

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

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по выполнению лабораторных работ по дисциплине
«Анализ данных»
для студентов направления подготовки
09.03.01 Информатика и вычислительная техника
Направленность (профиль)
«Программное обеспечение средств вычислительной техники и
автоматизированных систем»

Содержание

Лабораторная работа №1-3 Построение вариационных рядов. Расчет числовых характеристик. Точечное оценивание параметров распределений. Интервальное оценивание параметров распределений	3
Лабораторная работа №4 Изучение распределений непрерывных случайных величин.....	19
Лабораторная работа №5 Проверка параметрических статистических гипотез	30
Лабораторная работа №6 Проверка непараметрических статистических гипотез	41
Лабораторная работа №7 Двухфакторный параметрический дисперсионный анализ.....	69
Лабораторная работа №8 Линейные преобразования случайных функций	89
Лабораторная работа № 9 Построение модели линейной корреляции по не сгруппированным данным	102
Лабораторная работа №10 Построение выборочного уравнения линии регрессии по сгруппированным данным	113
Лабораторная работа №11 Кластерный анализ.	124
Приложения.....	139

Лабораторная работа №1-3 Построение вариационных рядов. Расчет числовых характеристик. Точечное оценивание параметров распределений. Интервальное оценивание параметров распределений

Ц е л ь р а б о т ы: овладение способами построения рядов распределения и методами расчета числовых характеристик.

Выполнение лабораторного занятия № 1 рассмотрим на примере следующей задачи.

З а д а ч а. Имеются данные об обводненности нефти из насосных скважин (в %):

61,2	61,4	60,2	61,2	61,3	60,4	61,4	60,8	61,2	60,6
61,6	60,2	61,3	60,3	60,7	60,9	61,2	60,5	61,0	61,4
61,1	60,9	61,5	61,4	60,6	61,2	60,1	61,3	61,1	61,3
60,3	61,3	60,6	61,7	60,6	61,2	60,8	61,3	61,0	61,2
60,5	61,4	60,7	61,3	60,9	61,2	61,1	61,3	60,9	61,4
60,7	61,2	60,3	61,1	61,0	61,5	61,3	61,9	61,4	61,3
61,6	61,0	61,7	61,1	60,9	61,5	61,6	61,4	61,5	61,2
61,6	61,3	61,8	61,1	61,7	60,9	62,2	61,1	62,1	61,0
61,5	61,7	62,3	62,3	61,7	62,9	62,5	62,8	62,6	61,5
62,1	62,6	61,6	62,5	62,4	62,3	62,1	62,3	62,2	62,1

С о д е р ж а н и е р а б о т ы: на основе совокупности данных опыта выполнить следующее:

1. Построить ряды распределения (интервальный и дискретный вариационные ряды). Изобразить их графики.

2. Построить график накопительных частот — кумуляту.

3. Составить эмпирическую функцию распределения и изобразить ее графически.

4. Вычислить моду, медиану, выборочную среднюю, выборочную дисперсию, выборочное среднее квадратическое отклонение, коэффициент вариации, асимметрию, эксцесс.

5. Построить доверительные интервалы для истинного значения измеряемой величины и среднего квадратического отклонения генеральной совокупности.

6. Раскрыть смысловую сторону каждой характеристики.

Методические указания по выполнению работы

1.1. Построить интервальный вариационный ряд. Для этого найти:

а) размах варьирования признака по формуле $R = x_{\max} - x_{\min}$, где $x_{\min}$ — наименьшая, $x_{\max}$ — наибольшая варианты в данной выборочной совокупности;

б) число интервалов вариационного ряда, пользуясь одним из приведенных ниже соотношений:

$k \approx \sqrt{n}$, $6 < k < 12$, $k \approx 1 + \log_2 n \approx 1 + 3,2 \lg n$, где n — объем выборки;

в) длину h частичных интервалов по формуле $h = \frac{R}{k}$ и, если необходимо, округлить это значение до некоторого числа;

г) записать полученный интервальный вариационный ряд, заполнив табл. 2, §1. Сделать контроль, убедившись, что $\sum n_i = n$.

1.2. Построить дискретный вариационный ряд, взяв в качестве вариантов середины вариант-интервалов непрерывного вариационного ряда, а в качестве частот — частоты непрерывного вариационного ряда (табл. 3).

1.3. Изобразить графически интервальный и дискретный вариационные ряды (построить гистограмму и полигон частот).

2. Построить график накопленных частот — *кумуляту*, т.е. ломаную, проходящую через точки с координатами x_i и соответствующими накопленными частотами. Предварительно составить табл. 5 § 1.

3. Найти эмпирическую функцию распределения и изобразить ее графически.

4.1. Найти моду $M_o X$ и медиану $M_e X$.

4.2. Для вычисления остальных статистик воспользоваться методом произведений. Ввести условные варианты $u_i = \frac{x_i - C}{h}$, где $C = M_o X$, h — шаг (длина интервала). Составить расчетную табл. 6, § 2.

Контроль вычислений произвести по формуле:

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

4.3. Пользуясь табл. 6, вычислить начальные моменты (11) — (14):

$$M_1^* = \frac{1}{n} \sum n_i u_i, \quad M_2^* = \frac{1}{n} \sum n_i u_i^2, \quad M_3^* = \frac{1}{n} \sum n_i u_i^3, \quad M_4^* = \frac{1}{n} \sum n_i u_i^4.$$

4.4. Найти выборочную среднюю $\bar{x} = M_1^* h + C$.

4.5. Найти выборочную дисперсию $D(X) = (M_2^* - M_1^{*2})h^2$.

4.6. Найти выборочное среднее квадратическое отклонение (17):

$$\sigma_{\text{в}} = S = \sqrt{D(X)}.$$

4.7. Найти коэффициент вариации (22): $V = S / \bar{x}$.

4.8. Найти центральные моменты (20), (21):

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

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

4.9. Вычислить асимметрию (18): $A_s = \frac{m_3}{S^3}$ и эксцесс (19): $E_x = \frac{m_4}{S^4} - 3$.

5. Доверительные интервалы для σ и σ найти, согласно (23), (24):

$$\bar{x} - t_{\gamma} \times \frac{S_x}{\sqrt{n}} < \sigma < \bar{x} + t_{\gamma} \times \frac{S_x}{\sqrt{n}} \text{ при } \gamma = 0,95.$$

Величину t_{γ} найти по приложению 3, а по приложению 4 найти величину q , удовлетворяющую одному из условий:

$$S(1 - q) < \sigma < S(1 + q) \text{ при } q < 1,$$

$$0 < \sigma < S(1 + q) \text{ при } q > 1.$$

6. Раскрыть смысловую сторону каждой характеристики.

Выполнение работы

Обозначим через X обводненность нефти из рассматриваемых насосных скважин.

1.1. По данным выборки строим интервальный вариационный ряд.

а) Поскольку, как легко выяснить, $x_{\max} = 62,8$, $x_{\min} = 60,1$, то размах варьирования признака X равен $R = x_{\max} - x_{\min} = 62,8 - 60,1 = 2,7$.

б) Определяя число k интервалов (число столбцов в таблице) вариационного ряда, положим $k = 10$.

в) Длина h каждого частичного интервала равна $h = \frac{R}{k} = \frac{2,7}{10} = 0,27$.

Так как исходные данные мало отличаются друг от друга и содержат один

десятичный знак, то величину h округляем до одного десятичного знака: $h = 0,3$.

г) Подсчитываем число вариантов, попадающих в каждый интервал, по данным выборки. Значение x_i , попадающее на границу интервала, относим к правому интервалу. За начало x_0 первого интервала берем величину $x_0 = x_{\min} - 0,5h = 60,1 - 0,5 \cdot 0,3 = 59,95 \approx 60$. Конец x_k последнего интервала находим по формуле $x_k = x_{\max} + 0,5h = 62,8 + 0,15 = 62,95 \approx 63,0$. Сформированный интервальный вариационный ряд записываем в виде табл. 7.

Т а б л и ц а 7

Варианты-интервалы	60-60,3	60,3-60,6	60,6-60,9	60,9-61,2	61,2-61,5	61,5-61,8	61,8-62,1	62,1-62,4	62,4-62,7	62,7-63,0
Частоты, n_i	3	6	9	18	29	16	2	10	5	2

Контроль: $\sum n_i = 100$, и объем выборки $n = 100$.

1.2. Записываем дискретный вариационный ряд (табл. 8). В качестве вариант x_i берем середины интервалов интервального вариационного ряда.

Т а б л и ц а 8

варианты, x_i	60,15	60,45	60,75	61,05	61,35	61,65	61,95	62,25	62,55	62,85
частоты, n_i	3	6	9	18	29	16	2	10	5	2

1.3. Изображаем интервальный и дискретный вариационные ряды графически, построив гистограмму и полигон частот в одной системе координат (рис. 2).

2. Строим график накопленных частот — кумуляту (рис. 3). Предварительно составляем расчетную табл. 9.

Т а б л и ц а 9

Варианты, x_i	60,15	60,45	60,75	61,05	61,35	61,65	61,95	62,25	62,55	62,85
относительные частоты, $w_i = n_i / n$	0,03	0,06	0,09	0,18	0,29	0,16	0,02	0,1	0,05	0,02
накопительные относительные частоты, $W_i = W_{i-1} + w_i$	0,03	0,09	0,18	0,36	0,65	0,81	0,83	0,93	0,98	1

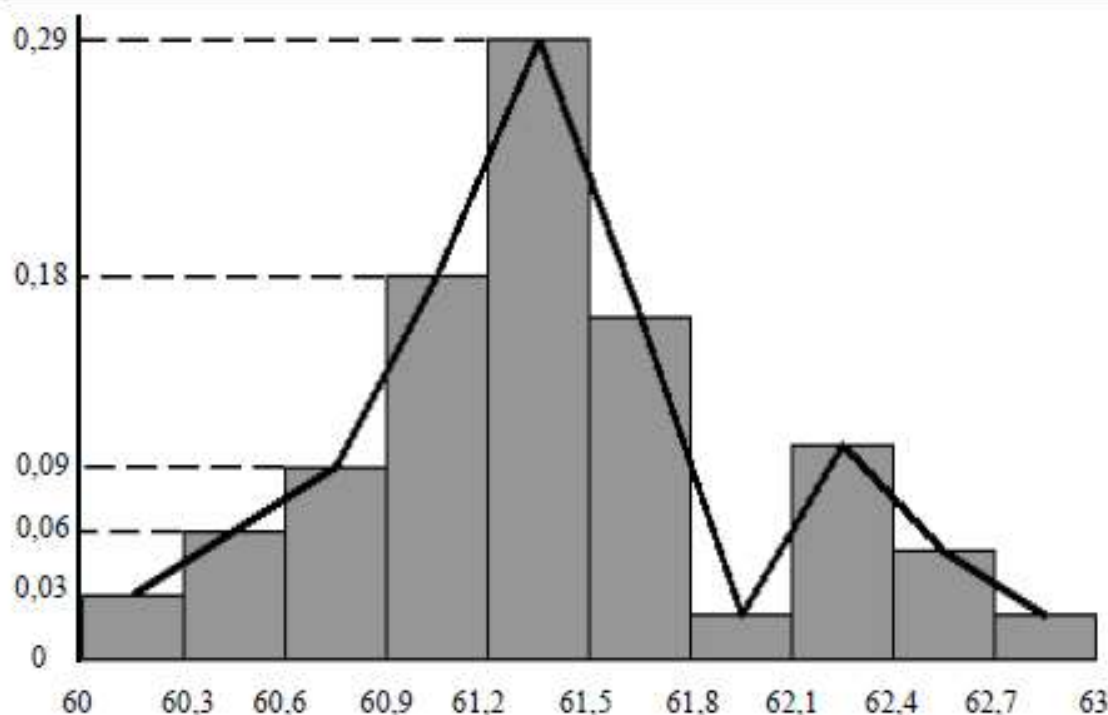


Рис.2. Гистограмма и полигон.

3. Находим эмпирическую функцию распределения. Воспользуемся формулой (1): $F_z(x) = \frac{n_x}{n}$.

Если $x \leq 60,15$, то $F_z(x) = 0$ — по свойству эмпирической функции распределения.

$$\text{Если } 60,15 < x \leq 60,45, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{3}{100} = 0,03.$$

$$\text{Если } 60,45 < x \leq 60,75, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{3+6}{100} = 0,09.$$

$$\text{Если } 60,75 < x \leq 61,05, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{3+6+9}{100} = 0,18.$$

$$\text{Если } 61,05 < x \leq 61,35, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{18+18}{100} = 0,36.$$

$$\text{Если } 61,35 < x \leq 61,65, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{36+29}{100} = 0,65.$$

$$\text{Если } 61,65 < x \leq 61,95, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{65+16}{100} = 0,81.$$

$$\text{Если } 61,95 < x \leq 62,25, \text{ то } F_z(x) = \frac{n_x}{n} = \frac{81+2}{100} = 0,83.$$

Если $62,25 < x \leq 62,55$, то $F_z(x) = \frac{n_x}{n} = \frac{83+10}{100} = 0,93$.

Если $62,55 < x \leq 62,85$, то $F_z(x) = \frac{n_x}{n} = \frac{93+5}{100} = 0,98$.

Если $x > 62,85$, то $F_z(x) = 1$ — по свойству эмпирической функции распределения.

Записываем полученную эмпирическую функцию в виде:

$$F_z(x) = \begin{cases} 0, & x \in (-\infty, 60,15], \\ 0,03, & x \in (60,15; 60,45], \\ 0,09, & x \in (60,45; 60,75], \\ 0,18, & x \in (60,75; 61,05], \\ 0,36, & x \in (61,05; 61,35], \\ 0,65, & x \in (61,35; 61,65], \\ 0,81, & x \in (61,65; 61,95], \\ 0,83, & x \in (61,95; 62,25], \\ 0,93, & x \in (62,25; 62,55], \\ 0,98, & x \in (62,55; 62,85], \\ 1, & x \in (62,85; +\infty). \end{cases}$$

График функции $F_z(x)$ представлен на рис.4.

Соединив середины вертикальных частей ступенчатой кусочно-постоянной кривой, являющейся графиком функции $F_z(x)$, получаем плавную кривую (на рис. 4 это штриховая линия). Абсциссами точек этой кривой служат значения обводненности нефти, добываемой насосным способом из скважин, а ординатами — значения эмпирической функции распределения, характеризующей оценку вероятности события $X \leq x_i$, т.е. вероятности попадания возможных значений обводненности нефти на промежуток $(-\infty, x_i]$.

Для нахождения числовых характеристик признака X — обводненности нефти (несмещенных оценок для $M(X) = a$, $D(X)$, а также $M_e X$, $M_o X$, A_x , E_x) воспользуемся табл. 8.

4.1. Так как варианта $x = 61,35$ в табл. 8 встречается с наибольшей частотой $n_5 = 29$, то $M_o X = 61,35$, т.е. это значение обводненности нефти, встречающееся в данной выборке с наибольшей частотой.

Находим $M_e X$. Так как табл. 8 содержит четное число столбцов, то $M_e X = \frac{61,35+61,65}{2} = 61,5$. Это значение обводненности нефти, которое делит данные выборки признака X на равные части.

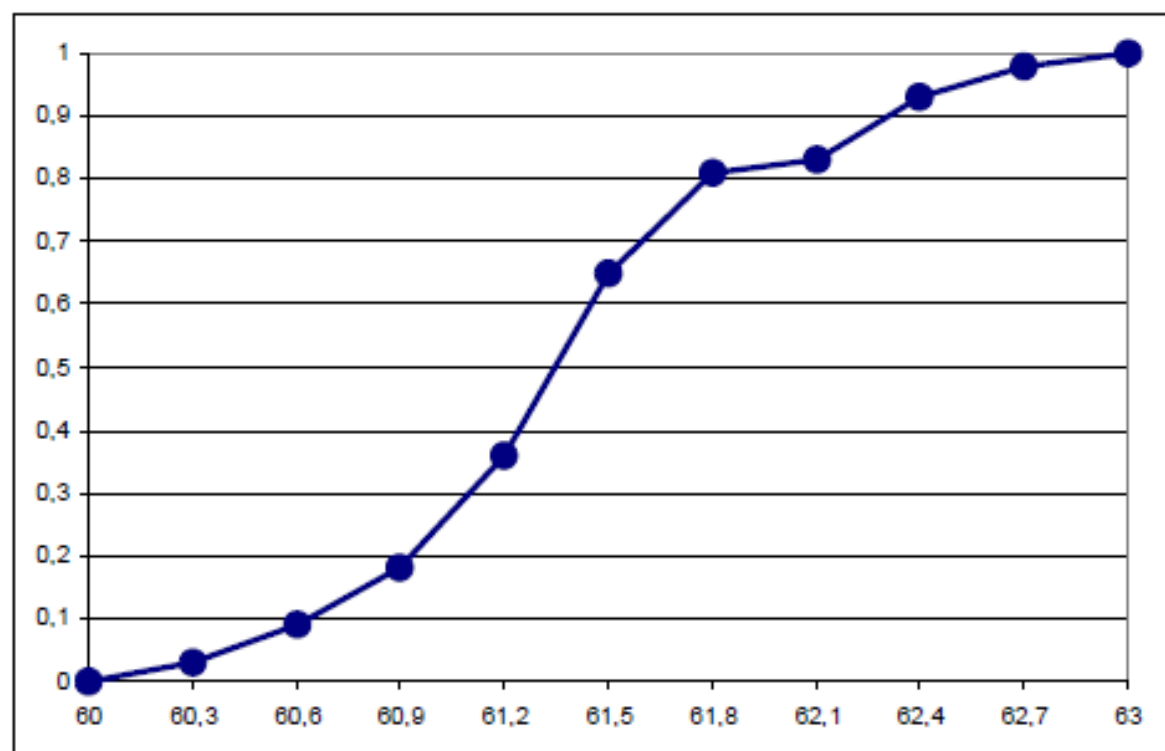


Рис. 3. Кумулятивная кривая.

4.2. Для нахождения остальных статистик, характеризующих обводненность нефти, воспользуемся методом произведений. Введем, согласно (9), условные варианты $u_i = \frac{x_i - C}{h}$; $C = M_o X = 61,35$, $h = 0,3$.

Составим расчетную табл. 10.

Таблица 10

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	контрольный столбец $n_i(u_i + 1)^2$
60,15	3	-4	-12	48	-192	768	27
60,45	6	-3	-18	54	-162	486	24
60,75	9	-2	-18	36	-72	144	9
61,05	18	-1	-18	18	-18	18	0
61,35	29	0	0	0	0	0	29
61,65	16	1	16	16	16	16	64
61,95	2	2	4	8	16	32	18
62,25	10	3	30	90	270	810	160
62,55	5	4	20	80	320	1280	125
62,85	2	5	10	50	250	1250	72
	100		14	400	428	4804	528

Контроль вычислений проводим по формуле:

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

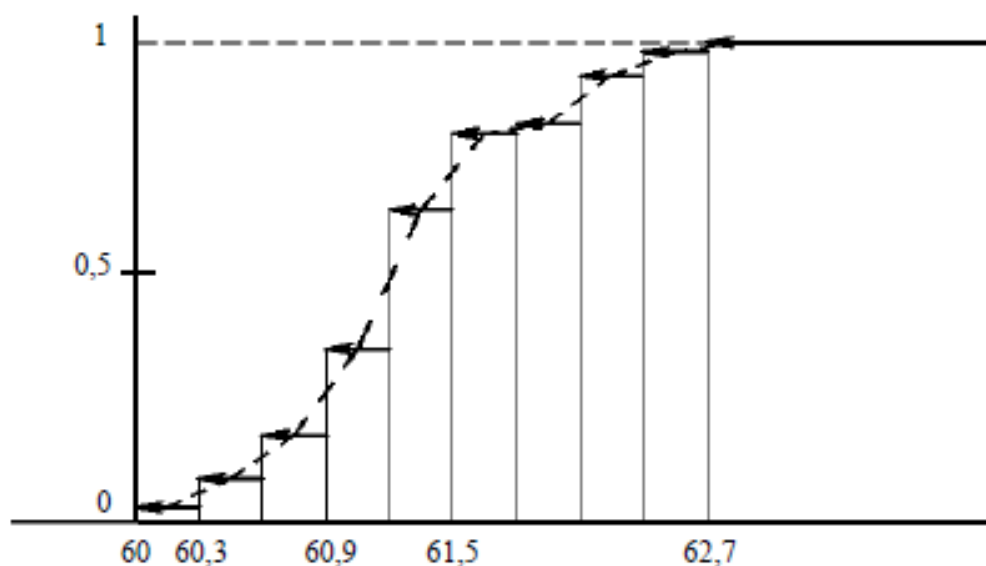


Рис.4. Кумулята и эмпирическая функция распределения.

т.е.

$$100 + 2 \cdot 14 + 400 = 528 = 27 + 24 + 9 + 29 + 64 + 18 + 160 + 125 + 72.$$

Следовательно, вычисления проведены верно.

4.3. Пользуясь результатами последней строки табл. 10, находим условные начальные моменты (11) — (14):

$$M_1^* = \frac{1}{n} \sum n_i u_i = 14/100 = 0,14,$$

$$M_2^* = \frac{1}{n} \sum n_i u_i^2 = 400/100 = 4,$$

$$M_3^* = \frac{1}{n} \sum n_i u_i^3 = 428/100 = 4,28,$$

$$M_4^* = \frac{1}{n} \sum n_i u_i^4 = 4804/100 = 48,04.$$

4.4. Находим выборочную среднюю (15):

$$\bar{x} = M_1^* h + C = 0,14 \cdot 0,3 + 61,35 = 61,392 \approx 61,39,$$

которая характеризует среднюю обводненность нефти из насосных скважин в данной выборке, составляющую 61,39 %.

4.5. Находим выборочную дисперсию (16):

$$S^2 = (M_2^* - M_1^{*2}) h^2 = (4 - 0,14^2) \cdot 0,09 = 0,358236$$

4.6. Вычисляем выборочное среднее квадратичное отклонение (17):

$$S = \sqrt{S^2} = \sqrt{0,358236} = 0,59853.$$

4.7. Величина $S = 0,599$ характеризует степень рассеяния значений обводненности нефти относительно средней обводненности. Для определения колеблемости значений обводненности нефти в процентном отношении вычисляем коэффициент вариации (22):

$$V = S / \bar{x} = \frac{0,599}{61,39} = 0,00976$$

Величина коэффициента вариации мала (составляет 0,01), что означает тесную сгруппированность значений обводненности нефти около центра рассеяния, т.е. около средней обводненности нефти.

4.8. Для предварительной оценки отклонения значений обводненности нефти от нормального распределения вычисляем асимметрию и эксцесс. Сначала находим центральные моменты третьего и четвертого порядков (20), (21):

$$m_3 = (M_3^* - 3M_2^*M_1^* + 2M_1^{*3})h^3 = (4,28 - 3 \times 4 \times 0,14 + 2 \times 0,14^3) \times 0,3^3 = 0,07035.$$

$$m_4 = (M_4^* - 4M_3^*M_1^* + 6M_2^*M_1^{*2} - 3M_1^{*4})h^4 = \\ = (48,04 - 4 \times 4,28 \times 0,14 + 6 \times 4 \times 0,14^2 - 3 \times 0,14^4) \times 0,3^4 = 0,37023.$$

4.9. Тогда в соответствии с (18), (19), находим:

$$A_s = \frac{m_3}{S^3} = \frac{0,07}{0,599^3} = 0,328,$$

$$E_x = \frac{m_4}{S^4} - 3 = \frac{0,37}{0,599^4} - 3 = -0,115.$$

Вывод. Значения A_s и E_x мало отличаются от нуля. Поэтому можно предположить близость данной выборки, характеризующей обводненность нефти, к нормальному распределению. Эта гипотеза будет проверяться в лабораторной работе № 2.

5. Произведем оценку генеральной средней $M(X) = a$ и генерального среднеквадратического отклонения $\sigma = S$ по выборочным статистикам $\bar{x}$ и S , используя теорию доверительных интервалов для нормального распределения.

Доверительный интервал для истинного значения обводненности нефти с надежностью $\gamma = 0,95$ находим, согласно (23):

$$\bar{x} - \frac{S}{\sqrt{n}} \times t_{\gamma} < a < \bar{x} + \frac{S}{\sqrt{n}} \times t_{\gamma}.$$

Согласно приложению 3, при $n=100$ и $\gamma=0,95$ находим $t_\gamma=1,984$.
Записываем доверительный интервал:

$$61,39 - \frac{0,599}{10} \times 1,984 < a < 61,39 + \frac{0,599}{10} \times 1,984,$$

или $61,27 < a < 61,51$.

Таким образом, средняя обводненность нефти из насосных скважин (в %) по данным выборки должна находиться в промежутке (61,27;61,51).

Запишем доверительный интервал для генерального среднеквадратического отклонения $\sigma = S$. При заданных $\gamma=0,95$ и $n=100$ по таблице приложения 4 находим $q=0,143$. Так как $q < 1$, то доверительный интервал записываем в виде (24):

$$S(1-q) < \sigma < S(1+q),$$

или

$$0,599(1-0,143) < \sigma < 0,599(1+0,143),$$

или

$$0,51 < \sigma < 0,68;$$

следовательно, отклонения истинных значений обводненности нефти из насосных скважин не должны выходить за пределы промежутка (0,51;0,68).

