ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

УПРАВЛЕНИЕ ЦИФРОВЫХ ОБРАЗОВАТЕЛЬНЫХ ТЕХНОЛОГИЙ

Кафедра «Прикладная математика»

Учебно-методическое пособие

по дисциплине

«Математика»

«Сравнительный и корреляционный анализ»

Авторы

Рябых Г.Ю. Фролова Н.В.

Ростов-на-Дону, 2025

Аннотация

Методические указания предназначены для студентов всех специальностей и форм обучения.

Авторы

Доцент
Рябых Г.Ю.

Старший преподаватель
Фролова Н.В.

Оглавление

1. АСИММЕТРИЯ И ЭКСЦЕСС ЭМПИРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Асимметрия и эксцесс эмпирического распределения определяется соответственно равенствами:

$$\alpha_s = \frac{m_3}{\sigma^3_B}, \qquad \mathbf{e_к} = \frac{\mathbf{m_4}}{\mathbf{\sigma^4_B}} - \mathbf{3}$$

здесь $\boldsymbol{\sigma_B}$ - выборочное среднее квадратическое отклонение; $\mathbf{m_3}$ и $\mathbf{m_4}$ - центральные эмпирические моменты третьего и четвертого порядков:

$$\mathbf{m_3} = \frac{\sum \mathbf{n_i}(\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{B})^3}{\mathbf{n}}, \quad \mathbf{m_4} = \frac{\sum \mathbf{n_i}(\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{B})^4}{\mathbf{n}}$$

Эти моменты в случае равноотстоящих вариант с шагом $\boldsymbol{h}$ (шаг равен разности между любыми двумя соседними вариантами) удобно вычислять по формулам:

$m_3 = [M_3^* - 3M_1^*M_2^* + 2(M_1^*)^3] \cdot h^3,$
$m_4 = [M_4^* - 4M_1^*M_2^* + 6(M_1^*)^2M_2^* - 3(M_1^*)^4] \cdot h^4,$

где $\boldsymbol{M_k^* = \frac{\sum n_i u_i^k}{n}}$ - условные моменты $\boldsymbol{k}$ **-го** порядка; $\mathbf{u_i} = \frac{\mathbf{x_i} - \mathbf{C}}{\mathbf{h}}$ - условные варианты. Здесь $\mathbf{x_i}$ - первоначальные варианты, **C** - ложный нуль, т. е. варианта, имеющая наибольшую частоту (либо любая варианта, расположенная примерно в середине вариационного ряда).

Итак, для отыскания асимметрии и эксцесса необходимо вычислить условные моменты, что можно сделать ***методом произведений или методом сумм***.

А. Метод произведения

Пример 1. Найти методом произведений асимметрию и эксцесс по заданному распределению выборки объема $\boldsymbol{n = 100}$:

варианта x_i :

12	14	16	18	20	22

частота n_i :

5	15	50	16	10	4

Решение. Воспользуемся методом произведений. Составим расчетную

таблицу.

Для заполнения столбца 6 удобно перемножить числа каждой строки столбцов 3 и 5.

Для заполнения столбца 7 удобно перемножать числа каждой строки столбцов 3 и 6.

Столбец 8 служит для контроля вычислений с помощью тождества

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

Приведем расчетную таблицу 3.

Таблица 3

1	2	3	4	5	6	7	8
x_i	n_i	u_i	n_iu_i	$n_iu_i^2$	$n_iu_i^3$	$n_iu_i^4$	$n_i(u_i+1)^4$
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		$\sum n_iu_i = 23$	$\sum n_iu_i^2 = 127$	$\sum n_iu_i^3 = 149$	$\sum n_iu_i^4 = 595$	$\sum n_i(u_i+1)^4 = 2145$

Контроль:

$$\sum n_i(u_i+1)^4 = 2145,$$

$$\sum n_iu_i^4 + 4\sum n_iu_i^3 + 6\sum n_iu_i^2 + 4\sum n_iu_i + n = 595 + 4\cdot 149 + 6\cdot 127 + 4\cdot 23 + 100 = 2145.$$

Совпадение контрольных сумм свидетельствует о правильности вычислений.

