

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное
образовательное учреждение высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

Методические указания
к практическим занятиям
по дисциплине
**«ОСНОВЫ РАБОТЫ В МАТЕМАТИЧЕСКИХ И
СТАТИСТИЧЕСКИХ ПАКЕТАХ»**
для студентов направления подготовки
01.03.02 Прикладная математика и информатика

Ставрополь, 2026

СОДЕРЖАНИЕ

Введение

Лабораторная работа №1. Первичная обработка статистических данных

Лабораторная работа №2. «Метод произведений» и «Метод сумм» для вычислений $\bar{X}_B, D_B$

Лабораторная работа №3. Элементы корреляционного анализа

Лабораторная работа №4. Регрессионный анализ

Лабораторная работа №5. Криволинейная (параболическая) корреляция

Лабораторная работа №6. Критерий согласия Пирсона

Лабораторная работа №7. Статистическая проверка статистических гипотез

Лабораторная работа №8. Однофакторный дисперсионный анализ

Введение

При изучении теории «Основы работы в математических и статистических пакетах» студенты сталкиваются со многими проблемами. Иногда они затрудняются правильно определить вид полученного уравнения, верно выбрать алгоритм решения. Студенты не всегда правильно делают математические преобразования, соблюдают последовательность в решении. В результате при самостоятельном изучении у студентов возникают многие вопросы и ряд проблем остается невыясненными.

Для того, чтобы студенты самостоятельно выполнили решение того или иного задания, при этом выявили недочеты и исправили свои ошибки, необходима постоянная помощь преподавателя. Однако эта задача может быть решена самостоятельно студентами, если они в своей работе воспользуются методическими указаниями, составленными преподавателем в помощь студентам по данной теме. В данных методических указаниях подробно изложена методика решения различных типов задач, связанных со статистикой, сформулированы алгоритмы решений со ссылками на теоретический материал.

Тема 1. Основы теории вероятностей и случайных событий

Лабораторная работа №1. Первичная обработка статистических данных

Цель работы: научиться обрабатывать экспериментальные данные, строить вариационный ряд, эмпирическую функцию распределения, полигон и гистограмму ряда, определять его характеристики.

Краткие теоретические сведения

Генеральной совокупностью называется вся исследуемая совокупность объектов.

Выборочной совокупностью или просто **выборкой** называют совокупность случайно отобранных объектов.

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

Пусть из генеральной совокупности извлечена выборка, причем значение x_1 наблюдалось n_1 раз, x_2 наблюдалось n_2 раз, ..., x_k наблюдалось n_k раз. Общий объем выборки можно определить как

$$n = \sum_{i=1}^k n_i$$

Наблюдаемые значения x_i называют **вариантами**, а последовательность вариантов, записанных в возрастающем порядке, – **вариационным рядом**.

Числа наблюдений называют **частотами**, а их отношения к объему выборки i

$$\frac{n_i}{n} = W_i$$

– **относительными частотами**.

Модой M_0 называется варианта, имеющая наибольшую частоту.

Медианой M_e называется варианта, которая делит вариационный ряд на две части с одинаковым числом вариантов в каждой. Для вариационного ряда с нечетным числом членов медиана равна срединной variante, а для ряда с четным числом членов – полусумме двух срединных вариантов.

Размахом выборки называется разность между максимальным и минимальным элементами выборки.

Статистическим распределением выборки называют перечень вариантов и соответствующих им частот или относительных частот.

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

 или

x_i	x_1	x_2	...	x_k
w_i	w_1	w_2	...	w_k

Статистическое распределение можно задать также в виде последовательности интервалов одинаковой длины и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал).

x	$x_0 - x_1$	$x_1 - x_2$	...	$x_{k-1} - x_k$
-----	-------------	-------------	-----	-----------------

n_i	n_1	n_2	$\dots$	n_k
-------	-------	-------	---------	-------

Алгоритм группировки выборочных данных при построении интервального ряда

1) Найти наименьшее и наибольшее значения признака в совокупности и определить размах варьирования: $R = x_{max} - x_{min}$.

2) Определить число интервалов k . Для этого используют формулу Стерджесса:

$$k = 1 + 3,322 \cdot \lg n$$

3) Найти постоянную величину интервала:

$$h = \frac{R}{k}$$

4) Определить границы интервалов. За начало первого интервала следует взять $x_0 = x_{min}$. Промежуточные интервалы получают, прибавляя к концу предыдущего интервала длину частичного интервала h . Конец последнего интервала должен удовлетворять условию: $x_{max} \leq x_k$.

5) Подсчитать число выборочных данных, которые попадут в каждый из полученных интервалов: $n_1, n_2, \dots, n_k$. При этом только один из промежутков будет замкнут с двух сторон: $[x_0; x_1]$, а остальные промежутки будут замкнуты только справа: $(x_i; x_{i+1}]$, $i = 1, \dots, k-1$.

6) Результаты вычислений занести в таблицу, которая называется «Статистическое распределение интервального ряда».

Эмпирической функцией распределения случайной величины (функцией распределения выборки) называют функцию F_x^* относительной частоты числа наблюдений n_x :

$$F_x^* = \frac{n_x}{n},$$

т.е. относительной частоты события $X < x$.

Полигоном частот называют ломаную, отрезки которой соединяют точки $(x_1; n_1), (x_2; n_2), \dots, (x_k; n_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i , а на оси ординат – соответствующие им частоты n_i . Точки $(x_i; n_i)$ соединяют отрезками прямых и получают полигон частот.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки $(x_1; W_1), (x_2; W_2), \dots, (x_k; W_k)$. Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i , а на оси ординат – соответствующие им относительные частоты W_i . Точки $(x_i; W_i)$ соединяют отрезками прямых и получают полигон относительных частот.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h , а высоты равны отношению n_i/h (плотность частоты).

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

интервалы длиной h , а высоты равны отношению W_i/h (плотность относительной частоты).

Выборочной средней $\bar{x}_B$ называют среднее арифметическое значение признака выборочной совокупности.

Если значения признака $x_1, x_2, \dots, x_k$ имеют соответственно частоты $n_1, n_2, \dots, n_k$, причем $n_1 + n_2 + \dots + n_k = n$, то

$$\bar{x}_B = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

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

Если значения признака $x_1, x_2, \dots, x_k$ имеют соответственно частоты $n_1, n_2, \dots, n_k$, причем $n_1 + n_2 + \dots + n_k = n$, то

$$D_B = \frac{\sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i}{n}.$$

Выборочным средним квадратическим отклонением называют квадратный корень из выборочной дисперсии:

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

Задание: Имеется выборка, содержащая 30 числовых значений некоторого признака случайной величины X :

19	25	22	16	22	14	17	19	18	20
22	26	24	18	16	19	22	14	18	14
25	17	18	14	20	18	24	25	16	18

Построить:

- 1) статистическое распределение выборки;
- 2) полигон частот;
- 3) эмпирическую функцию распределения;
- 4) интервальный ряд;
- 5) гистограмму частот;

вычислить:

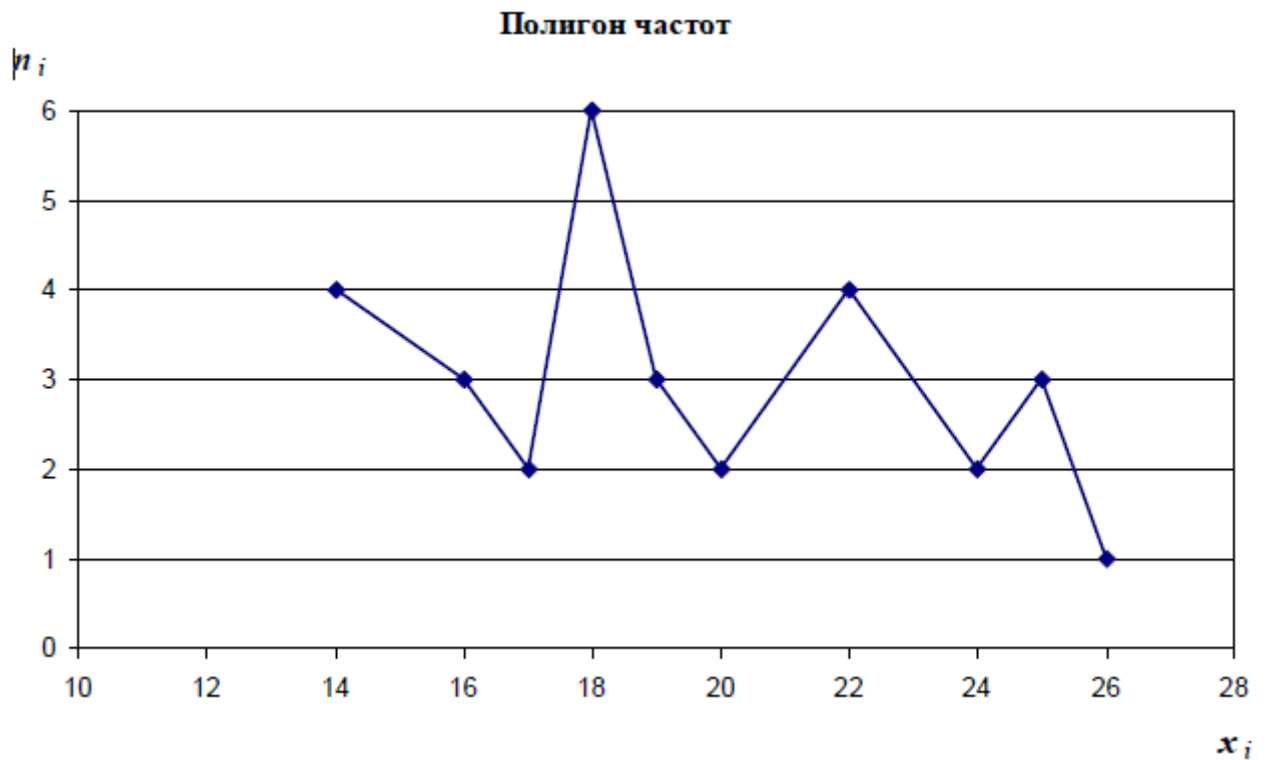
- 6) выборочную среднюю;
- 7) выборочную дисперсию;
- 8) выборочное среднее квадратическое отклонение;
- 9) моду;
- 10) медиану.

Решение.

1) Статистическое распределение выборки представляет собой таблицу, в которой первая строка содержит варианты (значения случайной величины, расположенные в порядке возрастания), а вторая – соответствующие частоты (сколько раз эти значения встречались).

x_i	14	16	17	18	19	20	22	24	25	26
n_i	4	3	2	6	3	2	4	2	3	1

2) Полигон частот – это ломаная, соединяющая точки с координатами $(x_i; n_i)$.



3) Проведем вычисления для построения эмпирической функции распределения:

$x \leq 14$	$F_x^* = 0$
$14 < x \leq 16$	$F_x^* = \frac{4}{30} \approx 0,13$
$16 < x \leq 17$	$F_x^* = \frac{4+3}{30} \approx 0,23$
$17 < x \leq 18$	$F_x^* = \frac{4+3+2}{30} = 0,3$
$18 < x \leq 19$	$F_x^* = \frac{4+3+2+6}{30} = 0,5$
$19 < x \leq 20$	$F_x^* = \frac{4+3+2+6+3}{30} = 0,6$
$20 < x \leq 22$	$F_x^* = \frac{4+3+2+6+3+2}{30} \approx 0,67$
$22 < x \leq 24$	$F_x^* = \frac{4+3+2+6+3+2+4}{30} = 0,8$
$24 < x \leq 25$	$F_x^* = \frac{4+3+2+6+3+2+4+2}{30} \approx 0,87$
$25 < x \leq 26$	$F_x^* = \frac{4+3+2+6+3+2+4+2+3}{30} \approx 0,97$
$x > 26$	$F_x^* = \frac{4+3+2+6+3+2+4+2+3+1}{30} = 1$

Получили эмпирическую функцию распределения:

$$F_x^* = \begin{cases} 0, & \text{при } x \leq 14, \\ 0,13 & \text{при } 14 < x \leq 16, \\ 0,23 & \text{при } 16 < x \leq 17, \\ 0,3 & \text{при } 17 < x \leq 18, \\ 0,5 & \text{при } 18 < x \leq 19, \\ 0,6 & \text{при } 19 < x \leq 20, \\ 0,67 & \text{при } 20 < x \leq 22, \\ 0,8 & \text{при } 22 < x \leq 24, \\ 0,87 & \text{при } 24 < x \leq 25, \\ 0,97 & \text{при } 25 < x \leq 26, \\ 1 & \text{при } x > 26, \end{cases}$$

4) Построим интервальный ряд.

Найдем наименьшее и наибольшее значения признака в совокупности и определим размах варьирования: $R = x_{\max} - x_{\min} = 26 - 14 = 12$.

Определим число интервалов k . Для этого воспользуемся формулой Стерджесса:

$$k = 1 + 3,322 \cdot \lg n = 1 + 3,322 \cdot \lg 30 \approx 6.$$

Найдем постоянную величину интервала: $h = \frac{R}{k} = \frac{12}{6} = 2$

Определим границы интервалов. За начало первого интервала следует взять $x_0 = x_{\min} = 14$. Промежуточные интервалы получаем, прибавляя к концу предыдущего интервала длину частичного интервала $h = 2$:

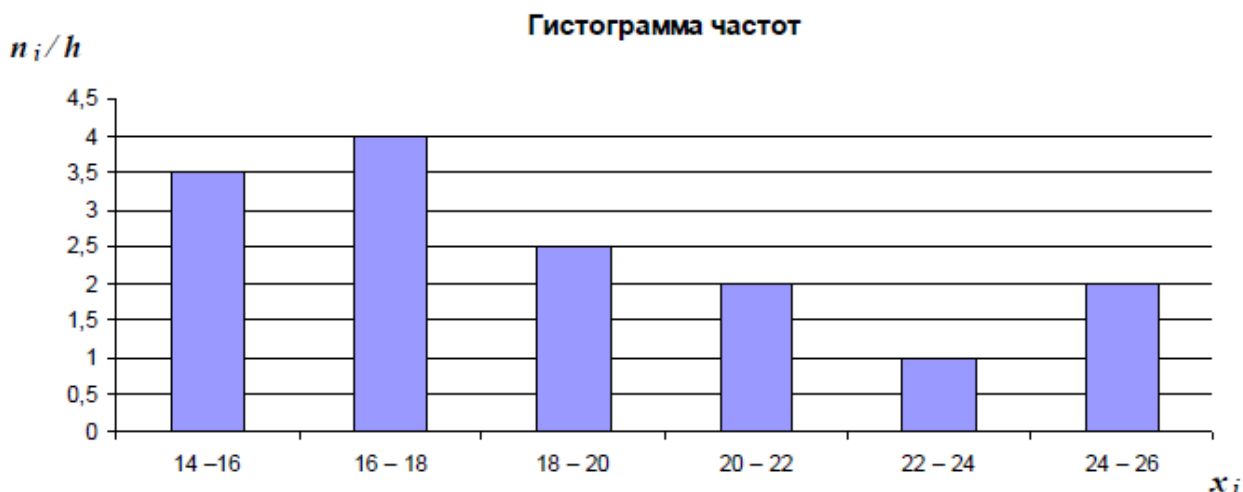
$$x_1 = 14 + 2 = 16, x_2 = 16 + 2 = 18, x_3 = 18 + 2 = 20, x_4 = 20 + 2 = 22, \\ x_5 = 22 + 2 = 24, x_6 = 24 + 2 = 26.$$

Будем рассматривать следующие промежутки: $[14; 16]$, $(16; 18]$, $(18; 20]$, $(20; 22]$, $(22; 24]$, $(24; 26]$ и посчитаем количество вариантов для каждого промежутка.

Получим статистическое распределение интервального ряда:

x_i	14-16	16-18	18-20	20-22	22-24	24-26
n_i	7	8	5	4	2	4

5) Построим гистограмму частот:



6) Вычислим характеристики ряда. Статистическое распределение ряда, вычисленное выше (пункт 1), имеет вид:

x_i	14	16	17	18	19	20	22	24	25	26
n_i	4	3	2	6	3	2	4	2	3	1

Выборочная средняя:

$$\bar{x}_B = \frac{\sum_{i=1}^k x_i n_i}{n} = \frac{14 \cdot 4 + 16 \cdot 3 + 17 \cdot 2 + 18 \cdot 6 + 19 \cdot 3 + 20 \cdot 2 + 22 \cdot 4 + 24 \cdot 2 + 25 \cdot 3 + 26 \cdot 1}{30} \approx 19,3$$

7) Выборочная дисперсия:

$$D_B = \frac{\sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i}{n} = \frac{4(14 - 19,3)^2 + 3(16 - 19,3)^2 + 2(17 - 19,3)^2 + 6(18 - 19,3)^2 + 3(19 - 19,3)^2 + 2(20 - 19,3)^2 + 4(22 - 19,3)^2 + 2(24 - 19,3)^2 + 3(25 - 19,3)^2 + 1(26 - 19,3)^2}{30} \approx 12,76$$

8) Среднее квадратическое отклонение: $\sigma_B = \sqrt{D_B} = \sqrt{12,76} \approx 3,57$.

9) Запишем вариационный ряд (все значения в порядке возрастания):

14, 14, 14, 14, 16, 16, 16, 17, 17, **18, 18, 18, 18, 18, 18**, 19, 19, 19, 20, 20, 22, 22, 22, 22, 24, 24, 25, 25, 25, 26

Мода – наиболее часто встречаемое значение, $M_0 = 18$.

10) Медиана – варианта, которая делит вариационный ряд на две части с одинаковым числом вариантов в каждой. Для нашего ряда с четным числом членов – медиана равна полусумме двух срединных вариантов.

14, 14, 14, 14, 16, 16, 16, 17, 17, 18, 18, 18, 18, 18, **18, 19**, 19, 19, 20, 20, 22, 22, 22, 22, 24, 24, 25, 25, 25, 26

Медиана $M_e = \frac{18 + 19}{2} = 18,5$

Задания для самостоятельной работы

Первичная обработка статистических данных

Задание: Имеется выборка, содержащая 15 числовых значений некоторого признака случайной величины X .

вычислить:

6) выборочную среднюю;

7) выборочную дисперсию;

Построить:

- 1) статистическое распределение выборки;
- 2) полигон частот;
- 3) эмпирическую функцию распределения;
- 4) интервальный ряд;
- 5) гистограмму частот;

8) выборочное среднее

квадратическое отклонение;

9) моду;

10) медиану.

1.	X	17	10	26	20	4	17	20	26	20	4	10	29	20	17	10
2.	X	13	8	3	17	8	20	13	8	3	23	8	13	17	8	13
3.	X	15	22	27	6	15	36	27	22	15	27	22	27	6	27	31
4.	X	14	12	7	14	12	17	14	12	17	22	20	12	14	20	12
5.	X	11	5	14	18	23	11	14	5	18	25	11	18	14	18	23
6.	X	12	16	23	30	12	5	28	12	16	23	5	12	16	12	16
7.	X	13	18	8	21	28	13	21	26	21	8	13	18	21	18	21
8.	X	10	6	13	10	17	10	13	20	17	10	20	13	10	21	13
9.	X	18	13	24	7	32	24	18	30	7	24	13	18	30	13	24
10.	X	15	10	4	15	17	22	10	15	24	10	15	4	10	17	10
11.	X	13	7	10	13	3	13	7	17	13	18	3	13	7	10	13
12.	X	12	19	12	5	19	24	12	19	35	24	19	30	12	30	12
13.	X	14	9	20	14	23	20	26	9	23	14	26	23	20	29	23
14.	X	15	10	3	19	15	10	28	15	3	10	15	19	10	24	10
15.	X	17	7	14	12	17	20	22	17	14	7	12	17	12	17	14
16.	X	16	11	6	25	16	21	11	26	11	16	11	25	16	21	11
17.	X	13	16	20	7	13	27	20	24	13	7	20	16	24	20	16
18.	X	12	25	8	12	8	25	8	12	3	12	3	8	33	8	30
19.	X	19	8	14	24	19	24	33	24	19	8	14	24	30	24	14
20.	X	10	14	8	10	14	8	3	10	8	16	8	10	8	18	16
21.	X	14	18	22	14	7	18	22	25	14	22	27	25	22	7	18
22.	X	15	11	4	11	15	22	15	11	4	15	25	11	29	22	11
23.	X	17	24	12	24	17	12	24	4	17	34	24	4	30	24	12
24.	X	16	13	8	16	13	18	21	16	13	21	18	13	16	23	13
25.	X	10	5	16	10	20	5	22	20	10	22	16	20	25	20	16

Лабораторная работа 2. «Метод произведений» и «Метод сумм» для вычислений $\bar{X}_B, D_B$

Цель работы: научиться «методом произведений» и «методом сумм» вычислять $\bar{X}_B$ и D_B в случае распределения равностоящих вариантов.