Варианты заданий по лабораторному занятию № 1

Вариант № 1. Имеются данные о производительности труда (количество деталей в смену):

71	76	79	86	78	76	84	78	74	76	99	87	82
78	84	81	76	75	82	85	80	76	79	76	86	86
86	89	77	80	74	86	87	74	79	84	75	85	81
88	77	74	93	85	83	80	75	93	95	91	88	85
85	83	85	82	86	79	84	88	74	92	95	76	

Вариант № 2. Имеются данные о пропускной способности 50 участков нефтепровода (м³/сут.):

19,8	19,1	19,3	18,8	20,2	20,8	20,7	19,7	19,6	19,2	20,9	20,9	20,2
19,6	20,4	20,4	20,2	20,4	18,9	19,7	19,8	20,6	20,7	19,7	20,3	19,8
20,4	20,3	20,6	20,5	20,4	20,5	20,3	20,5	20,2	20,5	20,7	21,0	20,4
20,8	20,5	20,4	20,6	21,0	20,4	20,4	20,3	19,7	19,9	20,1		

Вариант № 8. Имеются данные о числе рабочих дней без простоя для пятидесяти буровых бригад одного из районов страны:

261	260	258	263	257	260	259	264	261	260	264	261	265
261	260	263	260	260	259	260	258	265	259	265	261	258
259	259	258	262	264	258	259	263	266	259	261	266	262
259	262	261	266	262	259	262	261	259	262	262	261	266
259	262											

Вариант № 9. Приведено количество деталей, выработанных за смену различными рабочими:

75	88	74	80	76	82	86	76	93	74	72	82	71
82	87	81	87	79	78	87	82	87	82	74	77	83
86	85	86	76	81	86	76	71	80	85	73	75	88
89	84	85	85	81	82	85	83	76	87	87	76	76
73	78	87	80	78	72	83	91	82	93	76	83	80

Вариант № 10. Имеются данные о рабочих дебитах газовой скважины (тыс. м³/сут.):

550	550	551	550	551	562	550	562	561	530	542	535	542
539	537	543	540	556	546	556	556	534	548	533	558	560
558	548	540	541	551	549	551	550	552	568	538	551	547
552	559	557	546	552	550	557	547	552	554	547	554	567
558	563	562	569	552	554	549	545	560	539	549	539	

Вариант № 11. Имеются данные о коэффициенте эксплуатации насосных скважин в различных нефтеносных районах страны:

0,90	0,79	0,84	0,86	0,88	0,90	0,89	0,85	0,91	0,98	0,91	0,80	0,87
0,89	0,88	0,78	0,81	0,85	0,88	0,94	0,86	0,80	0,86	0,91	0,78	0,86
0,91	0,95	0,97	0,88	0,79	0,82	0,84	0,90	0,81	0,87	0,91	0,90	0,82
0,85	0,90	0,82	0,85	0,90	0,96	0,98	0,89	0,87	0,99	0,85		

Вариант № 12. 50 сверл были подвергнуты испытанию на твердость. При этом фиксировалась твердость лапки. Результаты испытания следующие:

14,5	14,6	15,1	15,5	16,3	16,8	17,9	16,3	14,5	14,9	13,6	15,4	16,9
15,4	14,3	15,5	11,3	15,5	17,1	16,8	12,2	15,2	15,7	11,6	16,9	15,7
17,7	16,6	16,2	15,5	12,8	14,2	15,5	16,1	14,3	16,5	14,5	17,9	17,8
16,9	11,7	13,2	14,9	19,8	16,6	17,9	14,9	15,2	17,3	16,9		

Вариант № 13. Даны значения обследуемого признака X — себестоимости единицы продукции (в руб.):

73	77	78	88	76	78	86	77	75	90	88	84	79
87	83	79	73	84	86	85	74	77	74	88	81	87
85	76	79	71	88	83	76	76	82	73	89	79	90
76	75	91	83	82	84	85	78	85	85	79	92	86
84	77	92	93	91	85	84	87	81	83	80	82	76
81	90	78	81	95	77	91	84	96	84	79	79	83
88	84	83	93	73	79	92	89	75	83	87	89	71
75	83	87	92	80	88	91	95	82				

Вариант № 14. Имеются данные о суточном дебите газа в наблюдаемой скважине:

39	19	21	28	26	27	29	28	28	27	23	26	32
34	26	24	22	19	23	27	30	29	25	18	18,5	20
22	24	28	31	33	25	18	21	26	30	32	34	29
26	21	20	23	25	27	30	32	29	27	23		

Вариант № 15. Даны замеры толщины резца (в мм):

24,5	26,8	23,6	25,5	22,2	26,9	25,3	24,1	28,5	25,3	24,1	28,5	25,3
24,6	27,9	25,4	21,3	25,2	27,7	23,6	25,2	26,8	25,9	25,1	26,3	25,4
21,3	25,2	25,5	25,7	26,6	28,2	25,4	23,2	26,6	25,7	24,3	26,8	25,8
27,1	26,2	25,9	21,6	25,3	25,1	24,8	26,3	24,9	24,3	26,8		

Вариант № 16. Имеются данные о расходах, связанных с монтажом и демонтажом оборудования на предприятии (в тыс. руб.):

4,7	7,2	6,2	6,7	7,2	5,7	7,7	8,2	6,2	5,2	7,2	5,7	6,2
5,7	8,2	5,7	6,7	6,2	5,7	6,2	6,7	5,2	7,7	6,2	7,2	7,7
6,7	7,2	8,2	6,2	5,7	6,2	7,7	6,7	7,2	5,7	6,7	8,2	7,7
8,2	4,7	8,7	4,2	8,7	6,2	6,7	6,2	7,2	4,9	5,5		

Вариант № 17. Даны значения обследуемого признака X — себестоимости одной детали (в руб.):

82	83	73	76	79	89	95	92	93	84	88	76	88
81	78	86	84	84	86	85	87	84	74	83	87	73
76	73	78	76	76	74	88	82	73	85	79	77	79
97	84	80	75	81	73	78	83	75	90	83	77	84
85	90	92	91	85	71	85	87	82	94	92	76	93
90	73	92	84	93	88	84	81	93	81	91	78	85
84	95	79	79	83	96	89	82	79	77	83	88	81
88	82	77	92	76	84	83	87	89				

Вариант № 18. Даны значения диаметров шестерен, обрабатываемых на станке:

21	29	27	29	27	29	31	29	31	29	29	23	39
31	29	31	29	31	29	31	33	31	31	31	27	23
27	33	29	25	29	19	29	31	23	31	29	27	33
29	31	29	31	23	35	27	29	29	27	29	29	21
29	27	29	29	29	33	29	25	25	27	31	29	29
27	33	29	31	29	29	29	35	27	29	35	29	33
29	27	31	31	27	29	35	27	33	29	27	29	25
27	31	37	25	31	27	27	29	25				

Вариант № 19. Даны значения израсходованных долот на 100 скважинах при механической скорости проходки 18 м/сек.:

28	30	28	27	28	29	29	29	31	28	26	25	33
35	27	31	31	30	28	33	23	30	31	33	31	27
30	28	30	29	30	26	25	31	33	26	27	33	29
30	30	36	26	25	28	30	29	27	32	29	31	30
31	26	25	29	31	33	27	32	30	31	34	28	26
38	29	31	29	27	31	30	28	34	30	26	30	32
30	29	30	28	32	30	29	34	32	35	29	27	28
30	30	29	32	29	34	30	32	24				

Вариант № 20. Даны значения внутреннего диаметра гайки (в мм):

4,25	4,38	4,48	4,53	4,54	4,41	4,52	4,39	4,16	4,27	4,59	4,48	4,56
4,13	4,51	4,31	4,27	4,87	4,32	4,49	4,74	4,17	4,66	4,92	4,48	4,68
4,45	4,12	4,69	4,28	4,74	4,55	4,28	4,54	4,51	4,77	4,71	4,78	4,13
4,51	4,42	4,36	4,45	4,32	4,17	4,79	4,13	4,52	4,73	4,95		

Вариант № 21. Даны значения ширины пера круглой плашки (в мм):

3,69	3,56	3,52	3,68	3,49	3,58	3,59	3,54	3,35	3,69	3,87	3,67	3,79
3,75	3,43	3,50	3,57	3,53	3,49	3,68	3,36	3,63	3,51	3,99	3,90	3,53
3,50	3,55	3,40	3,73	3,72	3,53	3,42	3,72	3,68	3,46	3,46	3,36	3,37
3,53	3,48	3,70	3,48	3,68	3,46	3,61	3,57	3,47	3,74	3,47		

Вариант № 22. Имеются данные об энергетических затратах на 1 м проходки при разведочном бурении нефтяных скважин в различных нефтяных районах страны (в тыс. руб.):

48	29	6	18	24	30	35	25	17	24	36	42	47
40	28	12	7	25	23	33	28	19	14	8	40	27
20	27	15	6	16	25	34	17	25	46	6	51	13
28	37	43	27	38	53	24	41	21	11	26		

Вариант № 23. Имеются данные о пластовом давлении (в атм.) при насосном способе эксплуатации 100 скважин:

95	57	15	26	35	46	52	55	59	47	42	48	58
55	102	96	45	54	56	60	10	16	20	49	48	43
12	19	51	103	62	61	38	29	10	39	40	18	14
41	58	63	59	60	63	68	70	71	75	82	87	92
99	65	68	78	91	94	77	65	79	67	74	80	89
69	81	83	100	90	36	64	97	50	76	72	31	55
28	57	85	69	13	53	11	61	90	76	17	37	

Вариант № 24. Имеются данные о продолжительности (в мес.) 50 фонтанирующих скважин:

19,2	18,1	18,4	18,2	18,6	18,9	19,0	18,4	18,5	19,3	18,3	18,7	18,8
19,1	18,9	19,3	18,4	19,2	18,2	18,7	19,5	18,7	19,1	18,7	19,1	19,6
18,6	18,8	19,3	18,8	19,0	19,5	18,9	19,0	19,8	19,7	19,4	19,3	19,1
19,8	18,9	19,7	18,5	19,0	19,9	19,2	19,1	18,6	19,5	19,6		

Вариант № 25. Имеются данные замеров температуры масла двигателя автомобиля ГАЗ-53А:

19	29	21	39	25	26	32	25	28	26	36	30	31
29	35	23	32	27	27	26	26	30	27	25	28	28
36	29	35	26	32	29	38	28	25	29	34	28	29
32	34	28	28	29	33	27	34	25	28	26	30	38
39	32	29	29	34	35	32	27	26	25	26	35	36
30	28	33	26	28	26	28	27	33	33	29	32	25
38	26	36	23	24	27	26	30	34	25	24	33	

Вариант № 26. Результаты измерения температуры раздела фракции бензин-авиакеросин на установке первичной переработке нефти (в °C):

133	133	142	135	145	144	145	147	146	134	130	134	138
144	141	141	134	141	136	140	143	139	141	137	140	145
145	141	144	138	139	143	141	141	146	143	140	139	143
143	139	140	139	138	138	135	141	141	140	138	145	135
148	136	139	142	143	143	137	138	138	139	138	144	143
138	142	138	140	140	137	139	140	139	137	136	136	135
135	141	142	136	140	136	137	138	138	137	139	139	140
139	140	140	139	139	139	140	140	146				

Вариант № 27. Имеются данные о суточном дебите нефти наблюдаемой скважины (в т/сут.):

16	13	11	15	18	19	21	18	17	15	13	16	18
17	19	15	13	12	14	16	17	20	17	17	20	19
18	22	24	1	15	14	10	12	16	18	18	19	21
23	20	22	24	17	16	14	15	18	15	11	16	17
15	13	16	17	18	14	15	19	17	18	16	13	15
17	21	23	26	19	22	24	25	20	21	24	19	22
23	20	25	21	20	22	26	19	22	25	28	23	20
21	27	19	15	22	23	18	22	22				

Вариант № 28. Имеются результаты испытания — твердости лапки сверла:

36,8	32,0	39,4	36,3	35,4	37,3	34,7	39,0	28,3	41,3	36,1	37,3	32,2
38,5	34,2	37,2	30,6	37,3	35,2	36,9	34,3	35,2	30,8	36,0	39,3	32,7
34,6	36,8	39,1	29,5	30,4	35,2	36,5	38,2	40,2	36,8	39,3	32,7	37,1
29,3	28,4	40,2	34,8	37,2	32,6	41,0	40,4	28,3	34,8	39,2		

Вариант № 29. Имеются данные о расходах, связанных с подготовительными работами, на 1 м проходки при разведочном бурении нефтяных скважин в различных нефтеносных районах страны (в тыс. руб.):

11	15	20	25	29	34	19	25	16	21	29	20	21
22	23	26	28	30	18	13	17	22	29	26	39	14
16	24	27	25	31	32	23	37	23	27	37	36	42
32	34	39	38	44	28	33	23	35	36	34		

Вариант № 30. Даны значения овальности валика (в мк):

25	29	33	21	29	25	29	28	31	23	31	27	29
27	27	29	31	27	29	29	29	31	25	29	29	27
29	31	29	27	25	28	27	31	31	29	27	27	33
29	33	31	33	25	27	35	37	35	27	27	29	27
29	31	29	27	29	31	29	21	23	29	37	29	31
29	31	29	31	29	39	29	39	39	27	31	37	29
31	29	27	23	29	27	31	29	29	31	29	35	29
19	29	27	29	29	31	33	29	25				

Лабораторная работа №4 Изучение распределений непрерывных случайных величин

Целью лабораторной работы является изучение наиболее распространенных непрерывных случайных величин в среде математического пакета Mathcad.

Как известно, каждая случайная величина полностью определяется своей функцией распределения.

Если ξ - случайная величина, то функция $F_{\xi}(x) = P(\xi \leq x)$ называется *функцией распределения* случайной величины ξ . Здесь $P(\xi \leq x)$ – вероятность того, что случайная величина ξ принимает значение меньше x . Если $F_{\xi}(x)$ непрерывна, то случайная величина ξ называется *непрерывной случайной величиной*. В случае, когда $F_{\xi}(x)$ непрерывно дифференцируема, то более наглядное представление о случайной величине дает *плотность вероятности* случайной величины $p_{\xi}(x)$, которая связана с функцией распределения $F_{\xi}(x)$ соотношениями

$$F_{\xi}(x) = \int_{-\infty}^x p_{\xi}(t) dt \text{ и } p_{\xi}(x) = \frac{dF_{\xi}(x)}{dx}.$$

Вероятность того, что значение случайной величины ξ попадет в интервал $[a, b]$, вычисляется для непрерывной случайной величины по формуле

$$P(a < \xi < b) = F_{\xi}(b) - F_{\xi}(a) = \int_a^b p_{\xi}(t) dt.$$

Как и в случае дискретных случайных величин изучение непрерывной случайной величины заключается в исследовании ее функции распределения, которая содержит полную информацию об этой величине. Поэтому обратимся к рассмотрению наиболее распространенных законов распределения непрерывной случайной величины.

1.Равномерное распределение

Непрерывная случайная величина ξ , принимающая значения на отрезке $[a, b]$, распределена равномерно на $[a, b]$, если плотность распределения $p_{\xi}(x)$ и функция распределения случайной величины $F_{\xi}(x)$ имеют вид:

$$p_{\xi}(x) = \begin{cases} 0, & x \notin [a, b], \\ \frac{1}{b-a}, & x \in [a, b] \end{cases} \quad F_{\xi}(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & x > b. \end{cases}$$

В Mathcad, в случае равномерного распределения, значения плотности распределения и функции распределения случайной величины в точке x вычисляются с помощью встроенных функций, соответственно, $\text{dunif}(x, a, b)$ и $\text{punif}(x, a, b)$. Пример

использования этих функций в рабочем документе показан на рисунке 1.

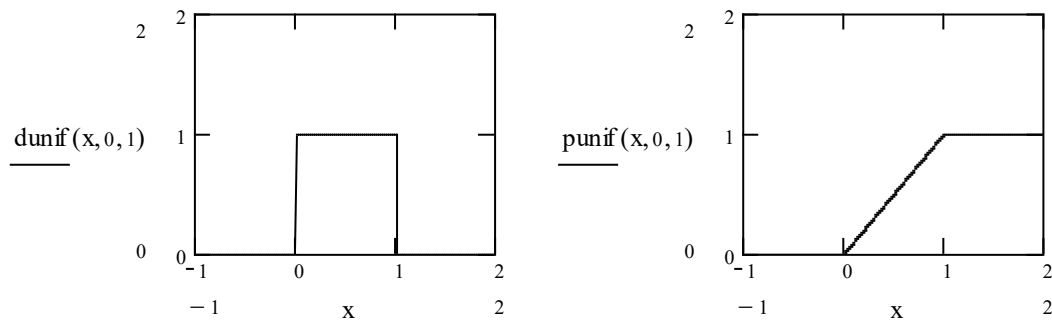


Рисунок 1. Графики плотности вероятностей и функции распределения случайной величины ξ , принимающей значения на отрезке $[0, 1]$ и имеющей равномерное распределение.

2. Экспоненциальное (показательное распределение)

Непрерывная случайная величина ξ имеет показательное распределение с параметром λ , если плотность распределения p_ξ имеет вид

$$p_\xi(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0. \end{cases}$$

Отсюда видно, что показательное распределенная случайная величина принимает только неотрицательные значения. Функция распределения такой случайной величины имеет вид

$$F_\xi(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$$

В Mathcad значения в точке x плотности распределения и функции распределения случайной величины, имеющей экспоненциальное распределение с параметром λ , вычисляются с помощью встроенных функций $\text{dexp}(x, \lambda)$ и $\text{rexp}(x, \lambda)$ соответственно.

На рисунке 2 приведен фрагмент рабочего документа, содержащего графики плотностей вероятностей и функций распределения случайных величин для параметров $\lambda=1$ и $\lambda=2$.

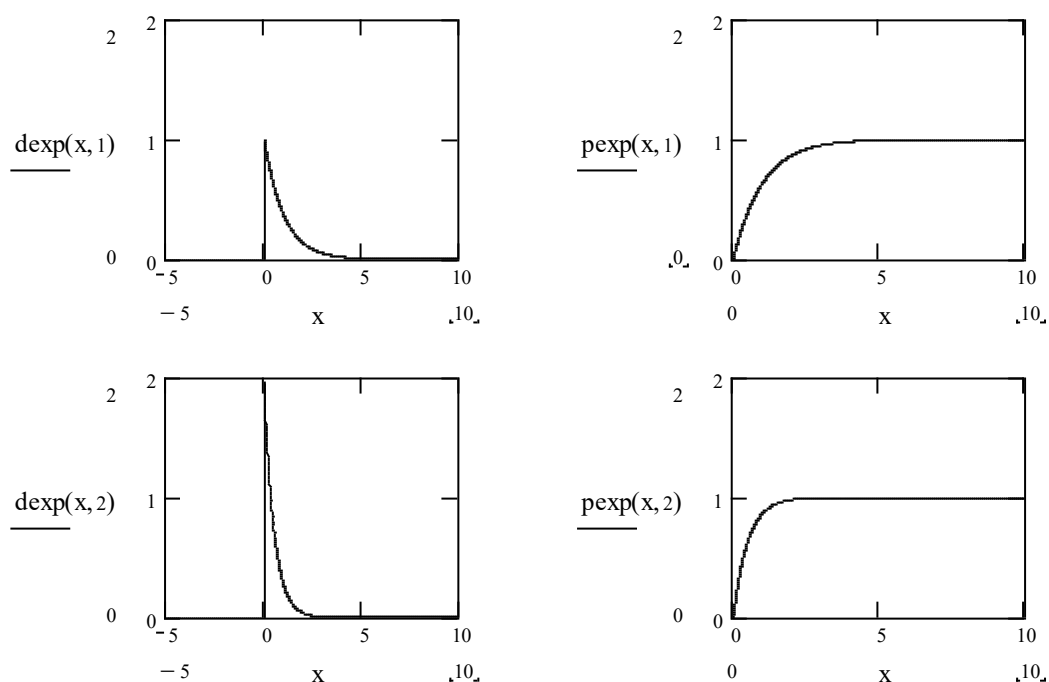


Рисунок 2. Графики плотностей вероятностей и функций распределения случайной величины ξ , имеющей показательное распределение с параметрами $\lambda=1$ и $\lambda=2$.

3. Нормальное распределение

Это распределение играет исключительно важную роль в теории вероятностей и математической статистике. Случайная величина ξ имеет нормальное распределение с параметрами a и σ , $\sigma > 0$, это записывается в форме $\xi \sim N(a, \sigma)$, если ее плотность распределения p_ξ имеет вид

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Случайная величина ξ имеет стандартное нормальное распределение, если $a=0$ и $\sigma=1$, $\xi \sim N(0, 1)$. Плотность стандартного нормального распределения имеет вид

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

а его функция распределения $F_\xi(x)=\Phi(x)$, где $\Phi(x)$ – функция Лапласа:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{z^2}{2}\right) dz.$$

Функция распределения нормальной величины $\eta \sim N(a, \sigma)$, также выражается через функцию Лапласа

$$F_{\eta}(x) = \Phi\left(\frac{x-a}{\sigma}\right).$$

Возможность выражения плотности распределения и функции распределения с помощью функции Лапласа обусловили широкое распространение нормального распределения при решении различных задач теории вероятности и математической статистики.

В Mathcad значения в точке x плотности распределения и функции распределения нормальной случайной величины с параметрами a и σ , находятся встроенными функциями $dnorm(x, a, \sigma)$ и $pnorm(x, a, \sigma)$.

На рисунке 3 приведены построенные в Mathcad графики плотности вероятностей и функций распределения случайных величин для $\xi \sim N(0, 1)$ и $\eta \sim N(1, 2)$.

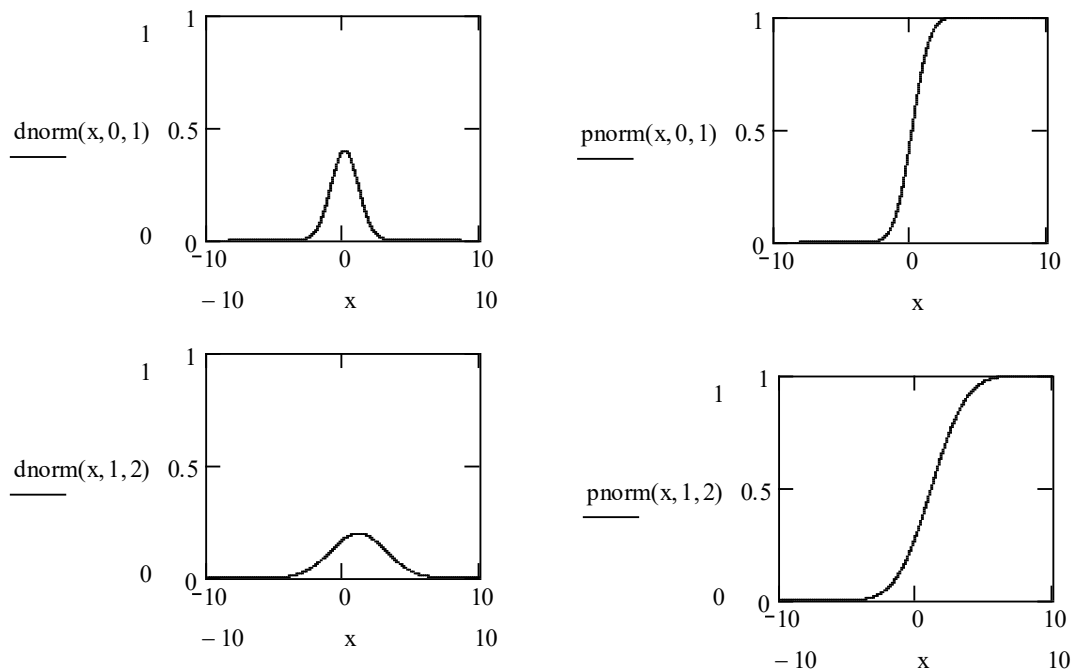


Рисунок 3. Графики плотности вероятностей и функций распределения случайной величины ξ , имеющей нормальное распределение для $\xi \sim N(0, 1)$ и $\eta \sim N(1, 2)$.

4. Распределение хи-квадрат (χ^2 -распределение)

Пусть $\xi_1, \xi_2, \dots, \xi_n$ – независимые случайные величины, каждая из которых имеет стандартное нормальное распределение $N(0, 1)$. Можно составить случайную величину $\chi^2_n = \xi_1^2 + \xi_2^2 + \dots + \xi_n^2$. Ее распределение называется χ^2 -распределением с n степенями свободы. Выражение плотности распределения этой случайной величины определяется следующим образом:

$$p_n(z) = \begin{cases} 0, & z < 0, \\ \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} z^{\frac{n-2}{2}} \exp(-\frac{z}{2}) & z \geq 0, \end{cases}$$

где $\Gamma(x)$ - гамма функция Эйлера $\Gamma(x) = \int_0^{\infty} x^{z-1} \exp(-z) dz$.

В среде Mathcad значения в точке x плотности распределения и функции распределения с n степенями свободы вычисляется встроенными функциями $dchisq(x, n)$ и $pchisq(x, n)$.

На рисунке 4 приведены графики плотности вероятностей и функций распределения для χ^2 -распределения с двумя, четырьмя и восьмью степенями свободы.

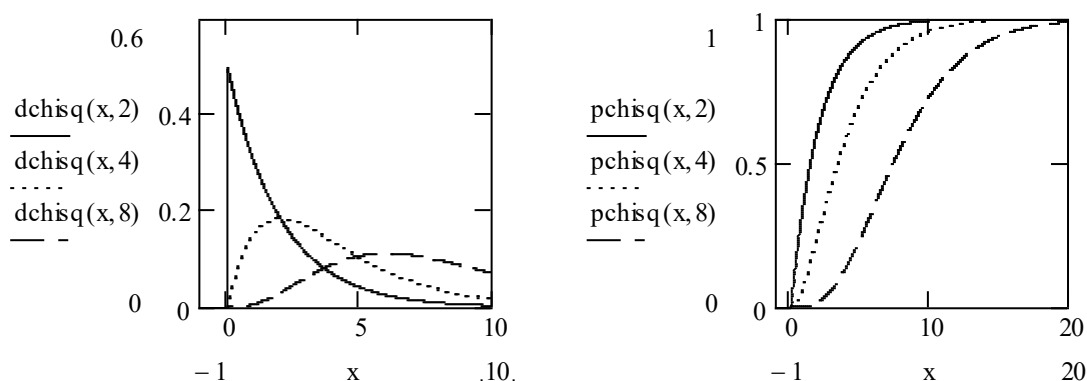


Рисунок 4. Графики плотности вероятностей и функций распределения для χ^2 -распределения с двумя, четырьмя и восьмью степенями свободы.

5. Распределение Стьюдента

Пусть случайная величина ξ имеет стандартное нормальное распределение, а случайная величина χ_n^2 - распределение с n степенями свободы. Если ξ и χ_n^2 независимы, то про случайную величину $\tau_n = \frac{\xi}{\sqrt{\frac{\chi_n^2}{n}}}$ говорят, что она имеет распределение

Стьюдента с числом степеней свободы n . Доказано, что плотность вероятности такой случайной величины вычисляется по формуле

$$p_{\tau_n}(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}$$

При больших значениях n распределение Стьюдента практически не отличается от $N(0,1)$.

В Mathcad значения в точке x плотности распределения и функции Стьюдента с n степенями свободы вычисляются встроенными функциями $dt(x, n)$ и $pt(x, n)$.

На рисунке 5 показан пример использования этих функций для построения соответствующих графиков.

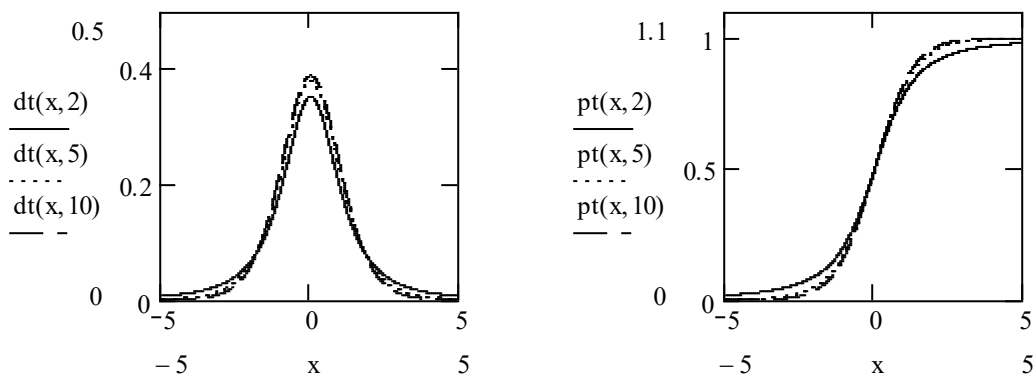


Рисунок 5. Графики плотности вероятностей и функций распределения при $n=2,5,10$.

6. Распределение Фишера

Пусть случайные величины χ_n^2 и χ_m^2 независимы и имеют распределение χ^2 с

n и m степенями свободы. Тогда случайная величина $F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$ имеет F-распределение с плотностью вероятности

$$p_F(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{n/2} \cdot \frac{x^{\frac{n-2}{2}}}{\left(1 + \frac{nx}{m}\right)^{\frac{n+m}{2}}}, \quad x > 0.$$

В Mathcad значения в точке x плотности распределения и функции Фишера вычисляются встроенными функциями $dF(x, n, m)$ и $pF(x, n, m)$.

На рисунке 6 приведены построенные в Mathcad графики плотности вероятностей и функций распределения для $n=2,5$ и $m=5,2$.

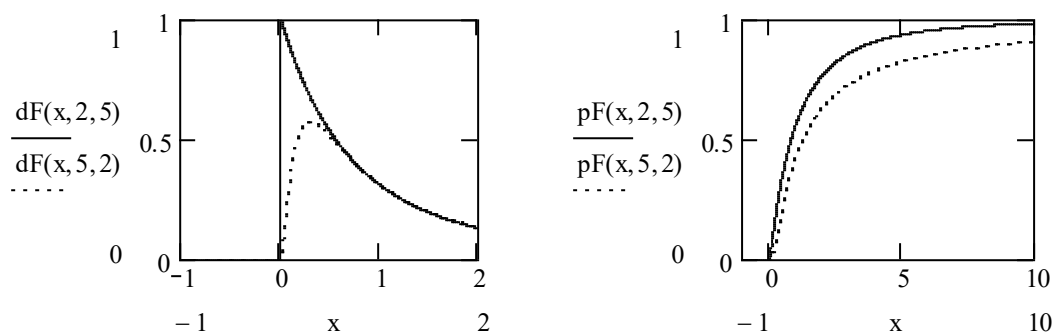


Рисунок 6. Графики плотности вероятностей и функций распределения в случае F-распределения для $n=2,5$ и $m=5,2$

7. Распределение Парето

Это распределение часто используется для экономических исследований плотность вероятностей для случайной величины при этом распределении имеет вид

$$p_{\xi}(x) = \begin{cases} 0, & x < a, \\ \rho a^{\rho} x^{-(\rho+1)}, & x \geq a, \end{cases}$$

где $a > 0, \rho > 0$.