Найдем условные моменты третьего и четвертого порядков $M_1^* = 0.23$, $M_2^* = 1.27$:

$$M_3^* = \frac{\sum n_i u_i^3}{n} = \frac{149}{100} = 1{,}49; \quad M_4^* = \frac{\sum n_i u_i^4}{n} = \frac{595}{100} = 5{,}95.$$

Найдем центральные эмпирические моменты третьего и четвертого порядков:

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

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

Подставляя $h = 2$ и $M_1^* = 0.23$, $M_2^* = 1.27$, $M_3^* = 1.49$, $M_4^* = 5.95$, получим

$$m_3 = 5.124, m_4 = 79.582.$$

Найдем искомые асимметрию и эксцесс, учитывая, что $D_B = 4.87$ (см. задачу 486):

$$a_s = \frac{m_3}{\sigma^3_B} = \frac{5.124}{(\sqrt{4.87})^3} = 0.49;$$

$$e_k = \frac{m_4}{\sigma^4_B} - 3 = \frac{79.582}{(\sqrt{4.87})^4} - 3 = 0.36.$$

Пример 2. Найти методом произведений асимметрию и эксцесс по заданному распределению выборки объема $n = 100$:

A.

x_i	2.6	3.0	3.4	3.8	4.2
n_i	8	20	45	15	12

B.

x_i	1	6	11	16	21
n_i	5	25	40	20	10

Б. Метод сумм

Найти методом сумм асимметрию и эксцесс по заданному распределению выборки объема $n = 100$:

x_i	48	52	56	60	64	68	72	76	80	84

n_i	2	4	6	8	12	30	18	8	7	5

Решение. Воспользуемся методом сумм, для этого составим расчетную таблицу 4.

Для заполнения столбца 5 запишем нуль в клетке строки, содержащей ложный нуль (68); над этим нулем и под ним поставим еще по два нуля.

В клетках над нулями запишем накопленные частоты, для чего просуммируем частоты столбца 4 ***сверху вниз***; в итоге будем иметь следующие накопленные частоты: 2; 2+8 = 10; 2+ 8 + 20 = 30. Сложив накопленные частоты, получим число $\mathbf{b_3} = 2 + 10 + 30 = 42$, которое поместим в верхнюю клетку пятого столбца.

В клетках под нулями запишем накопленные частоты, для чего просуммируем частоты столбца 4 ***снизу вверх***; в итоге будем иметь следующие накопленные частоты: 5; 5+ 17 = 22. Сложив накопленные частоты, получим число $\mathbf{a_3} = 5 + 22 = 27$, которое поместим в нижнюю клетку пятого столбца.

Аналогично заполняют столбец 6, причем суммируем частоты столбца 5. Сложив накопленные частоты, расположенные над нулями, получим число $\mathbf{b_4} = 2 + 12 = 14$, которое запишем в верхнюю клетку шестого столбца. Сложив числа, расположенные под нулями (в нашей задаче есть лишь одно слагаемое), получим число $\mathbf{a_4} = 5$, которое поместим в нижнюю клетку шестого столбца.

В итоге получим расчетную таблицу 4.

Контроль: сумма чисел, расположенных непосредственно над нулем третьего столбца, слева от него и под ним, должна быть равна объему выборки (32 + 30 + 38 = 100); сумма двух чисел, расположенных над двумя ступеньками ступенчатой линии (обведены жирными отрезками), должна быть равна соответственно числам $\mathbf{b_i}$, стоящим над предшествующей ступенькой (при движении по «лесенке» вверх): 32 + 40 = =72= b_1; $40 + 30 = 70 = b_2$; $30 + 12 = 42 = b_3$. Аналогично проверяется совпадение сумм двух чисел, стоящих под «ступеньками лесенки», ведущей вниз: $38 + 37 = 75 = a_1$, $37 + 22 = 59 = a_2$,

$22 + 5 = 27 = a_3$. При несовпадении хотя бы одной из указанных сумм следует искать ошибку в расчете.