Краткие теоретические сведения «Метод произведений»

Пусть выборка задана в виде распределения равноотстоящих вариантов и соответствующих им частот. В этом случае удобно находить $\bar{X}_B, D_B$ методом произведений по формулам:

$$\bar{X}_B = M_1^* h + C, \quad D_B = [M_2^* - (M_1^*)^2] h^2;$$

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

C – ложный ноль (варианта, которая имеет наибольшую частоту);

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

$$M_1^* = \frac{\sum_{i=1}^k u_i n_i}{n} \text{ – условный момент I порядка;}$$

$$M_2^* = \frac{\sum_{i=1}^k u_i^2 n_i}{n} \text{ – условный момент II порядка;}$$

$$M_3^* = \frac{\sum_{i=1}^k u_i^3 n_i}{n} \text{ – условный момент III порядка;}$$

$$M_4^* = \frac{\sum_{i=1}^k u_i^4 n_i}{n} \text{ – условный момент IV порядка;}$$

$$m_3 = \frac{\sum_{i=1}^k (x_i - \bar{X}_B)^3 n_i}{n} = [M_3^* - 3M_1^* M_2^* + 2(M_1^*)^3] h^3 \text{ – центральный эмпирический момент III порядка;}$$

$$m_4 = \frac{\sum_{i=1}^k (x_i - \bar{X}_B)^4 n_i}{n} = [M_4^* - 4M_1^* M_3^* + 6(M_1^*)^2 M_2^* - 3(M_1^*)^4] h^4 \text{ – центральный эмпирический момент IV порядка;}$$

$$a_s = \frac{m_3}{\sigma_B^3} \text{ – асимметрия эмпирического распределения;}$$

$e_k = \frac{m_4}{\sigma_B^4} - 3$ - эксцесс эмпирического распределения;

$\sigma_B = \sqrt{D_B}$ - выборочное среднеекватратическое отклонение.

«Метод сумм»

Пусть выборка задана в виде распределения равноотстоящих вариантов и соответствующих им частот. Для вычисления $\bar{X}_B, D_B$ воспользуемся методом сумм по формулам:

$$\bar{X}_B = M_1^* h + C, \quad D_B = M_2^* - (M_1^*)^2 h^2$$

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

C – ложный ноль (варианта, которая имеет наибольшую частоту);

$M_1^* = \frac{d_1}{n}$ – условный момент I порядка;

$M_2^* = \frac{S_1 + 2S_2}{n}$ – условный момент II порядка;

$M_3^* = \frac{d_1 + 6d_2 + 6d_3}{n}$ – условный момент III порядка;

$M_4^* = \frac{S_1 + 14S_2 + 36S_3 + 24S_4}{n}$ – условный момент IV порядка;

$$d_1 = a_1 - b_1, \quad d_2 = a_2 - b_2, \quad d_3 = a_3 - b_3$$

$$S_1 = a_1 + b_1, \quad S_2 = a_2 + b_2, \quad S_3 = a_3 + b_3, \quad S_4 = a_4 + b_4$$

Таким образом, в конечном счете, надо вычислить числа $a_i, b_i, i = 1, 2, 3, 4$. Принцип расчета $a_i, b_i, i = 1, 2, 3, 4$ описан в решении задания.

Задание 1: По данной выборке вычислить «методом произведений» $\bar{X}_B, D_B, a_s, e_k$

x_i	12	14	16	18	20	22
n_i	5	15	50	16	10	4

Составим расчётную таблицу:

- 1) запишем варианты в первый столбец;
- 2) запишем частоты во второй столбец, сумму частот (100) поместим в нижнюю клетку столбца;
- 3) в качестве ложного нуля C выберем варианту, которая имеет наибольшую частоту ($C = 16$). В клетке третьего столбца, которая принадлежит строке, содержащей ложный ноль, пишем 0; над нулем последовательно запишем: -1, -2; под нулём 1, 2, 3;
- 4) в четвёртый столбец записываем произведение частот n_i на условные варианты u_i , сумму произведений $u_i n_i$ (23) записываем в нижнюю клетку столбца;

5) произведение частот на квадраты условных вариантов, т.е. $u_i^2 n_i$ запишем в пятый столбец. Сумму чисел $u_i^2 n_i$ (127) записываем в нижнюю клетку столбца; аналогично заполняются столбцы для $u_i^3 n_i$, $u_i^4 n_i$;

6) произведения частот на квадраты условных вариантов, увеличенных на единицу, т.е. $(u_i + 1)^4 n_i$, запишем в шестой контрольный столбец; сумму чисел столбца (2145) помещаем в нижнюю клетку восьмого столбца.

Получим следующую таблицу:

x_i	n_i	u_i	$u_i n_i$	$u_i^2 n_i$	$u_i^3 n_i$	$u_i^4 n_i$	$(u_i + 1)^4 n_i$
12	5	-2	-10	20	-40	80	5
14	15	-1	-15	15	-15	15	-
16	50	0	-25	-	-55	-	50
18	16	1	16	16	16	16	256
20	10	2	20	40	80	160	810
22	4	3	12	36	108	324	1024
			48		204		
	$n=100$	$\frac{x_i - C}{h}$	$\sum_{i=1}^k u_i n_i$ = =23	$\sum_{i=1}^k u_i^2 n_i$ = =127	$\sum_{i=1}^k u_i^3 n_i$ = =149	$\sum_{i=1}^k u_i^4 n_i$ = =595	$\sum_{i=1}^k (u_i + 1)^4 n_i$ = =2145

Для контроля вычисления пользуются тождеством:

$$\sum_{i=1}^k (u_i + 1)^4 n_i = \sum_{i=1}^k u_i^4 n_i + 4 \sum_{i=1}^k u_i^3 n_i + 6 \sum_{i=1}^k u_i^2 n_i + 4 \sum_{i=1}^k u_i n_i + n$$

Контроль: $595 + 4 \cdot 149 + 6 \cdot 127 + 4 \cdot 23 + 100 = 2145$

Для вычисления $\bar{X}_B, D_B, a_s, e_k$ воспользуемся формулами:

$$M_1^* = 0,23; M_2^* = 1,27; M_3^* = 1,49; M_4^* = 5,95;$$

$$h = 2; m_3 = 5,124; m_4 = 79,582; C = 16;$$

$$\bar{X}_B = M_1^* h + C = 0,23 \cdot 2 + 16 = 16,46; D_B = [1,27 - 0,23^2] 4 = 4,87;$$

$$a_s = \frac{m_3}{\sigma_B^3} = \frac{5,124}{(\sqrt{4,87})^3} = 0,49; e_k = \frac{m_4}{\sigma_B^4} - 3 = \frac{79,582}{(\sqrt{4,87})^4} - 3 = 0,36.$$

Задание 2: По данной выборке вычислить «методом сумм» $\bar{X}_B, D_B, a_s, e_k$

x_i	48	52	56	60	64	68	72	76	80	84
n_i	2	4	6	8	12	30	18	8	7	5

Составим расчётную таблицу:

1) запишем варианты в первый столбец;

2) запишем частоты во второй столбец, сумму частот (100) поместим в нижнюю клетку столбца;

3) в качестве ложного нуля C выберем варианту, которая имеет наибольшую частоту ($C = 68$). В клетках строки, содержащей ложный нуль, запишем нули;

4) в четвёртом столбце над и под уже помещенным нулем запишем еще по одному нулю;

5) в оставшихся незаполненными над нулем клетках третьего столбца (исключая самую верхнюю) запишем последовательно накопленные частоты: 2; $2+4=6$; $2+4+6=12$; $2+4+6+8=20$; $2+4+6+8+12=32$; сложив все накопленные частоты, получим число $b_1 = 72$, которое поместим в верхнюю клетку третьего столбца. В оставшихся незаполненными под нулем клетках третьего столбца (исключая самую нижнюю) запишем последовательно накопленные частоты: 5; $5+7=12$; $5+7+8=20$; $5+7+8+18=38$; сложив все накопленные частоты, получим число $a_1 = 75$, которое поместим в нижнюю клетку третьего столбца;

6) аналогично заполняется четвертый (пятый или шестой) столбец, причем суммируют частоты третьего (четвертого или пятого) столбца; сложив все накопленные частоты, расположенные над нулем, получим число $b_2 = 70$, которое поместим в верхнюю клетку четвертого столбца; сумма накопленных частот, расположенных под нулем, равна числу $a_2 = 59$, которое поместим в нижнюю клетку четвертого столбца.

Получим следующую таблицу:

x_i	n_i	$b_1=72$	$b_2=70$	$b_3=42$	$b_4=14$
48	2	2	2	2	2
52	4	6	8	10	12
56	6	12	20	30	0
60	8	20	40	0	0
64	12	32	0	0	0
68	30	0	0	0	0
72	18	38	0	0	0
76	8	20	37	0	0
80	7	12	17	22	0
84	5	5	5	5	5
	$n=100$	$a_1 = 75$	$a_2 = 59$	$a_3 = 27$	$a_4 = 5$

Найдем $d_i, i=1,2,3, S_j, j=1,2,3,4$

$$d_1 = a_1 - b_1 = 75 - 72 = 3, \quad S_1 = a_1 + b_1 = 75 + 72 = 147,$$

$$d_2 = a_2 - b_2 = 59 - 70 = -11, \quad S_2 = a_2 + b_2 = 59 + 70 = 129,$$

$$d_3 = a_3 - b_3 = 27 - 42 = -15, \quad S_3 = a_3 + b_3 = 27 + 42 = 69$$

$$S_4 = a_4 + b_4 = 5 + 14 = 19$$

Найдем условные моменты:

$$M_1^* = \frac{d_1}{n} = \frac{3}{100} = 0,03, \quad M_2^* = \frac{S_1 + 2S_2}{n} = \frac{147 + 2 \cdot 149}{100} = 4,05,$$

$$M_3^* = \frac{d_1 + 6d_2 + 6d_3}{n} = \frac{3 + 6 \cdot (-11) + 6 \cdot (-15)}{100} = -1,53,$$

$$M_4^* = \frac{S_1 + 14S_2 + 36S_3 + 24S_4}{n} = 48,9.$$

Вычислим $\bar{X}_B, D_B$ (учитывая, что $h = 4; C = 68$):

$$\bar{X}_B = M_1^* h + C = 0,03 \cdot 4 + 68 = 68,12;$$

$$D_B = M_2^* - (M_1^* h)^2 = [4,05 - 0,03^2] 4^2 = 64,78;$$

Вычислим a_s, e_k :

$$m_3 = [M_3^* - 3M_1^* M_2^* + 2(M_1^*)^3] h^3 = -121,248;$$

$$m_4 = [M_4^* - 4M_1^* M_3^* + 6(M_1^*)^2 M_2^* - 3(M_1^*)^4] h^4 = 49,135;$$

$$a_s = \frac{m_3}{\sigma_B^3} = \frac{-121,248}{(\sqrt{64,78})^3} = -0,23;$$

$$e_k = \frac{m_4}{\sigma_B^4} - 3 = \frac{49,134}{(\sqrt{64,78})^4} - 3 = 0,01.$$

Задание: По данной выборке вычислить «методом произведений» и «методом сумм» $\bar{X}_B, D_B, a_s, e_k$:

1	x_i	12	14	16	18	20	22	24
	n_i	6	12	18	24	17	13	10
2	x_i	11	13	15	17	19	21	23
	n_i	8	14	21	25	13	10	9
3	x_i	25	30	35	40	45	50	55
	n_i	9	10	16	22	18	15	10
4	x_i	16	20	24	28	32	36	40
	n_i	10	15	16	26	13	11	9
5	x_i	20	22	24	26	28	30	32
	n_i	11	12	20	24	14	12	7
6	x_i	23	25	27	29	31	33	35
	n_i	8	15	18	23	15	12	9
7	x_i	10	13	16	19	22	25	28
	n_i	7	13	19	22	16	13	10
8	x_i	22	24	26	28	30	32	34
	n_i	9	16	19	23	14	11	8
9	x_i	15	20	25	30	35	40	45
	n_i	12	14	19	20	14	12	9
10	x_i	11	15	19	23	27	31	35
	n_i	8	11	18	26	15	12	10

11	x_i	24	26	28	30	32	34	36
	n_i	7	12	19	24	17	12	9
12	x_i	18	20	22	24	26	28	30
	n_i	10	13	16	21	19	13	8
13	x_i	17	20	23	26	29	32	35
	n_i	10	12	17	25	14	12	10
14	x_i	13	15	17	19	21	23	25
	n_i	9	13	17	24	14	13	10
15	x_i	18	22	26	30	34	38	42
	n_i	11	12	16	23	15	14	9
16	x_i	14	18	22	26	30	34	38
	n_i	10	11	19	21	16	12	11
17	x_i	16	18	20	22	24	26	28
	n_i	7	12	14	24	20	12	11
18	x_i	21	23	25	27	29	31	33
	n_i	8	13	19	21	16	13	10
19	x_i	17	21	25	29	33	37	41
	n_i	9	12	15	23	18	15	8
20	x_i	12	15	18	21	24	27	30
	n_i	9	12	17	24	17	12	9
21	x_i	10	14	18	22	26	30	34
	n_i	6	13	18	23	17	14	9
22	x_i	19	21	23	25	27	29	31
	n_i	9	10	14	26	16	15	10
23	x_i	22	25	28	31	34	37	40
	n_i	9	13	18	22	18	13	9
24	x_i	18	24	30	36	42	48	54
	n_i	9	11	13	26	16	15	10
25	x_i	19	23	27	31	35	39	43
	n_i	7	13	20	23	14	12	11

Лабораторная работа №3. Элементы корреляционного анализа

Цель: определить тесноту связи между признаками X и Y .

Краткие теоретические сведения

Все явления и процессы, которые существуют в природе и обществе, взаимосвязанные. Изучение взаимосвязей – важная задача статистики, которую она решает с помощью особых методов. Существуют *функциональные* и *корреляционные* связи.

При *функциональной* связи каждому значению x соответствует ровно одно возможное значение y , которое может быть вычислено по точной формуле $y = f(x)$. Например, в функции $y = 2x$ каждому значению x соответствует в два раза большее значение y .

Но такого рода функциональные (однозначные) связи между переменными величинами встречаются не всегда. Например, при одном и том же росте масса различных индивидуумов может быть различна. В данном случае, можно и целесообразно вести речь о *среднем значении* массы индивидуумов.

Корреляционная связь проявляется в том, что при изменении одной из величин изменяется среднее значение другой.

Виды корреляционных связей:

- 1) по форме может быть прямолинейной или криволинейной;
- 2) по направлению может быть положительной («прямой») или отрицательной («обратной»).

Линейной называется связь, которая графически изображается прямой линией.

Нелинейной (криволинейной) называется связь, которая графически изображается любой другой линией, кроме прямой (гиперболой, параболой и т.д.).

Связь называется *положительной* («прямой»), если при увеличении одного параметра второй тоже увеличивается. Связь называется *отрицательной* («обратной»), если при увеличении одного параметра второй уменьшается.

Математической мерой корреляции двух случайных величин служит коэффициент корреляции. Обозначается символом r . Коэффициент корреляции r характеризует величину, отражающую степень взаимосвязи двух переменных между собой. Он может варьировать в пределах от -1 (отрицательная корреляция) до +1 (положительная корреляция). В зависимости от того, насколько r приближается к 1, различают связи:

- $0,91 \leq |r| < 1$ очень сильная связь;
- $0,71 \leq |r| < 0,9$ сильная связь;
- $0,51 \leq |r| < 0,7$ значительная связь;
- $0,31 \leq |r| < 0,5$ умеренная связь;
- $|r| < 0,3$ связь очень слабая.

Если коэффициент корреляции равен 0, то это говорит об отсутствии корреляционных связей между переменными.

При линейной зависимости между признаками коэффициент корреляции определяется методом квадратов (метод Пирсона).

Метод квадратов (метод Пирсона)

- 1) определить для каждого вариационного ряда средние значения $\bar{x}$ и $\bar{y}$;
- 2) найти отклонения $x_i - \bar{x}$ и $y_i - \bar{y}$ каждого числового значения от среднего значения своего вариационного ряда;
- 3) полученные отклонения перемножить и суммировать;
- 4) каждое отклонение возвести в квадрат и суммировать по каждому ряду.

Подставить полученные значения в формулу расчета коэффициента корреляции:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

ОШИБКА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r :

$$m_r = \sqrt{\frac{1-r^2}{n-2}},$$

где n – число парных вариантов.

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Оценка значимости коэффициента корреляции проводится по критерию Стьюдента путем сравнения его значения с величиной случайной ошибки:

$$t_r = \frac{r}{m_r}$$

Статистика t_r подчиняется распределению Стьюдента с числом степени свободы $\nu = n - 2$.

Из таблицы распределения критерия Стьюдента находят критическое значение $t_{kp} = t(\alpha, \nu)$.

Если $|t_r| > t_{kp}$, то параметр r статистически значим. В противном случае ($|t_r| < t_{kp}$) параметр r статистически незначим.

Задание. Найти коэффициент корреляции по данным $n = 10$ наблюдений, которые получены при изучении зависимости между ростом (X , см) и массой (Y , кг) некоторых животных:

x	31	32	33	34	35	36	40	41	42	46
y	7,8	8,3	7,6	9,1	9,6	9,8	11,8	12,1	14,7	13

Решение.

Составим расчетную вспомогательную таблицу:

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
31	7,8	-6	26	-2,58	6,6564	15,48
32	8,3	-5	25	-2,08	4,3264	10,4
33	7,6	-4	16	-2,78	7,7284	11,12
34	9,1	-3	9	-1,28	1,6384	3,84
35	9,6	-2	4	-0,78	0,6084	1,56
36	9,8	-1	1	-0,58	0,3364	0,58
40	11,8	3	9	1,42	2,0164	4,26
41	12,1	4	16	1,72	2,9584	6,88
42	14,7	5	25	4,32	18,6624	21,6
46	13	9	81	2,62	6,8644	23,58
$\sum x_i = 370$	$\sum y_i = 103,8$		222		51,796	99,3
$\bar{x} = 37$	$\bar{y} = 10,38$					

Подставляем в формулу:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{99,3}{\sqrt{222 \cdot 51,796}} \approx 0,926$$

Вывод: связь положительная (прямая, т.е. чем больше рост животных, тем больше у них масса), очень тесная.

ОШИБКА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r :

$$m_r = \sqrt{\frac{1 - 0,962^2}{10 - 2}} \approx 0,1335$$

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Оценка значимости коэффициента корреляции проводится по критерию Стьюдента путем сравнения его значения с величиной случайной ошибки:

$$t_r = \frac{r}{m_r} = \frac{0,926}{0,1335} \approx 6,94$$

Статистика t_r подчиняется распределению Стьюдента с числом степени свободы $\nu=8$.

Из таблицы распределения критерия Стьюдента находят критическое значение при уровне надежности 95% уровень значимости - $\alpha=0,05$: $t_{kp} = t(\alpha, \nu) = 2,31$.

Так как $|t_r| > t_{kp}$, то параметр r статистически значим.

Задание. Вычислить коэффициент корреляции r Пирсона и оценить его значимость. Определить ошибку коэффициента корреляции m_r .