Ниже приведен фрагмент рабочего документа Mathcad, содержащий пример определения и построения графиков (рисунок 7) плотности вероятностей и функций распределения Парето. Поскольку в Mathcad нет встроенных функций для этого распределения, они определяются исходя из соглашения об именах для имеющихся статистических функций, как функции переменной x и параметров a и ρ : $dP(x, \rho, a)$ и $pP(x, \rho, a)$ - соответственно для плотности вероятностей и функции распределения Парето (параметр ρ обозначен как p).

$$dP(x, p, a) := \begin{cases} p \cdot a^p \cdot x^{-(p+1)} & \text{if } x \geq a \\ 0 & \text{if } x < a \end{cases} \quad pP(x, p, a) := \begin{cases} 0 & \text{if } x < a \\ \int_a^x p \cdot a^p \cdot t^{-(p+1)} dt & \text{if } x \geq a \end{cases}$$

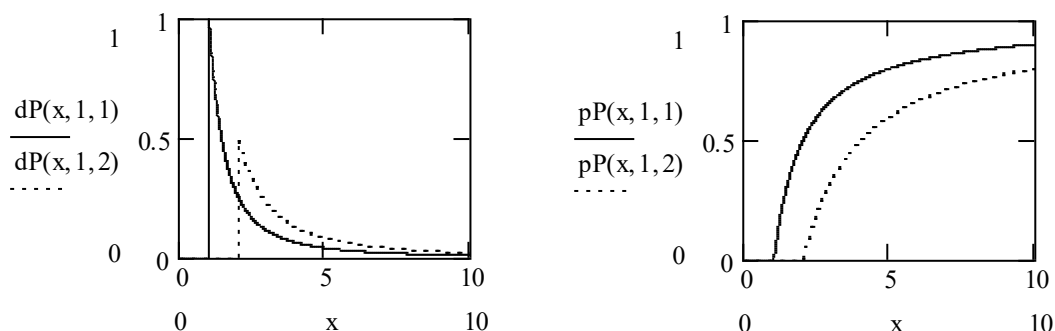


Рисунок 7. Графики плотности вероятностей и функций распределения в случае распределения Парето.

8. Логистическое распределение

Это еще одно распределение, широко применяемое в экономических исследованиях. Оно имеет следующую функцию распределения:

$$F_{\xi}(x) = \frac{1}{1 + \exp(-\frac{x - \alpha}{\beta})}, \quad x \in \mathbb{R},$$

где α и β - параметры распределения.

Плотность распределения вероятностей для логистического распределения вычисляется по формуле:

$$p_{\xi}(x) = \frac{1}{\beta} \cdot \frac{\exp(-\frac{x - \alpha}{\beta})}{1 + \exp(-\frac{x - \alpha}{\beta})^2}, \quad x \in \mathbb{R}.$$

По своим свойствам логистическое распределение очень похоже на нормальное. В среде Mathcad значения в точке x плотности распределения и функции распределения для логистического распределения вычисляются встроенными функциями соответственно $dlogis(x, \alpha, \beta)$ и $plogis(x, \alpha, \beta)$.

Графики плотности вероятностей и функций распределения для $\alpha=0$, $\beta=1, 2$ показаны на рисунке 8.

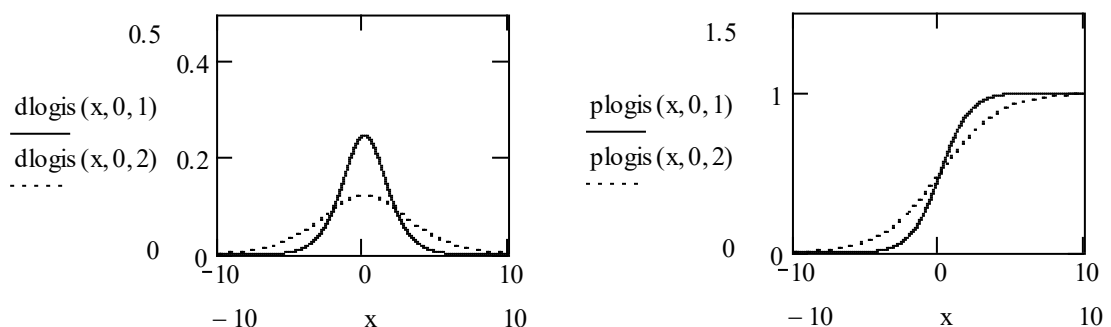


Рисунок 8. Графики плотности вероятностей и функций распределения в случае логистического распределения.

ЗАДАНИЕ

Построить графики плотности распределения и функции распределения для экспоненциального, нормального, хи-квадрат, Стьюдента, Фишера, Парето и

логистического распределений. Найти вероятность того, что значение случайной величины ξ попадет в заданный интервал $[a, b]$. Исследовать поведение $p_{\xi}(x)$ и $F_{\xi}(x)$ при заданном изменении параметров распределений.

ПОРЯДОК ВЫПОЛНЕНИЯ ЗАДАНИЯ

1. Построить графики плотности распределения и функции распределения для показательного распределения с параметром λ , равным N . Здесь N – номер варианта.
2. Построить графики плотности распределения и функции распределения в случае нормального закона распределения случайной величины. Параметры распределения задаются преподавателем.
3. Построить графики плотности распределения и функции распределения χ^2 с указанным числом степеней свободы $k=N$. Здесь N – номер варианта. При построении в тех же осях для сравнения построить графики стандартного нормального распределения.
4. Построить графики плотности распределения и функции распределения Стюдента с указанным числом степеней свободы, равным $k=N$. Здесь N – номер варианта. В тех же осях для сравнения построить графики стандартного нормального распределения.
5. Построить графики плотности распределения и функции распределения Фишера для указанных в Таблице 1 значений n и m .
6. Построить графики плотности распределения и функции распределения Парето для указанных в Таблице 2 значений α и ρ .
7. Построить графики плотности распределения и функции распределения для логистического распределения при значениях параметров $\alpha=\alpha$ и $\beta=\rho$. Значения α и ρ берутся из Таблицы 2.
8. В пп. 1÷7 для указанных преподавателем значений a и b найти вероятность того, что значение случайной величины ξ попадет в заданный интервал $[a, b]$.
9. В пп. 1÷7 построить $p_{\xi}(x)$ и $F_{\xi}(x)$ для 5 различных значений параметров распределений.

Таблица 1.

N	n	m
1	6	4
2	7	4
3	8	4
4	9	4
5	10	4
6	9	5
7	8	5
8	7	5
9	6	3
10	5	3
11	4	2

12	5	2
13	6	3
14	7	4
15	8	4
16	9	5
17	10	5
18	9	5
20	7	5

Таблица 2.

N	α	ρ
1	1	1
2	1.5	1
3	1.6	1
4	1.7	1
5	1.8	1
6	1.9	1
7	2.0	1.5
8	2.1	1.5
9	2.2	1.5
10	2.3	1.5
11	2.1	2
12	2.5	2
13	2.6	2
14	2.7	2
15	2.8	2
16	2.9	2
17	3	2.5
18	3.1	2.5
19	3.2	2.5
20	3.3	2.5

Контрольные вопросы

1. Что называется непрерывной случайной величиной?
2. Чему равна вероятность того, что случайная величина ξ принимает значение меньше x ?
3. При каких условиях случайная величина ξ имеет стандартное нормальное распределение?
4. Когда распределение Стьюдента не отличается от нормального распределения?

Лабораторная работа №5 Проверка параметрических статистических гипотез

Цель: освоить методику проверки параметрических статистических гипотез, научиться применять полученные знания в практических задачах

При изучении различных совокупностей перед исследователем могут возникать различные задачи. Если необходимо знать закон распределения генеральной совокупности и есть основание предполагать, что он имеет определенный вид (A), то выдвигают гипотезу: генеральная совокупность распределена по закону A . Если закон распределения известен, но неизвестен параметр распределения, однако есть основания предполагать, что неизвестный параметр θ равен a , то выдвигают гипотезу: $\theta=a$.

В первом случае речь идет о виде предполагаемого распределения, во втором – о предполагаемой величине параметра известного распределения. Возможны и другие гипотезы, например, о равенстве параметров двух или более распределений.

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 противоречит нулевой.

В итоге статистической проверки гипотезы в двух случаях могут быть приняты неправильные решения, то есть могут быть допущены ошибки двух родов. Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза. Ошибка второго рода состоит в том, что будет принята неправильная гипотеза. Последствия этих ошибок могут быть различными. Вероятность совершить ошибку первого рода принято обозначать через α ; ее называют уровнем значимости. Наиболее часто уровень значимости принимают равным 0,05 или 0,01. Если принять уровень значимости $\alpha=0,05$, то это означает, что в пяти случаях из 100 имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное и приближенное распределение которой известно.

Статистическим критерием (критерием) называют случайную величину K , которая служит для проверки нулевой гипотезы. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных распределений, то в качестве критерия K принимают отношение исправленных выборочных дисперсий: $F = \frac{S_1^2}{S_2^2}$. Эта величина случайная, так как в различных опытах дисперсии принимают различные значения, и распределена по закону Фишера-Снедекора.

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

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

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения, при которых нулевая гипотеза отвергается, а другое – при которых она принимается.

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

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

Критическими точками (границами) $k_{\text{кр}}$ называются точки, отделяющие критическую область от области принятия гипотезы. Правосторонней называют критическую область, определяемую неравенством $K > k_{\text{кр}}$, где $k_{\text{кр}} > 0$. Левосторонней называют критическую область, определенную неравенством $K < k_{\text{кр}}$, где $k_{\text{кр}} < 0$. Двусторонней называют критическую область, определяемую неравенствами $K < k_1$, $K > k_2$, где $k_2 > k_1$.

Правосторонняя критическая область определяется исходя из требования, что при справедливости нулевой гипотезы выполняется соотношение $P(K > k_{\text{кр}}) = \alpha$. Так как α – малая вероятность, то событие « $K > k_{\text{кр}}$ » при справедливости нулевой гипотезы в силу принципа практической невозможности маловероятных событий в единичном испытании не должно произойти. Если же оно произошло, то это можно объяснить тем, что нулевая гипотеза ложна.

Левосторонняя критическая область определяется соотношением $P(K < k_{\text{кр}}) = \alpha$. Двусторонняя критическая область определяется соотношением $P(K < k_1) + P(K > k_2) = \alpha$.

Если нулевая гипотеза принята, то это еще не значит, что она доказана, так как один пример этого не доказывает. Более правильно говорить, что данные наблюдений согласуются с нулевой гипотезой, и, следовательно, не дают оснований ее отвергнуть. Отвергают гипотезу более категорично, чем принимают, поскольку достаточно привести один пример, противоречащий некоторому общему утверждению, чтобы это утверждение отвергнуть.

Проверка статистических гипотез о параметрах распределения

Пусть генеральные совокупности X и Y распределены нормально. Из совокупности X извлечена выборка объемом n_x , а из совокупности Y – объемом n_y . Пусть для этих выборок вычислены исправленные дисперсии S_x^2 и S_y^2 соответственно. Требуется проверить при уровне значимости α нулевую гипотезу о равенстве дисперсий генеральных совокупностей: $H_0: D_x = D_y$.

Для проверки гипотезы о равенстве дисперсий двух нормальных распределений используют случайную величину, распределенную по закону Фишера-Снедекора $F = \frac{S_x^2}{S_y^2}$, где в числителе большая из вычисленных исправленных выборочных дисперсий.

Критическое значение $F_{\text{крит}}$ следует находить с помощью таблицы распределения Фишера-Снедекора (приложение 7) по уровню значимости α и числу степеней свободы $k_1 = n_x - 1$ и $k_2 = n_y - 1$,

где k_1 — число степеней свободы большей (по величине) дисперсии;

k_2 — число степеней свободы меньшей (по величине) дисперсии;

n_x — объем выборки большей (по величине) дисперсии;

n_y — объем выборки меньшей (по величине) дисперсии.

При этом, если в качестве альтернативной гипотезы выступает $H_1: D_x > D_y$, то $F_{\text{крит}} = F(\alpha, k_1, k_2)$, если $H_1: D_x \neq D_y$, то $F_{\text{крит}} = F\left(\frac{\alpha}{2}, k_1, k_2\right)$. Если вычисленное значение $F < F_{\text{крит}}$, то нет оснований отвергать нулевую гипотезу.

Пусть требуется проверить нулевую гипотезу о равенстве математических ожиданий нормально распределенных генеральных совокупностей $H_0: M_x = M_y$.

Здесь возможны случаи:

- 1) дисперсии генеральных совокупностей известны;
- 2) дисперсии генеральных совокупностей неизвестны и выборки имеют объем больший 30;
- 3) дисперсии генеральных совокупностей неизвестны и выборки имеют объем меньший 30.

В первом случае используется нормированная нормально распределенная случайная величина $Z = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x}{n_x} + \frac{D_y}{n_y}}}$. Если альтернативная гипотеза $H_1: M_x > M_y$ или

$H_1: M_x < M_y$, то $Z_{\text{крит}}$ находят по таблице функции Лапласа (приложение 2) из соотношения $\Phi(Z_{\text{крит}}) = \frac{1 - 2\alpha}{2}$. Если $H_1: M_x \neq M_y$, то по той же таблице $Z_{\text{крит}}$ находят из

соотношения $\Phi(Z_{\text{крит}}) = \frac{1 - \alpha}{2}$. При $|Z| < Z_{\text{крит}}$ нет оснований отвергать нулевую гипотезу.

Во втором случае используется приближенно нормально распределенная величина $Z' = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{(D_x)_B}{n_x} + \frac{(D_y)_B}{n_y}}}$, для вычисления которой находят выборочные дис-

персии $(D_x)_B$ и $(D_y)_B$. Вычисление критического значения, в зависимости от вида конкурирующей гипотезы, и правило принятия нулевой гипотезы аналогичны первому случаю.

В третьем случае предварительно требуется выполнить проверку гипотез о равенстве дисперсий генеральных совокупностей. Гипотеза о равенстве математических ожиданий генеральных совокупностей (в случае равенства дисперсий) проверяется с помощью случайной величины имеющей распределение t-

Стьюдента: $T = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}$. Критическое значение $t_{крит}$

следует находить с помощью таблицы распределения Стьюдента (приложение 3) по уровню значимости α и числу степеней свободы $k = n_x + n_y - 2$. Если альтернативная гипотеза $H_1: M_x > M_y$ или $H_1: M_x < M_y$, то α выбирают по нижней строке таблицы приложения 3. Если $H_1: M_x \neq M_y$, то α выбирают по верхней строке таблицы приложения 3. При $|T| < t_{крит}$ нет оснований отвергать нулевую гипотезу.

**Пример проверки статистических гипотез
о параметрах распределения**

Задание. 1. Из двух нормально распределенных генеральных совокупностей извлечены выборки:

x_i : 41 50 35 45 53 30 57 20 44 36 48 55 28

y_i : 40 55 52 38 20 25 47 53 48 54 51 39 49 46 45

Требуется при уровне значимости $\alpha=0,05$ проверить гипотезу о равенстве математических ожиданий генеральных совокупностей, если:

- а) дисперсии генеральных совокупностей $D_x=130$ и $D_y=115$;
- б) дисперсии генеральных совокупностей неизвестны.

В качестве конкурирующей гипотезы принять $H_1: M_x \neq M_y$.

2. По двум независимым выборкам объемов $n_x=100$, $n_y=120$ вычислены выборочные средние $\bar{x}_B=32,4$ и $\bar{y}_B=30,1$ и выборочные дисперсии $(D_x)_B=15$, $(D_y)_B=25$.

Проверить при уровне значимости $\alpha=0,05$ гипотезу о равенстве математических ожиданий генеральных совокупностей при конкурирующей гипотезе $H_1: M_x < M_y$.

Решение. 1. а) Вычислим выборочные средние:

$$\bar{x}_B = \frac{1}{13} \cdot 542 = 41,69;$$

$$\bar{y}_B = \frac{1}{15} \cdot 662 = 44,13.$$

Так как дисперсии генеральных совокупностей известны, наблюдаемое значение критерия найдем по формуле: $Z = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x}{n_x} + \frac{D_y}{n_y}}} = \frac{41,69 - 44,13}{\sqrt{\frac{130}{13} + \frac{115}{15}}} = -0,58$.

Критическое значение найдем из соотношения $\Phi(Z_{\text{крит}}) = \frac{1-0,05}{2} = 0,495$. По таблице (приложение 2) $Z_{\text{крит}} = 2,58$. Так как $|Z| < Z_{\text{крит}}$ нет оснований отвергать нулевую гипотезу.

б) Дисперсии генеральных совокупностей неизвестны и объемы выборок меньше 30. Поэтому проверим сначала гипотезу о равенстве дисперсий генеральных совокупностей, используя критерий Фишера-Снедекора.

Вычисляем исправленные выборочные дисперсии:

$$S_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i - \bar{x}_B)^2 = 128,06 \text{ — исправленная дисперсия первой выборки;}$$

$$S_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_i - \bar{y}_B)^2 = 105,98 \text{ — исправленная дисперсия второй выборки.}$$

Так как первая дисперсия больше, ее помещаем в числитель формулы $F = \frac{S_x^2}{S_y^2} = \frac{128,06}{105,98} = 1,21$.

Критическое значение $F_{\text{крит}}$ найдем с помощью таблицы распределения Фишера-Снедекора (приложение 7) по уровню значимости $\alpha=0,05$ и числу степеней свободы $k_1 = n_x - 1 = 13 - 1 = 12$ и $k_2 = n_y - 1 = 15 - 1 = 14$. $F_{\text{крит}} = F(\alpha, k_1, k_2) = F(0,05; 12; 14) = 2,53$.

Вычисленное значение критерия меньше критического, следовательно, гипотезы генеральных совокупностей можно считать равными.

Далее сравним средние:

$$T = \frac{41,69 - 44,13}{\sqrt{(13-1) \cdot 128,06 + (15-1) \cdot 105,98}} \cdot \sqrt{\frac{13 \cdot 15(13+15-2)}{13+15}} = -0,597$$

Критическое значение $t_{\text{крит}}$ находим с помощью таблицы приложения 3, выбирая α по верхней строке ($H_1: M_x \neq M_y$).

Число степеней свободы $k=n_1+n_2-2=13+15-2=26$. $t_{крит}=t(0,05;26)=2,06$.

Наблюдаемое значение по модулю меньше критического, поэтому нет оснований отвергать нулевую гипотезу.

2. Воспользуемся приближенно нормально распределенной величиной

$$Z' = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{(D_x)_B}{n_x} + \frac{(D_y)_B}{n_y}}} = \frac{32,4 - 30,1}{\sqrt{\frac{15}{100} + \frac{25}{120}}} = 3,59. \text{ Так как альтернативная гипотеза } H_1: M_x < M_y,$$

то $Z_{крит}$ найдем из соотношения $\Phi(Z_{крит}) = \frac{1-2\alpha}{2} = \frac{1-2 \cdot 0,05}{2} = 0,495$. По таблице функции Лапласа (приложение 2) $Z_{крит}=2,56$. Наблюдаемое значение критерия больше критического, следовательно, нулевую гипотезу отвергаем в пользу конкурирующей. Иными словами, средние генеральных совокупностей различаются значимо.

Задание №1.

1. Из двух нормально распределенных генеральных совокупностей извлечены выборки.

Требуется при уровне значимости $\alpha=0,05$ проверить гипотезу о равенстве математических ожиданий генеральных совокупностей, если:

- а) дисперсии генеральных совокупностей D_x и D_y ;
- б) дисперсии генеральных совокупностей неизвестны.

В качестве конкурирующей гипотезы принять $H_1: M_x \neq M_y$.

2. По двум независимым выборкам объемов n_x, n_y вычислить выборочные средние и выборочные дисперсии.

Проверить при уровне значимости $\alpha=0,05$ гипотезу о равенстве математических ожиданий генеральных совокупностей при конкурирующей гипотезе $H_1: M_x < M_y$.

Задание №1.1

Вариант 1

x_i :	17	18	19	20	20		$D_x=2$
y_i :	18	18	19	20	20	21	$D_y=1$

Вариант 2							
x_i	0,52	0,53	0,54	0,55	0,55		$D_x=0,00020$
y_i	0,53	0,53	0,54	0,55	0,55	0,56	$D_y=0,0001$
Вариант 3							
x_i	0,12	0,13	0,34	0,55	0,71		$D_x=0,07$
y_i	0,1	0,25	0,54	0,56	0,56	0,69	$D_y=0,05$
Вариант 4							
x_i	23	35	41	52	63		$D_x=230$
y_i	32	35	45	56	67	70	$D_y=260$
Вариант 5							
x_i	5,6	5,8	5,9	6,1	7,3		$D_x=0,4$
y_i	5,1	5,3	5,5	5,9	5,9	7	$D_y=0,5$
Вариант 6							
x_i	2,3	3,5	4,1	5,2	6,3		$D_x=2,3$
y_i	3,2	3,5	4,5	5,6	6,7	7,1	$D_y=2,5$
Вариант 7							
x_i	12	13	34	55	60		$D_x=500$
y_i	10	25	54	56	56	73	$D_y=550$
Вариант 8							
x_i	1,52	1,53	1,53	1,55	1,55		$D_x=0,00020$
y_i	1,52	1,53	1,54	1,55	1,55	1,55	$D_y=0,0001$

Задание №1.2

Вариант 1

x_i	8	11	14	17	20	23	26	29	32
n_{xi}	5	26	27	27	28	20	14	8	5
y_i	8	11	14	17	20	23	26	29	32
n_{yi}	6	14	17	22	25	20	18	10	8

Вариант 2									
x_i	3	5	7	9	11	13	15	17	19
n_{ii}	11	15	22	26	28	25	17	10	6
y_i	3	5	7	9	11	13	15	17	19
n_{ji}	6	10	12	13	15	14	10	6	4

Вариант 3									
x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,6
n_{ii}	10	12	16	18	24	20	10	6	4
y_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,6
n_{ji}	6	9	10	12	13	10	8	7	5

Вариант 4									
x_i	1	3	5	7	9	11	13	15	17
n_{ii}	10	12	16	18	24	20	10	6	4
y_i	1	3	5	7	9	11	13	15	17
n_{ji}	6	7	10	13	14	12	8	6	4

Вариант 5									
x_i	4	6	8	10	12	14	16	18	
n_{ii}	6	9	10	12	13	10	8	7	5
y_i	4	6	8	10	12	14	16	18	20
n_{ji}	5	9	14	16	18	15	10	9	4

Вариант 6									
x_i	5	8	11	14	17	20	23	26	
n_{ii}	6	9	10	12	13	10	8	7	5
y_i	5	8	11	14	17	20	23	26	29
n_{ji}	5	9	12	13	14	12	10	6	4

Вариант 7									
x_i	2	5	8	11	14	17	20	23	26
n_{ii}	2	6	10	13	22	20	13	10	4

y_i	2	5	8	11	14	17	20	23	26
n_{yi}	10	11	13	15	16	14	12	11	8

Вариант 8

x_i	1	2	3	4	5	6	7	8	9
n_{xi}	10	11	13	15	16	14	12	11	8
y_i	1	2	3	4	5	6	7	8	9
n_{yi}	2	6	10	13	22	20	13	10	4

Лабораторная работа №6 Проверка непараметрических статистических гипотез

Цель работы:

Освоить методику проверки непараметрических статистических гипотез, а также научиться применять полученные знания для решения задач.

Сравнение наблюдаемой относительной частоты с гипотетической вероятностью появления события

Пусть по достаточно большому числу n независимых испытаний, в каждом из которых вероятность p появления события постоянна, но неизвестна, найдена относительная частота m/n . Требуется при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что неизвестная вероятность p равна гипотетической вероятности p_0 .

Правило 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 при конкурирующей гипотезе $H_1: p \neq p_0$, надо вычислить наблюдаемое значение критерия

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

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

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

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

Правило 2. При конкурирующей гипотезе $H_1: p > p_0$ находят критическую точку правосторонней критической области из равенства

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

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

Правило 3. При конкурирующей гипотезе $H_1: p < p_0$ находят сначала «вспомогательную» критическую точку $u_{\text{кр}}$ по правилу 2, а затем полагают границу левосторонней критической области $u'_{\text{кр}} = -u_{\text{кр}}$. 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 $np_0q_0 > 9$.

586. По 100 независимым испытаниям найдена относительная частота $m/n = 0,14$. При уровне значимости 0,05 требуется проверить нулевую гипотезу $H_0: p = p_0 = 0,20$ при конкурирующей гипотезе $H_1: p \neq 0,20$.

Решение. Найдем наблюдаемое значение критерия, учитывая, что $q_0 = 1 - p_0 = 1 - 0,20 = 0,80$:

$$U_{\text{набл}} = \frac{(m/n - p_0) \cdot \sqrt{n}}{\sqrt{p_0 q_0}} = \frac{(0,14 - 0,20) \cdot \sqrt{100}}{\sqrt{0,20 \cdot 0,80}} = -1,5.$$

По условию, конкурирующая гипотеза имеет вид $p \neq p_0$, поэтому критическая область — двусторонняя. Найдем критическую точку $u_{\text{кр}}$ по равенству

$$\Phi(u_{\text{кр}}) = (1 - \alpha)/2 = (1 - 0,05)/2 = 0,475.$$

По таблице функции Лапласа (см. приложение 2) находим

$$u_{\text{кр}} = 1,96.$$

Так как $|U_{\text{набл}}| < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, наблюдаемая относительная частота 0,14 незначимо отличается от гипотетической вероятности 0,20.

587. Решить задачу 586 при конкурирующей гипотезе $H_1: p < p_0$.

Решение. По условию, конкурирующая гипотеза имеет вид $p < p_0$, поэтому критическая область — левосторонняя. Найдем сначала «вспомогательную» точку — границу правосторонней критической области из равенства (правило 2)

$$\Phi(u_{\text{кр}}) = (1 - 2\alpha)/2 = (1 - 2 \cdot 0,05)/2 = 0,45.$$

По таблице функции Лапласа находим $u_{\text{кр}} = 1,645$. Следовательно, граница левосторонней критической области $u'_{\text{кр}} = -1,645$.

Так как $U_{\text{набл}} > u'_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу (правило 3).

588. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превышает 0,02. Среди случайно отобранных 480 изделий оказалось 12 дефектных. Можно ли принять партию?

Решение. Нулевая гипотеза H_0 имеет вид $p = p_0 = 0,02$. Найдем относительную частоту брака:

$$m/n = 12/480 = 0,025.$$

Примем в качестве конкурирующей гипотезы $H_1: p > 0,02$ и уровень значимости $\alpha = 0,05$.

Найдем наблюдаемое значение критерия:

$$U_{\text{набл}} = \frac{(m/n - p_0) \cdot \sqrt{n}}{\sqrt{p_0 q_0}} = \frac{(0,025 - 0,02) \cdot \sqrt{480}}{\sqrt{0,02 \cdot 0,98}} = 0,78.$$

По условию, конкурирующая гипотеза имеет вид $p > p_0$, поэтому критическая область — правосторонняя. Найдем критическую точку $u_{\text{кр}}$ правосторонней критической области из равенства (правило 2)

$$\Phi(u_{\text{кр}}) = (1 - 2 \cdot 0,05)/2 = 0,45.$$

По таблице функции Лапласа (см. приложение 2) находим

$$u_{\text{кр}} = 1,645.$$

Так как $U_{\text{набл}} < u_{\text{кр}}$ — нет оснований отвергнуть гипотезу о том, что вероятность брака в партии не превышает 0,02. Таким образом, партию можно принять.

Сравнение двух вероятностей биномиальных распределений

Пусть в двух генеральных совокупностях производятся независимые испытания: в результате каждого испытания событие A может появиться в первой совокупности с неизвестной вероятностью p_1 , а во второй — с неизвестной вероятностью p_2 . По выборкам, извлеченным из первой и второй совокупностей, найдены соответственные частоты:

$$w_1(A) = m_1/n_1 \text{ и } w_2(A) = m_2/n_2,$$

где m_1, m_2 — числа появлений события A ; n_1, n_2 — количества испытаний.

В качестве оценок неизвестных вероятностей примем относительные частоты: $p_1 \simeq w_1$ и $p_2 \simeq w_2$. Требуется при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что вероятности p_1 и p_2 равны между собой: $H_0: p_1 = p_2$. Другими словами, требуется установить, значимо или незначимо различаются относительные частоты w_1 и w_2 .

Предполагается, что выборки имеют достаточно большой объем.

