0,085	0,082	0,067	0,050	0,036
60	0,090	0,080	0,072	0,067	0,063	0,059	0,057	0,046	0,034	0,025

Св. план 2011, поз. 54

Учебное издание

Волорова Наталья Алексеевна
Летохо Александр Сергеевич

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Методическое пособие
для студентов специальности 1-31 03 04 «Информатика»
дневной формы обучения

В 2-х частях

Часть 2

Редактор Т. П. Андрейченко
Корректор Е. Н. Батурчик

Подписано в печать 20.03.2012. Формат 60x84 1/16. Бумага офсетная.
Гарнитура «Таймс». Отпечатано на ризографе. Усл. печ. л. 3,95.
Уч.-изд. л. 3,5. Тираж 120 экз. Заказ 64.

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники»
ЛИ №02330/0494371 от 16.03.2009. ЛП №02330/0494175 от 03.04.2009.
220013, Минск, П. Бровки, 6