1	X	41	40	39	38	36	37	39	36	36	38
	Y	47	48	44	45	42	45	50	47	46	46
2	X	25	30	25	30	35	35	40	40	45	45
	Y	36	38	42	36	24	28	24	20	22	20
3	X	21	20	18	14	20	21	17	18	16	15
	Y	19	18	16	13	19	17	16	15	14	13
4	X	17	18	21	25	29	33	37	41	45	44
	Y	36	32	36	46	40	48	50	45	50	46
5	X	56	56	58	57	57	54	55	52	55	60
	Y	38	36	40	41	36	39	38	36	37	39
6	X	17	14	18	15	16	17	16	15	19	13
	Y	30	40	35	40	40	35	35	25	45	38
7	X	56	58	61	60	59	58	56	57	59	56
	Y	46	47	46	48	44	45	42	46	50	46
8	X	20	20	35	20	30	25	25	27	30	30
	Y	18	19	25	20	25	21	23	22	23	24
9	X	28	23	24	25	26	27	29	23	29	26
	Y	30	25	26	28	27	31	30	24	31	28
10	X	40	39	38	37	35	36	38	35	36	37
	Y	46	47	42	44	41	44	48	43	42	45
11	X	20	19	18	21	23	22	23	24	25	25
	Y	40	38	36	45	42	50	49	45	36	40
12	X	26	32	26	31	34	33	38	40	42	44
	Y	35	37	41	36	25	27	26	22	21	24
13	X	20	22	19	15	20	19	18	17	16	14
	Y	18	19	16	13	18	17	16	15	13	15
14	X	27	24	23	25	24	26	29	24	29	25

	<i>Y</i>	30	26	25	27	28	30	28	23	31	27
15	<i>X</i>	19	20	24	23	28	25	24	26	29	30
	<i>Y</i>	18	19	25	21	24	23	22	23	25	26
16	<i>X</i>	46	48	51	50	49	48	46	47	49	46
	<i>Y</i>	35	37	36	38	34	35	32	36	40	36
17	<i>X</i>	16	15	18	16	15	17	16	18	13	14
	<i>Y</i>	26	30	28	29	30	24	25	22	25	28
18	<i>X</i>	34	36	36	38	37	37	34	35	32	35
	<i>Y</i>	28	26	30	31	26	29	28	26	27	29
19	<i>X</i>	20	19	18	21	23	22	23	24	25	22
	<i>Y</i>	30	38	36	35	32	40	39	35	36	40
20	<i>X</i>	27	28	21	25	29	23	27	21	25	24
	<i>Y</i>	36	32	36	33	40	28	40	35	38	36

Тема 2. Общие представления об анализе данных

Лабораторная работа №4. Регрессионный анализ

- Цель:** разработать регрессионные модели: 1) линейную $\hat{y}_i = b_1 x_i + b_0$,
2) логарифмическую $\hat{y}_i = b_1 \ln x_i + b_0$,
3) гиперболическую $\hat{y}_i = \frac{b_1}{x_i} + b_0$

Краткие теоретические сведения

1) Для построения модели линейной регрессии воспользуемся методом наименьших квадратов. Составим систему уравнений:

$$b_0 n + b_1 \sum x_i = \sum y_i,$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i.$$

Из системы уравнений находим коэффициенты b_0 и b_1 по формулам:

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Получим регрессионную линейную модель $\hat{y}_i = b_1 x_i + b_0$.

2) Для построения логарифмической модели регрессии воспользуемся методом наименьших квадратов. Составим систему уравнений:

$$b_0 n + b_1 \sum \ln x_i = \sum y_i,$$

$$b_0 \sum \ln x_i + b_1 \sum (\ln x_i)^2 = \sum y_i \ln x_i.$$

Из системы уравнений находим коэффициенты b_0 и b_1 :

$$b_0 = \frac{\sum \ln x_i \sum y_i \ln x_i - \sum y_i \sum (\ln x_i)^2}{(\sum \ln x_i)^2 - n \sum (\ln x_i)^2}, \quad b_1 = \frac{\sum \ln x_i \sum y_i - n \sum y_i \ln x_i}{(\sum \ln x_i)^2 - n \sum (\ln x_i)^2}.$$

Получим уравнение $\hat{y}_i = b_1 \ln x_i + b_0$

3) Для построения гиперболической модели регрессии воспользуемся методом наименьших квадратов. Составим систему уравнений:

$$b_0 n + b_1 \sum \frac{1}{x_i} = \sum y_i,$$

$$b_0 \sum \frac{1}{x_i} + b_1 \sum \frac{1}{x_i^2} = \sum \frac{y_i}{x_i}.$$

Из системы уравнений находим коэффициенты b_0 и b_1 :

$$b_0 = \frac{\sum \frac{1}{x_i} \sum \frac{y_i}{x_i} - \sum y_i \sum \frac{1}{x_i^2}}{\left(\sum \frac{1}{x_i} \right)^2 - n \sum \frac{1}{x_i^2}}, \quad b_1 = \frac{\sum \frac{1}{x_i} \sum y_i - n \sum \frac{y_i}{x_i}}{\left(\sum \frac{1}{x_i} \right)^2 - n \sum \frac{1}{x_i^2}}$$

Получим уравнение $\hat{y}_i = \frac{b_1}{x_i} + b_0$

ОЦЕНКА КАЧЕСТВА МОДЕЛИ

Для оценки качества модели определяют среднюю ошибку аппроксимации – это среднее отклонение расчетных данных от фактических. Чем меньше эти отклонения, тем ближе теоретические значения подходят к эмпирическим данным, это лучшее качество модели.

Среднюю ошибку аппроксимации рассчитывают по формуле:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\%$$

Допустимый предел средней ошибки аппроксимации 8-10%. На практике для грубого приближения регрессии к реальной зависимости полагают, что значение средней ошибки аппроксимации не должно превышать 12-15%. Маленькая ошибка аппроксимации указывает на хорошие перспективы при использовании модели для прогнозных расчетов.

ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ

Оценка адекватности уравнения регрессии в целом производится на основе F -критерия Фишера:

$$F = \frac{\bar{S}_y^2}{\bar{S}_{ост}^2}$$

(отношение общей и остаточной дисперсий показывает, во сколько раз уравнение регрессии предсказывает результаты опыта лучше, чем среднее $\bar{y}$),

где $\bar{S}_y^2 = \frac{1}{k_1} \left(\sum y_i^2 - \frac{1}{n} \left[\sum y_i \right]^2 \right)$ - общая дисперсия; $\bar{S}_{ост}^2 = \frac{1}{k_2} \sum (y_i - \hat{y}_i)^2$ - остаточная дисперсия.

Вычисленное значение F -критерия Фишера сравнивается с табличным значением: $F_{табл}(\alpha, k_1, k_2)$ при уровне значимости α и степенях свободы $k_1 = n - 1$, $k_2 = n - m - 1$, где n – число наблюдений, m – число параметров при переменной x (в нашем случае $m = 1$).

Если вычисленное значение $F > F_{табл}$, то полученное уравнение регрессии адекватно (статистически значимо). В противном случае ($F < F_{табл}$) уравнение регрессии неадекватно (статистически незначимо).

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Оценка значимости коэффициентов регрессии проводится по критерию Стьюдента путем сравнения их значений с величиной случайной ошибки:

$$t_{b_0} = \frac{b_0}{m_{b_0}}, \quad t_{b_1} = \frac{b_1}{m_{b_1}}$$

Случайные ошибки коэффициентов регрессии определяются по формулам:

$$m_{b_0} = \frac{S_{ост} \sqrt{\sum x_i^2}}{n\sigma_x}, \quad m_{b_1} = \frac{S_{ост}}{\sqrt{n}\sigma_x}$$

где $\bar{S}_{ост}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$ – остаточная дисперсия, $\sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 - \bar{x}^2}$ – среднее квадратическое отклонение.

Статистики t_{b_0} и t_{b_1} подчиняются распределению Стьюдента с числом степени свободы $\nu = n - 2$.

Из таблицы распределения критерия Стьюдента находят критическое значение $t_{кр} = t(\alpha, \nu)$.

Если $t_{b_0} > t_{кр}$, то параметр b_0 статистически значим. В противном случае ($t_{b_0} < t_{кр}$) параметр b_0 статистически незначим.

Если $t_{b_1} > t_{кр}$, то параметр b_1 статистически значим. В противном случае ($t_{b_1} < t_{кр}$) параметр b_1 статистически незначим.

Пример 1. Найти уравнение прямой линии регрессии по данным $n = 8$ наблюдений, которые получены при изучении зависимости количества продаж товара y от затрат на рекламу этого товара x :

x	1,5	4,0	5,0	7,0	8,5	10,0	11,0	12,5
y	5,0	4,5	7,0	6,5	9,5	9,0	11,0	9,0

Решение. Составим расчетную вспомогательную таблицу:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i = b_1 x_i + b_0$	$(y_i - \hat{y}_i)^2$	$A_i \%$
	1,5	5,0	2,25	25	7,5	4,525	0,2256	9,50
	4,0	4,5	16	20,25	18	5,85	1,8225	30,00
	5,0	7,0	25	49	35	6,38	0,3844	8,86
	7,0	6,5	49	42,25	45,5	7,44	0,8836	14,46
	8,5	9,5	72,25	90,25	80,75	8,235	1,6002	13,32
	10,0	9,0	100	81	90	9,03	0,0009	0,33
	11,0	11,0	121	121	121	9,56	2,0736	13,09
	12,5	9,0	156,25	81	112,5	10,355	1,8360	15,06
Σ	59,5	61,5	541,75	509,75	510,25	-	8,8269	104,61

$$\bar{x} = \frac{1}{n} \sum x_i = 7,44, \quad \bar{A} = \frac{1}{n} \sum A_i = 13,08$$

Вычислим коэффициенты b_0 и b_1 по формулам:

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} = 3,73, \quad b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = 0,53$$

Получим уравнение вида: $\hat{y}_i = 0,53x_i + 3,73$.

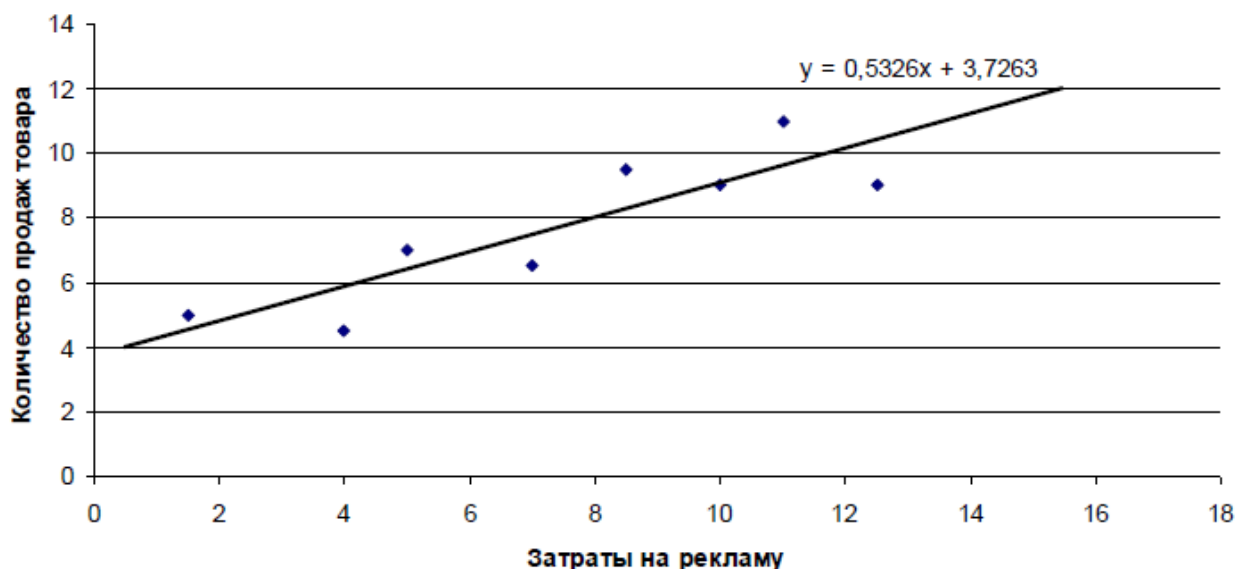
ОЦЕНКА КАЧЕСТВА МОДЕЛИ

Качество построенной модели оценим по средней ошибке аппроксимации:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\% = \frac{1}{n} \sum A_i = \frac{104,61}{8} = 13,08\%$$

Вывод: в среднем расчетные значения отклоняются от фактических на 13,08%.

Прямая, построенная по этому уравнению, показана на рисунке вместе с исходными данными.



ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ

Число степеней свободы $k_1 = n - 1 = 8 - 1 = 7$, $k_2 = n - 2 = 8 - 2 = 6$.
Общая дисперсия

$$\bar{S}_y^2 = \frac{1}{k_1} \left(\sum y_i^2 - \frac{1}{n} [\sum y_i]^2 \right) = 5,28$$

Остаточная дисперсия

$$\bar{S}_{ост}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = 1,47$$

F -критерий Фишера $F = \frac{\bar{S}_y^2}{\bar{S}_{ост}^2} = 3,59$.

Из таблицы значений F -критерия Фишера находим

$$F_{табл}(\alpha; k_1; k_2) = 4,21.$$

Вывод: $F < F_{табл}$ ($3,59 < 4,21$) *линейное уравнение регрессии адекватно описывает фактические значения количества продаж товара от затрат на рекламу этого товара*

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Случайные ошибки коэффициентов линейной регрессии:

$$m_{b_0} = \frac{S_{ocm} \sqrt{\sum x_i^2}}{n\sigma_x}, \quad m_{b_1} = \frac{S_{ocm}}{\sqrt{n}\sigma_x}$$

Вычислим среднее квадратическое отклонение:

$$\sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{8} 541,75 - 7,44^2} = 3,52,$$

$$m_{b_0} = \frac{S_{ocm} \sqrt{\sum x_i^2}}{n\sigma_x} = \frac{1,47 \sqrt{541,75}}{8 \cdot 3,52} = 1,22,$$

$$m_{b_1} = \frac{S_{ocm}}{\sqrt{n}\sigma_x} = \frac{1,47}{\sqrt{8} \cdot 3,52} = 0,15$$

Оценим значимости коэффициентов регрессии по критерию Стьюдента путем сравнения их значений с величиной случайной ошибки:

$$t_{b_0} = \frac{b_0}{m_{b_0}} = 3,06, \quad t_{b_1} = \frac{b_1}{m_{b_1}} = 3,53$$

Из таблицы распределения критерия Стьюдента находим критическое значение: $t_{кр} = t(\alpha, \nu) = t(\alpha, n - 2) = t(0,05; 6) = 2,45$.

Так как $t_{b_0} > t_{кр}$ и $t_{b_1} > t_{кр}$, то параметры b_0 и b_1 статистически значимы.

Вывод: модель по F-критерию Фишера в целом адекватна, и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений к осуществлению прогнозов.

Пример 2. Найти логарифмическое уравнение линии регрессии по данным $n = 8$ наблюдений, которые получены при изучении зависимости количества продаж товара уот затрат на рекламу этого товара x :

x	1,5	4,0	5,0	7,0	8,5	10,0	11,0	12,5
y	5,0	4,5	7,0	6,5	9,5	9,0	11,0	9,0

Решение. Составим расчетную вспомогательную таблицу:

	x_i	y_i	$\ln x_i$	$\ln^2 x_i$	$y_i \ln x_i$	$\hat{y}_i = b_1 \ln x_i + b_0$	$(y_i - \hat{y}_i)^2$	$A_i \%$
	1,5	5,0	0,4055	0,1644	2,0273	3,8040	1,4304	36,40
	4,0	4,5	1,3863	1,9218	6,2383	6,4650	3,8613	3,75
	5,0	7,0	1,6094	2,5903	11,2661	7,0704	0,0050	34,38
	7,0	6,5	1,9459	3,7866	12,6484	7,9833	2,2000	23,26
	8,5	9,5	2,1401	4,5799	20,3306	8,5100	0,9801	45,09
	10	9,0	2,3026	5,3019	20,7233	8,9509	0,0024	39,92
	11	11	2,3979	5,7499	26,3768	9,2095	3,2059	49,83
	12,5	9,0	2,5257	6,3793	22,7316	9,5563	0,3095	37,01
Σ	59,5	61,5	14,7134	30,4741	122,3424	-	11,9946	269,64

$$b_0 = \frac{\sum \ln x_i \sum y_i \ln x_i - \sum y_i \sum (\ln x_i)^2}{D} = 2,704, \quad b_1 = \frac{\sum \ln x_i \sum y_i - n \sum y_i \ln x_i}{D} = 2,713$$

Получим уравнение вида $\hat{y}_i = 2,713 \ln x_i + 2,704$

ОЦЕНКА КАЧЕСТВА МОДЕЛИ

Качество построенной модели оценим по средней ошибке аппроксимации:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\% = \frac{1}{n} \sum A_i = \frac{269,64}{8} = 33,71\%$$

Вывод: в среднем расчетные значения отклоняются от фактических на 33,71%.

ОЦЕНКА АДЕКВАТНОСТИ УРАВНЕНИЯ

Число степеней свободы $k_1 = n - 1 = 8 - 1 = 7$, $k_2 = n - 2 = 8 - 2 = 6$
Общая дисперсия

$$\bar{S}_y^2 = \frac{1}{k_1} \left(\sum y_i^2 - \frac{1}{n} [\sum y_i]^2 \right) = 5,28$$

Остаточная дисперсия

$$\bar{S}_{ост}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = \frac{11,9946}{6} = 2$$

Остаточная дисперсия 2,00.

$$F\text{-критерий Фишера } F = \frac{\bar{S}_y^2}{\bar{S}_{ост}^2} = \frac{5,28}{2} = 2,64$$

Из таблицы значений F -критерия Фишера находим

$$F_{табл}(\alpha; k_1; k_2) = 4,21.$$

Вывод: $F_{табл}(\alpha; k_1; k_2) < F_{k_1, k_2}$, $2,64 < 4,21$ логарифмическое уравнение регрессии адекватно описывает фактические значения количества продаж товара от затрат на рекламу этого товара

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Случайные ошибки коэффициентов регрессии:

$$m_{b_0} = \frac{S_{ост} \sqrt{\sum x_i^2}}{n \sigma_x}, \quad m_{b_1} = \frac{S_{ост}}{\sqrt{n} \sigma_x}$$

Вычислим среднее квадратическое отклонение:

$$\sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{8} 541,75 - 7,44^2} = 3,52$$

Тогда

$$m_{b_0} = \frac{S_{ост} \sqrt{\sum x_i^2}}{n \sigma_x} = \frac{2 \sqrt{541,75}}{8 \cdot 3,52} = 1,65,$$

$$m_{b_1} = \frac{S_{ост}}{\sqrt{n} \sigma_x} = \frac{2}{\sqrt{8} \cdot 3,52} = 0,06$$

Оценим значимости коэффициентов регрессии по критерию Стьюдента путем сравнения их значений с величиной случайной ошибки:

$$t_{b_0} = \frac{b_0}{m_{b_0}} = 1,64, \quad t_{b_1} = \frac{b_1}{m_{b_1}} = 45,05$$

Из таблицы распределения критерия Стьюдента находим критическое значение $t_{кр} = t(\alpha, \nu) = t(\alpha, n - 2) = t(0,05; 6) = 2,45$

Так как $t_{b_0} < t_{кр}$, то параметр b_0 статистически незначим.

Так как $t_{b_1} > t_{кр}$, то параметр b_1 статистически значим.

Вывод: модель по F-критерию Фишера в целом адекватна, но часть коэффициентов регрессии незначимы. В этом случае модель пригодна для принятия некоторых решений, но не для осуществления прогнозов.