Правило 1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: p_1 = p_2 = p$ о равенстве вероятностей появления события в двух генеральных совокупностях (имеющих биномиальные распределения) при конкурирующей гипотезе $H_1: p_1 \neq p_2$,

надо вычислить наблюдаемое значение критерия

$$U_{\text{набл}} = \frac{m_1/n_1 - m_2/n_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2} \left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

и по таблице функции Лапласа найти критическую точку $u_{\text{кр}}$ по равенству $\Phi(u_{\text{кр}}) = (1 - \alpha)/2$. Если $|U_{\text{набл}}| < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Если $|U_{\text{набл}}| > u_{\text{кр}}$ — нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: p_1 > p_2$ находят критическую точку правосторонней критической области по равенству $\Phi(u_{\text{кр}}) = (1 - 2\alpha)/2$. Если $U_{\text{набл}} < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Если $U_{\text{набл}} > u_{\text{кр}}$ — нулевую гипотезу отвергают.

Правило 3. При конкурирующей гипотезе $H_1: p_1 < p_2$ находят критическую точку $u_{\text{кр}}$ по правилу 2, а затем полагают границы левосторонней критической области $u_{\text{кр}}' = -u_{\text{кр}}$. Если $U_{\text{набл}} > -u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Если $U_{\text{набл}} < -u_{\text{кр}}$ — нулевую гипотезу отвергают.

606. За смену отказали 15 элементов устройства 1, состоящего из 800 элементов и 25 элементов устройства 2, состоящего из 1000 элементов. При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0: p_1 = p_2 = p$ о равенстве вероятностей отказа элементов обоих устройств при конкурирующей гипотезе $H_1: p_1 \neq p_2$.

Решение. По условию, конкурирующая гипотеза имеет вид $p_1 \neq p_2$, поэтому критическая область двусторонняя. Найдем наблюдаемое значение критерия:

$$U_{\text{набл}} = \frac{m_1/n_1 - m_2/n_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2} \cdot \left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Подставив $m_1 = 15$, $n_1 = 800$, $m_2 = 25$, $n_2 = 1000$, получим $U_{\text{набл}} = -0,89$.

Найдем критическую точку по равенству

$$\Phi(u_{\text{кр}}) = (1 - \alpha)/2 = (1 - 0,05)/2 = 0,475.$$

По таблице функции Лапласа (см. приложение 2) находим $u_{\text{кр}} = 1,96$. Так как $|U_{\text{набл}}| < u_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу. Другими словами, вероятности отказа элемента обоих устройств различаются незначимо.

Проверка гипотезы о показательном распределении генеральной совокупности

Задано эмпирическое распределение непрерывной случайной величины X в виде последовательности интервалов $x_i - x_{i+1}$ и соответствующих им частот n_i , причем $\sum n_i = n$ (объем выборки). Требуется, используя критерий Пирсона, проверить гипотезу о том, что случайная величина X имеет показательное распределение.

Правило. Для того чтобы при уровне значимости α проверить гипотезу о том, что непрерывная случайная величина распределена по показательному закону, надо:

1. Найти по заданному эмпирическому распределению выборочную среднюю $\bar{x}_B$. Для этого, приняв в качестве «представителя» i -го интервала его середину $x_i^* = (x_i + x_{i+1})/2$, составляют последовательность равноотстоящих вариантов и соответствующих им частот.

2. Принять в качестве оценки параметра λ показательного распределения величину, обратную выборочной средней:

$$\lambda^* = 1/\bar{x}_B.$$

3. Найти вероятности попадания X в частичные интервалы (x_i, x_{i+1}) по формуле

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}.$$

4. Вычислить теоретические частоты:

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

где $n = \sum n_i$ — объем выборки.

5. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $k = s - 2$, где s — число первоначальных интервалов выборки; если же было произведено объединение малочисленных частот, следовательно, и самих интервалов, то s — число интервалов, оставшихся после объединения.

647. Почему при проверке по критерию Пирсона гипотезы о показательном распределении генеральной совокупности число степеней свободы определяется равенством $k = s - 2$, где s — число интервалов выборки?

Решение. При использовании критерия Пирсона число степеней свободы $k = s - 1 - r$, где r — число параметров, оцениваемых по выборке. Показательное распределение определяется одним параметром λ . Так как этот параметр оценивается по выборке, то $r = 1$ и, следовательно, число степеней свободы $k = s - 1 - 1 = s - 2$.

648. В результате испытания 200 элементов на длительность работы получено эмпирическое распределение, приведенное в табл. 31 (в первом столбце указаны интервалы времени в часах, во втором столбце — частоты, т. е. количество элементов, проработавших время в пределах соответствующего интервала).

Таблица 31

$x_i - x_{i+1}$	n_i	$x_i - x_{i+1}$	n_i
0—5	133	15—20	4
5—10	45	20—25	2
10—15	15	25—30	1

Требуется, при уровне значимости 0,05, проверить гипотезу о том, что время работы элементов распределено по показательному закону.

Решение 1. Найдем среднее время работы всех элементов (в качестве среднего времени работы одного элемента примем середину интервала, которому принадлежит элемент):

$$\bar{x}_B = (133 \cdot 2,5 + 45 \cdot 7,5 + 15 \cdot 12,5 + 4 \cdot 17,5 + 2 \cdot 22,5 + 1 \cdot 27,5) / 200 = 1000 / 200 = 5.$$

2. Найдем оценку параметра предполагаемого показательного распределения:

$$\lambda = 1/\bar{x}_B = 1/5 = 0,2.$$

Таким образом, дифференциальная функция предполагаемого показательного распределения имеет вид

$$f(x) = 0,2e^{-0,2x} \quad (x > 0).$$

3. Найдем вероятности попадания X в каждый из интервалов по формуле

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}.$$

Например, для первого интервала

$$P_1 = P(0 < X < 5) = e^{-0,2 \cdot 0} - e^{-0,2 \cdot 5} = 1 - e^{-1} = 1 - 0,3679 = 0,6321.$$

Аналогично вычислим вероятности попадания X в остальные интервалы: $P_2 = 0,2326$; $P_3 = 0,0855$; $P_4 = 0,0315$; $P_5 = 0,0116$;

$$P_6 = 0,0042.$$

4. Найдем теоретические частоты:

$$n'_i = n \cdot P_i = 200 \cdot P_i,$$

где P_i — вероятность попадания X в i -й интервал.

Например, для первого интервала

$$n'_1 = 200 \cdot P_1 = 200 \cdot 0,6321 = 126,42.$$

Аналогично вычислим остальные теоретические частоты:

$$n'_2 = 46,52; \quad n'_3 = 17,10; \quad n'_4 = 6,30; \quad n'_5 = 2,32; \quad n'_6 = 0,84.$$

5. Сравним эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого составим расчетную таблицу 32, при-

чем объединим малочисленные частоты ($4+2+1=7$) и соответствующие им теоретические частоты ($6,30+2,32+0,84=9,46$).

Т а б л и ц а 32

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
1	133	126,42	6,58	43,2964	0,3425
2	45	46,52	-1,52	2,3104	0,0497
3	15	17,10	-2,10	4,4100	0,2579
4	7	9,46	-2,46	6,0516	0,6397
Σ	$n=200$				$\chi^2_{\text{набл}}=1,29$

З а м е ч а н и е. Для упрощения вычислений в случае объединения малочисленных частот целесообразно объединить и сами интервалы, которым принадлежат малочисленные частоты, в один интервал. Так, в рассматриваемой задаче, объединив последние три интервала, получим один интервал (15, 30). В этом случае теоретическая частота

$$n'_4 = n \cdot P(15 < X < 30) = 200 \cdot 0,0473 = 9,46$$

совпадает с суммой теоретических частот (9,46), приведенной в табл. 32.

Из табл. 32 находим: $\chi^2_{\text{набл}}=1,29$. По таблице критических точек распределения χ^2 (см. приложение 5), по уровню значимости $\alpha=0,05$ и числу степеней свободы $k=s-2=4-2=2$ находим критическую точку правосторонней критической области $\chi^2_{\text{кр}}(0,05; 2)=6,0$.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ — нет оснований отвергнуть гипотезу о распределении X по показательному закону. Другими словами, данные наблюдений согласуются с этой гипотезой.

Проверка гипотезы о распределении генеральной совокупности по биномиальному закону

Произведено n опытов. Каждый опыт состоит из N независимых испытаний, в каждом из которых вероятность появления события A одна и та же. Регистрируется число появлений события A в каждом опыте. В итоге получено следующее распределение дискретной случайной величины X —числа появлений события A (n первой строке указано число x_i появлений события A в одном опыте; во второй строке—частота n_i , т. е. число опытов, в которых зарегистрировано x_i появлений события A):

x_i	0	1	2	...	N
n_i	0	1	2	...	n_N

Требуется, используя критерий Пирсона, проверить гипотезу о распределении дискретной случайной величины X по биномиальному закону.

Правило. Для того чтобы при уровне значимости α проверить гипотезу о том, что дискретная случайная величина X (число появлений события A) распределена по биномиальному закону, надо:

1. Найти по формуле Бернулли вероятности P_i появления ровно i событий A в N испытаниях ($i=0, 1, 2, \dots, s$, где s —максимальное число наблюдавшихся появлений события A в одном опыте, т. е. $s \leq N$).

2. Найти теоретические частоты

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

где n —число опытов.

3. Сравнить эмпирические и теоретические частоты по критерию Пирсона, приняв число степеней свободы $k=s-1$ (при этом предполагается, что вероятность p появления события A задана, т. е. не оценивалось по выборке и не производилось объединение малочисленных частот).

Если же вероятность p была оценена по выборке, то $k=s-2$. Если, кроме того, было произведено объединение малочисленных частот, то s —число групп выборки, оставшихся после объединения частот.

652. Произведено $n=100$ опытов. Каждый опыт состоял из $N=10$ испытаний, в каждом из которых вероятность p появления события A равна 0,3. В итоге получено следующее эмпирическое распределение (в первой строке указано число x_i появлений события A в одном опыте; во второй строке—частота n_i , т. е. число опытов, в которых наблюдалось x_i появлений события A):

x_i	0	1	2	3	4	5
n_i	2	10	27	32	23	6

Требуется при уровне значимости 0,05 проверить гипотезу о том, что дискретная случайная величина X (число появлений события A) распределена по биномиальному закону.

Решение. 1. По формуле Бернулли

$$P_i = P_N(i) = C_N^i p^i q^{N-i}$$

найдем вероятность P_i ($i=0, 1, 2, 3, 4, 5$) того, что событие A появится в $N=10$ испытаниях ровно i раз.

Учитывая, что $p=0,3$, $q=1-0,3=0,7$, получим:

$$P_0 = P_{10}(0) = 0,7^{10} = 0,0282; P_1 = P_{10}(1) = 10 \cdot 0,3 \cdot 0,7^9 = 0,1211.$$

Аналогично вычислим: $P_2=0,2335$; $P_3=0,2668$; $P_4=0,2001$; $P_5=0,1029$.

2. Найдем теоретические частоты $n'_i = n \cdot P_i$. Учитывая, что $n=100$, получим: $n'_0=2,82$; $n'_1=12,11$; $n'_2=23,35$; $n'_3=26,68$;

$$n'_4=20,01; n'_5=10,29.$$

3. Сравним эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого составим расчетную табл. 36. Поскольку частота $n_0=2$ малочисленная (меньше пяти), объединим ее с частотой $n_1=10$ и в таблицу запишем $2+10=12$; в качестве теоретической частоты, соответствующей объединенной частоте 12, запишем сумму соответствующих теоретических частот: $n'_0 + n'_1 = 2,82 + 12,11 = 14,93$.

Т а б л и ц а 36

i	n_i	n'	$n_i - n'_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
1	12	14,93	-2,93	8,5849	0,5750
2	27	23,35	3,65	13,3225	0,5706
3	32	26,68	5,32	28,3024	1,0608
4	23	20,01	2,99	8,9401	0,4468
5	6	10,29	-4,29	18,4041	1,7886
Σ	$n=100$				$\chi^2_{\text{набл}} = 4,44$

Из табл. 36 находим $\chi^2_{\text{набл}} = 4,44$.

По таблице критических точек распределения χ^2 по уровню значимости $\alpha=0,05$ и числу степеней свободы $k=5-1=4$ находим критическую точку правосторонней критической области $\chi^2_{\text{кр}}(0,05; 4) = 9,5$.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ — нет оснований отвергнуть гипотезу о биномиальном распределении X .

Проверка гипотезы о равномерном распределении генеральной совокупности

Задано эмпирическое распределение непрерывной случайной величины X в виде последовательности интервалов $x_{i-1} - x_i$ и соответствующих им частот n_i , причем $\sum n_i = n$ (объем выборки). Требуется, используя критерий Пирсона, проверить гипотезу о том, что случайная величина X распределена равномерно.

Правило. Для того чтобы проверить гипотезу о равномерном распределении X , т. е. по закону

$$f(x) = \begin{cases} 1/(b-a) & \text{в интервале } (a, b), \\ 0 & \text{вне интервала } (a, b), \end{cases}$$

надо:

1. Оценить параметры a и b — концы интервала, в котором наблюдались возможные значения X , по формулам (через a^* и b^* обозначены оценки параметров):

$$a^* = \bar{x}_n - \sqrt{3} \sigma_n, \quad b^* = \bar{x}_n + \sqrt{3} \sigma_n.$$

2. Найти плотность вероятности предполагаемого распределения

$$f(x) = 1/(b^* - a^*).$$

3. Найти теоретические частоты:

$$n'_1 = nP_1 = n [f(x) \cdot (x_1 - a^*)] = n \cdot \frac{1}{b^* - a^*} (x_1 - a^*);$$

$$n'_2 = n'_3 = \dots = n'_{s-1} = n \cdot \frac{1}{b^* - a^*} (x_i - x_{i-1}), \quad (i = 2, 3, \dots, s-1);$$

$$n'_s = n \cdot \frac{1}{b^* - a^*} (b^* - x_{s-1}).$$

4. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $k = s - 3$, где s — число интервалов, на которые разбита выборка.

656. Почему параметры a и b равномерно распределенной случайной величины X оцениваются по формулам

$$a^* = \bar{x}_n - \sqrt{3} \sigma_n, \quad b^* = \bar{x}_n + \sqrt{3} \sigma_n?$$

Решение. Известно, что в качестве оценок математического ожидания и среднего квадратического отклонения случайной величины X можно принять соответственно выборочную среднюю $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B .

Известно также (см. гл. VI, задачи 313, 315), что для равномерного распределения математическое ожидание и среднее квадратическое отклонение соответственно равны:

$$M(X) = (a+b)/2, \quad \sigma(X) = \sqrt{D(X)} = \sqrt{(b-a)^2/12} = (b-a)/2 \sqrt{3}.$$

Поэтому для оценки параметров равномерного распределения получаем систему двух линейных уравнений

$$\begin{cases} (b^* - a^*)/2 = \bar{x}_B, \\ (b^* - a^*)/2 \sqrt{3} = \sigma_B, \end{cases} \quad \text{или} \quad \begin{cases} b^* + a^* = 2\bar{x}_B, \\ b^* - a^* = 2 \sqrt{3} \sigma_B. \end{cases}$$

Решив эту систему, получим $a^* = \bar{x}_B - \sqrt{3} \sigma_B$, $b^* = \bar{x}_B + \sqrt{3} \sigma_B$.

657. Почему при проверке с помощью критерия Пирсона гипотезы о равномерном распределении генеральной совокупности X число степеней свободы определяется из равенства $k = s - 3$, где s — число интервалов выборки?

Решение. При использовании критерия Пирсона число степеней свободы $k = s - 1 - r$, где r — число параметров, оцениваемых по выборке. Равномерное распределение определяется двумя параметрами a и b . Так как эти два параметра оцениваются по выборке, то $r = 2$ и, следовательно, число степеней свободы $k = s - 1 - 2 = s - 3$.

658. Произведено $n = 200$ испытаний, в результате каждого из которых событие A появлялось в различные моменты времени. В итоге было получено эмпирическое распределение, приведенное в табл. 37 (в первом столбце указаны интервалы времени в минутах, во втором столбце — соответствующие частоты, т. е. число появлений события A в интервале). Требуется при уровне значимости 0,05 проверить гипотезу о том, что время появления событий распределено равномерно.

Таблица 37

Интервал $x_{i-1} - x_i$	Частота n_i	Интервал $x_{i-1} - x_i$	Частота n_i
2—4	21	12—14	14
4—6	16	14—16	21
6—8	15	16—18	22
8—10	26	18—20	18
10—12	22	20—22	25

Решение. 1. Найдем оценки параметров a и b равномерного распределения по формулам:

$$a^* = \bar{x}_B - \sqrt{3} \sigma_B, \quad b^* = \bar{x}_B + \sqrt{3} \sigma_B.$$

Для вычисления выборочной средней $\bar{x}_B$ и выборочного среднего квадратического отклонения σ_B примем середины x_i^* интервалов в качестве вариантов (наблюдаемых значений X). В итоге получим эмпирическое распределение равноотстоящих вариантов:

x_i^*	3	5	7	9	11	13	15	17	19	21
n_i	21	16	15	26	22	14	21	22	18	25

Пользуясь, например, методом произведений, найдем: $\bar{x}_B = 12,31$, $\sigma_B = 5,81$. Следовательно,

$$a^* = 12,31 - 1,73 \cdot 5,81 = 2,26, \quad b^* = 12,31 + 1,73 \cdot 5,81 = 22,36.$$

2. Найдем плотность предполагаемого равномерного распределения:

$$f(x) = 1/(b^* - a^*) = 1/(22,36 - 2,26) = 0,05.$$

3. Найдем теоретические частоты:

$$n'_1 = n \cdot f(x) \cdot (x_1 - a^*) = 200 \cdot 0,05 \cdot (4 - 2,26) = 17,4;$$

$$n'_2 = 200 \cdot 0,05 \cdot (x_2 - x_1) = 10 \cdot (6 - 4) = 20.$$

Длины третьего—девятого интервалов равны длине второго интервала, поэтому теоретические частоты, соответствующие этим интервалам и теоретическая частота второго интервала одинаковы, т. е.

$$n'_3 = n'_4 = n'_5 = n'_6 = n'_7 = n'_8 = n'_9 = 20;$$

$$n'_{10} = 200 \cdot 0,05 \cdot (b^* - x_9) = 10 \cdot (22,36 - 20) = 23,6.$$

4. Сравним эмпирические и теоретические частоты, используя критерий Пирсона, приняв число степеней свободы $k = s - 3 = 10 - 3 = 7$. Для этого составим расчетную табл. 38.

Т а б л и ц а 38

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
1	21	17,3	3,7	13,69	0,79
2	16	20	-4	16	0,80
3	15	20	-5	25	1,25
4	26	20	6	36	1,80
5	22	20	2	4	0,20
6	14	20	-6	36	1,80
7	21	20	1	1	0,05
8	22	20	2	4	0,20
9	18	20	-2	4	0,20
10	25	23,6	1,4	1,96	0,08
					$\chi_{\text{набл}} = 7,17$

Из расчетной таблицы получаем $\chi^2_{\text{набл}} = 7,17$.

Найдем по таблице критических точек распределения χ^2 (см. приложение 5) по уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = s - 3 = 10 - 3 = 7$ критическую точку правосторонней критической области $\chi^2_{\text{кр}}(0,05; 7) = 14,1$.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ — нет оснований отвергнуть гипотезу о равномерном распределении X . Другими словами, данные наблюдений согласуются с этой гипотезой.

Проверка гипотезы о распределении генеральной совокупности по закону Пуассона

Задано эмпирическое распределение дискретной случайной величины X . Требуется, используя критерий Пирсона, проверить гипотезу о распределении генеральной совокупности по закону Пуассона.

Правило. Для того чтобы при уровне значимости α проверить гипотезу о том, что случайная величина X распределена по закону Пуассона, надо:

1. Найти по заданному эмпирическому распределению выборочную среднюю $\bar{x}_n$.

2. Принять в качестве оценки параметра λ распределения Пуассона выборочную среднюю $\lambda = \bar{x}_n$.

3. Найти по формуле Пуассона (или по готовым таблицам) вероятности P_i появления ровно i событий в n испытаниях ($i = 0, 1, 2, \dots, r$, где r — максимальное число наблюдавшихся событий; n — объем выборки).

4. Найти теоретические частоты по формуле $n'_i = n \cdot P_i$.

5. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $k = s - 2$, где s — число различных групп выборки (если производилось объединение малочисленных частот в одну группу, то s — число оставшихся групп выборки после объединения частот).

662. Отдел технического контроля проверил $n = 200$ партий одинаковых изделий и получил следующее эмпи-

рическое распределение (в первой строке указано количество x_i нестандартных изделий в одной партии; во второй строке — частота n_i , т. е. количество партий, содержащих x_i нестандартных изделий);

x_i	0	1	2	3	4
n_i	116	56	22	4	2

Требуется при уровне значимости 0,05 проверить гипотезу о том, что число нестандартных изделий X распределено по закону Пуассона.

Решение. 1. Найдем выборочную среднюю:

$$\bar{x}_v = (\sum n_i x_i) / n = (116 \cdot 0 + 56 \cdot 1 + 22 \cdot 2 + 4 \cdot 3 + 2 \cdot 4) / 200 = 0,6.$$

2. Примем в качестве оценки параметра λ распределения Пуассона выборочную среднюю: $\lambda = 0,6$. Следовательно, предполагаемый закон Пуассона

$$P_n(i) = \lambda^i \cdot e^{-\lambda} / i!$$

имеет вид

$$P_{200}(i) = (0,6)^i \cdot e^{-0,6} / i!$$

3. Положив $i = 0, 1, 2, 3, 4$, найдем вероятности P_i появления i нестандартных изделий в 200 партиях:

$$P_0 = P_{200}(0) = 0,5488; P_1 = P_{200}(1) = 0,3293; P_2 = P_{200}(2) = 0,0988;$$

$$P_3 = P_{200}(3) = 0,0198; P_4 = P_{200}(4) = 0,0030.$$

4. Найдем теоретические частоты по формуле

$$n'_i = n \cdot P_i = 200 P_i.$$

Подставив в эту формулу найденные в п. 3 значения вероятностей P_i , получим $n'_0 = 200 \cdot 0,5488 = 109,76$.

Аналогично найдем: $n'_1 = 65,86$; $n'_2 = 19,76$; $n'_3 = 3,96$; $n'_4 = 0,60$.

5. Сравним эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого составим расчетную табл. 42. Учитывая замечание 1 (см. § 16), объединим малочисленные частоты ($4 + 2 = 6$) и соответствующие им теоретические частоты ($3,96 + 0,60 = 4,56$), результаты объединения частот запишем в табл. 42.

Т а б л и ц а 42

1	2	3	4	5	6
i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
0	116	109,76	6,24	38,9376	0,3548
1	56	65,86	-9,86	97,2196	1,4762
2	22	19,76	2,24	5,0176	0,2539
3	6	4,56	1,44	2,0736	0,4547
Σ	200				$\chi^2_{\text{набл}} = 2,54$

Из расчетной таблицы находим наблюдаемое значение критерия Пирсона: $\chi^2_{\text{набл}} = 2,54$.

По таблице критических точек распределения χ^2 (см. приложение 5), по уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = 4 - 2 = 2$ находим критическую точку правосторонней критической области: $\chi^2_{\text{кр}}(0,05; 2) = 6,0$.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ — нет оснований отвергнуть гипотезу о распределении случайной величины X по закону Пуассона.

Варианты индивидуальных заданий

Задание №1. (588)

Вариант 1.

Партия микросхем принимается, если доля брака не превышает 0,03. При контроле случайно отобранных 500 микросхем обнаружено 20 бракованных. Можно ли принять партию? Проведите проверку гипотезы на уровне значимости 0,05.

Вариант 2.

Партия ткани считается пригодной, если доля дефектных метров не превышает 0,015. Из 600 метров, отобранных случайно, 11 оказались с дефектами. Можно ли принять партию? Проверьте гипотезу при уровне значимости 0,05.

Вариант 3.

Партия таблеток принимается, если вероятность брака не превышает 0,025. В выборке из 400 таблеток обнаружено 14 бракованных. Можно ли утвердить партию? Используйте уровень значимости 0,01.

Вариант 4.

Партия упаковочных коробок принимается, если доля брака не больше 0,04. Из 300 случайно отобранных коробок 15 оказались непригодными. Подлежит ли партия приёмке? Проведите проверку на уровне значимости 0,05.

Вариант 5.

Партия кабелей считается качественной, если доля изделий с нарушением изоляции не превышает 0,02. В выборке из 800 метров кабеля обнаружено 18 дефектных участков. Можно ли принять партию? Проверьте гипотезу при уровне значимости 0,025.

Вариант 6.

Партия лампочек принимается, если доля перегоревших при тестировании не превышает 0,035. Из 700 протестированных лампочек 30 оказались неисправными. Подлежит ли партия приёмке? Проведите проверку на уровне значимости 0,05.

Вариант 7.

Партия батареек считается пригодной, если доля с пониженным зарядом не превышает 0,01. В выборке из 1000 батареек обнаружено 14 с недостаточным зарядом. Можно ли принять партию? Используйте уровень значимости 0,01.

Вариант 8.

Партия деталей для автомобилей принимается, если доля брака не превышает 0,02. При проверке 900 деталей выявлено 21 бракованное изделие. Подлежит ли партия приёмке? Проверьте гипотезу на уровне значимости 0,05.

Задание № 2. (606)

Вариант 1.

За смену отказало 12 элементов в первом устройстве, состоящем из 600 элементов, и 18 элементов во втором устройстве — из 900 элементов. При уровне значимости 0,05 проверьте нулевую гипотезу о том, что вероятности отказа элементов в обоих устройствах одинаковы, против альтернативной гипотезы о их различии.

Вариант 2.

В ходе испытаний за смену вышли из строя 20 элементов из 1000 в первом блоке устройства и 35 элементов из 1400 во втором блоке. Проверьте на уровне значимости 0,05, можно ли считать, что вероятности отказа элементов в блоках одинаковы, или они статистически различаются.

Вариант 3.

На производственной линии А за смену отказали 8 из 400 электронных компонентов, на линии Б — 15 из 750. При уровне значимости 0,01 проверьте гипотезу о равенстве вероятностей отказа компонентов на обеих линиях против альтернативы о неравенстве.

Вариант 4.

Два типа реле тестируются в одинаковых условиях. Из 1200 реле первого типа отказали 18, из 1500 второго типа — 30. Проверьте при уровне значимости 0,05, одинакова ли надёжность (вероятность отказа) у обоих типов реле.

Вариант 5.

В двух партиях микросхем проведено тестирование: из 500 микросхем первой партии неисправными оказались 10, из 800 второй партии — 24. При уровне значимости 0,05 проверьте, отличается ли доля брака в партиях, или различия можно объяснить случайностью.

Вариант 6.

Два сервера в дата-центре обрабатывают запросы. За сутки из 2000 элементов обработки на первом сервере произошло 32 сбоя, на втором — 50 сбоев из 2500. Проверьте при уровне значимости 0,025, одинакова ли вероятность сбоя на серверах, или один из них менее надёжен.

Вариант 7.

При контроле качества двух поставок деталей: из 900 деталей первой поставки 13 оказались бракованными, из 1200 деталей второй — 22. При уровне значимости 0,05 проверьте, можно ли считать, что доля брака в поставках одинакова.

Вариант 8.

Два датчика работают в сложных условиях. За время наблюдения из 700 элементов первого датчика отказали 14, из 1100 элементов второго — 33. Проверьте при уровне значимости 0,05, отличаются ли датчики по надёжности (вероятности отказа элементов), или различия статистически не значимы.

Задание № 3. (648)

Вариант 1

Было испытано 150 элементов на надёжность. Результаты по времени отказов представлены следующим образом: 95 элементов проработали менее 4 часов, 35 — от 4 до 8 часов, 12 — от 8 до 12 часов, 6 — от 12 до 16 часов, и 2 — более 16 часов. При уровне значимости 0,05 проверьте гипотезу о том, что время безотказной работы элементов подчиняется показательному закону распределения.

Вариант 2

При испытании 180 приборов получены следующие данные: 110 приборов вышли из строя в первые 3 часа, 42 — между 3 и 6 часами, 18 — между 6 и 9 часами, 7 — между 9 и 12 часами, и 3 — после 12 часов. Проверьте на

уровне значимости 0,05, можно ли считать, что время до отказа распределено по показательному закону.