Найдем $d_i(i = 1,2,3)$ *и* $s_i(i = 1,2,3,4)$:

$$d_1 = a_1 - b_1 = 75 - 72 = 3\ ,\quad d_2 = a_2 - b_2 = 59 - 70 = -11$$

Таблица 4

1	2	3	4	5	6
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

$d_3 = a_3 - b_3 = 27 - 42 = -15$;

$s_1 = a_1 + b_1 = 75 + 72 = 147$; $s_2 = a_2 + b_2 = 59 + 70 = 129$;

$s_3 = a_3 + b_3 = 27 + 42 = 69$; $s_4 = a_4 + b_4 = 5 + 14 = 19$.

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

$$M_1^* = \frac{d_1}{n} = \frac{3}{100} = 0{,}03,$$

$$M_2^* = \frac{s_1+2s_2}{n} = \frac{147+21\cdot 129}{100} = 4{,}05,$$

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

$$M_4^* = \frac{s_1+14s_2+36s_3+24s_4}{n} = \frac{147+14\cdot 129+36\cdot 69+24\cdot 19}{100} = 48{,}93.$$

Найдем центральные эмпирические моменты третьего и четвертого порядков:

$$m_3 = [M_3^* - 3M_1^*M_2^* + 2(M_1^*)^3]\cdot h^3 = [-1{,}53 - 3\cdot 0{,}03\cdot 4{,}05 + 2\cdot(0{,}03)^3]\,4^3 = -121{,}248,$$

$$m_4 = [M_4^* - 4M_1^*M_3^* + 6(M_1^*)^2M_2^* - 3(M_1^*)^4]h^4 = [48{,}93 - 4\cdot 0{,}03\cdot(-1{,}53) + 6\cdot(0{,}03)^2\cdot 4{,}05 - 3(0{,}03)^4]\cdot 4^4 = 49{,}135.$$

Найдем искомые асимметрию и эксцесс, учитывая, что $\sigma_B = \sqrt{D_B} = \sqrt{64{,}78}$:

$$a_s = \frac{m_3}{\sigma_B^3} = \frac{-121{,}248}{(\sqrt{64{,}78})^3} = -0{,}23,\ e_R = \frac{m_4}{\sigma_B^4} = \frac{49{,}134}{(\sqrt{64{,}78})^4} = 0{,}01.$$

Пример 3. Найти методом сумм асимметрию и эксцесс по заданному распределению выборки объема n=100:

а)

x_i	10.2	10.4	10.6	10.8	11.0	11.2	11.4	11.6	11.8	12.0
n_i	2	8	13	25	20	12	10	6	1	3

б)

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

Отв. а) a_s = - 0,01, e_R = - 0,24; б) a_s = 0,49, e_R = 0,36.

2. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

1. Линейная корреляция

Если обе линии регрессии Y на X и X на Y – прямые, то корреляцию называют *линейной*.

Выборочное уравнение прямой линии регрессии Y на X имеет вид

$$\bar{y}_x - \bar{y} = r_B\frac{\sigma_y}{\sigma_x}(x - \bar{x}), \qquad (*)$$

где $\bar{y}$ – условная средняя; $\bar{x}$ *и* $\bar{y}$ – выборочные средние признаков X и Y; σ_x *и* σ_y – выборочные средние квадратические отклонения; r_B – выборочный коэффициент корреляции, причем $r_B = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\sigma_x\sigma_y}$.

Выборочное уравнение прямой линии регрессии X на Y имеет вид

$$\bar{x}_y - \bar{x} = r_B \frac{\sigma_x}{\sigma_y}(y - \bar{y}). \qquad (**)$$

3. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СИСТЕМЫ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН. КОРЕЛЯЦИОННЫЙ МОМЕНТ. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ.