Пример 3. Найти гиперболическое уравнение линии регрессии по данным $n = 8$ наблюдений, которые получены при изучении зависимости количества продаж товара y от затрат на рекламу этого товара x :

x	1,5	4,0	5,0	7,0	8,5	10,0	11,0	12,5
y	5,0	4,5	7,0	6,5	9,5	9,0	11,0	9,0

Решение. Составим расчетную вспомогательную таблицу:

	x_i	y_i	x_i^2	$\frac{y_i}{x_i}$	$\frac{1}{x_i}$	$\frac{1}{x_i^2}$	$\hat{y}_i = \frac{b_1}{x_i} + b_0$	$(y_i - \hat{y}_i)$	$A_i\%$
	1,5	5,0	2,25	3,33	0,667	0,4444	16,957	142,962	239,14
	4,0	4,5	16	1,13	0,250	0,0625	12,203	59,329	171,18
	5,0	7,0	25	1,40	0,200	0,0400	11,632	21,455	66,17
	7,0	6,5	49	0,93	0,143	0,0204	10,980	20,070	68,92
	8,5	9,5	72,25	1,12	0,118	0,0138	10,692	1,422	12,55
	10	9,0	100	0,90	0,100	0,0100	10,491	2,223	16,57
	11	11	121	1,00	0,091	0,0083	10,387	0,375	5,57
	12,5	9,0	156,25	0,72	0,080	0,0064	10,263	1,595	14,03
Σ	59,5	61,5	541,75	10,52	1,648	0,6059	-	249,431	594,13

$$b_0 = \frac{\sum \frac{1}{x_i} \sum \frac{y_i}{x_i} - \sum y_i \sum \frac{1}{x_i^2}}{D} = \frac{1,648 \cdot 10,52 - 61,5 \cdot 0,6059}{1,648^2 - 8 \cdot 0,6059} = 9,35,$$

$$b_1 = \frac{\sum \frac{1}{x_i} \sum y_i - n \sum \frac{y_i}{x_i}}{D} = \frac{1,648 \cdot 61,5 - 8 \cdot 10,52}{1,648^2 - 8 \cdot 0,6059} = 11,41$$

Получим уравнение вида $\hat{y}_i = \frac{11,41}{x_i} + 9,35$

ОЦЕНКА КАЧЕСТВА МОДЕЛИ

Качество построенной модели оценим по средней ошибке аппроксимации:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\% = \frac{1}{n} \sum A_i = \frac{594,13}{8} = 74,27\%$$

Вывод: в среднем расчетные значения отклоняются от фактических на 74,27%.

ОЦЕНКА АДЕКВАТНОСТИ УРАВНЕНИЯ

Число степеней свободы $k_1 = n - 1 = 8 - 1 = 7$, $k_2 = n - 2 = 8 - 2 = 6$
Общая дисперсия

$$\bar{S}_y^2 = \frac{1}{k_1} \left(\sum y_i^2 - \frac{1}{n} [\sum y_i]^2 \right) = \frac{509,75 - \frac{1}{8} \cdot 61,5^2}{7} = 5,28$$

Остаточная дисперсия

$$\bar{S}_{ост}^2 = \frac{1}{n - 2} \sum (y_i - \hat{y}_i)^2 = \frac{249,431}{6} = 41,57$$

Остаточная дисперсия 41,57.

$$F\text{-критерий Фишера } F = \frac{\bar{S}_y^2}{\bar{S}_{ост}^2} = \frac{5,28}{41,57} = 0,13$$

Из таблицы значений F -критерия Фишера находим $F_{табл}(\alpha; k_1; k_2) = 4,21$ регрессии адекватно описывает фактические значения количества продаж товара от затрат на рекламу этого товара.

Вывод: $F_{табл}(\alpha; k_1; k_2) < F_{k_1, k_2}$, $0,13 < 4,21$ *гиперболическое уравнение регрессии адекватно описывает фактические значения количества продаж товара от затрат на рекламу этого товара*

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Случайные ошибки коэффициентов регрессии:

$$m_{b_0} = \frac{S_{ост} \sqrt{\sum x_i^2}}{n \sigma_x}, \quad m_{b_1} = \frac{S_{ост}}{\sqrt{n} \sigma_x}$$

Вычислим среднее квадратическое отклонение:

$$\sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{8} 541,75 - 7,44^2} = 3,52$$

Тогда

$$m_{b_0} = \frac{S_{ост} \sqrt{\sum x_i^2}}{n \sigma_x} = \frac{41,57 \sqrt{541,75}}{8 \cdot 3,52} = 34,36,$$

$$m_{b_1} = \frac{S_{ост}}{\sqrt{n} \sigma_x} = \frac{41,57}{\sqrt{8} \cdot 3,52} = 1,19$$

Оценим значимости коэффициентов регрессии по критерию Стьюдента путем сравнения их значений с величиной случайной ошибки:

$$t_{b_0} = \frac{b_0}{m_{b_0}} = \frac{9,35}{34,36} = 0,27, \quad t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{11,41}{1,19} = 9,59$$

Из таблицы распределения критерия Стьюдента находим критическое значение $t_{кр} = t(\alpha, \nu) = t(\alpha, n - 2) = t(0,05; 6) = 2,45$

Так как $t_{b_0} < t_{кр}$, то параметр b_0 статистически незначим.

Так как $t_{b_1} > t_{кр}$, то параметр b_1 статистически значим.

Вывод: модель по F-критерию Фишера в целом адекватна, но часть коэффициентов регрессии незначимы. В этом случае модель пригодна для принятия некоторых решений, но не для осуществления прогнозов.

ВЫБОР ОПТИМАЛЬНОГО УРАВНЕНИЯ

Метод наименьших квадратов позволяет построить уравнение практически любой формы парной связи и выбрать оптимальное уравнение. Его сущность заключается в минимизации отклонений фактических значений y_i от предсказанных $\hat{y}_i$ по одному из уравнений регрессии. Из множества уравнений исследователю необходимо выбрать одно, которое бы хорошо описывало изучаемое явление и логически было обосновано.

Наилучшие результаты опыта дает **уравнение линейной регрессии**, так как оно характеризуется меньшей остаточной дисперсией ($\bar{S}_{ост}^2 = 1,47$) и наибольшим фактическим F-критерием ($F_{табл} = 3,59$), а также имеет самую маленькую ошибку аппроксимации – среднее отклонение расчетных данных от фактических (13,08%). По критерию Стьюдента оба коэффициента в уравнении регрессии значимы ($t_1 > t_{кр}$ и $t_2 > t_{кр}$).

Вид уравнения регрессии	Степень свободы		Дисперсия		Критерий Фишера		Ср. ошиб аппр	Критерий Стьюдента		
	k_1	k_2	общ	ост.	$F_{табл}$	F_{k_1, k_2}	$\bar{A}_i \%$	t_1	t_2	$t_{кр}$
$\hat{y}_i = 0,54x_i + 3,73$	7	6	5,28	1,47	3,59	4,21	13,08	3,06	3,53	2,45
$\hat{y}_i = 2,703 \ln x_i + 2,704$	7	6	5,28	2,00	2,64	4,21	33,71	1,64	45,05	2,45
$\hat{y}_i = \frac{11,41}{x_i} + 9,35$	7	6	5,28	41,57	0,13	4,21	74,27	0,27	9,59	2,45

Задание. Разработать регрессионные модели: 1) линейную $\hat{y}_i = b_1 x_i + b_0$, 2) логарифмическую $\hat{y}_i = b_1 \ln x_i + b_0$, 3) гиперболическую $\hat{y}_i = \frac{b_1}{x_i} + b_0$.

Построить заданный разброс точек и полученные графики.

Оценить качество и адекватность моделей, значимость коэффициентов регрессионных моделей, выбрать оптимальное уравнение.

1	X	18	19	17	18	16	19	18	22	17	18	15	20	18	17	21
	Y	14	19	16	14	14	16	14	21	18	16	19	18	18	18	12
2	X	22	23	20	22	21	20	21	18	16	22	18	25	13	23	17
	Y	24	21	21	19	23	23	25	21	20	19	24	23	18	23	18
3	X	17	20	19	22	18	16	23	25	14	16	22	18	24	22	23
	Y	19	22	18	24	22	17	20	22	19	21	19	21	19	19	22
4	X	16	13	16	14	19	17	21	17	21	10	20	19	11	13	20
	Y	15	16	17	14	18	16	18	16	21	11	17	21	12	12	17
5	X	22	15	19	21	23	18	23	19	18	21	17	20	16	22	20
	Y	18	13	15	16	19	17	19	18	15	18	13	17	15	20	19
6	X	18	13	16	19	20	20	17	17	18	15	13	16	15	15	16
	Y	17	14	17	18	22	19	18	21	20	17	19	17	14	21	15
7	X	14	17	17	15	19	16	18	17	16	18	21	20	15	19	21
	Y	10	2	3	8	5	7	6	4	6	4	2	3	9	2	5
8	X	19	20	15	16	15	14	16	18	19	17	17	18	17	23	11
	Y	17	19	17	15	14	18	19	20	14	13	19	20	17	16	15
9	X	19	22	16	14	21	20	18	17	15	22	15	18	16	20	17
	Y	19	21	18	15	25	21	19	16	17	19	22	24	15	17	20
10	X	15	21	16	14	19	20	18	17	20	22	19	22	23	20	12
	Y	16	18	18	15	16	21	19	15	18	20	20	21	21	21	18
11	X	16	22	18	18	14	20	19	18	20	17	16	15	19	17	20
	Y	17	20	16	17	16	22	19	20	20	21	19	17	17	18	18
12	X	18	19	21	18	16	19	18	21	17	18	16	22	16	20	21
	Y	14	17	18	14	14	15	18	20	15	16	17	19	17	19	17
13	X	22	18	19	20	21	19	17	22	18	16	15	21	23	21	17
	Y	18	17	17	19	20	17	17	20	16	12	16	17	16	19	18
14	X	20	21	17	19	19	16	23	23	15	16	20	18	14	18	22
	Y	22	20	18	20	22	19	21	19	21	19	21	19	19	21	15
15	X	16	14	15	12	19	16	21	17	21	18	20	19	17	18	20
	Y	15	15	17	18	16	18	18	23	16	19	18	15	20	15	14
16	X	14	15	19	21	23	21	23	17	20	25	18	20	14	14	16
	Y	15	18	20	18	20	22	13	22	15	17	18	20	15	21	19
17	X	22	13	22	12	23	21	17	20	18	20	17	13	23	12	15
	Y	17	20	19	22	18	16	23	25	14	16	22	18	24	22	23
18	X	17	20	16	17	16	22	19	20	20	21	19	17	17	18	18
	Y	11	24	13	25	15	17	12	14	16	17	11	12	11	16	12
19	X	19	21	18	15	25	21	19	16	17	19	22	24	15	17	20
	Y	25	15	21	17	15	23	24	22	14	18	17	22	25	15	16
20	X	36	17	31	20	19	30	18	31	15	15	18	30	21	19	15
	Y	16	22	18	18	14	20	19	18	20	17	16	15	19	17	20
21	X	33	17	25	30	26	27	27	30	31	17	19	20	25	25	32
	Y	18	19	17	18	16	19	18	22	17	18	15	20	18	17	21

Лабораторная работа №5. Криволинейная (параболическая) корреляция

Цель: разработать криволинейные регрессионные модели.

Краткие теоретические сведения

Если график регрессии изображается кривой линией, то корреляцию называют криволинейной. Проведем разработку параболической корреляции второго порядка. Выборочное уравнение регрессии Y на X имеет вид

$$\bar{y}_x = Ax^2 + Bx + C$$

Неизвестные параметры A, B, C найдем из системы трех уравнений

$$\begin{cases} A \sum n_{xi} x_i^4 + B \sum n_{xi} x_i^3 + C \sum n_{xi} x_i^2 = \sum n_{xi} x_i^2 \bar{y}_{xi}, \\ A \sum n_{xi} x_i^3 + B \sum n_{xi} x_i^2 + C \sum n_{xi} x_i = \sum n_{xi} x_i \bar{y}_{xi}, \\ A \sum n_{xi} x_i^2 + B \sum n_{xi} x_i + Cn = \sum n_{xi} \bar{y}_{xi}, \end{cases} \quad (5.1)$$

Аналогично находится выборочное уравнение регрессии X на Y

$$\bar{x}_y = A_1 y^2 + B_1 y + C_1.$$

Для оценки силы корреляции Y на X служит выборочное корреляционное отношение (отношение межгруппового среднего квадратичного отклонения к общему среднему квадратичному отклонению признака Y)

$$\eta_{yx} = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}}, \quad \text{или (в других обозначениях)} \quad \eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y},$$

$$\text{здесь } \sigma_{\bar{y}_x} = \sqrt{D_{\text{межгр}}} = \sqrt{\frac{\sum n_{xi} (\bar{y}_{xi} - \bar{y})^2}{n}}, \quad \sigma_y = \sqrt{D_{\text{общ}}} = \sqrt{\frac{\sum n_{yi} (y_i - \bar{y})^2}{n}},$$

где n - объем выборки (сумма всех частот), n_{xi} - частота значения x_i признака X , n_{yi} - частота значения y_i признака Y , $\bar{y}_x$ - условная средняя признака Y , $\bar{y}$ - общая средняя признака Y .

Аналогично определяется выборочное корреляционное отношение X к Y

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}.$$

1. Выборочное корреляционное отношение удовлетворяет двойному неравенству $0 \leq \eta_{yx} \leq 1$.

2. Если $\eta_{yx} = 0$, то признак Y не связан с признаком X корреляционной зависимостью.

3. Если $\eta_{yx} = 1$, то признак Y связан с признаком X функциональной зависимостью. Справедливо и обратное утверждение: если признак Y связан с признаком X функциональной зависимостью, то $\eta_{yx} = 1$.

При возрастании η_{yx} значения признака Y внутри всех групп, соответствующих различным наблюдаемым значениям признака X , будут все меньше различаться между собой, корреляционная связь между Y и X будет становиться все теснее и при $\eta_{yx} = 1$ превратится в функциональную.

4. Выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции: $\eta_{yx} \geq |r_k|$.

5. Если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то между признаками Y и X имеет место точная линейная корреляционная зависимость. Верно и обратное утверждение: если между признаками Y и X имеет место точная линейная корреляционная зависимость, то $\eta_{yx} = |r_k|$,

$$r_k = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}.$$

Достоинством выборочного корреляционного отношения является то, что оно служит мерой тесноты связи любой формы, в том числе и нелинейной. Таким образом, η_{yx} является более общей характеристикой связи, чем выборочный коэффициент корреляции, который служит мерой тесноты лишь линейной связи. Недостатком выборочного корреляционного отношения может считаться то, что оно никак не указывает, какая именно (линейная, параболическая, гиперболическая, ..) форма связи имеет место между признаками.

Вычислять корреляционное отношение имеет смысл, если связь между признаками далека от линейной ($|r_k| \leq 0,7$). Более строгая оценка линейности

связи может быть получена с помощью критерия $t = \frac{k}{\delta_k}$, где $k = \eta_{yx}^2 - r_k^2$ -

мера криволинейности, δ_k - основная ошибка меры k , вычисляемая по

$$\text{формуле } \delta_k = \frac{2\sqrt{k + k^2(2 - \eta_{yx}^2 - r_k^2)}}{\sqrt{n}}.$$

Пример. Данные приведены в таблице. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ Y на X . Оценить силу корреляционной связи по выборочному корреляционному соотношению.

$X \backslash Y$	2	3	5	n_y
25	20		1	
45		29	1	
110		2	47	
n_x				

Решение. Пополним данные таблицы

$X \backslash Y$	2	3	5	n_{yi}
------------------	---	---	---	----------

$y \backslash x$				
25	20		1	21
45		29	1	30
110		2	47	49
n_{xi}	20	31	49	$n=100$

Составим расчетную таблицу

$\bar{x}_i$	n_{xi}	$\bar{y}_{xi}$	$n_{xi}\bar{x}_i$	$n_{xi}\bar{x}_i^2$	$n_{xi}\bar{x}_i^3$	$n_{xi}\bar{x}_i^4$	$n_{xi}\bar{y}_{xi}$	$n_{xi}\bar{y}_{xi}\bar{x}_i$	$n_{xi}\bar{y}_{xi}\bar{x}_i^2$
2	20	25	40	80	160	320	500	1000	2000
3	31	49,19	93	279	837	2511	1524,89	4574,67	13724,01
5	49	106,9	245	1225	6125	30625	5238,1	26190,5	130952,5
Σ	100		378	1584	7122	33456	7262,99	31765,17	146676,51

Подставим полученные данные в систему (5.1)

$$\begin{cases} 33456A + 7122B + 1584C = 146676,51 \\ 7122A + 1584B + 378C = 31765,17 \\ 1584A + 378B + 100C = 7262,99 \end{cases}$$

Воспользуемся для решения системы программным пакетом Mathcad

$$G := \begin{pmatrix} 33456 & 7122 & 1584 \\ 7122 & 1584 & 378 \\ 1584 & 378 & 100 \end{pmatrix} \quad Q := \begin{pmatrix} 146676.51 \\ 31765.17 \\ 7262.99 \end{pmatrix} \quad X := \begin{pmatrix} A \\ B \\ C \end{pmatrix}$$

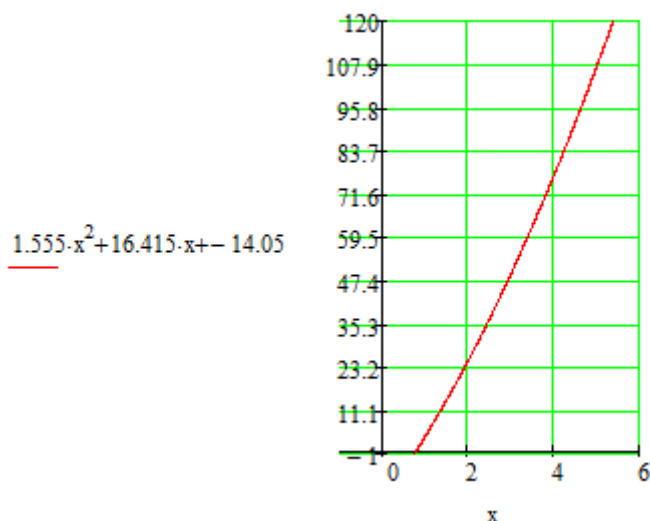
$$X := G^{-1} \cdot Q$$

$$X = \begin{pmatrix} 1.555 \\ 16.415 \\ -14.05 \end{pmatrix}$$

Выборочное уравнение регрессии Y на X имеет вид

$$\bar{y}_x = 1,555x^2 + 16,415x - 14,05$$

Построим график уравнения регрессии



Для того чтобы оценить силу корреляционной связи вычислим выборочное корреляционное отношение η_{yx} . Найдем предварительно общую

среднюю $\bar{y}$

$$\bar{y} = \frac{21 \cdot 25 + 30 \cdot 45 + 110 \cdot 49}{100} = 72,65;$$

общее среднее квадратичное отклонение признака Y

$$\begin{aligned}\sigma_y &= \sqrt{\frac{\sum n_{yi}(y_i - \bar{y})^2}{n}} = \\ &= \sqrt{\frac{21(25 - 72,65)^2 + 30(45 - 72,65)^2 + (110 - 72,65)^2 49}{100}} = 37,07;\end{aligned}$$

межгрупповое среднее квадратичное отклонение

$$\begin{aligned}\sigma_{\bar{y}_x} &= \sqrt{\frac{\sum n_{xi}(\bar{y}_{xi} - \bar{y})^2}{n}} = \\ &= \sqrt{\frac{21(25 - 72,65)^2 + 30(49,19 - 72,65)^2 + (106,9 - 72,65)^2 49}{100}} = 33,06;\end{aligned}$$

вычислим выборочное корреляционное отношение

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y} = \frac{33,06}{37,07} = 0,89.$$

Вычислим выборочный коэффициент корреляции, для этого найдем:
общую среднюю $\bar{x}$

$$\bar{x} = \frac{20 \cdot 2 + 31 \cdot 3 + 5 \cdot 49}{100} = 3,78;$$

общее среднее квадратичное отклонение признака X

$$\begin{aligned}\sigma_x &= \sqrt{\frac{\sum n_{xi}(x_i - \bar{x})^2}{n}} = \\ &= \sqrt{\frac{20(2 - 3,78)^2 + 31(3 - 3,78)^2 + (5 - 3,78)^2 49}{100}} = 1,245;\end{aligned}$$

$$\begin{aligned}r_k &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} = \\ &= \frac{(25 - 72,65)(2 - 3,78) + (45 - 72,65)(3 - 3,78) + (110 - 72,65)(5 - 3,78)}{100 \cdot 1,245 \cdot 37,07} = 0,033\end{aligned}$$

Вывод:

значение $\eta_{yx} = 0,89$ близко к 1, поэтому признак Y связан с признаком X функциональной зависимостью;

корреляционная связь между Y и X тесная;

так как коэффициент корреляции $r_k = 0,033 < 0,7$, следовательно, связь между признаками нелинейная;

выборочное корреляционное отношение больше абсолютной величины выборочного коэффициента корреляции: $\eta_{yx} \geq |r_k|$.

n_x									
-------	--	--	--	--	--	--	--	--	--

5. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и $\bar{x}_y = Ay^2 + By + C$ на X . По данным, приведенным в следующей таблице оценить силу корреляционной связи

$y \backslash X$	11	15	19	23	27	31	35	39	n_y
125					2	1	4	3	
130				5	6	10	5	2	
145			2	7	6	9	1		
160		4	8	3	1				
175	3	2							
190	2	4							
n_x									

6. Найти выборочные уравнения регрессии $\bar{x}_y = Ay^2 + By + C$ и выборочное корреляционное отношение η_{xy} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи.