Вариант 3

Из 250 испытанных деталей 140 проработали менее 2 часов, 65 — от 2 до 4 часов, 28 — от 4 до 6 часов, 12 — от 6 до 8 часов, и 5 — более 8 часов. При уровне значимости 0,01 проверьте гипотезу о том, что время безотказной работы деталей имеет показательное распределение.

Вариант 4

В ходе испытаний 300 электронных компонентов зафиксировано: 190 отказов произошло в первые 5 часов, 60 — между 5 и 10 часами, 30 — между 10 и 15 часами, 15 — между 15 и 20 часами, и 5 — после 20 часов. Проверьте при уровне значимости 0,05, соответствует ли распределение времени до отказа показательному закону.

Вариант 5

Было протестировано 120 элементов. Из них 75 проработали менее 6 часов, 28 — от 6 до 12 часов, 10 — от 12 до 18 часов, 5 — от 18 до 24 часов, и 2 — более 24 часов. При уровне значимости 0,025 проверьте гипотезу о показательном распределении времени безотказной работы.

Вариант 6

Испытания 220 приборов показали: 130 отказов за первые 4 часа, 50 — между 4 и 8 часами, 25 — между 8 и 12 часами, 10 — между 12 и 16 часами, и 5 — после 16 часов. Проверьте на уровне значимости 0,05, можно ли считать, что время до отказа подчиняется показательному закону.

Вариант 7

При контроле 160 элементов выявлено: 100 проработали менее 3 часов, 38 — от 3 до 6 часов, 14 — от 6 до 9 часов, 6 — от 9 до 12 часов, и 2 — более 12 часов. При уровне значимости 0,05 проверьте гипотезу о том, что распределение времени безотказной работы является показательным.

Вариант 8

В ходе испытаний 240 деталей получено: 155 отказов произошло в первые 2 часа, 50 — между 2 и 4 часами, 20 — между 4 и 6 часами, 10 — между 6 и 8 часами, и 5 — после 8 часов. Проверьте при уровне значимости 0,01, соответствует ли распределение времени до отказа показательному закону.

Задание № 4. (652)

Вариант 1

Было проведено 120 опытов. Каждый опыт включал 8 независимых испытаний, в каждом из которых вероятность наступления события A составляла 0,4. В результате получено следующее распределение: в 3 опытах событие A не появилось ни разу, в 12 опытах — 1 раз, в 25 — 2 раза, в 30 — 3 раза, в 28 — 4 раза, в 15 — 5 раз, в 6 — 6 раз и в 1 опыте — 7 раз. При

уровне значимости 0,05 проверьте гипотезу о том, что число появлений события А в опыте подчиняется биномиальному закону распределения.

Вариант 2

Проведено 150 опытов, каждый из которых состоял из 6 испытаний с вероятностью успеха 0,5. В итоге было зафиксировано: 2 опыта с 0 успехами, 8 — с 1, 20 — с 2, 35 — с 3, 40 — с 4, 30 — с 5 и 15 — с 6 успехами. Проверьте при уровне значимости 0,01, можно ли считать, что число успехов в опыте распределено по биномиальному закону.

Вариант 3

Выполнено 200 серий испытаний. В каждой серии — 5 независимых испытаний, в каждом из которых вероятность события А равна 0,6. Эмпирическое распределение числа появлений события А таково: 0 раз — в 2 опытах, 1 раз — в 8, 2 раза — в 20, 3 раза — в 50, 4 раза — в 70, 5 раз — в 50. При уровне значимости 0,05 проверьте гипотезу о биномиальном распределении числа успехов.

Вариант 4

Проведено 80 опытов, каждый из которых включал 12 испытаний с вероятностью успеха 0,2. В результатах: в 5 опытах успехов не было, в 12 — 1 успех, в 18 — 2, в 20 — 3, в 15 — 4, в 7 — 5, в 3 — 6. При уровне значимости 0,005 проверьте, соответствует ли распределение числа успехов биномиальному закону.

Вариант 5

Было проведено 90 опытов, каждый — из 7 испытаний с вероятностью успеха 0,3. Зафиксировано: 0 успехов — в 5 опытах, 1 — в 10, 2 — в 20, 3 — в 25, 4 — в 18, 5 — в 8, 6 — в 3, 7 — в 1. Проверьте при уровне значимости 0,05 гипотезу о том, что число успехов в опыте распределено по биномиальному закону.

Вариант 6

Выполнено 180 опытов, каждый из которых состоял из 4 испытаний с вероятностью события А, равной 0,7. Распределение числа успехов: 0 раз — в 2 опытах, 1 раз — в 8, 2 раза — в 25, 3 раза — в 60, 4 раза — в 85. При уровне значимости 0,01 проверьте, можно ли считать, что данные подчиняются биномиальному распределению.

Вариант 7

Проведено 160 опытов, каждый — из 9 испытаний с вероятностью успеха 0,25. Результаты: 0 успехов — в 5 опытах, 1 — в 18, 2 — в 30, 3 — в 40, 4 — в 35, 5 — в 20, 6 — в 8, 7 — в 4. При уровне значимости 0,05 проверьте гипотезу о биномиальном характере распределения числа успехов.

Вариант 8

Было проведено 250 опытов, каждый из которых включал 6 испытаний с вероятностью успеха 0,4. В итоге получено: 0 успехов — в 3 опытах, 1 — в 15, 2 — в 40, 3 — в 70, 4 — в 65, 5 — в 40, 6 — в 17. При уровне значимости 0,005 проверьте, соответствует ли наблюдаемое распределение биномиальному закону.

Задание № 5. (658)

Вариант 1

Было проведено 150 наблюдений за появлением событий в течение дня. Время появления фиксировалось в интервалах от 0 до 24 часов. Получены следующие данные: в интервале 0–3 часов событие появилось 18 раз, 3–6 — 12 раз, 6–9 — 14 раз, 9–12 — 16 раз, 12–15 — 22 раза, 15–18 — 20 раз, 18–21 — 25 раз, 21–24 — 23 раза. При уровне значимости 0,05 проверьте гипотезу о том, что события появляются равномерно в течение суток.

Вариант 2

В ходе 180 испытаний фиксировалось время появления сигналов на приёмнике в течение 20-секундного цикла. Данные распределены так: 0–2 с — 22 раза, 2–4 с — 18 раз, 4–6 с — 15 раз, 6–8 с — 20 раз, 8–10 с — 17 раз, 10–12 с — 19 раз, 12–14 с — 21 раз, 14–16 с — 16 раз, 16–18 с — 14 раз, 18–20 с — 18 раз. Проверьте при уровне значимости 0,05, можно ли считать, что сигналы поступают равномерно во времени.

Вариант 3

Проведено 250 наблюдений за вызовами в службу поддержки в течение 8-часового рабочего дня. Частоты появления вызовов: 8–9 — 28 раз, 9–10 —

32 раза, 10–11 — 25 раз, 11–12 — 30 раз, 13–14 — 35 раз, 14–15 — 33 раза, 15–16 — 31 раз, 16–17 — 26 раз, 17–18 — 30 раз. (Данные за обеденный перерыв 12–13 не учитывались.) При уровне значимости 0,01 проверьте гипотезу о равномерном распределении вызовов в рабочее время.

Вариант 4

При моделировании работы сервера зафиксировано 300 запросов в течение 60-минутного цикла. Частоты по 5-минутным интервалам: 0–5 — 24, 5–10 — 28, 10–15 — 22, 15–20 — 30, 20–25 — 26, 25–30 — 29, 30–35 — 27, 35–40 — 25, 40–45 — 31, 45–50 — 28, 50–55 — 24, 55–60 — 26. Проверьте при уровне значимости 0,05, распределены ли запросы равномерно по времени.

Вариант 5

В ходе 120 испытаний измерялось время срабатывания автоматической системы. Все события зафиксированы в интервале от 1 до 7 секунд. Данные: 1–2 — 18 раз, 2–3 — 15 раз, 3–4 — 20 раз, 4–5 — 22 раза, 5–6 — 24 раза, 6–7 — 21 раз. При уровне значимости 0,025 проверьте гипотезу о том, что момент срабатывания распределён равномерно на этом интервале.

Вариант 6

Проведено 220 наблюдений за поступлением клиентов в кафе в течение дня. Интервалы и частоты: 10–11 — 19, 11–12 — 22, 12–13 — 26, 13–14 — 24, 14–15 — 20, 15–16 — 23, 16–17 — 27, 17–18 — 29, 18–19 — 21, 19–20 — 19. (Кафе работает с 10:00 до 20:00.) Проверьте при уровне значимости 0,05, можно ли считать, что клиенты приходят равномерно в течение рабочего дня.

Вариант 7

Было зафиксировано 160 сбоев в работе системы за 32-часовой период мониторинга. Данные по 4-часовым интервалам: 0–4 — 18, 4–8 — 22, 8–12 — 16, 12–16 — 20, 16–20 — 24, 20–24 — 22, 24–28 — 18, 28–32 — 20. При уровне значимости 0,05 проверьте гипотезу о равномерном распределении времени возникновения сбоев.

Вариант 8

В ходе 240 испытаний фиксировалось время появления случайных событий в интервале от 5 до 15 секунд. Получены частоты: 5–6 — 22, 6–7 — 18, 7–8 — 24, 8–9 — 26, 9–10 — 20, 10–11 — 23, 11–12 — 25, 12–13 — 19, 13–14 — 21, 14–15 — 22. Проверьте при уровне значимости 0,01, можно ли считать, что события появляются равномерно на указанном временном отрезке.

Задание №6. (662)

Вариант 1

Отдел технического контроля проверил 180 партий однотипных деталей и зафиксировал количество бракованных изделий в каждой. Данные следующие: в 105 партиях брак не обнаружен, в 48 — по одной бракованной детали, в 18 — по две, в 7 — по три, и в 2 партиях — по четыре бракованных

изделия. При уровне значимости 0,05 проверьте гипотезу о том, что число бракованных изделий в партии подчиняется распределению Пуассона.

Вариант 2

При анализе 250 партий электронных компонентов было установлено: в 130 партиях не было ни одного дефектного элемента, в 75 — по одному, в 30 — по два, в 10 — по три, в 5 — по четыре. Проверьте на уровне значимости 0,01, можно ли считать, что количество бракованных элементов в партии распределено по закону Пуассона.

Вариант 3

Инспекция качества проверила 300 упаковок с мелкими изделиями. В 160 упаковках дефектов не обнаружено, в 90 — по одной бракованной единице, в 35 — по две, в 10 — по три, в 5 — по четыре. При уровне значимости 0,05 проверьте гипотезу о том, что число бракованных изделий в упаковке распределено по закону Пуассона.

Вариант 4

При контроле 150 партий продукции выявлено: в 85 партиях не было брака, в 40 — по одному бракованному изделию, в 15 — по два, в 7 — по три, в 3 — по четыре. Проверьте при уровне значимости 0,025, соответствует ли распределение числа бракованных изделий в партии закону Пуассона.

Вариант 5

Отдел контроля проанализировал 220 поставок деталей. В 120 поставках брак отсутствовал, в 60 — обнаружен один дефектный элемент, в 25 — два, в 10 — три, в 5 — четыре. При уровне значимости 0,05 проверьте гипотезу о том, что количество бракованных изделий в поставке распределено по закону Пуассона.

Вариант 6

Было исследовано 160 партий приборов. В 90 из них не было ни одного неисправного прибора, в 45 — один, в 16 — два, в 6 — три, в 3 — четыре. Проверьте на уровне значимости 0,05, можно ли считать, что число дефектных приборов в партии подчиняется распределению Пуассона.

Вариант 7

При проверке 190 упаковок продукции обнаружено: в 110 упаковках — без брака, в 50 — по одной бракованной единице, в 20 — по две, в 7 — по три, в 3 — по четыре. При уровне значимости 0,01 проверьте гипотезу о том, что распределение числа бракованных изделий в упаковке является пуассоновским.

Вариант 8

Контроль качества затронул 240 партий однотипных узлов. В 140 партиях брак не выявлен, в 65 — по одному дефектному узлу, в 25 — по два, в 7 — по три, в 3 — по четыре. Проверьте при уровне значимости 0,05, подчиняется ли число дефектных узлов в партии закону Пуассона.

Вопросы:

1. Что называется непараметрической статистической гипотезой? Приведите примеры задач, где они применяются, и чем они отличаются от параметрических гипотез.
2. По каким формулам вычисляется наблюдаемое значение критерия Пирсона ($\chi^2_{\text{набл}}$)? Как интерпретируется величина разности между эмпирическими и теоретическими частотами?
3. Как оцениваются параметры показательного распределения при проверке гипотезы согласия? По какой формуле вычисляются теоретические вероятности попадания в интервал?
4. Как формулируется и проверяется гипотеза о равенстве неизвестной вероятности события её гипотетическому значению? Какой критерий используется и как вычисляется наблюдаемое значение статистики?
5. В каких случаях при проверке гипотез объединяют интервалы с малыми частотами? Как это влияет на число степеней свободы и достоверность проверки?

Лабораторная работа №7 Двухфакторный параметрический дисперсионный анализ

Цель работы: освоить методы использования двухфакторного параметрического анализа.

Ход работы:

1. Задана таблица наблюдений, содержащая значения двух независимых факторов и одну зависимую переменную.
2. Произвести проверку условий применимости: проверить, что распределение является нормальным, проверить гомогенность дисперсий.
3. Определить среднее значение зависимой переменной для каждой комбинации факторов, рассчитать общее среднее значение.
4. Вычислить суммы квадратов отклонений: общую сумму квадратов, сумму квадратов для фактора А, сумму квадратов для фактора В, сумму квадратов взаимодействия, остаточную сумму квадратов.
5. Рассчитать среднеквадратичное отклонение фактора А, среднеквадратичное отклонение фактора В, среднеквадратичное отклонение взаимодействия, среднеквадратичное отклонение ошибки.
6. Рассчитать F-критерии для фактора А, для фактора В, для взаимодействия А и В.
7. Оценить значимость факторов, сравнив рассчитанные значения F-критериев с табличными значениями при уровне значимости $\alpha = 0,05$.
8. Рассчитать меры тесноты связи η^2 для фактора А, для фактора В, для взаимодействия А и В.
9. Сформулировать практические выводы и интерпретацию полученных результатов.

Теоретическая часть

Понятие дисперсионного анализа

Дисперсионный анализ – метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. В отличие t-

критерия (общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента), он позволяет сравнивать средние значения трёх и более групп. Разработан Р. Фишером для анализа результатов экспериментальных исследований. В литературе также встречается обозначение ANOVA (от англ. *ANalysis Of VAriance*).

Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную. Зависимые переменные представлены значениями абсолютных шкал (шкала отношений). Независимые переменные являются номинативными (шкала наименований), то есть отражают групповую принадлежность, и могут иметь два или более значения (типа, градации, или уровня). Примерами независимой переменной X_i с двумя значениями могут служить пол (женский: X_1 , мужской: X_2) или тип экспериментальной группы (контрольная: X_1 , экспериментальная: X_2). Градации, соответствующие независимым выборкам объектов, называются межгрупповыми, а градации, соответствующие зависимым выборкам, — внутригрупповыми.

Классификация методов дисперсионного анализа

По количеству анализируемых признаков:

- Однофакторный (анализ различий групп по одному признаку)
- Многофакторный (анализ различий групп одновременно по двум и более признакам)

По принципам анализа:

- Параметрический (для анализа нормально распределённых признаков в группах)
- Непараметрический (для анализа количественного признака независимо от вида его распределения в группах)

По анализируемым данным:

- Данные, полученные в несвязанных (независимых) выборках (в частности данные однократных наблюдений)

- Данные, полученные в связанных (зависимых) выборках (в частности данные повторных наблюдений)

Понятие параметрического анализа

Параметрический дисперсионный анализ используется, когда нужно сравнить три или более группы по количественному нормально распределённому признаку.

В процедуре параметрического анализа вариаций общая вариация данных рассматривается как сумма двух видов вариаций:

1) Межгрупповая вариация – вариация между средним каждой группы и общим средним значением всей выборки.

2) Внутригрупповая вариация – вариация между каждым объектом исследования группы и средним значением соответствующей группы.

Если межгрупповая вариация оказывается статистически значимо больше внутригрупповой вариации, то можно полагать, что различия между средними значениями групп существуют.

Условия применимости метода:

- Анализируемый признак является количественным
- Анализируемый признак нормально распределён в каждой из групп
- Дисперсии анализируемого признака равны
- Группы определяются качественным признаком (группирующий признак является качественным)

Понятие двухфакторного параметрического анализа

Двухфакторный параметрический дисперсионный анализ — это статистический метод, предназначенный для изучения влияния двух независимых факторов (переменных) на одну зависимую переменную, а также для выявления возможного взаимодействия между этими факторами.

Для корректного применения двухфакторного параметрического дисперсионного анализа должны выполняться следующие параметрические предположения:

- Нормальность распределения данных в каждой группе (проверяется тестом Шапиро-Уилка или Q-Q plot).

- Гомогенность дисперсий (равенство дисперсий групп, проверяется тестом Левена).
- Независимость наблюдений

Критерий Шапиро-Уилка

Критерий Шапиро-Уилка — это статистический тест, используемый для проверки нормальности распределения данных. Он эффективен для небольших выборок (обычно до 50 наблюдений), но может применяться и к большим выборкам.

Критерий сравнивает выборочное распределение с нормальным, проверяя гипотезы:

Нулевая гипотеза H_0 : Данные имеют нормальное распределение.

Альтернативная гипотеза H_1 : Данные не распределены нормально.

Чем ближе значение статистики теста W к 1, тем больше данных соответствует нормальному распределению.

Данный критерий применяется, когда объём выборки $n \geq 3$.

Статистика Шапиро-Уилка вычисляется по формуле:

$$W = \frac{(\sum a_i x_i)^2}{\sum (x_i - \bar{x})^2}$$

где:

x_i — упорядоченные по возрастанию выборочные значения,

a_i — коэффициенты, рассчитанные на основе математического ожидания порядковых статистик нормального распределения. В приложении приведены значения данных коэффициентов в зависимости от n ,

$\bar{x}$ — выборочное среднее.

Если статистика $W > W_{\text{крит}}$, то нет оснований отвергать гипотезу о том, что данные имеют нормальное распределение. Если же $W < W_{\text{крит}}$, то гипотеза отвергается. Таблица $W_{\text{крит}}$ представлена в приложении.

Тест Левена

Тест Левена — это статистический тест, используемый для проверки однородности (гомогенности) дисперсий в разных группах данных. Он является альтернативой тесту Бартлетта и менее чувствителен к

отклонениям от нормальности распределения.

Основная идея теста:

Нулевая гипотеза H_0 : Дисперсии во всех группах равны.

Альтернативная гипотеза H_1 : Дисперсии хотя бы в двух группах различаются.

Тест Левена основан на ANOVA абсолютных отклонениях от центральной тенденции (медианы или среднего):

$$L = \frac{N - k}{k - 1} * \frac{\sum n_i (\bar{Z}_i - \bar{Z})^2}{\sum \sum (Z_{ij} - \bar{Z}_i)^2}$$

где:

N – общее количество наблюдений,

k – количество групп,

n_i – количество наблюдений в группе i ,

$Z_{ij} = |x_{ij} - \tilde{x}_i|$ – отклонение от медианы/среднего группы i ,

$\tilde{x}_i$ – медиана или среднее группы i

Вычисленная статистика L сравнивается с критическим значением F -распределения. Если $L \leq F_{\text{крит}}$, то нет оснований отвергать гипотезу о том, что дисперсии во всех группах равны. Если $L > F_{\text{крит}}$, то гипотеза о равенности дисперсий во всех группах отвергается. Таблица значений $F_{\text{крит}}$ представлена в приложении.

Оценка силы связи

Эта-квадрат (η^2) — это стандартизированный показатель силы связи между независимой переменной (фактором) и зависимой переменной в ANOVA. Он показывает, какая доля общей изменчивости данных объясняется групповыми различиями.

Эта-квадрат вычисляется как отношение межгрупповой изменчивости к общей изменчивости:

$$\eta^2 = \frac{SS_{\text{между}}}{SS_{\text{общая}}}$$

где:

$SS_{\text{между}}$ – сумма квадратов между группами (объяснённая вариация),

$SS_{\text{общая}}$ – общая сумма квадратов (полная вариация данных).

Диапазон значений: $0 \leq \eta^2 \leq 1$

0 – отсутствие связи,

1 – полная зависимость (все различия объясняются факторами).

Интерпретация значений η^2 :

0,01 и ниже – очень слабая связь,

0,01–0,05 – слабая связь,

0,06–0,13 – умеренная связь,

0,14 и выше – сильная связь.

Пример выполнения работы

Задача: выяснить, как тип освещения (фактор А) и температура (фактор В) влияют на скорость роста растений. Считать, что наблюдения независимы.

Фактор А: А1 – LED освещение, А2 – солнечный свет.

Фактор В: В1 – 20°С, В2 – 25°С.

Зависимая переменная Y: рост растений (в см).

1. Задана таблица с исходными данными

Таблица 1 – Исходные данные

№	Освещение (А)	Температура (В)	Рост растения (см)
1	А1	В1	12
2	А1	В1	12,3
3	А1	В1	12,7
4	А1	В1	13
5	А1	В2	15
6	А1	В2	15,3
7	А1	В2	15,6
8	А1	В2	15,9
9	А2	В1	10,2
10	А2	В1	10,5
11	А2	В1	10,8
12	А2	В1	11,1
13	А2	В2	13,2
14	А2	В2	13,5

15	A2	B2	13,8
16	A2	B2	14,1

2. Проверим, что распределение является нормальным. Для этого воспользуемся тестом Шапиро-Уилка:

$$W = \frac{(\sum a_i y_i)^2}{\sum (y_i - \bar{y})^2}$$

Рассчитаем W-критерий для всех групп и сравним с $W_{\text{крит}}$.

Для всех групп $n = 4$. При уровне значимости $\alpha = 0,05$ критическое значение $W_{\text{крит}} = 0,767$ (критические значения представлены в таблице П.2, которая находится в приложении). Важно, чтобы данные были упорядочены. В нашем случае данные во всех групп уже упорядочены. Определим критерий W для всех групп.

W-критерий для группы A1B1:

Данные: [12; 12,3; 12,7; 13]

Выборочное среднее: $\bar{y} = \frac{12+12,3+12,7+13}{4} \approx 12,5$

Коэффициенты a_i для $n = 4$: $a_1 = 0,6872$, $a_2 = 0,1677$, $a_3 = -0,1677$, $a_4 = -0,6872$ (значения коэффициентов представлены в таблице П.1, приложения)

$$W = \frac{(0,6872 * 12 + 0,1677 * 12,3 - 0,1677 * 12,7 - 0,6872 * 13)^2}{(12 - 12,5)^2 + (12,3 - 12,5)^2 + (12,7 - 12,5)^2 + (13 - 12,5)^2} \approx 0,981$$

Так как $W = 0,981 > W_{\text{крит}} = 0,767$, то нет оснований отвергать гипотезу о том, что распределение для данной группы является нормальным.

W-критерий для группы A1B2:

Данные: [15; 15,3; 15,6; 15,9]

Выборочное среднее: $\bar{y} = \frac{15+15,3+15,6+15,9}{4} \approx 15,45$

Коэффициенты a_i для $n = 4$: $a_1 = 0,6872$, $a_2 = 0,1677$, $a_3 = -0,1677$, $a_4 = -0,6872$ (значения коэффициентов представлены в таблице П.1, приложения)

$$W = \frac{(0,6872 * 15 + 0,1677 * 15,3 - 0,1677 * 15,6 - 0,6872 * 15,9)^2}{(15 - 15,45)^2 + (15,3 - 15,45)^2 + (15,6 - 15,45)^2 + (15,9 - 15,45)^2} \approx 0,994$$

Так как $W = 0,994 > W_{\text{крит}} = 0,767$, то нет оснований отвергать гипотезу о том, что распределение для данной группы является нормальным.

W-критерий для группы A2B1:

Данные: [10,2; 10,5; 10,8; 11,1]

Выборочное среднее: $\bar{y} = \frac{10,2+10,5+10,8+11,1}{4} \approx 10,65$

Коэффициенты a_i для $n = 4$: $a_1 = 0,6872$, $a_2 = 0,1677$, $a_3 = -0,1677$, $a_4 = -0,6872$ (значения коэффициентов представлены в таблице П.1, которая находится в приложении)

$$W = \frac{(0,6872 * 10,2 + 0,1677 * 10,5 - 0,1677 * 10,8 - 0,6872 * 11,1)^2}{(10,2 - 10,65)^2 + (10,5 - 10,65)^2 + (10,8 - 10,65)^2 + (11,1 - 10,65)^2} \approx 0,994$$

Так как $W = 0,994 > W_{\text{крит}} = 0,767$, то нет оснований отвергать гипотезу о том, что распределение для данной группы является нормальным.

W-критерий для группы A2B2:

Данные: [13,2; 13,5; 13,8; 14,1]

Выборочное среднее: $\bar{y} = \frac{13,2+13,5+13,8+14,1}{4} \approx 13,65$

Коэффициенты a_i для $n = 4$: $a_1 = 0,6872$, $a_2 = 0,1677$, $a_3 = -0,1677$, $a_4 = -0,6872$ (значения коэффициентов представлены в таблице П.1, которая находится в приложении)

$$W = \frac{(0,6872 * 13,2 + 0,1677 * 13,5 - 0,1677 * 13,8 - 0,6872 * 14,1)^2}{(13,2 - 13,65)^2 + (13,5 - 13,65)^2 + (13,8 - 13,65)^2 + (14,1 - 13,65)^2} \approx 0,994$$

Так как $W = 0,994 > W_{\text{крит}} = 0,767$, то нет оснований отвергать гипотезу о том, что распределение для данной группы является нормальным.

С использованием теста Левена проверим гомогенность дисперсий:

$$L = \frac{N - k}{k - 1} * \frac{\sum n_i (\bar{Z}_i - \bar{Z})^2}{\sum \sum (Z_{ij} - \bar{Z}_i)^2}$$

Вычислим медианы для каждой группы:

$$\text{Группа A1B1: } M_e = \frac{12,3+12,7}{2} = 12,5$$

$$\text{Группа A1B2: } M_e = \frac{15,3+15,6}{2} = 15,45$$

$$\text{Группа A2B1: } M_e = \frac{10,5+10,8}{2} = 10,65$$

$$\text{Группа A2B2: } M_e = \frac{13,5+13,8}{2} = 13,65$$

Рассчитываем абсолютные отклонения от медиан:

Группа A1B1:

$$|y_1 - M_e| = |12 - 12,5| = 0,5$$

$$|y_2 - M_e| = |12,3 - 12,5| = 0,2$$

$$|y_3 - M_e| = |12,7 - 12,5| = 0,2$$

$$|y_4 - M_e| = |13 - 12,5| = 0,5$$

Группа A1B2:

$$|y_1 - M_e| = |15 - 15,45| = 0,45$$

$$|y_2 - M_e| = |15,3 - 15,45| = 0,15$$

$$|y_3 - M_e| = |15,6 - 15,45| = 0,15$$

$$|y_4 - M_e| = |15,9 - 15,45| = 0,45$$

Группа A2B1:

$$|y_1 - M_e| = |10,2 - 10,65| = 0,45$$

$$|y_2 - M_e| = |10,5 - 10,65| = 0,15$$

$$|y_3 - M_e| = |10,8 - 10,65| = 0,15$$

$$|y_4 - M_e| = |11,1 - 10,65| = 0,45$$

Группа A2B2:

$$|y_1 - M_e| = |13,2 - 13,65| = 0,45$$

$$|y_2 - M_e| = |13,5 - 13,65| = 0,15$$

$$|y_3 - M_e| = |13,8 - 13,65| = 0,15$$

$$|y_4 - M_e| = |14,1 - 13,65| = 0,45$$

Вычисляем средние отклонений по группам $\bar{Z}_i$

$$\text{Группа A1B1: } \bar{Z}_i = \frac{0,5+0,2+0,2+0,5}{4} = 0,35$$

$$\text{Группа A1B2: } \bar{Z}_i = \frac{0,45+0,15+0,15+0,45}{4} = 0,30$$

$$\text{Группа A2B1: } \bar{Z}_i = \frac{0,45+0,15+0,15+0,45}{4} = 0,30$$

Группа A2B2: $\bar{Z}_i = \frac{0,45+0,15+0,15+0,45}{4} = 0,30$

Общее среднее отклонение: $\bar{Z} = \frac{0,35+0,30+0,30+0,30}{4} = 0,3125$

Вычисляем числитель статистики Левена:

$$\begin{aligned} & \sum n_i(\bar{Z}_i - \bar{Z})^2 \\ &= 4 * (0,35 - 0,3125)^2 + 4 * (0,30 - 0,3125)^2 + 4 \\ & * (0,30 - 0,3125)^2 + 4 * (0,30 - 0,3125)^2 = 0,0075 \end{aligned}$$

Вычисляем знаменатель статистики Левена:

Группа A1B1: $(0,5 - 0,35)^2 + (0,2 - 0,35)^2 + (0,2 - 0,35)^2 + (0,5 - 0,35)^2 = 0,09$

Группа A1B2: $(0,45 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 = 0,09$

Группа A2B1: $(0,45 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 = 0,09$

Группа A2B2: $(0,45 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 + (0,15 - 0,30)^2 = 0,09$

$$\sum \sum (Z_{ij} - \bar{Z}_i)^2 = 0,09 + 0,09 + 0,09 + 0,09 = 0,36$$

Вычисляем статистику Левена:

$$L = \frac{N - k}{k - 1} * \frac{\sum n_i(\bar{Z}_i - \bar{Z})^2}{\sum \sum (Z_{ij} - \bar{Z}_i)^2} = \frac{16 - 4}{4 - 1} * \frac{0,0075}{0,36} \approx 0,0833$$

Сравниваем с критическим значением F-распределения (критические значения представлены в таблице П.3, которая находится в приложении):

$$F_{\text{крит}}(3, 12; \alpha = 0,05) \approx 3,49$$

Так как $L = 0,0833 < F_{\text{крит}} = 3,49$, то нет оснований отвергать гипотезу о равенстве (гомогенности) дисперсий.