Математические ожидания и дисперсия составляющих системы двух случайных величин служат для описания этой системы. Кроме них используют и другие характеристики – это корреляционный момент $\boldsymbol{\mu_{xy}}$ и коэффициент корреляции $\mathbf{r_{xy}}$

Определение 1: Корреляционным моментом $\boldsymbol{\mu_{xy}}$ случайных величин **X** *и* ***Y*** называют *математическое ожидание произведения отклонений этих величин:*

$$\mu_{xy} = M\{[X - M(X)][Y - M(Y)]\} \qquad (1)$$

(это второй смешанный центральный момент, момент связи).

При вычислении корреляционного момента дискретных величин используют формулу:

$$\mu_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m}[x_i - M(X)]\left[y_j - M(Y)\right]p(x_i, y_j) \qquad (2)$$

А для непрерывных величин – формулу:

$$\mu_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x - M(X)][y - M(Y)]\, f(x,y)\, \boldsymbol{dx\, dy} \qquad (3)$$

Корреляционный момент служит для характеристики связи между величинами **X** *и* ***Y***

Теорема: Корреляционный момент двух независимых случайных величин **X** *и* ***Y*** равен нулю.

Доказательство: По условию случайные величины **X** *и* ***Y*** независимы, значит, независимы и их отклонения $\mathbf{X} - \mathbf{M(X)}$ *и* $\boldsymbol{Y} - \mathbf{M}(\boldsymbol{Y})$. На основе свойств математического ожидания имеем:

$$\mu_{xy} = M\{[X - M(X)][Y - M(Y)]\} = M[x - M(X)][y - M(Y)] \qquad (4)$$

Известно, что математическое ожидание отклонения равно нулю и, следовательно, корреляционный момент $\boldsymbol{\mu_{xy} = 0}$

Если корреляционный момент отличен от нуля, то это есть признак наличия связи между этими двумя случайными величинами. Эта связь не только в их зависимости, но и в рассеивании. Если одна из величин мало отличается от своего математического ожидания (почти не случайна), то $\boldsymbol{\mu_{xy}}$ будет мал, какой бы тесной зависимостью ни были связаны величины **X** *и* ***Y***.

Очевидно, что размерность $\boldsymbol{\mu_{xy}}$ равна произведению размерностей величин **X** *и* ***Y***, и поэтому величина корреляционного момента может иметь различные значения для одних и тех же двух величин в зависимости от того, в каких единицах они были измерены. Эта особенность корреляционного момента является недостатком этой числовой характеристики. Для его устранения вводят еще одну характеристику – коэффициент корреляции.

Определение 2: Коэффициентом корреляции $\mathbf{r_{xy}}$ случайных величин **X** *и* ***Y*** называют отношение корреляционного момента к произведению средних квадратичных отклонений этих величин:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y} \qquad (5)$$

$\sigma_x\sigma_y$ имеют размерность величин **X** *и* ***Y*** соответственно. Значит r_{xy} - безразмерная величина, в этом ее преимущество перед корреляционным моментом. Очевидно, что коэффициент корреляции r_{xy} обращается в ноль одновременно с корреляционным моментом μ_{xy}. Величины для которых $r_{xy} = 0$ называют некоррелироваными. Коэффициент корреляции характеризует не всякую зависимость, а только так называемую линейную зависимость. Линейная вероятностная зависимость двух случайных величин **X** *и* ***Y*** заключается в том, что при возрастании одной величины другая имеет тенденцию возрастать по линейному закону.

Теорема 1: Абсолютная величина корреляционного момента двух случайных величин **X** *и* ***Y*** не превосходит произведения их средних квадратичных отклонений:

$$|\mu_{xy}| \leq \sqrt{D_x D_y}$$

Теорема 2: Абсолютная величина коэффициента корреляции не превышает единицу:

$$|r_{xy}| \leq 1$$

4. ПРОВЕРКА ГИПОТЕЗЫ О ЗНАЧИМОСТИ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ.

Пусть двумерная генеральная совокупность (X, Y) распределена нормально. Из этой совокупности извлечена выборка объема **n** и по ней найден выборочный коэффициент корреляции $r \neq 0$. Требуется проверить нулевую гипотезу H0: $r_r = 0$ о генеральном коэффициенте корреляции.