$y \backslash X$	18	23	28	33	38	43	48	n_y
125		1						
150	1	2	5					
175		3	2	12				
200			1	8	7			
225					3	3		
250						1	1	
n_x								

7. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы.

$y \backslash X$	0	1	2	3	4	n_y
0	18	1	1			
3	1	20				
5	3	5	10	2		
10			7	12		
17					20	
n_x						

8. Найти выборочные уравнения регрессии параболического типа X на Y по данным, приведенным в следующей корреляционной таблице. Оценить силу корреляционной связи по выборочному корреляционному соотношению.

$y \backslash X$	11	15	19	23	27	31	35	39	n_y
125					2	1	4	3	
130				5	6	10	5	2	
145			2	7	6	9	1		
160		5	8	3	1				

175	3	2							
n_x									

9. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице.

$y \backslash X$	0	4	6	7	10	n_y
7	19	1	1			
13	2	14				
40		3	22	2		
80				15		
200					21	
n_x						

10. Найти выборочные уравнения параболической регрессии X на Y . По данным, приведенным в следующей корреляционной таблице оценить силу корреляционной связи.

$y \backslash X$	10	20	30	40	50	60	70	80	n_y
12						1	4	3	
20					3	10	5	2	
28			2	7	6	9	1		
36		4	8	3	1				
44	4	2							
52	1	4							
n_x									

11. Найти выборочные уравнения параболической регрессии Y на X по данным, приведенным в следующей корреляционной таблице. Найти выборочное корреляционное отношение η_{yx} и сделать выводы.

$y \backslash X$	10	20	30	40	50	60	70	80	n_y
12						1	4	3	
20					3	10	5	2	
28			2	7	6	9	1		
36		4	8	3	1				
44	5	2	2	2					
n_x									

12. Найти выборочные уравнения регрессии $\bar{x}_y = Ay^2 + By + C$ и выборочное корреляционное отношение η_{xy} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи.

$y \backslash X$	2	4	6	8	10	12	14	16	n_y
12							4	13	
20					3	11	5	2	
28			2	2	6	9	1		
36		4	18	3	1				
44	5	12	2	2					

n_x									
-------	--	--	--	--	--	--	--	--	--

13. Найти выборочные уравнения регрессии вида $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционно отношение η_{yx} по данным, приведенным в следующей таблице

$y \backslash X$	0	4	5	7	12	17	24	n_y
1	13		1	1				
35		12	2	15	1			
50				1	4	5	20	
n_x								

14. Найти выборочные уравнения регрессии $\bar{x}_y = Ay^2 + By + C$ и выборочное корреляционное отношение η_{xy} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

$y \backslash X$	0	2	4	6	8	n_y
10	20	5				
11	7	15	3	1		
20		3	17	14		
35			8	23	7	
50				5	22	
n_x						

15. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

$y \backslash X$	0	3	6	9	12	15	r
2	9	1					
13	12	13					
30		3	22	5			
60				15	2	3	
100					21	44	
n_x							

16. Найти выборочные уравнения регрессии $\bar{x}_y = Ay^2 + By + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

$y \backslash X$	7	8	9	10	11	n_y
200	41	7				
300	1	52	1			
400		8	40	5		
500				15	30	

n_x						
-------	--	--	--	--	--	--

17. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

X \ y	1	9	19	29	32	n_y
0	13					
2	2	10				
3	1	1	23	11	4	
5				5	20	
n_x						

18. Найти выборочные уравнения регрессии $\bar{x}_y = Ay^2 + By + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

X \ y	6	30	50	70	90	n_y
1	15					
3	1	24				
4		2	18	5	10	
5				15	20	
n_x						

19. Найти выборочные уравнения регрессии $\bar{y}_x = Ax^2 + Bx + C$ и выборочное корреляционное отношение η_{yx} по данным, приведенным в следующей таблице. Сделать выводы о корреляционной связи

X \ y	0	4	5	7	12	17	24	n_y
1	23	1						
3	1	12	2	5	1			
5			1	12	4	27	20	
n_x								

Лабораторная работа №6. Критерий согласия Пирсона

Цель работы: овладеть критерием Пирсона.

Краткие теоретические сведения

Пусть эмпирическое распределение задано в виде последовательности равностоящих вариантов и соответствующих им частот.

Критерием согласия Пирсона называется критерий проверки гипотезы о нормальном распределении генеральной совокупности.

Для того, чтобы при заданном уровне значимости α проверить гипотезу о нормальном распределении генеральной совокупности необходимо:

а) вычислить выборочную среднюю $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B ;

б) вычислить теоретические частоты

$$n'_i = \frac{nh}{\sigma_B} \varphi(u_i)$$

где n – объем выборки (сумма всех частот), h – шаг (разность между двумя соседними вариантами),

$$u_i = \frac{x_i - \bar{x}_B}{\sigma_B}, \quad \varphi(u_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

– вычисляется по таблице.

в) вычислить наблюдаемое значение критерия

$$\chi^2_{набл} = \sum_{i=1}^s \frac{(n_i - n'_i)^2}{n'_i}$$

где n_i – эмпирические (опытные) частоты, n'_i – теоретические частоты, s – число групп выборки;

г) найти $\chi^2_{кр}(\alpha, k)$, где $k = s - 3$, по таблице критических точек распределения χ^2 ;

д) если $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть нулевую гипотезу о нормальном распределении генеральной совокупности. Другими словами, эмпирические и теоретические частоты различаются незначимо.

Если $\chi^2_{набл} > \chi^2_{кр}$ – гипотезу отвергаем, то есть, эмпирические и теоретические частоты различаются значимо.

З а м е ч а н и е . Малочисленные частоты ($n_i < 5$) следует объединить; в этом случае и соответствующие им теоретические частоты также надо сложить. Если производилось объединение частот, то при определении числа степеней свободы по формуле $k = s - 3$ следует в качестве s принять число групп выборки, оставшихся после объединения частот.

Пример. По данному распределению выборки проверить гипотезу о нормальном распределении генеральной совокупности при уровне значимости $\alpha = 0,01$.

x_i	10,6	10,8	11,0	11,2	11,4	11,6	11,8
n_i	5	10	17	30	20	12	6

Решение.

а) Выборочная средняя

$$\bar{x}_B = \frac{\sum_{i=1}^s x_i n_i}{n} = \frac{10,6 \cdot 5 + 10,8 \cdot 10 + 11,0 \cdot 17 + 11,2 \cdot 30 + 11,4 \cdot 20 + 11,6 \cdot 12 + 11,8 \cdot 6}{5 + 10 + 17 + 30 + 20 + 12 + 6} = 11,22$$

и выборочная дисперсия

$$D_B = \frac{\sum_{i=1}^s (x_i - \bar{x}_B)^2 n_i}{n} = \frac{5(10,6 - 11,22)^2 + 10(10,8 - 11,22)^2 + 17(11,0 - 11,22)^2 + 30(11,2 - 11,22)^2 + 20(11,4 - 11,22)^2 + 12(11,6 - 11,22)^2 + 6(11,8 - 11,22)^2}{100} = 0,0892$$

тогда выборочное среднее квадратическое отклонение

$$\sigma_B = \sqrt{D_B} = 0,2987.$$

б) Вычислим теоретические частоты, учитывая, что $n = 100$, $h = 0,2$, $\sigma_B = 0,2987$, по формуле

$$n'_i = \frac{nh}{\sigma_B} \varphi(u_i) = \frac{100 \cdot 0,2}{0,2987} \varphi(u_i) = 67 \varphi(u_i).$$

Составим расчетную таблицу:

x_i	$u_i = \frac{x_i - \bar{x}_B}{\sigma_B}$	$\varphi(u_i)$	$n'_i = 67 \varphi(u_i)$
10,6	-2,08	0,0459	3
10,8	-1,41	0,1476	10
11,0	-0,74	0,3034	21
11,2	-0,07	0,3980	27
11,4	0,60	0,3332	23
11,6	1,27	0,1781	2
11,8	1,94	0,0608	4

в) Для вычисления

$$\chi^2_{набл} = \sum_{i=1}^7 \frac{(n_i - n'_i)^2}{n'_i},$$

составим расчетную таблицу:

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
1	5	3	2	4	1,333

2	10	10	0	0	0
3	17	21	-4	16	0,762
4	30	27	3	9	0,333
5	20	23	-3	9	0,391
6	12	12	0	0	0
7	6	4	2	4	1
Σ	100				3,819

Таким образом, $\chi^2_{набл} = 3,819$

г) По таблице критических точек распределения найдем:
 $\chi^2_{кр}(\alpha, k) = \chi^2_{кр}(0,01; 7 - 3 = 4) = 13,3$

д) Так как $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть нулевую гипотезу о нормальном распределении генеральной совокупности. Другими словами, эмпирические и теоретические частоты различаются незначимо.

Задание

Используя критерий Пирсона, при уровне значимости α проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки:

1	α	x_i	12	14	16	18	20	22	24
	0,01	n_i	6	8	12	24	22	17	11
2	α	x_i	26	28	30	32	34	36	38
	0,025	n_i	10	15	16	18	20	13	8
3	α	x_i	25	30	35	40	45	50	55
	0,05	n_i	13	14	16	25	12	11	9
4	α	x_i	11	13	15	17	19	21	23
	0,95	n_i	9	10	17	21	16	15	12
5	α	x_i	14	17	20	23	26	29	32
	0,975	n_i	10	12	15	20	18	16	9
6	α	x_i	17	19	21	23	25	27	29
	0,99	n_i	9	12	16	24	15	13	11
7	α	x_i	10	14	18	22	26	30	34
	0,01	n_i	6	8	12	18	23	19	14
8	α	x_i	23	25	27	29	31	33	35
	0,025	n_i	8	14	17	23	15	14	9
9	α	x_i	6	10	14	18	22	26	30
	0,05	n_i	2	15	27	21	14	13	8
10	α	x_i	12	15	18	21	24	27	30
	0,95	n_i	13	15	17	20	14	13	8
11	α	x_i	11	15	19	23	27	31	35
	0,975	n_i	12	14	18	20	16	11	9
12	α	x_i	8	12	16	20	24	28	32

	0,99	n_i	8	13	18	22	15	10	9
13	α	x_i	13	16	19	22	25	28	31
	0,01	n_i	11	15	30	23	12	9	1
14	α	x_i	10	12	14	16	18	20	22
	0,025	n_i	12	13	18	20	14	13	10
15	α	x_i	14	18	22	26	30	34	38
	0,05	n_i	8	22	7	19	13	20	11
16	α	x_i	10	13	16	19	22	25	28
	0,95	n_i	12	17	20	16	24	3	8
17	α	x_i	15	20	25	30	35	40	45
	0,975	n_i	7	9	11	29	23	18	3
18	α	x_i	20	22	24	26	28	30	32
	0,99	n_i	11	14	17	29	16	10	3
19	α	x_i	17	20	23	26	29	32	35
	0,01	n_i	10	13	19	22	13	12	11
20	α	x_i	20	26	32	38	44	50	56
	0,025	n_i	2	17	28	21	16	10	6

Тема 3. Методы описательной статистики

Лабораторная работа №7 Статистическая проверка статистических гипотез

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений.

Нулевой(основной) называют выдвинутую гипотезу H_0 . *Конкурирующей (альтернативной)* называют гипотезу H_1 , которая противоположна нулевой.

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

Простой называют гипотезу, содержащую только одно предположение.

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

Выдвинутая гипотеза может быть правильной или неправильной; поэтому возникает необходимость ее проверки.

Поскольку проверку производят статистическими методами, ее называют статистической. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза. Вероятность ошибки первого рода называют уровнем значимости и обозначают через α .

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза. Вероятность ошибки второго рода обозначают через β .

Последствия этих ошибок могут оказаться различными.

Статистическим критерием (или просто *критерием*) называют случайную величину K , которая служит для проверки нулевой гипотезы.

Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин, и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением $K_{\text{набл.}}$ называют значение критерия, вычисленное по выборкам.

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

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

Рассмотрим *основной принцип проверки статистических значений*: если наблюдаемое значение критерия принадлежит критической области, то гипотезу отвергают; если наблюдаемое значение критерия принадлежит области принятия гипотезы, то гипотезу принимают.

Поскольку критерий K – случайная одномерная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы также являются интервалами и, следовательно, существуют точки, которые их разделяют.

Критическими точками (границами) $k_{кр.}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую неравенством $K > K_{кр.}$, где $k_{кр.}$ – положительное число (рис.1).

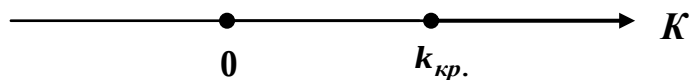


Рис. 1

Левосторонней называют критическую область, определяемую неравенством $K < K_{кр.}$, где $k_{кр.}$ – отрицательное число (рис. 2).



Рис. 2

Односторонней называют правостороннюю или левостороннюю критическую область.

Двусторонней называют критическую область, определяемую неравенствами

$$K < k_1, K > k_2, \text{ где } k_2 > k_1.$$

В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами (в предположении, что $k_{кр.} > 0$) (рис.3):

$$K < -k_{кр.}, K > k_{кр.} \Leftrightarrow |K| > k_{кр.}$$

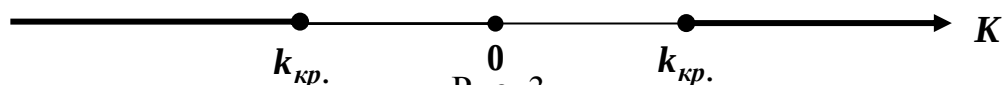


Рис. 3

Для отыскания критической области задаются достаточно малой вероятностью – уровнем значимости α . Затем ищут критические точки $k_{кр.}$, исходя из требования, чтобы при условии справедливости нулевой гипотезы, вероятность того, что критерий K примет значение, большее $k_{кр.}$, была равна принятому уровню значимости:

А) для **правосторонней** критической области

$$P(K > k_{кр.}) = \alpha, k_{кр.} > 0;$$

Б) для **левосторонней** критической области

$$P(K < k_{кр.}) = \alpha, k_{кр.} < 0;$$

В) для **двусторонней** симметричной критической области

$$P(K > k_{\text{кр.}}) = \frac{\alpha}{2}, \quad k_{\text{кр.}} > 0;$$

$$P(K < -k_{\text{кр.}}) = \frac{\alpha}{2}.$$

I. Сравнение двух дисперсий нормальных генеральных совокупностей

По независимым выборкам, объемы которых n_1 и n_2 , извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии s_x^2 и s_y^2 . Требуется сравнить эти дисперсии.

Правило 1.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(X) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H_1: D(X) > D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей исправленной выборочной дисперсии к меньшей)

$$F_{\text{набл}} = \frac{s_{\text{max}}^2}{s_{\text{min}}^2}$$

и по таблице критических точек распределения Фишера – Снедекора, по заданному уровню значимости α и числом степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ (k_1 - число степеней свободы большей исправленной выборочной дисперсии), найти критическую точку $F_{\text{кр}}(\alpha, k_1, k_2)$.

Если $F_{\text{набл}} < F_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 1.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(X) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H_1: D(X) \neq D(Y)$, надо вычислить наблюдаемое значение критерия по уровню значимости $\frac{\alpha}{2}$ (вдвое меньшему заданного) и числом степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ (k_1 - число степеней свободы большей исправленной выборочной дисперсии), найти критическую точку $F_{\text{кр}}\left(\frac{\alpha}{2}, k_1, k_2\right)$.

Если $F_{\text{набл}} < F_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу отвергают.

II. Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

По независимой выборке объема n найдена исправленная дисперсия s^2

Правило 2.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2$ о равенстве неизвестной генеральной дисперсии σ^2 гипотетическому (предполагаемому) значению σ_0^2 , при конкурирующей гипотезе $H_1: \sigma^2 > \sigma_0^2$, надо вычислить наблюдаемое значение критерия

$$\chi_{\text{набл}}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

и по таблице критических точек распределения χ^2 , по заданному уровню значимости α и числом степеней свободы $k = n-1$, найти критическую точку $\chi_{\text{кр}}^2(\alpha, k)$.

Если $\chi_{\text{набл}}^2 < \chi_{\text{кр}}^2$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{набл}}^2 > \chi_{\text{кр}}^2$ - нулевую гипотезу отвергают.

Правило 2.2. При конкурирующей гипотезе $H_1: \sigma^2 \neq \sigma_0^2$, надо вычислить левую $\chi_{\text{лев.кр}}^2\left(1 - \frac{\alpha}{2}, k\right)$ и правую $\chi_{\text{пр.кр}}^2\left(\frac{\alpha}{2}, k\right)$ критические точки критерия по уровню значимости $\frac{\alpha}{2}$ (вдвое меньшему заданного).

Если $\chi_{\text{лев.кр}}^2 < \chi_{\text{набл}}^2 < \chi_{\text{пр.кр}}^2$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{лев.кр}}^2 > \chi_{\text{набл}}^2$ или $\chi_{\text{набл}}^2 > \chi_{\text{пр.кр}}^2$ - нулевую гипотезу отвергают.

Правило 2.3. При конкурирующей гипотезе $H_1: \sigma^2 < \sigma_0^2$, надо вычислить $\chi_{\text{кр}}^2(1 - \alpha, k)$ критическую точку критерия.

Если $\chi_{\text{набл}}^2 > \chi_{\text{кр}}^2(1-\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{набл}}^2 < \chi_{\text{кр}}^2(1-\alpha, k)$ - нулевую гипотезу отвергают.

Замечание. Если число степеней свободы $k > 30$, то критическую точку $\chi_{\text{кр}}^2(\alpha, k)$ можно найти из равенства Уилсона-Гильферти

$$\chi_{\text{кр}}^2(\alpha, k) = k \left(1 - \frac{2}{9k} + z_{\alpha} \sqrt{\frac{2}{9k}} \right)^3,$$

где z_{α} находят, используя функцию Лапласа

$$\Phi(z_{\alpha}) = \frac{1}{2} - \alpha.$$

III. Сравнение двух средних генеральных совокупностей, дисперсии которых известны (большие независимые выборки)

По независимым выборкам, объемы которых $n_1 > 30$ и $n_2 > 30$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}$, $\bar{y}$, генеральные дисперсии известны.

Правило 3.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве математических ожиданий (генеральных средних) двух нормальных совокупностей при конкурирующей гипотезе, $H_1: M(X) \neq M(Y)$ надо вычислить наблюдаемое значение критерия

$$Z_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}}$$

и по таблице функции Лапласа найти критическую точку $z_{\text{кр}}$ из равенства

$$\Phi(z_{\text{кр}}) = \frac{1-\alpha}{2}$$

Если $|Z_{\text{набл}}| < z_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|Z_{\text{набл}}| > z_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 3.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 при конкурирующей гипотезе $H_1: M(X) > M(Y)$,

надо найти критическую точку $z_{кр}$ по таблице функции Лапласа из равенства

$$\Phi(z_{кр}) = \frac{1}{2} - \alpha$$

Если $Z_{набл} < z_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $Z_{набл} > z_{кр}$ - нулевую гипотезу отвергают.

Правило 3.3. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 при конкурирующей гипотезе $H_1: M(X) < M(Y)$, надо найти критическую точку $z_{кр}$ по правилу 2

Если $Z_{набл} > -z_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $Z_{набл} < -z_{кр}$ - нулевую гипотезу отвергают.

IV. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки)

По независимым выборкам, объемы которых $n_1 < 30$ и $n_2 < 30$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}$, $\bar{y}$, исправленные выборочные дисперсии s_x^2 , s_y^2 неизвестны, но равны.