3. Определим среднее значение зависимой переменной для каждой комбинации факторов, рассчитаем общее среднее значение.

В предыдущем пункте были рассчитаны средние значения для каждой группы, поэтому перейдём к расчёту общего среднего значения.

$$\bar{y}_{\text{общ}} = \frac{12,5 + 15,45 + 10,65 + 13,65}{4} = \frac{52,25}{4} = 13,0625$$

4. Вычислим суммы квадратов отклонений: общую сумму квадратов, сумму квадратов для фактора А, сумму квадратов для фактора В, сумму квадратов взаимодействия, остаточную сумму квадратов.

Вычисляем общую сумму квадратов (SST)

$$SST = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{\text{общ}})^2$$

$$\text{Группа A1B1: } (12 - 13,06)^2 + (12,3 - 13,06)^2 + (12,7 - 13,06)^2 + (13 - 13,06)^2 = 1,8344$$

$$\text{Группа A1B2: } (15 - 13,06)^2 + (15,3 - 13,06)^2 + (15,6 - 13,06)^2 + (15,9 - 13,06)^2 = 23,2984$$

$$\text{Группа A2B1: } (10,2 - 13,06)^2 + (10,5 - 13,06)^2 + (10,8 - 13,06)^2 + (11,1 - 13,06)^2 = 23,6824$$

$$\text{Группа A2B2: } (13,2 - 13,06)^2 + (13,5 - 13,06)^2 + (13,8 - 13,06)^2 + (14,1 - 13,06)^2 = 1,8424$$

$$SST = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{\text{общ}})^2 = 1,8344 + 23,2984 + 23,6824 + 1,8424 \\ = 50,6576$$

Вычисляем сумму квадратов для фактора А (SSA)

$$SSA = n * b * \sum_{i=1}^a (\bar{y}_i - \bar{y}_{\text{общ}})^2$$

где $a = 2$ уровня освещения (A1, A2), $b = 2$ уровня температуры (B1, B2), $n = 4$ повторения

Средние по фактору А:

$$A1: \frac{12,5+15,45}{2} = 13,975$$

$$A2: \frac{10,65+13,65}{2} = 12,15$$

$$SSA = n * b * \sum_{i=1}^a (\bar{y}_i - \bar{y}_{\text{общ}})^2$$

$$= 4 * 2 * [(13,975 - 13,06)^2 + (12,15 - 13,06)^2] = 13,322$$

Вычисляем сумму квадратов для фактора В (SSB)

$$SSB = n * a * \sum_{i=1}^b (\bar{y}_i - \bar{y}_{\text{общ}})^2$$

где $a = 2$ уровня освещения (A1, A2), $b = 2$ уровня температуры (B1, B2), $n = 4$ повторения

Средние по фактору В:

$$B1: \frac{12,5+10,65}{2} = 11,575$$

$$B2: \frac{15,45+13,65}{2} = 14,55$$

$$SSB = n * a * \sum_{i=1}^b (\bar{y}_i - \bar{y}_{\text{общ}})^2$$

$$= 4 * 2 * [(11,575 - 13,06)^2 + (14,55 - 13,06)^2] = 35,406$$

Вычисляем сумму квадратов взаимодействия ($SSAB$)

$$SSAB = n * \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}_{\text{общ}})^2$$

$$\text{Для группы A1B1: } (12,5 - 13,975 - 11,575 + 13,06)^2 = 0,9801$$

$$\text{Для группы A1B2: } (15,45 - 13,975 - 14,55 + 13,06)^2 = 0,0002$$

$$\text{Для группы A2B1: } (10,65 - 12,15 - 11,575 + 13,06)^2 = 0,0002$$

$$\text{Для группы A2B2: } (13,65 - 12,15 - 14,55 + 13,06)^2 = 0,9801$$

$$SSAB = n * \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}_{\text{общ}})^2$$

$$= 4 * (0,9801 + 0,0002 + 0,0002 + 0,9801) = 7,8424$$

Вычисляем остаточную сумму квадратов (SSE)

SSE вычисляем как сумму квадратов внутри групп.

$$SSE = \sum_{\text{группы}} \sum_{\text{набл.}} (y_{ij} - \bar{y}_{ij})^2$$

$$\text{Для группы A1B1: } (12 - 12,5)^2 + (12,3 - 12,5)^2 + (12,7 - 12,5)^2 + (13 - 12,5)^2 = 0,58$$

$$\text{Для группы A1B2: } (15 - 15,45)^2 + (15,3 - 15,45)^2 + (15,6 - 15,45)^2 + (15,9 - 15,45)^2 = 0,45$$

Для группы A2B1: $(10,2 - 10,65)^2 + (10,5 - 10,65)^2 + (10,8 - 10,65)^2 + (11,1 - 10,65)^2 = 0,45$

Для группы A2B2: $(13,2 - 13,65)^2 + (13,5 - 13,65)^2 + (13,8 - 13,65)^2 + (14,1 - 13,65)^2 = 0,45$

$$SSE = \sum_{\text{группы набл.}} \sum (y_{ij} - \bar{y}_{ij})^2 = 0,58 + 0,45 + 0,45 + 0,45 = 1,93$$

Вычисляем сумму квадратов взаимодействия ($SSAB$)

$$SSAB = SST - SSA - SSB - SSE \\ = 50,6576 - 13,322 - 35,406 - 1,93 = -0,0004 \approx 0$$

Таким образом, взаимодействие практически отсутствует и при дальнейшем расчёте F-критерия будет не значимым.

5. Рассчитаем среднеквадратичное отклонение фактора А, среднеквадратичное отклонение фактора В, среднеквадратичное отклонение взаимодействия, среднеквадратичное отклонение ошибки.

Вычислим степени свободы:

$$df_A = a - 1 = 2 - 1 = 1 \\ df_B = b - 1 = 2 - 1 = 1 \\ df_{AB} = (a - 1)(b - 1) = 1 \\ df_E = N - ab = 16 - 4 = 12 \\ df_T = N - 1 = 16 - 1 = 15$$

Среднеквадратичное отклонение для фактора А:

$$MSA = \frac{SSA}{df_A} = \frac{13,3224}{1} = 13,3224$$

Среднеквадратичное отклонение для фактора В:

$$MSB = \frac{SSB}{df_B} = \frac{35,4056}{1} = 35,4056$$

Среднеквадратичное отклонение для взаимодействия АВ:

$$MSAB = \frac{SSAB}{df_{AB}} = \frac{0,0004}{1} = 0,0004$$

Среднеквадратичное отклонение для ошибки:

$$MSE = \frac{SSE}{df_E} = \frac{1,93}{12} \approx 0,1608$$

6. Рассчитаем F-критерии для фактора А, для фактора В, для взаимодействия А и В.

F-критерий для фактора А:

$$F_A = \frac{MSA}{MSE} = \frac{13,3224}{0,1608} \approx 82,85$$

F-критерий для фактора В:

$$F_B = \frac{MSB}{MSE} = \frac{35,4056}{0,1608} \approx 220,18$$

F-критерий для взаимодействия АВ:

$$F_{AB} = \frac{MSAB}{MSE} = \frac{0,0004}{0,1608} \approx 0,0025$$

7. Оценим значимость факторов, сравнив рассчитанные значения F-критериев с табличными значениями при уровне значимости $\alpha = 0,05$.

Вычислим $F_{\text{крит}} = F_{1-\alpha}(df_1, df_2) = F_{0,95}(1, 12) = 4,75$ (критические значения представлены в таблице П.3, которая находится в приложении)

Фактор А значим, так как $F_A = 82,85 > F_{\text{крит}} = 4,75$.

Фактор В значим, так как $F_B = 220,18 > F_{\text{крит}} = 4,75$.

Взаимодействие факторов АВ не значимо, так как $F_{AB} = 0,0025 < F_{\text{крит}} = 4,75$, о чём мы и сказали в 4 пункте.

8. Рассчитаем меры тесноты связи η^2 для фактора А, для фактора В, для взаимодействия А и В.

Для фактора А:

$$\eta^2_A = \frac{SSA}{SST} = \frac{13,3224}{50,6576} \approx 0,263$$

Для фактора В:

$$\eta^2_B = \frac{SSB}{SST} = \frac{35,4056}{50,6576} \approx 0,699$$

Для взаимодействия факторов АВ:

$$\eta^2_{AB} = \frac{SSAB}{SST} = \frac{0,0004}{50,6576} \approx 0,000008$$

Проверка полноты объяснённой вариации:

$$\eta^2_A + \eta^2_B + \eta^2_{AB} + \eta^2_E = 1$$

$$0,263 + 0,699 + 0,000008 + \frac{1,93}{50,6576} \approx 1$$

9. Сформулируем практическое значение и интерпретацию полученных результатов.

Температура (фактор В) – ключевой фактор, объясняющий 69,9% вариации результатов. Это наиболее сильное и статистически значимое влияние. Повышение температуры с 20°C (фактор В1) до 25°C (фактор В2) даёт значительный прирост растений независимо от типа освещения.

Освещение (фактор А) – оказывает существенное, но менее выраженное влияние, составляющее 26,3% вариации. LED-освещение (фактор А1) стабильно обеспечивает лучший рост по сравнению с солнечным светом (фактор А2).

Взаимодействие факторов практически отсутствует. Это означает, что преимущество LED-освещения сохраняется при любой температуре и эффект повышения температуры одинаково выражен как при LED, так и при солнечном свете.

Варианты заданий к лабораторной работе

Вариант 1. Влияние вида тренировок и питания на снижение веса.

Считать, что наблюдения независимы.

Фактор А: А1 – кардио, А2 – силовая.

Фактор В: В1 – низкоуглеводное, В2 – сбалансированное.

Зависимая переменная Y: Потеря веса (кг за 8 недель).

№	Тренировка (А)	Питание (В)	Потеря веса (кг)
1	А1	В1	4,2
2	А1	В1	4,5
3	А1	В1	4
4	А1	В1	4,3
5	А1	В2	3
6	А1	В2	3,2
7	А1	В2	2,8
8	А1	В2	3,1
9	А2	В1	3,8
10	А2	В1	3,5
11	А2	В1	3,9
12	А2	В1	3,6
13	А2	В2	2,5
14	А2	В2	2,7
15	А2	В2	2,3
16	А2	В2	2,6

Вариант 2. Влияние сорта пшеницы и типа полива на урожайность.

Считать, что наблюдения независимы.

Фактор А: А1 – традиционный, А2 – ГМО.

Фактор В: В1 – капельный, В2 – дождевание.

Зависимая переменная Y: Урожай (ц/га).

№	Сорт (А)	Полив (В)	Урожай (ц/га)
1	A1	B1	42,5
2	A1	B1	43,1
3	A1	B1	41,8
4	A1	B1	42,3
5	A1	B2	38
6	A1	B2	37,5
7	A1	B2	38,2
8	A1	B2	37,8
9	A2	B1	52,0
10	A2	B1	51,5
11	A2	B1	52,3
12	A2	B1	51,8
13	A2	B2	46
14	A2	B2	45,5
15	A2	B2	46,2
16	A2	B2	45,8

Вариант 3. Влияние освещения и влажности на рост микрозелени.

Считать, что наблюдения независимы.

Фактор А: А1 – LED, А2 – естественный свет.

Фактор В: В1 – 60%, В2 – 80%.

Зависимая переменная Y: Высота ростков (см).

№	Освещение (А)	Влажность (В)	Высота ростков (см)
1	A1	B1	10,2
2	A1	B1	10,5
3	A1	B1	9,8
4	A1	B1	10
5	A1	B2	12,3
6	A1	B2	12,6
7	A1	B2	12,1
8	A1	B2	12,4
9	A2	B1	7,5
10	A2	B1	7,8
11	A2	B1	7,3
12	A2	B1	7,6
13	A2	B2	9
14	A2	B2	9,3
15	A2	B2	8,8
16	A2	B2	9,1

Вариант 4. Влияние типа почвы и удобрения на урожай картофеля.
Считать, что наблюдения независимы.

Фактор А: А1 – чернозём, А2 – песчаная.

Фактор В: В1 – органическое, В2 – минеральное.

Зависимая переменная Y: Урожай (кг/м²).

№	Тип почвы (А)	Удобрение (В)	Урожай (кг/м ²)
1	A1	B1	4,8
2	A1	B1	5
3	A1	B1	4,7
4	A1	B1	4,9
5	A1	B2	5,5
6	A1	B2	5,7
7	A1	B2	5,4
8	A1	B2	5,6
9	A2	B1	2,9
10	A2	B1	3,1
11	A2	B1	2,8
12	A2	B1	3
13	A2	B2	3,8
14	A2	B2	4
15	A2	B2	3,7
16	A2	B2	3,9

Вариант 5. Влияние добавок и температуры на прочность бетона.
Считать, что наблюдения независимы.

Фактор А: А1 –пластификатор, А2 – без добавок.

Фактор В: В1 – 20°С, В2 – 30°С.

Зависимая переменная Y: Прочность на сжатие (МПа).

№	Добавка (А)	Температура (В)	Прочность (МПа)
1	A1	B1	42,5
2	A1	B1	43
3	A1	B1	41,8
4	A1	B1	42,2
5	A1	B2	45
6	A1	B2	45,5
7	A1	B2	44,8
8	A1	B2	45,2
9	A2	B1	35
10	A2	B1	35,5
11	A2	B1	34,8
12	A2	B1	35,2
13	A2	B2	38
14	A2	B2	38,5
15	A2	B2	37,8
16	A2	B2	38,2

Контрольные вопросы

1. Что такое дисперсионный анализ?
2. В чём заключается суть дисперсионного анализа?
3. Что понимается под параметрическим анализом?
4. Что такое двухфакторный параметрический анализ?
5. Что позволяет узнать критерий Шапиро-Уилка?
6. Поясните суть теста Левена.
7. Что такое мера η^2 и как она используется для оценки влияния факторов?

Лабораторная работа №8 Линейные преобразования случайных функций

Цель работы:

Исследовать влияние различных линейных преобразований на случайные функции, изучить изменение их статистических характеристик.

Теоретическая часть

1. Основные понятия

Случайная функция (случайный процесс) — это функция, значения которой в каждый момент времени являются случайными величинами. Для анализа таких процессов используются следующие статистические характеристики:

- Математическое ожидание (среднее значение)

$m_X = E[X_n]$ — характеризует средний уровень сигнала.

- Дисперсия

$\sigma_X^2 = E[(X_n - m_X)^2]$ — мера разброса значений относительно среднего.

- Корреляционная функция

$K_X(k) = E[(X_n - m_X)(X_{n+k} - m_X)]$ — описывает статистическую связь между значениями процесса в разные моменты времени.

Случайный процесс называется стационарным, если его статистические характеристики не зависят от времени. Условия стационарности:

- Математическое ожидание постоянно $m_X(t) = \text{const}$

- Корреляционная функция зависит только от лага k : $K_X(t, t + k) = K_X(k)$

2. Линейные преобразования случайных процессов

Линейное преобразование случайного процесса — это математическая операция, которая изменяет значения процесса согласно линейному закону. Оно выражается формулой:

$$Y_n = a \cdot X_n + b$$

где a — коэффициент масштабирования (определяет усиление / ослабление амплитуды),

b — коэффициент сдвига (смещает сигнал по уровню).

Влияние на статистические характеристики:

- Математическое ожидание:

$$m_Y = a \cdot m_X + b$$

- Дисперсия:

$$\sigma_Y^2 = a^2 \cdot \sigma_X^2$$

- Корреляционная функция:

$$K_Y(k) = a^2 \cdot K_X(k)$$

Масштабирование и сдвиг используются для нормализации данных, приведения сигнала к нужному диапазону или центрированию (например, к нулевому среднему).

Особенности линейного преобразования:

- Линейность сохраняет гауссовость процесса: если X_n гауссовский, то Y_n тоже.
- Преобразование не меняет форму распределения, только его параметры.
- При $a < 0$ происходит инверсия сигнала (отражение относительно оси времени).

Линейное преобразование — это простой, но мощный инструмент для управления свойствами случайных процессов без изменения их фундаментальной структуры.

Виды линейного преобразования:

1) Масштабирование и сдвиг.

Формула: $Y_n = a \cdot X_n + b$.

Назначение: нормализация данных, адаптация сигналов к требуемому диапазону.

2) Скользящее среднее.

Формула: $Y_n = \frac{1}{m} \sum_{i=0}^{m-1} X_{n-i}$

Назначение: сглаживание временных рядов, подавление высокочастотного шума в сигналах

3) Дифференцирование (разностный оператор)

Формула: $Y_n = X_{n+1} - X_n$

Назначение: выделение резких изменений, обнаружение аномалий в данных

4) Накопление (интегрирование)

Формула: $Y_n = \sum_{k=0}^n X_k$

Назначение: восстановление сигнала из его разностей, анализ трендов.

5) Взвешенное скользящее среднее

Формула: $Y_n = \sum_{i=0}^{m-1} w_i X_{n-i}$

Назначение: технический анализ, адаптивная фильтрация сигналов

6) Авторегрессионное преобразование (линейный фильтр)

Формула: $Y_n = \sum_{k=1}^p \alpha_k Y_{n-k} + X_n$

Назначение: прогнозирование временных рядов, анализ речевых и аудиосигналов

7) Линейная комбинация нескольких процессов

Формула: $Y_n = \sum_{k=1}^N a_k X_n^{(k)} + b$

Назначение: синтез сложных сигналов, адаптивные системы управления

3. Фильтрация (скользящее среднее)

Скользящее среднее — это метод сглаживания случайного процесса путём усреднения его значений в заданном окне m . Для процесса X_n скользящее среднее Y_n вычисляется по формуле:

$$Y_n = \frac{1}{m} \sum_{i=0}^{m-1} X_{n-i}$$

Окно (или окно сглаживания) m — это фиксированный интервал данных, который используется для усреднения значений случайного процесса. Окно определяет количество соседних точек (значений процесса), участвующих в вычислении каждого нового сглаженного значения.

Влияние на статистические характеристики:

- Математическое ожидание:

$$m_Y = m_X \text{ (не изменяется).}$$

- Дисперсия:

$$\sigma_Y^2 = \frac{1}{m^2} (m\sigma_X^2 + 2 \sum_{k=1}^{m-1} (m-k) K_X(k))$$

Чем больше окно m , тем сильнее уменьшается дисперсия.

- Корреляционная функция:

$$K_Y(l) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{j=0}^{m-1} K_X(k+l-j)$$

Применение:

Скользящее среднее действует как фильтр низких частот, подавляя высокочастотные колебания и шумы. Используется для сглаживания данных, например, в экономике, обработке сигналов и анализе временных рядов.

4. Дифференцирование (разностный оператор)

Дифференцирование случайного процесса X_n — это линейное преобразование, которое вычисляет разность между соседними значениями процесса. Формула разностного оператора первого порядка:

$$Y_n = X_{n+1} - X_n$$

Влияние на статистические характеристики:

- Математическое ожидание:

$$m_Y = 0$$

- Дисперсия:

$$\sigma_Y^2 = 2(\sigma_X^2 - K_X(1))$$

Если корреляция между соседними значениями отрицательна ($K_X(1) < 0$), дисперсия увеличивается.

- Корреляционная функция:

$$K_Y(l) = 2K_X(l) - K_X(l+1) - K_X(l-1)$$

Применение:

Дифференцирование работает как фильтр высоких частот, усиливая резкие изменения сигнала и подавляя медленные тренды. Используется в техническом анализе, обработке изображений и обнаружении аномалий.

Выбор преобразования зависит от цели обработки: фильтрация шумов, выделение трендов или анализ быстрых изменений.

Контрольные вопросы:

1. Дайте определение случайного процесса. Какие статистические характеристики используются для его описания?
2. Как вычисляются математическое ожидание, дисперсия и корреляционная функция для дискретного процесса?
3. Что такое стационарный случайный процесс? Какие условия он должен удовлетворять?
4. Запишите формулу линейного преобразования случайного процесса. Как влияют коэффициенты a и b на сигнал?
5. Приведите примеры практического применения масштабирования и сдвига в обработке сигналов.
6. Дайте определение скользящего среднего. Как выбор размера окна m влияет на результат?
7. Какие преимущества и недостатки имеет скользящее среднее по сравнению с другими методами сглаживания?
8. Как работает разностный оператор первого порядка? Какую информацию о процессе он позволяет получить?
9. Как изменяются дисперсия и корреляционная функция после дифференцирования?

Содержание работы:

1. Рассчитайте основные статистические характеристики исходного процесса: математическое ожидание, дисперсию, корреляционную функцию.
2. Примените линейное преобразование при помощи масштабирования и сдвига. Вычислите новые статистические характеристики. Сравните их с исходными данными.
3. Примените преобразование скользящим средним. Вычислите новые статистические характеристики. Постройте графики X_n и Y_n . Объясните, почему дисперсия уменьшается. Сделайте выводы.
4. Примените преобразование дифференцированием. Вычислите новые статистические характеристики. Объясните, почему увеличивается доля высокочастотных составляющих. Сделайте выводы.

Порядок выполнения работы:

1. Вычисление основных характеристик дискретного процесса

Рассмотрим дискретный случайный процесс X_n (например, отсчёты шума с гармонической составляющей):

n	1	2	3	4	5	6	7	8	9	10
X_n	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1

Рисунок 1. Дискретный случайный процесс

Рассчитаем его основные математические характеристики:

$$\text{Математическое ожидание: } m_X = \frac{1}{N} \sum_{n=1}^N X_n$$

$$m_X = \frac{0.5 - 0.2 + 1.1 + 0.3 - 0.8 + 0.6 + 1.2 - 0.5 + 0.9 - 1.0}{10} = \frac{2.1}{10} = 0.21$$

$$\text{Дисперсия: } \sigma_X^2 = \frac{1}{N} \sum_{n=1}^N (X_n - m_X)^2$$

Для расчёта составим расчётную таблицу:

n	1	2	3	4	5	6	7	8	9	10
X_n	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1
$X_n - m_X$	0,29	-0,41	0,89	0,09	-1,01	0,39	0,99	-0,71	0,69	-1,21
$(X_n - m_X)^2$	0,0841	0,1681	0,7921	0,0081	1,0201	0,1521	0,9801	0,5041	0,4761	1,4641
										5,649

Рисунок 2. Расчётная таблица

$$\sigma_X^2 = \frac{1}{N} \sum (X_n - m_X)^2 = \frac{0.0841 + 0.1681 + \dots + 1.4641}{10} = 0.5648$$

Корреляционная функция для стационарного процесса с нулевым средним ($m_X = 0$):

$$K_X(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} X_n X_{n+k}$$

В нашем случае $m_X \neq 0$, поэтому используем несмещённую оценку:

$$K_X(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} (X_n - m_X)(X_{n+k} - m_X)$$

Пример расчёта для лага ($k = 1$):

$$K_X(1) = \frac{1}{9} \sum_{n=1}^9 (X_n - 0.21)(X_{n+1} - 0.21)$$

Составим расчётную таблицу:

n	1	2	3	4	5	6	7	8	9	10	
Xn	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1	
Xn-Mx	0,29	-0,41	0,89	0,09	-1,01	0,39	0,99	-0,71	0,69	-1,21	
X(n+1) - Mx	-0,41	0,89	0,09	-1,01	0,39	0,99	-0,71	0,69	-1,21		
Xn-Mx*(X(n+1)-Mx)	-0,1189	-0,3649	0,0801	-0,0909	-0,3939	0,3861	-0,7029	-0,4899	-0,8349		-0,28112

Рисунок 3. Расчётная таблица

Таким образом, $K_X(1) = -0,28112$

2. Линейное преобразование при помощи масштабирования и сдвига

$$Y_n = 2 \cdot X_n + 1$$

Коэффициенты $a = 2$, $b = 1$.

Составим расчётную таблицу:

n	1	2	3	4	5	6	7	8	9	10
Xn	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1
Yn	2	0,6	3,2	1,6	-0,6	2,2	3,4	0	2,8	-1

Рисунок 4. Расчётная таблица

Характеристики нашего процесса изменились соответствующим образом:

- Математическое ожидание $m_Y = am_X = 2m_X + 1 = 2 \cdot 0.21 + 1 = 1,42$

- Дисперсия $\sigma_Y^2 = a^2 \sigma_X^2 = 4\sigma_X^2 = 4 \cdot 0.5648 = 2,2592$

- Корреляционная функция $K_Y(k) = a^2 K_X(k)$

$$K_Y(1) = 4 \cdot K_X(1) = 4 \cdot -0,28112 = -1,12448$$

Как мы видим из формул, при помощи изменения коэффициентов a и b мы можем изменять и характеристики процесса. При помощи изменения коэффициента a мы увеличиваем / уменьшаем наши характеристики в несколько раз, а меняя сдвиг b мы смещаем средний уровень сигнала, меняя математическое ожидание.

Это линейное преобразование мы можем применять для масштабирования сигнала или приведения к необходимым характеристикам (например, нулевому среднему и единичной дисперсии).

3. Линейное преобразование скользящим средним (фильтрацией)

$$Y_n = \frac{X_{n-1} + X_n + X_{n+1}}{3}$$

Окно $m = 3$.

Составим расчётную таблицу:

n	1	2	3	4	5	6	7	8	9	10
Xn	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1
Yn	-	0,466667	0,4	0,2	0,033333	0,333333	0,433333	0,533333	-0,2	-

Рисунок 5. Расчётная таблица

Расчёты для крайних членов рядов не производятся.