Если нулевая гипотеза принимается, то это означает, что X и Y некоррелированы; в противном случае – коррелированы.

Правило: Для того, чтобы при уровне значимости **a** проверить нулевую гипотезу о равенстве нулю генерального коэффициента

корреляции нормальной двумерной случайной величины при конкурирующей гипотезе Н1: **r ≠ 0,** надо вычислить наблюдаемое значение критерия

$$T_{набл} = r_B \frac{\sqrt{n-2}}{\sqrt{1-r_B^2}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости **a** и числу степеней свободы найти критическую точку **tкр** двусторонней критической области. Если **|Тнабл| < tкр** – нет оснований отвергнуть нулевую гипотезу. Если **|Тнабл| > tкр** – нулевую гипотезу отвергают.

Задания для лабораторной работы.

1)Построить законы распределения для X и Y;
2)Найти числовые характеристики $\bar{x}$, y, σ_x, σ_y, r_B ;

3)Написать уравнения линий прямой регрессии, построить их и найти точку пересечения;

4)Проверить гипотезу о значимости выборочного коэффициента корреляции.

(1)

Y	X					n_y
	20	25	30	35	40	
16	4	6	-	-	-	10
26	-	8	10	-	-	18
36	-	-	32	3	9	44
46	-	-	4	12	6	22
56	-	-	-	1	5	6
n_x	4	14	46	16	20	n=100

(2)

Y	X								n_y
	5	10	15	20	25	30	35	40	
100	2	1	-	-	-	-	-	-	3
120	3	4	3	-	-	-	-	-	10
140	-	-	5	10	8	-	-	-	23
160	-	-	-	1	-	6	1	1	9
180	-	-	-	-	-	-	4	1	5
n_x	5	5	8	11	8	6	5	2	n=50

(3)

Y	X							n_y
	18	23	28	33	38	43	48	
125	-	1	-	-	-	-	-	1
150	1	2	5	-	-	-	-	8
175	-	3	2	12	-	-	-	17
200	-	-	1	8	7	-	-	16
225	-	-	-	-	3	3	-	6
250	-	-	-	-	-	1	1	2
n_x	1	6	8	20	10	4	1	n=50

(4)

Y	X							n_y
	5	10	15	20	25	30	35	
100	-	-	-	-	-	6	1	7
120	-	-	-	-	-	4	2	6
140	-	-	8	10	5	-	-	23
160	3	4	3	-	-	-	-	10
180	2	1	-	1	-	-	-	4
n_x	5	5	11	11	5	10	3	n=50

(5)

Y	X					n_y
	0	4	6	7	10	
7	19	1	1	-	-	21
13	2	14	-	-	-	16
40	-	3	22	2	-	27
80	-	-	-	15	-	15
200	-	-	-	-	21	21
n_x	21	18	23	17	21	n=100

(6)

Y	X					n_y
	0	1	2	3	4	
10	20	5	-	-	-	25
11	7	15	3	1	-	26
20	-	3	17	4	-	24
35	-	-	8	13	7	28
50	-	-	-	5	42	47
n_x	27	23	28	23	49	n=150

(7)

Y	X			n_y
	7	8	9	
200	41	7	-	48
300	1	52	1	54
400	-	8	40	48
n_x	42	67	41	n=150

(8)

Y	X			n_y
	0	4	5	
1	50	5	1	56
35	-	44	-	44
50	-	5	45	50
n_x	50	54	46	n=150

(9)

Y	X					n_y
	0	1	2	3	4	
0	18	1	1	-	-	20
3	1	20	-	-	-	21
5	3	5	10	2	-	20
10	-	-	7	12	-	19
17	-	-	-	-	20	20
n_x	22	26	18	14	20	n=100

(10)

Y	X			n_y
	2	3	5	
25	20	-	-	20
45	-	30	1	31
110	-	1	48	49
n_x	20	31	49	n=100