Правило 4.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве математических ожиданий (генеральных средних) двух нормальных совокупностей с неизвестными, но равными дисперсиями при конкурирующей гипотезе, $H_1: M(X) \neq M(Y)$ надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n_1 + n_2 - 2$ найти критическую точку $t_{двуст.кр}(\alpha, k)$

Если $|T_{набл}| < t_{двуст.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 4.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: M(X) > M(Y)$, надо найти критическую точку $t_{\text{правост.кр}}(\alpha, k)$ по таблице по заданному уровню значимости α , помещенной в нижней строке таблицы, и числу степеней свободы $k = n_1 + n_2 - 2$

Если $T_{\text{набл}} < t_{\text{правост.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 4.3. При конкурирующей гипотезе $H_1: M(X) < M(Y)$, надо найти критическую точку $t_{\text{правост.кр}}(\alpha, k)$ по правилу 2 и полагают $t_{\text{левост.кр}} = -t_{\text{правост.кр}}(\alpha, k)$

Если $T_{\text{набл}} > -t_{\text{правост.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} < -t_{\text{правост.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

V. Сравнение выборочной средней с гипотетической средней нормальной совокупности

а) Дисперсия генеральной совокупности известна

По независимой выборке объема n найдена исправленная дисперсия s^2

Правило 5.1а. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: a = a_0$ о равенстве генеральной средней a нормальной совокупности с известной дисперсии σ^2 гипотетическому(предполагаемому) значению a_0 , при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$$

и по таблице функции Лапласа найти критическую точку $u_{\text{кр}}$ из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - \alpha}{2}$$

Если $|U_{\text{набл}}| < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|U_{\text{набл}}| > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 5.2а. При конкурирующей гипотезе $H_1: a > a_0$, надо найти критическую точку $u_{\text{кр}}$ по таблице функции Лапласа из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - 2\alpha}{2}$$

Если $U_{\text{набл}} < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 5.3а. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: a < a_0$, надо найти критическую точку $u_{\text{кр}}$ по правилу 2

Если $U_{\text{набл}} > -u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} < -u_{\text{кр}}$ - нулевую гипотезу отвергают.

б) Дисперсия генеральной совокупности неизвестна

Если дисперсия генеральной совокупности неизвестна, то в качестве критерия проверки нулевой гипотезы принимают случайную величину

$$T = \frac{(\bar{X} - a_0)\sqrt{n}}{S}, \quad \text{где} \quad S = \frac{\sum n_i x_i^2 - \frac{1}{n}(\sum n_i x_i)^2}{n-1} - \text{исправленное среднее}$$

квадратическое отклонение. Величина T имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Правило 5.1б. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: a = a_0$ о равенстве генеральной средней а нормальной совокупности с неизвестной дисперсии гипотетическому (предполагаемому) значению a_0 , при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{(\bar{x} - a_0)\sqrt{n}}{S}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n - 1$ найти критическую точку $t_{\text{двуст.кр}}(\alpha, k)$

Если $|T_{\text{набл}}| < t_{\text{двуст.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{\text{набл}}| > t_{\text{двуст.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 5.26. При конкурирующей гипотезе $H_1: a > a_0$, надо найти критическую точку $t_{\text{правост.кр}}(\alpha, k)$ по таблице по заданному уровню значимости α , помещенной в нижней строке таблицы, и числу степеней свободы $k = n - 1$

Если $T_{\text{набл}} < t_{\text{правост.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 5.36. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: a < a_0$, надо найти критическую точку $u_{\text{кр}}$ по правилу 2 и полагают $t_{\text{левост.кр}} = -t_{\text{правост.кр}}(\alpha, k)$

Если $T_{\text{набл}} > -t_{\text{правост.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} < -t_{\text{правост.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

VI. Сравнение двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки)

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. Из этих совокупностей извлечены зависимые выборки одинакового объема n , варианты которых соответственно равны x_i, y_i . Введены обозначения:

$d_i = x_i - y_i$ - разности вариант с одинаковыми номерами,

$\bar{d} = \frac{\sum d_i}{n}$ - средняя разностей вариант с одинаковыми номерами,

$s_d = \sqrt{\frac{1}{n-1} \left(\sum d_i^2 - \frac{1}{n} [\sum d_i]^2 \right)}$ - «исправленное» среднее

квадратическое отклонение.

Правило 6.1. Для того чтобы при заданном уровне значимости α

проверить нулевую гипотезу $H_0: M(\bar{X}) = M(\bar{Y})$ о равенстве двух средних нормальных совокупностей X и Y с неизвестными дисперсиями, при конкурирующей гипотезе $H_1: M(\bar{X}) \neq M(\bar{Y})$, надо вычислить наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{\bar{d}\sqrt{n}}{s_d}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n - 1$ найти критическую точку $t_{\text{двуст.кр}}(\alpha, k)$

Если $|T_{\text{набл}}| < t_{\text{двуст.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{\text{набл}}| > t_{\text{двуст.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

VII. Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события

По достаточно большому числу n независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, найдена относительная частота $\frac{m}{n}$.

Требуется, при заданном уровне значимости α , проверить нулевую гипотезу, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0

Правило 7.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: p = p_0$ о равенстве неизвестной вероятности p гипотетической вероятности p_0 , надо вычислить наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}$$

и по таблице функции Лапласа найти критическую точку $u_{\text{кр}}$ из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - \alpha}{2}$$

Если $|U_{\text{набл}}| < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|U_{\text{набл}}| > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 7.2. При конкурирующей гипотезе $H_1: p > p_0$, надо найти критическую точку $u_{\text{кр}}$ по таблице функции Лапласа из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - 2\alpha}{2}$$

Если $U_{\text{набл}} < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 7.3. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: p < p_0$, надо найти критическую точку $u_{\text{кр}}$ по правилу 2

Если $U_{\text{набл}} > -u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} < -u_{\text{кр}}$ - нулевую гипотезу отвергают.

Замечание. Удовлетворительные результаты обеспечивает выполнение неравенства

$$np_0q_0 > 9.$$

Пример 1. Двумя методами проведены измерения одной и той же физической величины. Получены следующие физические величины

x_i	9,6	10	9,8	10,2	10,6
y_i	10,4	9,7	10	10,3	

Может ли считать, что оба метода обеспечивают одинаковую точность измерений, если принять уровень значимости $\alpha = 0,1$. Предполагается, что результаты измерений распределены нормально и выборки независимы.

Решение. Так как рассматриваются независимые выборки, объемы которых $n_1 = 5$ и $n_2 = 4$, извлеченным из нормальных генеральных совокупностей, то необходимо воспользоваться сравнением двух дисперсий нормальных генеральных совокупностей – правило I.

Будем судить о точности методов по исправленным выборочным дисперсиям.

Выдвинем нулевую гипотезу $H_0: D(X) = D(Y)$ в качестве конкурирующей гипотезы примем $H_1: D(X) \neq D(Y)$.

Найдем выборочные дисперсии

$$s_x^2 = \frac{1}{n-1} \left(\sum n_i x_i^2 - \frac{1}{n} \left[\sum n_i x_i \right]^2 \right).$$

Найдем исправленные выборочные дисперсии

$$\begin{aligned} s_x^2 &= \frac{1}{5-1} \left(9,6^2 + 100 + 9,8^2 + 10,2^2 + 10,6^2 - \frac{1}{5} [9,6 + 10 + 9,8 + 10,2 + 10,6]^2 \right) = \\ &= \frac{1}{4} (504,6 - 0,2 \cdot 50,2) = 123,64 \\ s_y^2 &= \frac{1}{4-1} \left(9,7^2 + 100 + 10,4^2 + 10,3^2 - \frac{1}{4} [9,7 + 10 + 10,4 + 10,3]^2 \right) = \\ &= \frac{1}{3} (408,34 - 0,25 \cdot 40,4) = 132,75 \end{aligned}$$

Сравним дисперсии. Найдем отношение большей исправленной дисперсии к меньшей

$$F_{набл} = \frac{s_y^2}{s_x^2} = \frac{132,75}{123,64} = 1,48.$$

Так как конкурирующая гипотеза $D(X) \neq D(Y)$, поэтому критическая область двусторонняя, в соответствии с правилом 1.2 следует брать уровень значимости вдвое меньше заданного.

По таблице распределения критических точек Фишера-Снедекора по уровню значимости $\frac{\alpha}{2} = 0,05$ и числам степеней свободы $k_1 = n_1 - 1 = 4$, $k_2 = n_2 - 1 = 3$ находим критическую точку

$$F_{кр}(\alpha, k_1, k_2) = F_{кр}(0,05; 4, 3) = 9,12.$$

Так как $F_{набл} < F_{кр}$ - нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий.

Вывод: исправленные дисперсии различаются незначительно, и, следовательно, оба метода обеспечивают одинаковую точность измерений.

Пример 2. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Исправленная

выборочная дисперсия, найденная по выборке объема 121, оказалась равной $s_x^2 = 0,3$. Можно ли принять партию при уровне значимости $\alpha = 0,01$.

Р е ш е н и е. По условию независимой выборке объема n найдена исправленная дисперсия s^2 , поэтому воспользуемся сравнением исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности - правила 2.

Выдвинем нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2 = 0,2$ в качестве конкурирующей гипотезы примем $H_1: \sigma^2 > 0,2$.

Найдем наблюдаемое значение критерия

$$\chi_{набл}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{120 \cdot 0,3}{0,2} = 180.$$

Конкурирующая гипотеза $\sigma^2 > 0,2$, следовательно, критическая область правосторонняя. Так как число степеней свободы $k = 120$, $k > 30$, то критическую точку $\chi_{кр}^2(\alpha, k)$ можно найти, используя равенство Уилсона-Гильферти

$$\chi_{кр}^2(\alpha, k) = k \left(1 - \frac{2}{9k} + z_\alpha \sqrt{\frac{2}{9k}} \right)^3,$$

где z_α находим, используя функцию Лапласа

$$\Phi(z_\alpha) = \Phi(z_{0,01}) = \frac{1}{2} - 0,01 = 0,49.$$

Используя функцию Лапласа, находим $z_{0,01} = 2,326$.

$$\chi_{кр}^2(0,01; 120) = 120 \left(1 - \frac{2}{9 \cdot 120} + 2,326 \sqrt{\frac{2}{9 \cdot 120}} \right)^3 = 158,85.$$

По правилу 2.1, так как $\chi_{набл}^2 > \chi_{кр}^2$ - нулевую гипотезу отвергают.

Вывод: партию принять нельзя.

Пример 3. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,02. Среди случайно отобранных 4890 изделий оказалось 12 дефектных. Можно ли принять партию.

Р е ш е н и е. В условии достаточно большому числу $n=4890$ независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, можно определить относительную частоту брака $\frac{m}{n} = \frac{12}{4890}$, поэтому следует использовать сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события - правила 7.

При заданном уровне значимости $\alpha=0,05$, сформулируем нулевую гипотезу $H_0: p = p_0 = 0,02$, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0 , при конкурирующей гипотезе $H_1: p > 0,02$.

Надо вычислить наблюдаемое значение критерия

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}},$$

где $n=4890$, $\frac{m}{n} = \frac{12}{4890}$

и по таблице функции Лапласа найти критическую точку $u_{кр}$ из равенства

$$\Phi(u_{кр}) = \frac{1-\alpha}{2}$$

ЗАДАНИЯ

1. По выборке объема $n=50$ найден средний размер диаметров валиков $\bar{x}=20,1\text{мм}$, изготовленных на первом станке; по выборке объема $m=50$ найден средний размер диаметров валиков $\bar{y}=19,8\text{мм}$, изготовленных на втором станке. Генеральные дисперсии известны: $D(X)=1,75\text{мм}^2$, $D(Y)=1,375\text{мм}^2$. Требуется, при уровне значимости $\alpha=0,05$, проверить нулевую гипотезу $H_0: M(X)=M(Y)$ в качестве конкурирующей гипотезы $H_1: M(X) \neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально и выборки независимы.

2. Точность работы станка проверяется по и контролируемого размера изделий, которая не должна превышать $\sigma_0^2 = 0,1$. Взята проба из случайно отобранных изделий, причем получены следующие результаты

x_i	3	3,5	3,8	4,4	4,5
n_i	2	6	9	7	1

Требуется, при уровне значимости $\alpha = 0,05$, проверить, обеспечивает ли станок требуемую точность

3. В результате длительного хронометража времени сборки узла различными сборщиками установлено, что дисперсия этого времени $\sigma_0^2 = 2 \text{ мин}^2$. Результаты 20 наблюдений за работой новичка таковы:

x_i	56	58	60	62	64
n_i	1	4	10	3	2

Можно ли, при уровне значимости $\alpha = 0,05$, считать, что новичок работает ритмично (в том смысле, что дисперсия затрачиваемого им времени существенно не отличается от дисперсии времени остальных сборщиков).

4. При уровне значимости $\alpha = 0,05$, требуется проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве генеральных средних нормальных совокупностей X и Y при конкурирующей гипотезе $H_1: M(X) > M(Y)$ по малым независимым выборкам $n = 10$, $m = 16$. Получены следующие результаты

x_i	12,3	12,5	12,8	13	13,5
n_i	1	2	4	2	1

y_i	12,2	12,3	13
m_i	6	8	2

5. На двух аналитических весах, в одном и том же порядке, взвешены 10 проб химического вещества и получены следующие результаты (в мг)

x_i	25	30	28	50	20	40	32	36	42	38
y_i	28	31	26	52	24	36	33	35	45	40

При уровне значимости $\alpha = 0,01$, проверить значимо или незначимо различаются результаты взвешиваний, в предположении, что они распределены нормально.

6 Установлено, что средний вес таблетки лекарства сильного токсического действия должен быть равен $a_0 = 0,5 \text{ мг}$. Выборочная проверка 121 таблетки полученной партии лекарства показала, что средний вес таблетки этой партии $\bar{x} = 0,53 \text{ мг}$. Требуется, при уровне значимости $\alpha = 0,01$, проверить нулевую гипотезу $H_0: a = a_0 = 0,5$ в качестве конкурирующей гипотезы $H_1: a \neq 0,5$. Многократными предварительными опытами по взвешиванию таблеток, поставляемых заводом, было найдено, что вес таблеток лекарства распределен нормально со средним квадратическим отклонением $\sigma = 0,11 \text{ мг}$.

7. По выборке объема $n = 30$ найден средний вес $\bar{x} = 130 \text{ г}$ изделий, изготовленных на первом станке; по выборке объема $m = 40$ найден средний вес $\bar{y} = 125 \text{ г}$ изделий, изготовленных на втором станке. Генеральные дисперсии известны: $D(X) = 60 \text{ г}^2$, $D(Y) = 80 \text{ г}^2$. Требуется, при уровне значимости $\alpha = 0,05$, проверить нулевую гипотезу $H_0: M(X) = M(Y)$ в качестве конкурирующей гипотезы $H_1: M(X) \neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально и выборки независимы.

8. Для сравнения точности двух станков взяты две пробы, объемы которых $n_1 = 10$ и $n_2 = 8$. В результате измерения контролируемого размера отобранных изделий получены результаты

x_i	1,08	1,1	1,12	1,14	1,15	1,25	1,36	1,38	1,4	1,42
y_i	1,1	1,12	1,18	1,22	1,33	1,35	1,36	1,38		

Можно ли считать, что станки обладают одинаковой точностью $H_0: D(X) = D(Y)$, если принять уровень значимости $\alpha = 0,1$ и в качестве конкурирующей гипотезы $H_1: D(X) \neq D(Y)$.

9 Проектный контролируемый размер изделий, изготовленных на станке $a = a_0 = 35 \text{ мм}$. Измерения 20 случайно отобранных изделий дали следующие результаты:

x_i	34,8	34,9	35	35,1	35,3
n_i	2	3	4	6	5

Требуется, при уровне значимости $\alpha = 0,05$, проверить нулевую гипотезу $H_0: a = a_0 = 35$ в качестве конкурирующей гипотезы $H_1: a \neq 35$.

10. Физическая подготовка 9 спортсменов была проведена при поступлении в спортивную школу, а затем после недели тренировок. Итог проверки в баллах оказался следующим (первая строка – число баллов, полученных при поступлении, вторая – после обучения)

x_i	76	71	57	49	70	69	26	65	59
y_i	81	85	52	52	70	63	33	83	62

Требуется, при уровне значимости $\alpha = 0,05$, проверить значимо или незначимо улучшилась физическая подготовка спортсменов, в предположении, что число баллов распределены нормально.

11. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,03. Среди случайно отобранных 400 изделий оказалось 18 бракованных. Можно ли принять партию.

12. По малым независимым выборкам $n=10$, $m=8$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}=142,3$; $\bar{y}=145,3$ и исправленные дисперсии $s_x^2=2,7$; $s_y^2=3,2$. При уровне значимости $\alpha=0,01$, требуется проверить нулевую гипотезу $H_0: M(X)=M(Y)$ при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$.

13. Химическая лаборатория провела в одно и том же порядке 8 проб двумя методами. В результате получены следующие результаты (первая строка – содержание вещества в процентах в каждой пробе, полученной первым методом, вторая – вторым методом)

x_i	15	20	16	22	24	14	18	20
y_i	15	22	14	25	29	16	20	24

Требуется, при уровне значимости $\alpha = 0,05$, проверить значимо или незначимо различаются средние результаты анализов, в предположении, что число баллов распределены нормально.

14. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Исправленная выборочная дисперсия, найденная по выборке объема $n = 121$, оказалась равной $s_x^2 = 0,3$. Можно ли принять партию при уровне значимости $\alpha = 0,05$.

15. Завод рассылает рекламные каталоги возможным заказчикам. Как показал опыт, вероятность того, что организация, получившая каталог, закажет рекламу, равна 0,08. Завод разослал 100 каталогов новой улучшенной формы и получил 100 заказов. Можно ли считать, что новая форма рекламы оказалась значимо эффективней первой.

16. По двум малым независимым выборкам, объемы которых $n_1 = 9$, $n_2 = 6$, извлеченным из нормальных генеральных совокупностей, найдены выборочные дисперсии $D_B(X) = 14,4$; $D_B(Y) = 20,5$. При уровне значимости $\alpha = 0,1$, требуется проверить нулевую гипотезу $H_0: D(X) = D(Y)$ при конкурирующей гипотезе $H_1: D(X) \neq D(Y)$.

17. В результате длительного наблюдения установлено, что вероятность полного выздоровления больного, принимавшего лекарство А, равно 0,8. Новое лекарство В назначено 800 больным, причем 660 из них полностью выздоровели. Можно ли считать, что новое лекарство значимо эффективнее лекарства А на пятипроцентном уровне значимости.

Лабораторная работа №8 Ошибка принятия гипотез и ее исправление

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений.

Нулевой(основной) называют выдвинутую гипотезу H_0 .

Конкурирующей (альтернативной) называют гипотезу H_1 , которая противоположна нулевой.

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

Простой называют гипотезу, содержащую только одно предположение.

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

Выдвинутая гипотеза может быть правильной или неправильной; поэтому возникает необходимость ее проверки.

Поскольку проверку производят статистическими методами, ее называют статистической. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза. Вероятность ошибки первого рода называют уровнем значимости и обозначают через α .

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза. Вероятность ошибки второго рода обозначают через β .

Последствия этих ошибок могут оказаться различными.

Статистическим критерием (или просто *критерием*) называют случайную величину K , которая служит для проверки нулевой гипотезы.

Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин, и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением $K_{\text{набл.}}$ называют значение критерия, вычисленное по выборкам.

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

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

Рассмотрим *основной принцип проверки статистических значений*: если наблюдаемое значение критерия принадлежит критической области, то гипотезу отвергают; если наблюдаемое значение критерия принадлежит области принятия гипотезы, то гипотезу принимают.

Поскольку критерий K – случайная одномерная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы также являются интервалами и, следовательно, существуют точки, которые их разделяют.

Критическими точками (границами) $k_{\text{кр.}}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую

неравенством $K > K_{кр.}$, где $k_{кр.}$ – положительное число (рис.1).

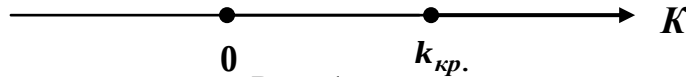


Рис. 1

Левосторонней называют критическую область, определяемую неравенством $K < K_{кр.}$, где $k_{кр.}$ – отрицательное число (рис. 2).

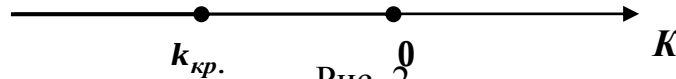


Рис. 2

Односторонней называют правостороннюю или левостороннюю критическую область.