Характеристики нашего процесса изменились соответствующим образом:

- Математическое ожидание $m_Y = m_X = 0.21$

- Дисперсия

$$\sigma_Y^2 = \frac{1}{m^2} (m\sigma_X^2 + 2 \sum_{k=1}^{m-1} (m-k) K_X(k))$$

$$\sigma_Y^2 = \frac{1}{9} (3\sigma_X^2 + 2 \sum_{k=1}^{3-1} (3-k) K_X(k))$$

Составляем расчётную таблицу для $K_X(2)$:

n	1	2	3	4	5	6	7	8	9	10
Xn	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1
Xn-Mx	0,29	-0,41	0,89	0,09	-1,01	0,39	0,99	-0,71	0,69	-1,21
X(n+2) - Mx	0,89	0,09	-1,01	0,39	0,99	-0,71	0,69	-1,21		
Xn-Mx*(X(n+2)-Mx)	0,2581	-0,0369	-0,8989	0,0351	-0,9999	-0,2769	0,6831	0,8591		
	-0,04715									

Рисунок 6. Расчётная таблица

$$\sigma_Y^2 = \frac{1}{9} (3 \cdot 0.5648 + 2 \cdot (2 \cdot (-0,28112)) + (-0,04715)) = 0.058$$

- Корреляционная функция

$$K_Y(l) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{j=0}^{m-1} K_X(k+l-j)$$

$$K_Y(1) = \frac{-0.282+0.565-0.282+0-0.282+0.565+0+0-0.282}{9} = -0.0313$$

Построим графики X_n и Y_n

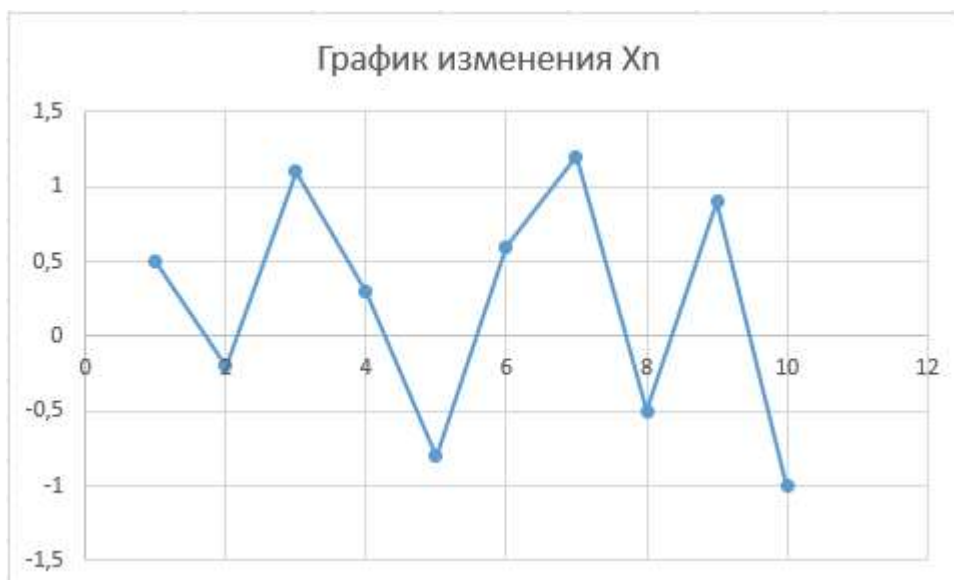


Рисунок 7. График изменения X_n

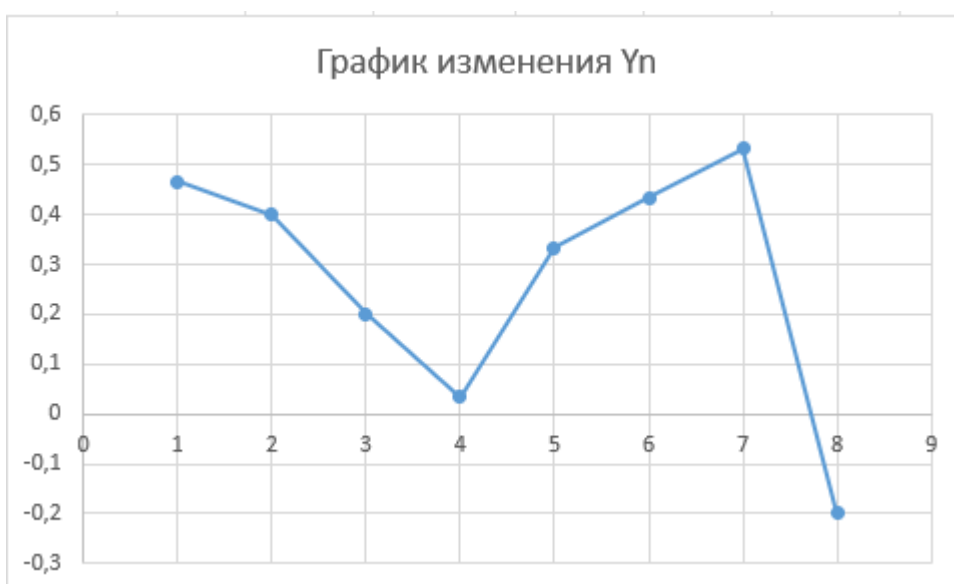


Рисунок 8. График изменения Y_n

Глядя на график X_n , мы видим хаотичные колебания вокруг среднего значения с высокой амплитудой.

После линейного преобразования скользящим средним значения Y_n усредняются. График выглядит более сглаженным, амплитуда колебаний заметно уменьшилась.

Всё это связано с уменьшением дисперсии нашего процесса. Дисперсия до линейного преобразования составляла $\sigma_X^2 = 0.5648$. После линейного преобразования дисперсия вычисляется по формуле $\sigma_Y^2 = \frac{1}{m^2} (m\sigma_X^2 + 2 \sum_{k=1}^{m-1} (m-k) K_X(k))$ и составляет 0.058. Из формулы можно видеть, что чем выше окно m , тем больше уменьшается дисперсия нашего процесса.

Таким образом, скользящее среднее действует как фильтр низких частот, подавляя быстрые колебания.

4. Линейное преобразование дифференцированием (разностный оператор)

$$Y_n = X_{n+1} - X_n$$

Составим расчётную таблицу:

n	1	2	3	4	5	6	7	8	9	10
X _n	0,5	-0,2	1,1	0,3	-0,8	0,6	1,2	-0,5	0,9	-1
Y _n	-0,7	1,3	-0,8	-1,1	1,4	0,6	-1,7	1,4	-1,9	

Рисунок 9. Расчётная таблица

Характеристики нашего процесса изменились соответствующим образом:

- Математическое ожидание $m_Y = 0$
- Дисперсия $\sigma_Y^2 = 2(\sigma_X^2 - K_X(1)) = 2 \cdot (0.5648 - (-0,28112)) = 1,69184$
- Корреляционная функция: $K_Y(l) = 2K_X(l) - K_X(l+1) - K_X(l-1)$

$$K_Y(1) = 2 \cdot (-0,28112) - (-0,04715) - 0.5648 = -1.08$$

Построим графики изменения X_n и Y_n

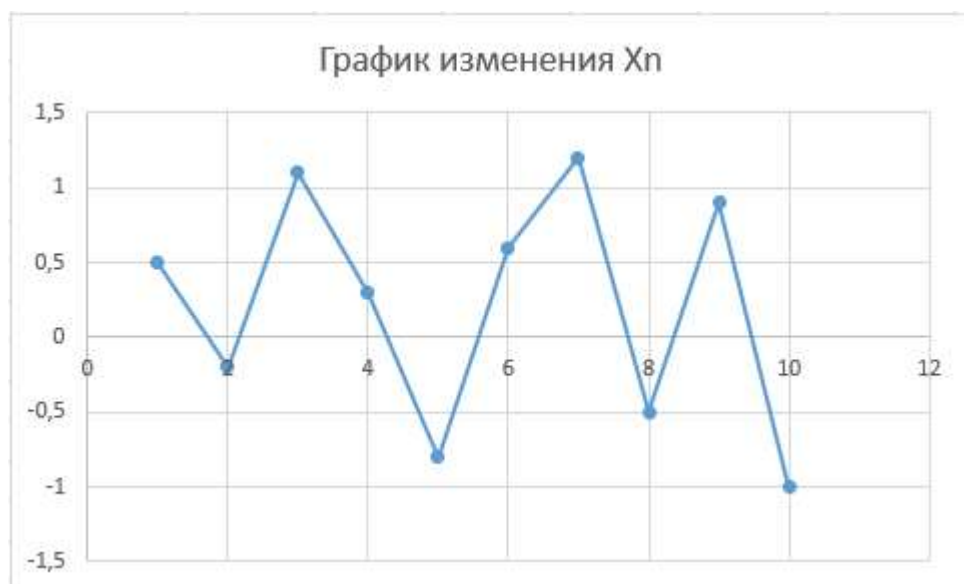


Рисунок 10. График изменения X_n



Рисунок 11. График изменения Y_n

Смотря на графики, видим, что после линейного преобразования амплитуда колебаний увеличилась ещё сильнее.

Дисперсия, изначальное значение которой составляло $\sigma_X^2 = 0.5648$, тоже увеличилась. После линейного преобразования она вычисляется по формуле $\sigma_Y^2 = 2(\sigma_X^2 - K_X(1))$, т. е. практически увеличивается в 2 раза. Если же корреляционная функция отрицательна, как в нашем случае, то увеличение дисперсии происходит ещё сильнее. В нашем случае она увеличилась почти в 3 раза, с 0.5648 до 1,69184.

Дифференцирование послужило фильтром высоких частот, усилив быстрые изменения сигнала и подавив медленные.

Вывод:

В ходе лабораторной работы были исследованы различные линейные преобразования случайных функций и их влияние на статистические характеристики. Масштабирование и сдвиг позволили изменить математическое ожидание и дисперсию процесса, что полезно для приведения сигнала к нужным параметрам. Скользящее среднее продемонстрировало эффект сглаживания, уменьшив дисперсию и подавив высокочастотные колебания, что характерно для фильтра низких частот. Напротив, дифференцирование усилило высокочастотные компоненты, увеличив дисперсию и амплитуду колебаний, что соответствует фильтру высоких частот. Эти преобразования показали, как линейные операции могут существенно влиять на свойства случайных процессов.

Варианты заданий

Вариант 1

1. Исходный процесс:

$$X_n = \{1.2, -0.3, 0.8, -1.1, 0.5, -0.7, 1.0, -0.4, 0.9, -1.2\}$$

2. Линейное преобразование:

$$Y_n = 1.5 \cdot X_n + 0.5$$

3. Окно для скользящего среднего: $m = 4$.

Вариант 2

1. Исходный процесс:

$$X_n = \{-0.6, 1.1, -0.9, 0.4, -1.3, 0.7, -0.2, 1.4, -0.8, 0.3\}$$

2. Линейное преобразование:

$$Y_n = 0.8 \cdot X_n - 0.2$$

3. Окно для скользящего среднего: $m = 2$.

Вариант 3

1. Исходный процесс:

$$X_n = \{0.9, -1.5, 0.3, -0.7, 1.2, -0.4, 0.6, -1.0, 0.8, -0.5\}$$

2. Линейное преобразование:

$$Y_n = 2.0 \cdot X_n + 1.0$$

3. Окно для скользящего среднего: $m = 3$.

Вариант 4

1. Исходный процесс:

$$X_n = \{-1.0, 0.5, -0.8, 1.3, -0.6, 0.9, -1.2, 0.4, -0.7, 1.1\}$$

2. Линейное преобразование:

$$Y_n = 1.2 \cdot X_n - 0.3$$

3. Окно для скользящего среднего: $m = 5$.

Вариант 5

1. Исходный процесс:

$$X_n = \{0.4, -0.9, 1.5, -0.2, 0.7, -1.3, 0.6, -0.8, 1.2, -0.5\}$$

2. Линейное преобразование:

$$Y_n = 0.5 \cdot X_n + 0.1$$

3. Окно для скользящего среднего: $m = 4$.

Вариант 6

1. Исходный процесс:

$$X_n = \{-0.7, 1.0, -0.3, 0.8, -1.4, 0.6, -0.9, 1.2, -0.5, 0.7\}$$

2. Линейное преобразование:

$$Y_n = 1.8 \cdot X_n + 0.6$$

3. Окно для скользящего среднего: $m = 3$.

Вариант 7

1. Исходный процесс:

$$X_n = \{1.1, -0.4, 0.9, -1.2, 0.3, -0.8, 1.3, -0.6, 0.5, -1.0\}$$

2. Линейное преобразование:

$$Y_n = 0.7 \cdot X_n - 0.5$$

3. Окно для скользящего среднего: $m = 2$.

Вариант 8

1. Исходный процесс:

$$X_n = \{-0.5, 1.2, -0.6, 0.9, -1.1, 0.4, -0.7, 1.0, -0.3, 0.8\}$$

2. Линейное преобразование:

$$Y_n = 1.0 \cdot X_n + 0.0$$

3. Окно для скользящего среднего: $m = 4$.

Вариант 9

1. Исходный процесс:

$$X_n = \{0.6, -1.0, 0.5, -0.9, 1.4, -0.3, 0.7, -1.2, 0.8, -0.4\}$$

2. Линейное преобразование:

$$Y_n = 1.3 \cdot X_n + 0.2$$

3. Окно для скользящего среднего: $m = 3$.

Вариант 10

1. Исходный процесс:

$$X_n = \{-0.8, 0.7, -1.3, 0.6, -0.2, 1.1, -0.5, 0.9, -1.4, 0.3\}$$

2. Линейное преобразование:

$$Y_n = 0.9 \cdot X_n - 0.1$$

3. Окно для скользящего среднего: $m = 5$.

Вариант 11

1. Исходный процесс:

$$X_n = \{1.3, -0.6, 0.4, -1.1, 0.8, -0.2, 1.0, -0.7, 0.5, -1.2\}$$

2. Линейное преобразование:

$$Y_n = 1.6 \cdot X_n + 0.4$$

3. Окно для скользящего среднего: $m = 4$.

Вариант 12

1. Исходный процесс:

$$X_n = \{-0.4, 0.8, -1.2, 0.5, -0.9, 1.3, -0.6, 0.7, -1.0, 0.2\}$$

2. Линейное преобразование:

$$Y_n = 0.6 \cdot X_n + 0.3$$

3. Окно для скользящего среднего: $m = 3$.

Вариант 13

1. Исходный процесс:

$$X_n = \{0.7, -1.1, 0.2, -0.5, 1.0, -0.3, 0.9, -1.3, 0.6, -0.8\}$$

2. Линейное преобразование:

$$Y_n = 1.4 \cdot X_n - 0.2$$

3. Окно для скользящего среднего: $m = 2$.

Вариант 14

1. Исходный процесс:

$$X_n = \{-0.3, 0.9, -1.4, 0.7, -0.1, 1.2, -0.6, 0.8, -1.1, 0.4\}$$

2. Линейное преобразование:

$$Y_n = 1.1 \cdot X_n + 0.7$$

3. Окно для скользящего среднего: $m = 5$.

Вариант 15

1. Исходный процесс:

$$X_n = \{0.5, -0.7, 1.4, -0.4, 0.6, -1.0, 0.3, -0.9, 1.2, -0.6\}$$

2. Линейное преобразование:

$$Y_n = 0.8 \cdot X_n + 0.5$$

3. Окно для скользящего среднего: $m = 3$.

Указания к работе:

- Для каждого варианта требуется выполнить все этапы лабораторной работы:

1. Расчёт статистических характеристик исходного процесса.
2. Применение линейного преобразования и анализ изменений.
3. Применение скользящего среднего и построение графиков.
4. Применение дифференцирования и анализ результатов.

- Размер окна m для скользящего среднего указан в каждом варианте.

- Все расчёты должны сопровождаться подробными пояснениями и выводами.

Лабораторная работа № 9 Построение модели линейной корреляции по не сгруппированным данным

Цель работы: овладение способами построения моделей линейной корреляции для несгруппированных данных, выработка умения и навыков оценки надежности коэффициента корреляции, уравнения регрессии и его коэффициентов.

Содержание работы: на основании опытных данных требуется:

1. Построить корреляционное поле. По характеру расположения точек в корреляционном поле выбрать общий вид регрессии.
2. Вычислить числовые характеристики $\bar{x}$, $\bar{y}$, S_x , S_y , r , σ_r .
3. Определить значимость коэффициента корреляции r и найти для него доверительный интервал с надежностью $\gamma = 0,95$.
4. Написать эмпирические уравнения линий регрессий y на x и x на y .
5. Вычислить коэффициент детерминации R^2 и объяснить его смысловое значение.
6. Проверить адекватность уравнения регрессии y на x .
7. Провести оценку величины погрешности уравнения регрессии y на x и его коэффициентов.
8. Построить уравнение регрессии y на x в первоначальной системе координат.

Суть лабораторной работы отражает следующая задача.

Задача. Результаты наблюдений зависимости средней заработной платы Y (тыс. руб.) от производительности труда X (тыс. руб.) по цеху технологической связи ТПЭУС № 1 по кварталам приведены в табл. 23.

Т а б л и ц а 23

X	24,3	24,9	28,1	30,5	31,5	39,3	40,2	43,5	45,4	45,9
Y	8,2	8,6	8,7	8,9	9,1	10,6	11,3	11,8	12,9	13,1

Выполнение работы

Для решения поставленной задачи методами корреляционного анализа определим, какой из указанных в условии показателей выбрать за факторный признак, а какой за результативный. На основании, например, экономического анализа производственной деятельности и взаимосвязи производительности труда и средней заработной платы следует, что за факторный признак X следует принять производительность труда, а среднюю зарплату за результативный признак Y .

Для определения формы связи между признаками X и Y строим на координатной плоскости точки (x_i, y_i) , пользуясь табл. 23. Около построенных точек проводим так называемую *линию тренда* (на рис. 8 — пунктирная линия) так, чтобы построенные точки были расположены как можно ближе к ней. По расположению точек около линии тренда делаем вывод о том, что связь между производительностью труда и средней зарплатой может носить линейный характер. Произведем расчет статистик $\bar{x}$, $\bar{y}$, S_x , S_y , r , которые войдут в уравнения линий регрессий. Составим расчетную табл. 24.

Таблица 24

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	x^2	xy
24,3	- 11,06	122,3236	8,2	- 2,12	4,4944	590,49	199,26
24,9	- 10,46	109,4116	8,6	- 1,72	2,9584	620,01	214,14
28,1	- 7,26	52,7076	8,7	- 1,62	2,6244	789,61	244,47
30,5	- 4,86	23,6196	8,9	- 1,42	2,0164	930,25	271,45
31,5	- 3,86	14,8996	9,1	- 1,22	1,4884	992,25	286,65
39,3	3,94	15,5236	10,6	0,28	0,0784	1544,49	416,58
40,2	4,84	23,4256	11,3	0,98	0,9604	1616,04	454,26
43,5	8,14	66,2596	11,8	1,48	2,1904	1892,25	513,3
45,4	10,04	100,8016	12,9	2,58	6,6564	2061,16	585,66
45,9	10,54	11,0916	13,1	2,78	7,7284	2106,81	601,29
353,6		640,064	103,2		31,196	13143,36	3787,06

Пользуясь результатами последней строки табл. 24, согласно (2),

(49), (50) — (53), находим:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} \times 353,6 = 35,36 \text{ — средняя производительность труда.}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{10} \times 103,2 = 10,32 \text{ — средняя зарплата сотрудников цеха технологической связи,}$$

$$\hat{S}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9} \times 640,064 = 71,118 \Rightarrow \hat{S}_x = 8,43,$$

$$\hat{S}_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{9} \times 31,196 = 3,466 \Rightarrow \hat{S}_y = 1,87,$$

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i = \frac{1}{10} \times 3787,06 = 378,71,$$

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\hat{S}_x \hat{S}_y} = \frac{378,71 - 35,36 \times 10,32}{8,43 \times 1,87} = 0,87.$$

Проверяем значимость коэффициента корреляции. Вычислим статистику t_p по формуле (59):

$$t_p = \frac{|r| \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,87 \times \sqrt{10-2}}{\sqrt{1-0,87^2}} = 5,02.$$

По таблице критических точек распределения Стьюдента (приложение 6) по уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = n - 2 = 10 - 2 = 8$ находим $t_T = t_{\alpha; k} = t_{0,05; 8} = 2,31$. Так как $t_p = 5,02 > t_T$, то выборочный коэффициент корреляции значимо отличается от нуля. Следовательно, можно предположить, что средняя зарплата Y и производительность X труда в цехе связаны линейной корреляционной зависимостью, и провести приблизительную прямую s на числовой плоскости (рис. 8).

Находим доверительный интервал для выборочного коэффициента корреляции r с надежностью $\gamma = 0,95$. Так как объем выборки $n = 10 < 50$, то доверительный интервал находим по формуле (61):

$$r - t_\gamma \times \sigma_r \leq \hat{r} \leq r + t_\gamma \times \sigma_r.$$

Так как по условию надежность (доверительная вероятность) равна

$\gamma = 0,95$, то по таблице функции Лапласа (приложение 2) находим $t_\gamma = 1,96$.
Вычисляем среднюю квадратическую ошибку σ_r по формуле (60):

$$\sigma_r = \frac{1-r^2}{\sqrt{n-2}} = \frac{1-0,87^2}{\sqrt{10-2}} = 0,08.$$

Записываем доверительный интервал:

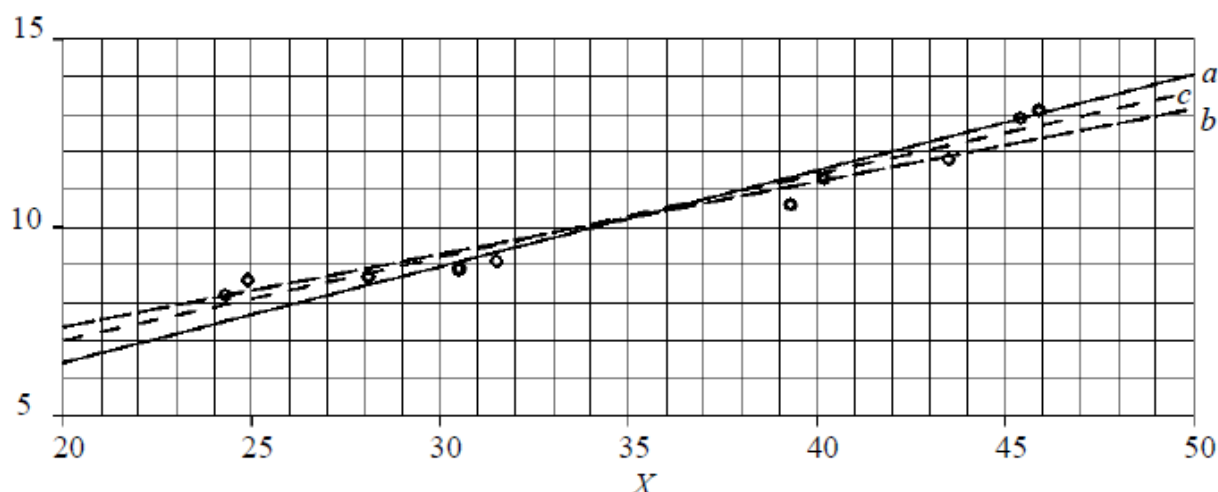


Рис. 8. Линейная регрессия Y на X — прямая a с уравнением $y = 0,193x + 3,5$ и линейная регрессия X на Y — прямая b с уравнением $x = 3,922y + 5,1$. Пунктированная прямая c проведена «от руки».

$$0,87 - 1,96 \times 0,08 \leq \hat{r} \leq 0,87 + 1,96 \times 0,08$$

или $\hat{r} \in [0,71; 1]$. Следовательно, с вероятностью 0,95 линейный коэффициент корреляции генеральной совокупности находится в пределах от 0,71 до 1. Применительно к решаемой задаче полученный результат означает, что по имеющейся выборке следует ожидать влияние производительности труда на рост средней зарплаты работников цеха технологической связи не менее чем на 71 %.

Найдем эмпирические линейные уравнения регрессии y на x и x на y , которые являются приближенными уравнениями для истинных уравнений регрессий.

Уравнение регрессии y на x :

$$\hat{y}_x = \bar{y} + r \frac{\hat{s}_y}{\hat{s}_x} (x - \bar{x}) \Rightarrow \hat{y}_x = 10,32 + 0,87 \times \frac{1,87}{8,43} (x - 35,36) \Rightarrow$$

$$\hat{y}_x = 0,192989x + 3,495898.$$

Уравнение регрессии x на y :

$$\hat{x}_y = \bar{x} + r \frac{s_x}{s_y} (y - \bar{y}) \Rightarrow \hat{x}_y = 35,36 + 0,87 \times \frac{8,43}{1,87} (y - 10,32) \Rightarrow$$

$$\hat{x}_y = 3,921979y - 5,114819.$$

Контроль вычислений:

$$a_1 b_1 = 0,192989 \times 3,921979 = 0,7569,$$

$$r^2 = 0,87 \times 0,87 = 0,7569.$$

Так как условие $a_1 b_1 = r^2$ выполняется, то вычисления выполнены верно.

Из уравнения $\hat{y}_x = 0,192989x + 3,495898$ следует, что при увеличении производительности труда на 1 тыс. руб. средняя зарплата работников цеха технологической связи возрастает на 192,989 рублей. Этот результат следует учесть на предприятии при разработке мероприятий по стимулированию производственной деятельности работников цеха в условиях рыночных отношений.

Подставляя $\bar{x} = 35,36$, $\bar{y} = 10,32$ в уравнения регрессий, получаем точки, координаты которых совпадают с координатами центра распределения $C(\bar{x}, \bar{y})$. Следовательно, линии регрессий пересекаются в точке $C(\bar{x}, \bar{y})$.

Найдем коэффициент детерминации. Для линейной корреляции при вычисленном коэффициенте r он равен r^2 . У нас $r^2 = 0,76$. Это означает, что 76 % рассеивания средней зарплаты работников технологического цеха связи объясняется линейной корреляционной зависимостью между средней зарплатой и производительностью труда, и только 24 % рассеивания средней зарплаты работников технологического цеха остались необъяснимыми. Такое положение могло произойти из-за того, что в модель не включены другие факторы, влияющие на изменение средней зарплаты работников технологического цеха связи, либо опытных данных в данной выборке не достаточно, чтобы построить более надежное уравнение регрессии.

Проверим адекватность уравнения линейной регрессии y на x по критерию Фишера — Снедекора. Вычислим статистику F_H по формуле (64):

$$F_H = \frac{R^2(n-2)}{1-R^2}, \text{ где } R^2 = 1 - \frac{\sum (y_i - \hat{y}_{x_i})^2}{\sum (y_i - \bar{y})^2}.$$

Для нахождения суммы $\sum (y_i - \hat{y}_{x_i})^2$ составляем табл. 25. Из табл. 24 и 25 находим: $\sum (y_i - \bar{y})^2 = 31,196$, $\sum (y_i - \hat{y}_{x_i})^2 = 1,8729$. Тогда

$$R^2 = 1 - \frac{1,8729}{31,196} = 0,94, \quad F_H = \frac{0,94 \times (10-2)}{1-0,94} = 125,3.$$

При уровне значимости $\alpha = 0,05$ и числах степеней свободы $k_1 = 1$, $k_2 = n - 2 = 10 - 2 = 8$ по таблице критических точек распределения Фишера — Снедекора (приложение 7) находим $F_T = F_{\alpha; k_1; k_2} = F_{0,05; 1; 8} = 5,32$. Так как $F_H = 125,3 > 5,32$, то заключаем, что уравнение линейной регрессии

Т а б л и ц а 2 5

y_i	$\hat{y}_x$	$y_i - \hat{y}_x$	$(y_i - \hat{y}_x)^2$
8,2	8,18	0,02	0,0004
8,6	8,3	0,3	0,09
8,7	8,9	-0,2	0,04
8,9	9,4	-0,5	0,25
9,1	9,6	-0,5	0,25
10,6	11,1	-0,5	0,25
11,3	11,25	0,05	0,0025
11,8	11,9	-0,1	0,01
12,9	12,2	0,7	0,49
13,1	12,4	0,7	0,49
			1,8729

$\hat{y}_x = 0,192989x + 3,495898$ статистически значимо описывает результаты эксперимента.

Проведем оценку величины погрешности уравнения регрессии $\hat{y}_x = 0,192989x + 3,495898$. Найдем относительную погрешность δ уравнения по формуле (67):

$$\delta = \frac{\sigma_u}{\bar{y}} \times 100\%,$$

где

$$\sigma_u = \sqrt{Du} = \sqrt{\frac{\sum (u_i - \bar{u})^2}{n-2}}, \quad u_i = y_i - \hat{y}_x, \quad \bar{u} = \frac{1}{n} \sum (y_i - \hat{y}_x)^2.$$

Так как $\sum (y_i - \hat{y}_{x_i})^2 = 1,8729$, то $\bar{u} = 0,19$. Для нахождения суммы $\sum (u_i - \bar{u})^2$ составляем табл. 26.

Тогда $\sigma_u = \sqrt{\frac{2,2588}{8}} = 0,53$, $\delta = \frac{0,53}{10,32} \times 100\% = 5\%$. Так как величина δ мала, то уравнение линейной регрессии $\hat{y}_x = 0,192989x + 3,495898$ хорошо описывает опытные данные.