Двусторонней называют критическую область, определяемую неравенствами

$$K < k_1, K > k_2, \text{ где } k_2 > k_1.$$

В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами (в предположении, что $k_{кр.} > 0$) (рис.3):

$$K < -k_{кр.}, K > k_{кр.} \Leftrightarrow |K| > k_{кр.}$$

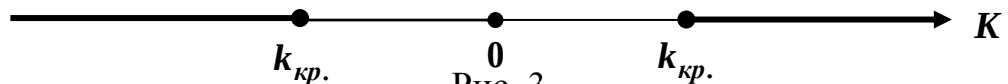


Рис. 3

Для отыскания критической области задаются достаточно малой вероятностью – уровнем значимости α . Затем ищут критические точки $k_{кр.}$, исходя из требования, чтобы при условии справедливости нулевой гипотезы, вероятность того, что критерий K примет значение, большее $k_{кр.}$, была равна принятому уровню значимости:

А) для *правосторонней* критической области

$$P(K > k_{кр.}) = \alpha, k_{кр.} > 0;$$

Б) для *левосторонней* критической области

$$P(K < k_{кр.}) = \alpha, k_{кр.} < 0;$$

В) для *двусторонней* симметричной критической области

$$P(K > k_{кр.}) = \frac{\alpha}{2}, k_{кр.} > 0;$$

$$P(K < -k_{кр.}) = \frac{\alpha}{2}.$$

VIII. Сравнение двух дисперсий нормальных генеральных совокупностей

По независимым выборкам, объемы которых n_1 и n_2 , извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии s_x^2 и s_y^2 . Требуется сравнить эти дисперсии.

Правило 1.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(X) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H_1: D(X) > D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей исправленной выборочной дисперсии к меньшей)

$$F_{\text{набл}} = \frac{s_{\text{max}}^2}{s_{\text{min}}^2}$$

и по таблице критических точек распределения Фишера – Снедекора, по заданному уровню значимости α и числом степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ (k_1 - число степеней свободы большей исправленной выборочной дисперсии), найти критическую точку $F_{\text{кр}}(\alpha, k_1, k_2)$.

Если $F_{\text{набл}} < F_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 1.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(X) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H_1: D(X) \neq D(Y)$, надо вычислить наблюдаемое значение критерия по уровню значимости $\frac{\alpha}{2}$ (вдвое меньшему заданного) и числом степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ (k_1 - число степеней свободы большей исправленной выборочной дисперсии), найти критическую точку $F_{\text{кр}}\left(\frac{\alpha}{2}, k_1, k_2\right)$.

Если $F_{\text{набл}} < F_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ - нулевую гипотезу отвергают.

IX. Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

По независимой выборке объема n найдена исправленная дисперсия s^2

Правило 2.1. Для того чтобы при заданном уровне значимости α

проверить нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2$ о равенстве неизвестной генеральной дисперсии σ^2 гипотетическому (предполагаемому) значению σ_0^2 , при конкурирующей гипотезе $H_1: \sigma^2 > \sigma_0^2$, надо вычислить наблюдаемое значение критерия

$$\chi_{\text{набл}}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

и по таблице критических точек распределения χ^2 , по заданному уровню значимости α и числом степеней свободы $k = n-1$, найти критическую точку $\chi_{\text{кр}}^2(\alpha, k)$.

Если $\chi_{\text{набл}}^2 < \chi_{\text{кр}}^2$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{набл}}^2 > \chi_{\text{кр}}^2$ - нулевую гипотезу отвергают.

Правило 2.2. При конкурирующей гипотезе $H_1: \sigma^2 \neq \sigma_0^2$, надо вычислить левую $\chi_{\text{лев.кр}}^2\left(1-\frac{\alpha}{2}, k\right)$ и правую $\chi_{\text{пр.кр}}^2\left(\frac{\alpha}{2}, k\right)$ критические точки критерия по уровню значимости $\frac{\alpha}{2}$ (вдвое меньшему заданного).

Если $\chi_{\text{лев.кр}}^2 < \chi_{\text{набл}}^2 < \chi_{\text{пр.кр}}^2$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{лев.кр}}^2 > \chi_{\text{набл}}^2$ или $\chi_{\text{набл}}^2 > \chi_{\text{пр.кр}}^2$ - нулевую гипотезу отвергают.

Правило 2.3. При конкурирующей гипотезе $H_1: \sigma^2 < \sigma_0^2$, надо вычислить $\chi_{\text{кр}}^2(1-\alpha, k)$ критическую точку критерия.

Если $\chi_{\text{набл}}^2 > \chi_{\text{кр}}^2(1-\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi_{\text{набл}}^2 < \chi_{\text{кр}}^2(1-\alpha, k)$ - нулевую гипотезу отвергают.

Замечание. Если число степеней свободы $k > 30$, то критическую точку $\chi_{\text{кр}}^2(\alpha, k)$ можно найти из равенства Уилсона-Гильферти

$$\chi_{\text{кр}}^2(\alpha, k) = k \left(1 - \frac{2}{9k} + z_{\alpha} \sqrt{\frac{2}{9k}} \right)^3,$$

где z_α находят, используя функцию Лапласа

$$\Phi(z_\alpha) = \frac{1}{2} - \alpha.$$

Х. Сравнение двух средних генеральных совокупностей, дисперсии которых известны (большие независимые выборки)

По независимым выборкам, объемы которых $n_1 > 30$ и $n_2 > 30$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}$, $\bar{y}$, генеральные дисперсии известны.

Правило 3.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X)=M(Y)$ о равенстве математических ожиданий (генеральных средних) двух нормальных совокупностей при конкурирующей гипотезе, $H_1: M(X) \neq M(Y)$ надо вычислить наблюдаемое значение критерия

$$Z_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}}$$

и по таблице функции Лапласа найти критическую точку $z_{\text{кр}}$ из равенства

$$\Phi(z_{\text{кр}}) = \frac{1 - \alpha}{2}$$

Если $|Z_{\text{набл}}| < z_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|Z_{\text{набл}}| > z_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 3.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 при конкурирующей гипотезе $H_1: M(X) > M(Y)$, надо найти критическую точку $z_{\text{кр}}$ по таблице функции Лапласа из равенства

$$\Phi(z_{\text{кр}}) = \frac{1}{2} - \alpha$$

Если $Z_{\text{набл}} < z_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $Z_{\text{набл}} > z_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 3.3. Для того чтобы при заданном уровне значимости α

проверить нулевую гипотезу H_0 при конкурирующей гипотезе $H_1: M(X) < M(Y)$, надо найти критическую точку $z_{кр}$ по правилу 2

Если $Z_{набл} > -z_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $Z_{набл} < -z_{кр}$ - нулевую гипотезу отвергают.

XI. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки)

По независимым выборкам, объемы которых $n_1 < 30$ и $n_2 < 30$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}$, $\bar{y}$, исправленные выборочные дисперсии s_x^2 , s_y^2 неизвестны, но равны.

Правило 4.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве математических ожиданий (генеральных средних) двух нормальных совокупностей с неизвестными, но равными дисперсиями при конкурирующей гипотезе, $H_1: M(X) \neq M(Y)$ надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n_1 + n_2 - 2$ найти критическую точку $t_{двуст.кр}(\alpha, k)$

Если $|T_{набл}| < t_{двуст.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 4.2. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: M(X) > M(Y)$, надо найти критическую точку $t_{правост.кр}(\alpha, k)$ по таблице по заданному уровню значимости α , помещенной в нижней строке таблицы, и числу степеней свободы $k = n_1 + n_2 - 2$

Если $T_{набл} < t_{правост.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} > t_{правост.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 4.3. При конкурирующей гипотезе $H_1: M(X) < M(Y)$, надо найти критическую точку $t_{\text{правост.кр}}(\alpha, k)$ по правилу 2 и полагают $t_{\text{левост.кр}} = -t_{\text{правост.кр}}(\alpha, k)$

Если $T_{\text{набл}} > -t_{\text{правост.кр}}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} < -t_{\text{правост.кр}}(\alpha, k)$ - нулевую гипотезу отвергают.

XI. Сравнение выборочной средней с гипотетической средней нормальной совокупности

б) Дисперсия генеральной совокупности известна

По независимой выборке объема n найдена исправленная дисперсия s^2

Правило 5.1а. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: a = a_0$ о равенстве генеральной средней a нормальной совокупности с известной дисперсии σ^2 гипотетическому (предполагаемому) значению a_0 , при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$$

и по таблице функции Лапласа найти критическую точку $u_{\text{кр}}$ из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - \alpha}{2}$$

Если $|U_{\text{набл}}| < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|U_{\text{набл}}| > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 5.2а. При конкурирующей гипотезе $H_1: a > a_0$, надо найти критическую точку $u_{\text{кр}}$ по таблице функции Лапласа из равенства

$$\Phi(u_{\text{кр}}) = \frac{1 - 2\alpha}{2}$$

Если $U_{\text{набл}} < u_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}} > u_{\text{кр}}$ - нулевую гипотезу отвергают.

Правило 5.3а. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: a < a_0$, надо найти критическую точку $u_{кр}$ по правилу 2

Если $U_{набл} > -u_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} < -u_{кр}$ - нулевую гипотезу отвергают.

б) Дисперсия генеральной совокупности неизвестна

Если дисперсия генеральной совокупности неизвестна, то в качестве критерия проверки нулевой гипотезы принимают случайную величину

$$T = \frac{(\bar{X} - a_0)\sqrt{n}}{S}, \quad \text{где} \quad S = \frac{\sum n_i x_i^2 - \frac{1}{n}(\sum n_i x_i)^2}{n-1} - \text{исправленное среднее}$$

квадратическое отклонение. Величина T имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Правило 5.1б. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: a = a_0$ о равенстве генеральной средней a нормальной совокупности с неизвестной дисперсией гипотетическому (предполагаемому) значению a_0 , при конкурирующей гипотезе $H_1: a \neq a_0$, надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{S}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n - 1$ найти критическую точку $t_{двуст.кр}(\alpha, k)$

Если $|T_{набл}| < t_{двуст.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 5.2б. При конкурирующей гипотезе $H_1: a > a_0$, надо найти критическую точку $t_{правост.кр}(\alpha, k)$ по таблице по заданному уровню значимости α , помещенной в нижней строке таблицы, и числу степеней свободы $k = n - 1$

Если $T_{набл} < t_{правост.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} > t_{правост.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

Правило 5.36. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу при конкурирующей гипотезе $H_1: a < a_0$, надо найти критическую точку $u_{кр}$ по правилу 2 и полагают $t_{левост.кр} = -t_{правост.кр}(\alpha, k)$

Если $T_{набл} > -t_{правост.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $T_{набл} < -t_{правост.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

ХII. Сравнение двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки)

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. Из этих совокупностей извлечены зависимые выборки одинакового объема n , варианты которых соответственно равны x_i, y_i . Введены обозначения:

$d_i = x_i - y_i$ - разности вариант с одинаковыми номерами,

$\bar{d} = \frac{\sum d_i}{n}$ - средняя разностей вариант с одинаковыми номерами,

$s_d = \sqrt{\frac{1}{n-1} \left(\sum d_i^2 - \frac{1}{n} [\sum d_i]^2 \right)}$ - «исправленное» среднее

квадратическое отклонение.

Правило 6.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: M(\bar{X}) = M(\bar{Y})$ о равенстве двух средних нормальных совокупностей X и Y с неизвестными дисперсиями, при конкурирующей гипотезе $H_1: M(\bar{X}) \neq M(\bar{Y})$, надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{\bar{d} \sqrt{n}}{s_d}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α , помещенной в верхней строке таблицы, и числу степеней свободы $k = n - 1$ найти критическую точку $t_{двуст.кр}(\alpha, k)$

Если $|T_{набл}| < t_{двуст.кр}(\alpha, k)$ - нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{двуст.кр}(\alpha, k)$ - нулевую гипотезу отвергают.

ХІІІ. Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события

По достаточно большому числу n независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, найдена относительная частота $\frac{m}{n}$.

Требуется, при заданном уровне значимости α , проверить нулевую гипотезу, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0

Правило 7.1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: p = p_0$ о равенстве неизвестной вероятности p гипотетической вероятности p_0 , надо вычислить наблюдаемое значение критерия

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}$$

и по таблице функции Лапласа найти критическую точку $u_{кр}$ из равенства

$$\Phi(u_{кр}) = \frac{1 - \alpha}{2}$$

Если $|U_{набл}| < u_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $|U_{набл}| > u_{кр}$ - нулевую гипотезу отвергают.

Правило 7.2. При конкурирующей гипотезе $H_1: p > p_0$, надо найти критическую точку $u_{кр}$ по таблице функции Лапласа из равенства

$$\Phi(u_{кр}) = \frac{1 - 2\alpha}{2}$$

Если $U_{набл} < u_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} > u_{кр}$ - нулевую гипотезу отвергают.

Правило 7.3. Для того чтобы при заданном уровне значимости α

проверить нулевую гипотезу при конкурирующей гипотезе $H_1: p < p_0$, надо найти критическую точку $u_{кр}$ по правилу 2

Если $U_{набл} > -u_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $U_{набл} < -u_{кр}$ - нулевую гипотезу отвергают.

Замечание. Удовлетворительные результаты обеспечивает выполнение неравенства

$$np_0q_0 > 9.$$

Пример 1. Двумя методами проведены измерения одной и той же физической величины. Получены следующие физические величины

x_i	9,6	10	9,8	10,2	10,6
y_i	10,4	9,7	10	10,3	

Может ли считать, что оба метода обеспечивают одинаковую точность измерений, если принять уровень значимости $\alpha = 0,1$. Предполагается, что результаты измерений распределены нормально и выборки независимы.

Р е ш е н и е. Так как рассматриваются независимые выборки, объемы которых $n_1 = 5$ и $n_2 = 4$, извлеченным из нормальных генеральных совокупностей, то необходимо воспользоваться сравнением двух дисперсий нормальных генеральных совокупностей – правило I.

Будем судить о точности методов по исправленным выборочным дисперсиям.

Выдвинем нулевую гипотезу $H_0: D(X) = D(Y)$ в качестве конкурирующей гипотезы примем $H_1: D(X) \neq D(Y)$.

Найдем выборочные дисперсии

$$s_x^2 = \frac{1}{n-1} \left(\sum n_i x_i^2 - \frac{1}{n} \left[\sum n_i x_i \right]^2 \right).$$

Найдем исправленные выборочные дисперсии

$$\begin{aligned}
s_X^2 &= \frac{1}{5-1} \left(9,6^2 + 100 + 9,8^2 + 10,2^2 + 10,6^2 - \frac{1}{5} [9,6 + 10 + 9,8 + 10,2 + 10,6]^2 \right) = \\
&= \frac{1}{4} (504,6 - 0,2 \cdot 50,2) = 123,64 \\
s_Y^2 &= \frac{1}{4-1} \left(9,7^2 + 100 + 10,4^2 + 10,3^2 - \frac{1}{4} [9,7 + 10 + 10,4 + 10,3]^2 \right) = \\
&= \frac{1}{3} (408,34 - 0,25 \cdot 40,4) = 132,75
\end{aligned}$$

Сравним дисперсии. Найдем отношение большей исправленной дисперсии к меньшей

$$F_{\text{набл}} = \frac{s_Y^2}{s_X^2} = \frac{132,75}{123,64} = 1,48.$$

Так как конкурирующая гипотеза $D(X) \neq D(Y)$, поэтому критическая область двусторонняя, в соответствии с правилом 1.2 следует брать уровень значимости вдвое меньше заданного.

По таблице распределения критических точек Фишера-Снедекора по уровню значимости $\frac{\alpha}{2} = 0,05$ и числам степеней свободы $k_1 = n_1 - 1 = 4$, $k_2 = n_2 - 1 = 3$ находим критическую точку

$$F_{кр}(\alpha, k_1, k_2) = F_{кр}(0,05; 4, 3) = 9,12.$$

Так как $F_{\text{набл}} < F_{кр}$ - нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий.

Вывод: исправленные дисперсии различаются незначительно, и, следовательно, оба метода обеспечивают одинаковую точность измерений.

Пример 2. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Исправленная выборочная дисперсия, найденная по выборке объема 121, оказалась равной $s_X^2 = 0,3$. Можно ли принять партию при уровне значимости $\alpha = 0,01$.

Решение. По условию независимой выборке объема n найдена исправленная дисперсия s^2 , поэтому воспользуемся сравнением исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности - правила 2.

Выдвинем нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2 = 0,2$ в качестве

конкурирующей гипотезы примем $H_1: \sigma^2 > 0,2$.

Найдем наблюдаемое значение критерия

$$\chi_{набл}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{120 \cdot 0,3}{0,2} = 180.$$

Конкурирующая гипотеза $\sigma^2 > 0,2$, следовательно, критическая область правосторонняя. Так как число степеней свободы $k=120$, $k > 30$, то критическую точку $\chi_{кр}^2(\alpha, k)$ можно найти, используя равенство Уилсона-Гильферти

$$\chi_{кр}^2(\alpha, k) = k \left(1 - \frac{2}{9k} + z_\alpha \sqrt{\frac{2}{9k}} \right)^3,$$

где z_α находим, используя функцию Лапласа

$$\Phi(z_\alpha) = \Phi(z_{0,01}) = \frac{1}{2} - 0,01 = 0,49.$$

Используя функцию Лапласа, находим $z_{0,01} = 2,326$.

$$\chi_{кр}^2(0,01;120) = 120 \left(1 - \frac{2}{9 \cdot 120} + 2,326 \sqrt{\frac{2}{9 \cdot 120}} \right)^3 = 158,85.$$

По правилу 2.1, так как $\chi_{набл}^2 > \chi_{кр}^2$ - нулевую гипотезу отвергают.

Вывод: партию принять нельзя.

Пример 3. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,02. Среди случайно отобранных 4890 изделий оказалось 12 дефектных. Можно ли принять партию.

Решение. В условии достаточно большому числу $n=4890$ независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, можно определить относительную частоту брака $\frac{m}{n} = \frac{12}{4890}$, поэтому следует использовать сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события - правила 7.

При заданном уровне значимости $\alpha=0,05$, сформулируем нулевую

гипотезу $H_0: p = p_0 = 0,02$, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0 , при конкурирующей гипотезе $H_1: p > 0,02$.

Надо вычислить наблюдаемое значение критерия

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}},$$

где $n=4890$, $\frac{m}{n} = \frac{12}{4890}$

и по таблице функции Лапласа найти критическую точку $u_{кр}$ из равенства

$$\Phi(u_{кр}) = \frac{1-\alpha}{2}$$

ЗАДАНИЯ

1. По выборке объема $n=50$ найден средний размер диаметров валиков $\bar{x} = 20,1 \text{ мм}$, изготовленных на первом станке; по выборке объема $m=50$ найден средний размер диаметров валиков $\bar{y} = 19,8 \text{ мм}$, изготовленных на втором станке. Генеральные дисперсии известны: $D(X) = 1,75 \text{ мм}^2$, $D(Y) = 1,375 \text{ мм}^2$. Требуется, при уровне значимости $\alpha = 0,05$, проверить нулевую гипотезу $H_0: M(X) = M(Y)$ в качестве конкурирующей гипотезы $H_1: M(X) \neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально и выборки независимы.

2. Точность работы станка проверяется по и контролируемого размера изделий, которая не должна превышать $\sigma_0^2 = 0,1$. Взята проба из случайно отобранных изделий, причем получены следующие результаты

x_i	3	3,5	3,8	4,4	4,5
n_i	2	6	9	7	1

Требуется, при уровне значимости $\alpha = 0,05$, проверить, обеспечивает ли станок требуемую точность

3. В результате длительного хронометража времени сборки узла различными сборщиками установлено, что дисперсия этого времени $\sigma_0^2 = 2 \text{ мин}^2$. Результаты 20 наблюдений за работой новичка таковы:

x_i	56	58	60	62	64
n_i	1	4	10	3	2

Можно ли, при уровне значимости $\alpha = 0,05$, считать, что новичок работает ритмично (в том смысле, что дисперсия затрачиваемого им времени существенно не отличается от дисперсии времени остальных сборщиков).

4. При уровне значимости $\alpha = 0,05$, требуется проверить нулевую гипотезу $H_0: M(X) = M(Y)$ о равенстве генеральных средних нормальных совокупностей X и Y при конкурирующей гипотезе $H_1: M(X) > M(Y)$ по малым независимым выборкам $n = 10$, $m = 16$. Получены следующие результаты

x_i	12,3	12,5	12,8	13	13,5
n_i	1	2	4	2	1

y_i	12,2	12,3	13
m_i	6	8	2

5. На двух аналитических весах, в одном и том же порядке, взвешены 10 проб химического вещества и получены следующие результаты (в мг)

x_i	25	30	28	50	20	40	32	36	42	38
y_i	28	31	26	52	24	36	33	35	45	40

При уровне значимости $\alpha = 0,01$, проверить значимо или незначимо различаются результаты взвешиваний, в предположении, что они распределены нормально.

6 Установлено, что средний вес таблетки лекарства сильного токсического действия должен быть равен $a_0 = 0,5 \text{ мг}$. Выборочная проверка 121 таблетки полученной партии лекарства показала, что средний вес таблетки этой партии $\bar{x} = 0,53 \text{ мг}$. Требуется, при уровне значимости $\alpha = 0,01$, проверить нулевую гипотезу $H_0: a = a_0 = 0,5$ в качестве конкурирующей гипотезы $H_1: a \neq 0,5$. Многократными предварительными

опытами по взвешиванию таблеток, поставляемых заводом, было найдено, что вес таблеток лекарства распределен нормально со средним квадратическим отклонением $\sigma = 0,11\text{мг}$.

7. По выборке объема $n=30$ найден средний вес $\bar{x}=130\text{г}$ изделий, изготовленных на первом станке; по выборке объема $m=40$ найден средний вес $\bar{y}=125\text{г}$ изделий, изготовленных на втором станке. Генеральные дисперсии известны: $D(X)=60\text{ г}^2$, $D(Y)=80\text{ г}^2$. Требуется, при уровне значимости $\alpha=0,05$, проверить нулевую гипотезу $H_0: M(X)=M(Y)$ в качестве конкурирующей гипотезы $H_1: M(X)\neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально и выборки независимы.

8. Для сравнения точности двух станков взяты две пробы, объемы которых $n_1=10$ и $n_2=8$. В результате измерения контролируемого размера отобранных изделий получены результаты

x_i	1,08	1,1	1,12	1,14	1,15	1,25	1,36	1,38	1,4	1,42
y_i	1,1	1,12	1,18	1,22	1,33	1,35	1,36	1,38		

Можно ли считать, что станки обладают одинаковой точностью $H_0: D(X)=D(Y)$, если принять уровень значимости $\alpha=0,1$ и в качестве конкурирующей гипотезы $H_1: D(X)\neq D(Y)$.

9 Проектный контролируемый размер изделий, изготовленных на станке $a=a_0=35\text{мм}$. Измерения 20 случайно отобранных изделий дали следующие результаты:

x_i	34,8	34,9	35	35,1	35,3
n_i	2	3	4	6	5

Требуется, при уровне значимости $\alpha=0,05$, проверить нулевую гипотезу $H_0: a=a_0=35$ в качестве конкурирующей гипотезы $H_1: a\neq 35$.

10. Физическая подготовка 9 спортсменов была проведена при поступлении в спортивную школу, а затем после недели тренировок. Итог

проверки в баллах оказался следующим (первая строка – число баллов, полученных при поступлении, вторая – после обучения)

x_i	76	71	57	49	70	69	26	65	59
y_i	81	85	52	52	70	63	33	83	62

Требуется, при уровне значимости $\alpha = 0,05$, проверить значимо или незначимо улучшилась физическая подготовка спортсменов, в предположении, что число баллов распределены нормально.

11. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,03. Среди случайно отобранных 400 изделий оказалось 18 бракованных. Можно ли принять партию.

12. По малым независимым выборкам $n=10$, $m=8$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}=142,3$; $\bar{y}=145,3$ и исправленные дисперсии $s_x^2=2,7$; $s_y^2=3,2$. При уровне значимости $\alpha=0,01$, требуется проверить нулевую гипотезу $H_0: M(X)=M(Y)$ при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$.

13. Химическая лаборатория провела в одно и том же порядке 8 проб двумя методами. В результате получены следующие результаты (первая строка – содержание вещества в процентах в каждой пробе, полученной первым методом, вторая – вторым методом)

x_i	15	20	16	22	24	14	18	20
y_i	15	22	14	25	29	16	20	24

Требуется, при уровне значимости $\alpha = 0,05$, проверить значимо или незначимо различаются средние результаты анализов, в предположении, что число баллов распределены нормально.

14. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Исправленная выборочная дисперсия, найденная по выборке объема $n=121$, оказалась равной $s_x^2=0,3$. Можно ли принять партию при уровне значимости $\alpha = 0,05$.

15. Завод рассылает рекламные каталоги возможным заказчикам. Как показал опыт, вероятность того, что организация, получившая каталог, закажет рекламу, равна 0,08. Завод разослал 100 каталогов новой улучшенной формы и получил 100 заказов. Можно ли считать, что новая форма рекламы оказалась значимо эффективней первой.

16. По двум малым независимым выборкам, объемы которых $n_1=9$, $n_2=6$, извлеченным из нормальных генеральных совокупностей, найдены выборочные дисперсии $D_B(X)=14,4$; $D_B(Y)=20,5$. При уровне значимости $\alpha=0,1$, требуется проверить нулевую гипотезу $H_0: D(X)=D(Y)$ при конкурирующей гипотезе $H_1: D(X) \neq D(Y)$.

17. В результате длительного наблюдения установлено, что вероятность полного выздоровления больного, принимавшего лекарство А, равно 0,8. Новое лекарство В назначено 800 больным, причем 660 из них полностью выздоровели. Можно ли считать, что новое лекарство значимо эффективнее лекарства А на пятипроцентном уровне значимости.

Лабораторная работа №9

Однофакторный дисперсионный анализ

Цель работы: овладеть методами однофакторного анализа для одинакового числа испытаний на всех уровнях.

Краткие теоретические сведения

Дисперсионный анализ – это статистический метод, позволяющий анализировать влияние различных факторов на результат эксперимента.

Под **фактором** понимают различные, независимые, качественные показатели, влияющие на изучаемые признаки. Факторы обозначаются прописными начальными буквами латинского алфавита: $A, B, C, \dots$.

Факторы, контролируемые и измеряемые в процессе исследования, называются **регулируемыми**.

Признаки, изменяющиеся под воздействием тех или иных факторов, называют **результативными**. Для их обозначения используются конечные буквы латинского алфавита X, Y, Z .

Основная идея дисперсионного анализа состоит в сравнении факторной дисперсии, определяемой влиянием регулируемого фактора и остаточной дисперсии, обусловленной действием неконтролируемых случайных причин.

С помощью F – критерия устанавливается влияние фактора на признак.

Условия применения дисперсионного анализа к выборочным данным:

- 1) выборочные данные должны быть взяты из совокупностей с нормальным законом распределения;
- 2) дисперсии всех совокупностей должны быть равны.

Однофакторный дисперсионный анализ решается по следующей схеме:

- 1) Вычислить общую среднюю:

$$\bar{x}_B = \sum_{j=1}^p \frac{\bar{x}_{2p j}}{p}.$$

- 2) Вычислить сумму квадратов отклонений групповых средних от общей средней:

$$S_{\text{общ}} = \sum_{j=1}^p \sum_{i=1}^q (x_{ij} - \bar{x}_B)^2$$

Имеет место формула: $S_{\text{общ}} = \sum_{j=1}^p \sum_{i=1}^q x_{ij}^2 - pq\bar{x}_B^2$.

- 3) Вычислить факторную сумму квадратов отклонений групповых средних от общей средней (рассеяние между группами):

$$S_{\text{факт}} = q \sum_{j=1}^p (x_{zpj} - \bar{x}_B)^2$$

Имеет место формула: $S_{\text{факт}} = q \sum_{j=1}^p \bar{x}_{ij}^2 - pq\bar{x}_B^2$

4) Вычислить остаточную сумму квадратов отклонений значений группы от своей групповой средней (рассеяние внутри групп):

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

5) Вычислить факторную, остаточную дисперсии:

$$S_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1}, \quad S_{\text{ост}}^2 = \frac{S_{\text{ост}}}{p(q-1)}$$

и величину:

$$F_{\text{набл}} = \frac{S_{\text{факт}}^2}{S_{\text{ост}}^2}$$

6) Находим $F_{\text{кр}}(\alpha, p-1, p(q-1))$ используя таблицу значений (приложение: значения F -критерия Фишера при уровне надежности 95% ($\alpha = 0,05$)).

Если $F_{\text{набл}} < F_{\text{кр}}$, то различие групповых средних **незначимое** (фактор оказывает незначительное воздействие на результаты испытаний и его не следует учитывать).

Если $F_{\text{набл}} > F_{\text{кр}}$, то различие групповых средних **значимое** (фактор оказывает существенное влияние на результаты испытаний и его следует учитывать).

Замечание: Если x_{ij} – десятичные дроби с m знаками после запятой, то целесообразно перейти к нормированным и центрированным вариантам:

$$y_{ij} = 10^m (x_{ij} - c)$$

При этом факторная и остаточная дисперсии увеличатся в 10^{2m} раз каждая.

Пример 1. В трех филиалах одного банка были организованы три уровня различных услуг для клиентов. После этого в течение шести месяцев измерялись объемы вкладов (тыс.руб). Методом дисперсионного анализа проверить нулевую гипотезу о существенном влиянии организации услуг на объем вкладов при уровне значимости 0,05. Данные измерений приведены в таблице:

Номер измерения	Уровни фактора		
	F_1	F_2	F_3
1	16	17	18
2	15	25	27
3	10	15	30
4	20	28	14
5	18	30	40

6	14	22	34
---	----	----	----

Решение.

1) Имеем $p = 3$, $q = 6$, значения вариант – целые числа. Находим групповые и общую средние:

$$\bar{x}_{gp1} = \frac{16 + 15 + 10 + 20 + 18 + 14}{6} = 15,5$$

$$\bar{x}_{gp2} = \frac{17 + 25 + 15 + 28 + 30 + 22}{6} = 22,8$$

$$\bar{x}_{gp3} = \frac{18 + 27 + 30 + 30 + 40 + 34}{6} = 27,2$$

$$\bar{x}_B = \frac{\bar{x}_{gp1} + \bar{x}_{gp2} + \bar{x}_{gp3}}{3} = 21,8$$

2) Для расчета $S_{общ}$ составим таблицу квадратов:

Номер измерения	Уровни фактора		
	F_1^2	F_2^2	F_3^2
1	256	289	324
2	225	625	729
3	100	225	900
4	400	784	496
5	324	900	1600
6	196	484	1156
Сумма	1501	3307	4905

Получаем

$$S_{общ} = \sum_{j=1}^p \sum_{i=1}^q x_{ij}^2 - pq\bar{x}_B^2 = 1501 + 3307 + 4905 - 3 \cdot 6 \cdot 21,8^2 = 1158,68$$

3) Вычислим

$$S_{факт} = q \sum_{j=1}^p \bar{x}_{ij}^2 - pq\bar{x}_B^2 = 6(15,5^2 + 22,8^2 + 27,2^2 - 3 \cdot 21,8^2) = 445,26$$

4) Находим $S_{ост} = S_{общ} - S_{факт} = 1158,65 - 445,26 = 713,42$.

5) Факторная, остаточная дисперсии равны:

$$S_{факт}^2 = \frac{S_{факт}}{p-1} = \frac{445,23}{2} = 222,63, \quad S_{ост}^2 = \frac{S_{ост}}{p(q-1)} = \frac{731,42}{15} = 47,56$$

$$\text{тогда } F_{набл} = \frac{S_{факт}^2}{S_{ост}^2} = \frac{222,63}{47,56} = 4,68.$$

6) По таблице значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ находим

$$F_{кр}(\alpha, p-1, p(q-1)) = F_{кр}(0,05; 2; 15) = 3,6$$

Так как $F_{набл} > F_{кр}$, то нулевая гипотеза о существенном влиянии организации услуг на объем вкладов принимается.

Пример 2. Произведено по 4 испытания на каждом из трех уровней фактора F . Методом дисперсионного анализа при уровне значимости 0,05 проверить гипотезу о равенстве групповых средних.

Номер измерения	Уровни фактора		
	F_1	F_2	F_3
1	0,38	0,2	0,21
2	0,36	0,24	0,22
3	0,35	0,26	0,31
4	0,31	0,3	0,34
Групповая средняя	0,35	0,25	0,27

Решение.

Поскольку данные варианты являются дробными числами, то для упрощения вычислений переходим к новым, целым и центрированным вариантам, полагая

$$y_{ij} = 10^m(x_{ij} - c) = 10^2(x_{ij} - 0,29) = 100x_{ij} - 29$$

(общее среднее арифметическое значений x_{ij} равно

$$c = \frac{0,35 + 0,25 + 0,27}{3} = 0,29$$

Новые варианты заносим в расчетную таблицу:

Номер измерения	F_1		F_2		F_3	
	U_{i1}	U_{i1}^2	U_{i2}	U_{i2}^2	U_{i3}	U_{i3}^2
1	9	81	-9	81	-8	64
2	7	49	-5	25	-7	49
3	6	36	-3	9	2	4
4	2	4	1	1	5	25
Суммы	24	170	-16	116	-8	142

1) Имеем $p=3, q=4$

Находим групповые и общие средние:

$$\bar{U}_{\text{гп1}} = \frac{24}{4} = 6, \quad \bar{U}_{\text{гп2}} = -\frac{16}{4} = -4, \quad \bar{U}_{\text{гп3}} = -\frac{8}{4} = -2$$

$$\bar{U}_B = \frac{6 - 4 - 2}{4} = 0$$

2) Находим

$$S_{\text{Уобщ}} = \sum_{j=1}^p \sum_{i=1}^q (U_{ij} - \bar{U}_B)^2 = 170^2 + 116^2 + 142^2 = 428$$

3) Находим

$$S_{\text{Уфакт}} = q \sum_{j=1}^p (U_{\text{гп}j} - \bar{U}_B)^2 = 4(6^2 + (-4)^2 + (-2)^2) = 224$$

4) Вычислим

$$S_{Уост} = S_{Уобщ} - S_{Уфакт} = 428 - 224 = 204$$

5) Факторная, остаточная дисперсии и $F_{набл}$ равны:

$$S_{факт}^2 = \frac{S_{Уфакт}}{p-1} = \frac{224}{2} = 112, \quad S_{ост}^2 = \frac{S_{Уост}}{p(q-1)} = \frac{204}{6} = 22,67$$

$$F_{набл} = \frac{S_{факт}^2}{S_{ост}^2} = \frac{112}{22,67} = 4,48$$

6) По таблице значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ находим

$$F_{кр}(\alpha, p-1, p(q-1)) = F_{кр}(0,05; 2; 9) = 4,26$$

Так как $F_{набл} > F_{кр}$, то гипотезу о равенстве групповых средних данных совокупностей отвергаем.

Задания для самостоятельной работы

Задание. В течение шести лет использовались три различных технологии по выращиванию сельскохозяйственной культуры. Данные по эксперименту (в ц/га) приведены в таблице. При уровне значимости $\alpha = 0,05$ методом дисперсионного анализа установить влияние различных технологий на урожайность культуры.

Вариант 1	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,28	0,32	0,36
2	0,34	0,36	0,38
3	0,3	0,29	0,31
4	0,36	0,35	0,34
5	0,34	0,32	0,35
6	0,32	0,3	0,33
Вариант 2	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,38	0,32	0,36
2	0,44	0,41	0,45
3	0,32	0,33	0,40
4	0,31	0,36	0,38
5	0,44	0,42	0,37
6	0,36	0,34	0,33
Вариант 3	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,18	0,24	0,36
2	0,28	0,36	0,12
3	0,12	0,28	0,22

4	0,14	0,40	0,45
5	0,32	0,16	0,4
6	0,25	0,3	0,28
Вариант 4	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,38	0,2	0,21
2	0,36	0,24	0,22
3	0,35	0,26	0,31
4	0,31	0,3	0,34
5	0,29	0,28	0,3
6	0,3	0,31	0,29
Вариант 5	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,1	0,17	0,14
2	0,15	0,15	0,18
3	0,14	0,25	0,3
4	0,18	0,22	0,27
5	0,2	0,3	0,34
6	0,16	0,28	0,4
Вариант 6	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,2	0,16	0,14
2	0,25	0,15	0,2
3	0,26	0,25	0,28
4	0,28	0,22	0,25
5	0,2	0,3	0,32
6	0,19	0,25	0,38
Вариант 7	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,26	0,32	0,29
2	0,31	0,31	0,26
3	0,27	0,31	0,3
4	0,29	0,28	0,31
5	0,27	0,33	0,3
6	0,26	0,3	0,28
Вариант 8	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,36	0,3	0,29
2	0,34	0,34	0,28
3	0,32	0,35	0,3
4	0,33	0,32	0,28
5	0,35	0,29	0,31
6	0,33	0,31	0,35

Вариант 9	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,36	0,34	0,3
2	0,38	0,35	0,32
3	0,34	0,32	0,36
4	0,38	0,37	0,29
5	0,35	0,29	0,31
6	0,36	0,31	0,34
Вариант 10	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,27	0,31	0,34
2	0,34	0,35	0,33
3	0,31	0,29	0,31
4	0,36	0,34	0,33
5	0,33	0,32	0,31
6	0,32	0,3	0,32
Вариант 11	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,32	0,34	0,36
2	0,41	0,4	0,44
3	0,3	0,29	0,38
4	0,32	0,31	0,36
5	0,31	0,32	0,32
6	0,32	0,33	0,33
Вариант 12	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,17	0,2	0,26
2	0,24	0,26	0,28
3	0,19	0,29	0,31
4	0,15	0,21	0,24
5	0,32	0,31	0,35
6	0,26	0,28	0,26
Вариант 13	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,36	26	0,24
2	0,38	0,28	0,22
3	0,3	0,29	0,31
4	0,35	0,3	0,34
5	0,31	0,28	0,3
6	0,29	0,31	0,28
Вариант 14	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,18	0,12	0,16

2	0,14	0,16	0,18
3	0,2	0,29	0,21
4	0,16	0,15	0,14
5	0,24	0,22	0,25
6	0,18	0,2	0,23
Вариант 15	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,24	0,22	0,26
2	0,31	0,3	0,32
3	0,22	0,29	0,31
4	0,26	0,25	0,24
5	0,34	0,32	0,31
6	0,31	0,28	0,3
Вариант 16	Технология (фактор)		
Год	F_1	F_2	F_3
1	0,18	0,16	0,2
2	0,24	0,26	0,28
3	0,3	0,29	0,31
4	0,31	0,3	0,29
5	0,24	0,22	0,25
6	0,32	0,3	0,33

ЛИТЕРАТУРА

1. Mathcad в инженерных расчетах / Б. Максфилд ; пер. с англ. Н. Ю. Устьян. - СПб. : "КОРОНА-ВЕК", 2010. – 365 с.
2. Excel 2010 на примерах/А. Н. Васильев. - СПб.: БХВ-Петербург, 2010. - 422 с.
3. Математическая статистика : учебник / А. А. Боровков. - 4-е изд., стер. - СПб. : Лань, 2010. - 703 с.
4. Прогнозирование в системе STATISTICA в среде Windows: Основы теории и интенсивная практика на компьютере [Текст] : учеб.пособие для студ. вузов / В. П. Боровиков, Г. И. Ивченко. - М. : Финансы и статистика, 2000. - 382 с.
5. Теория вероятностей и математическая статистика : учеб.пособие / А. А. Туганбаев, В. Г. Крупин. - СПб. : Лань, 2011. - 223 с.
6. Теория планирования эксперимента и анализ статистических данных : учеб.пособие / Н. И. Сидняев. - М. : Юрайт, 2011. - 399 с.
7. Прикладная математика в системе MATHCAD : учеб.пособие / В. А. Охорзин. - 3-изд., стер. - СПб. : Лань, 2009. - 348 с.
8. Регрессионный анализ данных в пакете Mathcad : учеб.пособие / Ю. Е. Воскобойников. - СПб. : Лань, 2011. - 223 с.
9. Excel : сб. примеров и задач / С. М. Лавренов. - М. : Финансы и статистика, 2001. - 336 с.
10. Кэмпбелл М. Excel, ответы. Пер с англ. – М.: БИНОМ, 1996.