Оценим коэффициенты уравнения регрессии. У нас $a_0 = 3,495898$,

$a_1 = 0,192989$. Для нахождения отношений $S_{a_0}/|a_0|$ и $S_{a_1}/|a_1|$ вычислим средние квадратические ошибки коэффициентов по формулам (68) и (69):

Т а б л и ц а 2 6

$u_i = y_i - \hat{y}_x$	$u_i - \bar{u}$	$(u_i - \bar{u})^2$
0,02	- 0,17	0,0289
0,03	- 0,16	0,0256
- 0,2	- 0,39	0,1521
- 0,5	- 0,69	0,4761
- 0,5	- 0,69	0,4761
- 0,5	- 0,69	0,4761
0,05	- 0,14	0,0196
- 0,1	- 0,29	0,0841
0,7	0,51	0,2601
0,7	0,51	0,2601
		2,2588

$$S_{a_0} = S_{y/x} \times \sqrt{\frac{[x^2]}{n[x^2] - ([x])^2}}, S_{a_1} = S_{y/x} \sqrt{\frac{n}{n[x^2] - ([x])^2}}, S_{y/x} = \hat{S}_y \sqrt{1 - r^2}.$$

По табл. 24 находим: $[x] = 353,6$, $[x^2] = 13143,36$. Учитывая, что $n = 10$, $r^2 = 0,7569$ и $\hat{S}_y = 1,87$, находим:

$$S_{y/x} = 1,87 \times \sqrt{1 - 0,7569} = 0,4931,$$

$$S_{a_0} = 0,4931 \times \sqrt{\frac{13143,36}{10 \times 13143,36 - (353,6)^2}} = 0,71,$$

$$S_{a_1} = 0,4931 \times \sqrt{\frac{10}{10 \times 13143,36 - (353,6)^2}} = 0,04.$$

Так как $S_{a_0}/|a_0| = \frac{0,71}{3,495898} < 0,5$ и $S_{a_1}/|a_1| = \frac{0,04}{0,192989} < 0,5$, то коэффициенты a_0 и a_1 уравнения регрессии y на x значимы. Графиками найденных регрессий являются прямые a , b , представленные на рис. 8.

Таким образом, уравнение регрессии $\hat{y}_x = 0,192989x + 3,495898$, описывающее зависимость средней зарплаты работников цеха технологической связи от производительности труда, значимо описывает опытные данные и может быть принято для практического руководства.

Варианты заданий

Вариант № 1. Имеются данные, характеризующие зависимость между стоимостью X основных производственных фондов и объемом Y валовой продукции по десяти однотипным предприятиям:

X	1	2	3	4	5	6	7	8	9	10
Y	20	25	31	31	40	56	52	60	60	70

Вариант № 2. Зависимость между стоимостью X (тыс. млн. руб.) основных средств предприятий и месячным выпуском Y (тыс. руб.) продукции характеризуется следующими данными:

X	1	2	3	4	5	6	7
Y	10	12	28	40	42	52	54

Найти эмпирическую формулу функциональной зависимости месячного выпуска продукции от стоимости основных средств предприятий. Построить эмпирическую и теоретическую линии.

Вариант № 3. Имеются данные наблюдений изменения средней заработной платы Y (руб.) в зависимости от изменения производительности труда X (шт.) за 4 месяца 1992 года по девяти токарям цеха № 23 электро-механического завода:

X	406	660	914	1168	1422	1676	1930	2184	2438
Y	518,5	813,5	1108,5	1403,5	1698,5	1993,5	2288,5	2583,5	2878,5

Вариант № 4. Дано распределение заводов по производственным средствам X (млн. руб.) и по суточной выработке Y (тыс. руб.):

X	50	49	48	51	52	53	54	57	59	60	61	55	60	62	63
Y	10	8	10	9	10	12	13	15	16	18	20	17	21	25	24

Вариант № 5. Данные нормы расхода моторного масла на угар и замену Y (л/100 л.т.) от максимальной мощности двигателя X (л.с.) приведены в таблице:

X	39	42	53	70	73,5	75	90	98	110	115
Y	1,3	1,3	0,8	2,2	1,8	2	2,2	1,8	2,8	2,1

Вариант № 6. Результаты наблюдений изменения диаметра Y (мм) вала и износа X (мм) резца приведены в следующей таблице:

X	0,01	0,02	0,024	0,03	0,032	0,035	0,037	0,042	0,048	0,057
Y	20	20,01	20,014	20,022	20,024	20,027	20,029	20,034	20,04	20,049

Найти зависимость диаметра вала от износа резца.

Вариант № 7. Компанию по прокату автомобилей интересует зависимость между пробегом X автомобилей и стоимостью Y ежемесячного технического обслуживания. Для выяснения характера этой связи было отобрано 15 автомобилей. Данные приведены в таблице:

X	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	13	16	15	20	19	21	26	24	30	32	30	35	34	40	39

Вариант № 8. При исследовании зависимости между выпуском готовой продукции Y (тыс. руб.) и энерговооруженностью труда X (кВт-час) получены следующие данные:

X	1201	1300	1375	1412	1443	1500	1526	1516	1718	1783	1819	1877
Y	14	21	27	29	30	31,5	35	34	41	38	39	46

Вариант № 9. В таблице приведены данные, характеризующие зависимость израсходованных долот Y (шт.) при бурении 8 скважин в зависимости от механической скорости X (м/с) проходки:

X	10	15	8	12	16	18	22	25
Y	55	40	60	50	40	30	25	20

Вариант № 10. Скорость Y (м/час) бурения в твердых породах от нагрузки X (атм.) на долото характеризуется следующими данными:

X	10	10,5	11	11,5	12	12,5	13	13,5
Y	3,5	2,5	2,5	1,5	1	0,5	1	0,5

Найти эмпирическую формулу зависимости Y от X .

Вариант № 11. Зависимость между выработкой продукции X (тыс. руб.) и затратами топлива Y в условных единицах характеризуется следующими данными:

X	3	4	5	6	7	8	9	10	11	12
Y	2	2,3	2,2	2,3	2,5	2,6	2,5	2,6	2,8	2,9

Вариант № 12. Имеются данные распределения заводов по производственным средствам X (млн. руб.) и по суточной выработке Y (млн. руб.):

X	48	49	50	52	53	54	57	60	63	65	68	80
Y	11	8	10	10	12	13	15	18	24	26	25	38

Вариант № 13. Найти формулу, устанавливающую зависимость между коэффициентом Y сменности техники и ее средним возрастом X по предприятию ПМК-7 объединения Сибкомплектмонтаж на основании следующих данных:

Y	1,18	1,21	1,25	1,26	1,3	1,32	1,33	0,69	0,72	0,8
X	6,31	5,8	5,1	5,6	6,1	6,5	6,55	3,8	3,41	4

Вариант № 14. Имеются данные о реализации продукции X (млн. руб.) и накладных расходах Y (тыс. руб.) на реализацию:

X	9	13	17	22	29	36	44	51	60	65
Y	27	36	29	41	54	71	65	81	90	95

Вариант № 15. Зависимость линейной нормы расхода топлива Y (л) от максимальной мощности двигателя автомобиля X (л.с.) характеризуется следующими данными:

X	39	53	70	75	90	98	110	115	120	150
Y	12	11	21,5	22,8	18	21	31	25	30	36

Вариант № 16. Имеются данные нормы расхода моторных масел на угар и замену Y (л/100 л.т.) в зависимости от максимального крутящего момента X :

X	7,4	7,6	8,2	10,8	17	21	35	41	42	44
Y	1,3	1,3	1,5	1,8	2,2	2,1	2,8	2,6	2,3	2,4

Вариант № 17. Фазовая проницаемость нефти Y и насыщенность породы нефтью X характеризуются следующими данными:

X	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95
Y	0,35	0,45	0,55	0,65	0,75	0,8	0,85	0,95	1	1,25

Вариант № 18. Имеются данные нормы расхода моторных масел на угар и замену Y (л/100 л.т.) в зависимости от максимальной мощности двигателя автомобиля X (л.с.):

X	39	42	53	70	75	90	110	115	150	170
Y	1,3	1,3	0,9	2,2	2,2	2,2	2,8	2,4	2,5	2,6

Вариант № 19. Зависимость между среднегодовой стоимостью основных производственных фондов X (млн. руб.) и стоимостью товарной продукции Y (млн. руб.) характеризуется следующими данными:

X	1,94	2,68	3,47	4,12	4,77	5,34	5,85	6,65
Y	0,82	0,97	1,06	1,08	1,1	1,14	1,21	1,25

Вариант № 20. Найти формулу, устанавливающую зависимость себестоимости одной тонны нефти Y (в руб.) от затрат X (в тыс. руб.) на одну тонну по следующим данным:

X	1,44	1,6	1,85	2,1	2,25	2,42	2,55	2,65
Y	161,5	165	170	175	178	182	186	190

Вариант № 21. Ниже приводятся данные о производительности труда Y (м) на одного чел/час и стаже рабочих X (в годах):

X	1	2	3	4	5	6	7	8
Y	9,8	15	16	19	20	22	23	27

Вариант № 22. Найти формулу зависимости электрического сопротивления R (Ом) проводника от температуры Θ °C по следующим данным:

Θ	19,1	25,0	30,1	36,0	40,0	45,1	50,0
R	76,30	77,80	79,75	80,80	82,35	83,90	85,10

Построить теоретическую и эмпирическую линии.

Вариант № 23. Найти зависимость израсходованных долот Y при бурении 10 скважин в зависимости от механической скорости X проходки на основании следующих данных:

X	8	10	12	14	15	16	18	19	20	21
Y	60	55	50	48	45	40	30	31	29	32

Вариант № 24. Зависимость между пористостью Y пород и их газонасыщенностью X задана таблицей:

X	10	18	25	36	43	54	62	70	75	80
Y	12	16	21	24	27	29	31	34	37	42

Лабораторная работа №10 Построение выборочного уравнения линии регрессии по сгруппированным данным

Ц е л ь р а б о т ы: овладение способами построения моделей линейной корреляции для сгруппированных данных по методу наименьших квадратов с использованием коэффициента линейной корреляции, выработка умения и навыков оценки надежности уравнения регрессии и его коэффициентов.

С о д е р ж а н и е р а б о т ы: по опытным данным требуется:

1. Построить корреляционное поле. По характеру расположения точек в корреляционном поле выбрать общий вид регрессии.

2. Написать уравнение линии регрессии y на x по методу наименьших квадратов и с использованием коэффициента корреляции r . Сравнить полученные уравнения и сделать вывод о выборе одного из них.

3. Оценить тесноту связи между признаками X и Y с помощью выборочного коэффициента корреляции r и его значимость.

4. Проверить адекватность модельного уравнения регрессии y на x , записанного через коэффициент корреляции r .

5. Проверить надежность уравнения регрессии y на x , записанного через коэффициент корреляции r и его коэффициентов.

6. Построить уравнения регрессий в первоначальной системе координат.

З а д а ч а. Валики при черновой обработке на станке № 1 передаются последовательно на станок № 2 для чистовой обработки. Экспериментатор, изучающий зависимость между отклонениями размеров валиков от номинала при черновой обработке (мкм), от номинала при чистовой обработке (мкм) произвел измерения отклонений у 50 случайно отобранных валиков. Результаты измерений сведены в табл. 27.

Т а б л и ц а 27

$\begin{matrix} x \\ y \end{matrix}$	- 30	- 20	- 10	0	n_y
- 8	1				1
- 4	4	1			5
0	1	15	1		17
4		2	13		15
8			2	1	3
12				9	9
n_x	6	18	16	10	50

Выполнение работы

Пусть признак X характеризует отклонение размеров валиков от номинала при черновой обработке, а признак Y отклонение размеров валиков от номинала при чистовой обработке. Используя данные табл. 27, строим корреляционное поле (рис. 9).

Проведя линию тренда (пунктирная линия), видим, что число точек, расположенных над и под ней, практически одинаково, причем расстояния этих точек до линии тренда одинаковые. Это дает основание предположить наличие линейной зависимости между признаками X и Y . Для подтверждения этой гипотезы перейдем от данного распределения к новому, найдя для каждого значения признака X условное среднее признака Y по формуле (42):

$$\bar{y}_{x_j} = \frac{\sum n_{ij} y_i}{n_{x_j}}.$$

$$\text{При } x_1 = -30, \bar{y}_{x_1} = \frac{1 \times (-8) + 4 \times (-4) + 1 \times 0}{6} = -4.$$

$$\text{При } x_2 = -20, \bar{y}_{x_2} = \frac{1 \times (-4) + 15 \times 0 + 2 \times 4}{18} = 0,2.$$

При $x_3 = -10$, $\bar{y}_{x_3} = \frac{1 \cdot 0 + 13 \cdot 4 + 2 \cdot 8}{16} = 4,3$.

При $x_4 = 0$, $\bar{y}_{x_4} = \frac{1 \cdot 8 + 9 \cdot 12}{10} = 11,6$.

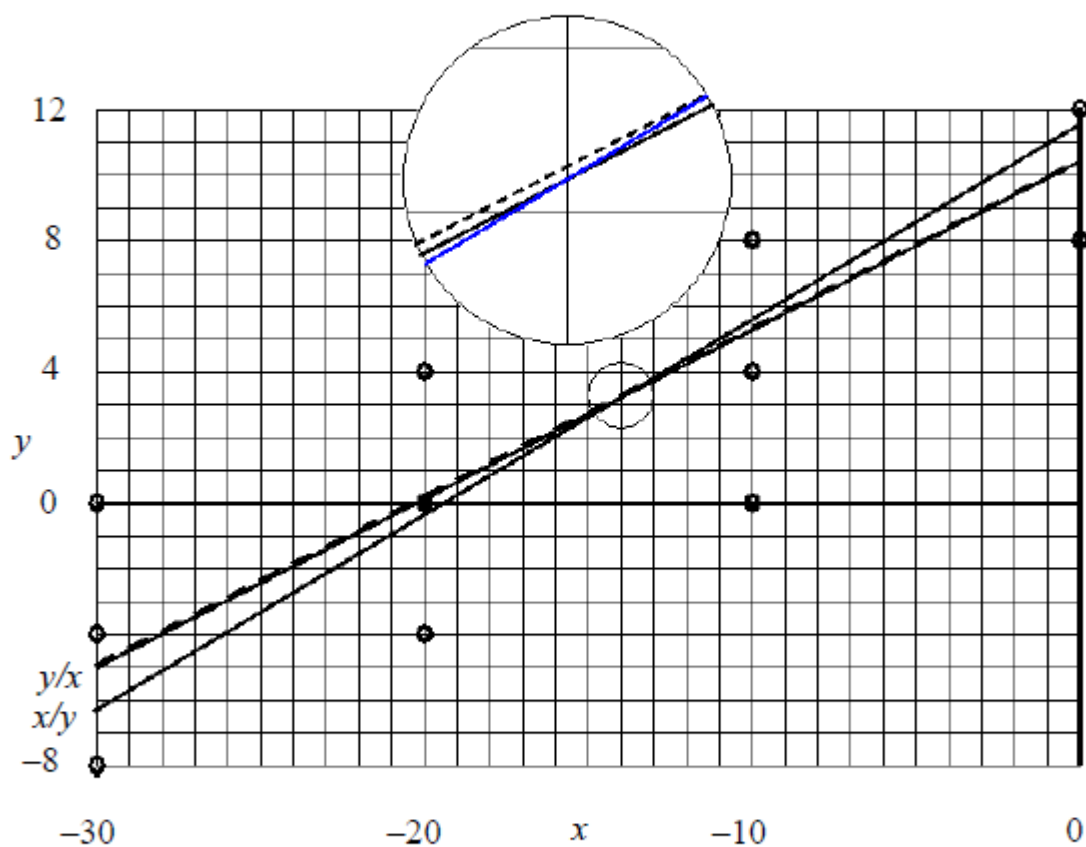


Рис. 9. Выявление линейной регрессии. «От руки» проводится линия тренда (штриховая прямая). Прямая (y/x) — регрессия y на x . Прямая (x/y) — регрессия x на y . 5-кратный «микроскоп» показывает картину в окрестности точки $(\bar{x}, \bar{y})$.

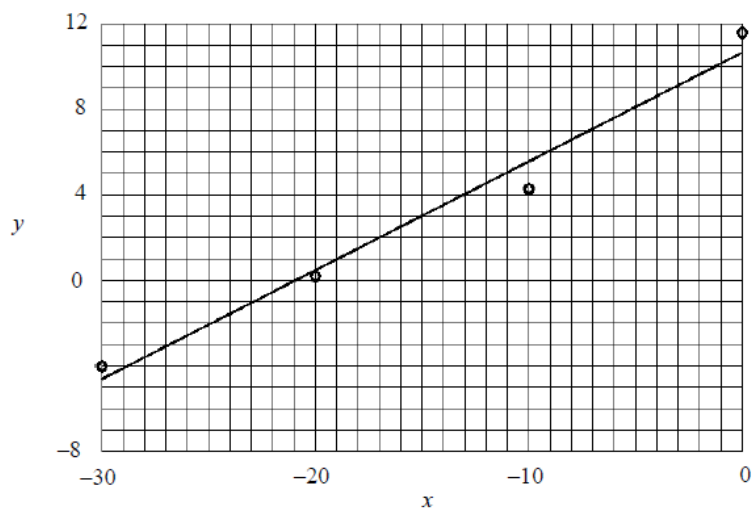


Рис. 10.

Строим точки с координатами $(x_j; \bar{y}_{x_j})$ (рис. 10).

Из рис. 10 видно, что отклонения точек от построенной прямой незначительны. Следовательно, связь между признаками X и Y может носить линейный характер. Составим уравнения линий регрессий y на x по методу наименьших квадратов и через коэффициент линейной корреляции r .

Применим метод наименьших квадратов к нахождению коэффициентов a_0 и a_1 уравнения линейной регрессии $\hat{y}_x = a_1 x + a_0$. Решаем систему нормальных уравнений (45):

$$\begin{cases} na_0 + [n_x x]a_1 = [n_y y], \\ [n_x x]a_0 + [n_x x^2]a_1 = [n_{xy} xy]. \end{cases}$$

Для нахождения сумм, входящих в систему, составляем табл. 28.

Т а б л и ц а 28

$y \backslash x$	-30	-20	-10	0	n_y	$n_y y$
-8	1				1	-8
-4	4	1			5	-20
0	1	15	1		17	0
4		2	13		15	60
8			2	1	3	24
12				9	9	108
n_x	6	18	16	10	50	164
$n_x x$	-180	-360	-160	0	-700	
$n_x x^2$	5400	7200	1600	0	14200	
$n_{xy} xy$	720	-80	-680	0	-40	

Полученная из табл. 28 система

$$\begin{cases} 50a_0 - 700a_1 = 164, \\ -700a_0 + 14200a_1 = -40 \end{cases}$$

имеет решение $(a_0, a_1) = (0,512728; 10,458192)$. Тогда уравнение линейной регрессии запишется в виде:

$$\bar{y}_x = 0,512728x + 10,458192 \quad (70)$$

Найдем уравнение линейной регрессии y на x по формуле (47), используя коэффициент линейной корреляции:

$$\hat{y}_x = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x}).$$

Так как данные выборки для признаков X и Y заданы в виде корреляционной таблицы и объем выборки $n = 50$, то для нахождения величин, входящих в уравнение регрессии, переходим к вспомогательному распределению с условными вариантами u_i и v_j . По корреляционной табл. 27 находим наибольшую частоту совместного появления признаков X и Y : $n_{32} = 15$. Тогда $C_1 = M_0X = -20$, $C_2 = M_0Y = 0$, $h_1 = 10$, $h_2 = 4$. Составляем корреляционную табл. 29 в условных вариантах.

Т а б л и ц а 29

$v \backslash u$	-1	0	1	2	n_v
-2	1				1
-1	4	1			5
0	1	15	1		17
1		2	13		15
2			2	1	3
3				9	9

По табл. 29 находим:

$$\bar{u} = \frac{1}{n} \sum n_{u_j} u_j = \frac{1}{50} (6 \times (-1) + 18 \times 0 + 16 \times 1 + 10 \times 2) = 0,6,$$

$$\bar{v} = \frac{1}{n} \sum n_{v_i} v_i = \frac{1}{50} (1 \times (-2) + 5 \times (-1) + 17 \times 0 + 15 \times 1 + 3 \times 2 + 9 \times 3) = 0,8,$$

$$\overline{u^2} = \frac{1}{n} \sum n_{u_j} u_j^2 = \frac{1}{50} (6 \times (-1)^2 + 18 \times 0^2 + 16 \times 1^2 + 10 \times 2^2) = 1,24,$$

$$\overline{v^2} = \frac{1}{n} \sum n_{v_i} v_i^2 = \frac{1}{50} (1 \times (-2)^2 + 5 \times (-1)^2 + 17 \times 0^2 + 15 \times 1^2 + 3 \times 2^2 + 9 \times 3^2) = 2,34.$$

Тогда

$$S_u = \sqrt{\overline{u^2} - (\bar{u})^2} = \sqrt{1,24 - 0,36} = 0,94,$$

$$S_v = \sqrt{\overline{v^2} - (\bar{v})^2} = \sqrt{2,34 - 0,64} = 1,3.$$

Для нахождения суммы $\sum n_{u_j v_i} u_j v_i$ составляем табл. 30.

Т а б л и ц а 30

$v \backslash u$	-1	0	1	2	
-2	1 2				2
-1	4 1				4
0					
1			13 1		13
2			2 2	1 4	8
3				9 6	54
	6		17	58	81

Тогда, согласно (56), (58),

$$r_e = \frac{81 - 50 \cdot 0,6 \cdot 0,8}{50 \cdot 0,94 \cdot 1,3} = 0,93,$$

$$\bar{x} = \bar{u}h_1 + C_1 = 0,6 \cdot 10 - 20 = -14,$$

$$\bar{y} = \bar{v}h_2 + C_2 = 0,8 \cdot 4 + 0 = 3,2,$$

$$S_x = S_u \cdot h_1 = 0,94 \cdot 10 = 9,4,$$

$$S_y = S_v h_2 = 1,3 \cdot 4 = 5,2.$$

Отсюда и из (47), (48) следуют уравнение линии регрессии y на x :

$$\hat{y}_x = 3,2 + 0,93 \frac{5,2}{9,4} (x + 14),$$

или

$$\hat{y}_x = 0,514468x + 10,40255,$$

и уравнение линии регрессии x на y :

$$\hat{x}_y = -14 + 0,93 \frac{9,4}{5,2} (y - 3,2),$$

или

$$\hat{x}_y = 1,681154y - 19,37969.$$

Проверяем тесноту связи между признаками X и Y . Для этого, используя критерий Стьюдента, вычисляем статистику

$$t_n = \frac{|r_e| \sqrt{n-2}}{\sqrt{1-r_e^2}} = \frac{0,93 \sqrt{50-2}}{\sqrt{1-0,93^2}} = 17,5.$$

При уровне значимости $\alpha = 0,05$ и числе степеней свободы $k = n - 2 = 50 - 2 = 48$ находим по таблице распределения Стьюдента $t_{\alpha; k} = t_{0,05; 48} = 2,01$. Так как $t_n = 17,5 > 2,01$, то выборочный коэффициент линейной корреляции r_e значимо отличается от нуля. Следовательно, можно считать, что отклонение размеров валиков от номинала при черновой обработке на станке № 1 и отклонение размеров валиков от номинала при чистовой обработке на станке № 2 связаны линейной корреляционной зависимостью. Дадим интерпретацию, например, уравнению регрессии y

на x . Из уравнения регрессии видно, что при отклонении от нормальных размеров валиков при черновой обработке (x_j), например, на 10 мкм на станке № 1 отклонение от нормального размера валиков при последующей чистовой обработке на станке № 2 составит $y_i = 10 \times 0,514468 = 5,14468 \approx 5,1$ мкм. Это результат воздействия отклонений при черновой обработке валиков на станке № 1. Фактически отклонение может составить $y = 5,14468 + 10,40255 = 15,54723$ мкм, что является результатом воздействия неучтенных в модели факторов, не зависящих от отклонений при черновой обработке. Уравнения линий регрессии построены на рис. 9.

Проверим полученное уравнение регрессии y на x на адекватность по критерию Фишера-Снедекора. Вычислим статистику

$$F_H = \frac{Q_R(n-2)}{Q_e(k-1)}.$$

Составим расчетные табл. 31 и 32.

Т а б л и ц а 31

y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
- 8	- 11,2	125,44
- 4	- 7,2	51,84
0	- 3,2	10,24
4	0,8	0,64
8	4,8	23,04
12	8,8	77,44
		$Q = 288,64$

Т а б л и ц а 32

$\bar{y}_{x_i}$	$\bar{y}_{x_i} - \bar{y}$	$(\bar{y}_{x_i} - \bar{y})^2$
- 4	- 7,2	51,84
0,2	- 3	9
4,3	1,1	1,21
11,6	8,4	70,56
		$Q_R = 132,61$

Находим $Q_e = Q - Q_R = 288,64 - 132,61 = 156,03$. По условию $n = 50$, $k = 4$. Тогда

$$F_H = \frac{132,61(50-2)}{156,03(4-1)} = 13,03.$$

При уровне значимости $\alpha = 0,05$ и числах степеней свободы $k_1 = 1$, $k_2 = n - 2 = 50 - 2 = 48$ по таблице критических точек распределения Фишера-Снедекора находим $F_T = 4,05$. Так как $F_H = 13,03 > 4,05$, то модель линейной регрессии $\hat{y}_x = 0,514468x + 10,40255$ согласуется с опытными данными.

Итак, мы получили два уравнения линий регрессии y на x , описывающих зависимость между признаками X и Y . При подстановке в каждое из них опытных значений признака X убеждаемся в том, что уравнение (70), полученное по методу наименьших квадратов без использования коэффициента линейной корреляции r , дает лучшие значения признака Y , чем уравнение (71). Однако практика решения инженерных задач, связан-

ных с построением моделей парной линейной корреляции показывает, что не всегда уравнения, полученные без применения коэффициента корреляции r , лучше уравнений с его применением. Поэтому, проводя моделирование в случае сгруппированных экспериментальных данных, из двух линейных регрессий следует выбирать лучшую из них.

Варианты заданий

Вариант № 1. Зависимость объема Y (см³) разрушенной породы от глубины X (мм) внедрения зуба при постоянном давлении приведены в корреляционной таблице:

$Y \backslash X$	0,5	1,5	2,5	3,5	4,5	n_y
0,05	4	3				7
0,15		5	3			8
0,25			8	1		9
0,35			4	6		10
0,45				5	1	6
0,55				6		6
0,65				1	3	4
n_x	4	8	15	19	4	50

Вариант № 2. Распределение предприятий по объему продукции X (тыс. руб.) и по ее себестоимости Y (руб.) приведено в корреляционной таблице:

$Y \backslash X$	1	2	3	4	5	n_y
2				3	4	7
2,5			5	3	5	13
3		3	6	4		13
3,5	4	4	2	2		12
4	1	4				5
n_x	5	11	13	12	9	50

Вариант № 3. Результаты исследования зависимости среднегодового перевыполнения нормы Y (%) от стажа X (в годах) работы приведены в корреляционной таблице:

$Y \backslash X$	2	3	4	5	6	7	n_y
5	1	1	1				3
6	4	2	1				7
7		2	8	6			16
8			3	9	2		14
9				1	5	2	8
10					1	1	2
n_x	5	5	13	16	8	3	50

Вариант № 4. Фонтанную скважину исследовали на приток Y нефти при различных режимах работы с замерами забойных давлений X глубинным манометром. Данные замеров приведены в корреляционной таблице:

$Y \backslash X$	125	135	145	155	165	175	185	195	205	n_y
11	3	4								7
12		5	4							9
13			3	5						8
14			5	6						11
15				2	18					20
16					4	14				18
17						7	2			9
18							4	6		10
19								2	6	8
n_x	3	9	12	13	22	21	6	8	6	100

Вариант № 5. Зависимость коэффициента обрабатываемости Y от ударной вязкости X (кг/мм²) инструментальных быстродействующих сталей задана корреляционной таблицей:

$Y \backslash X$	0,7	0,9	1,1	1,3	1,5	n_y
0,5	1	2				3
0,6	7	10	3			20
0,7		1	12	1		14
0,8				4	3	7
0,9				2	4	6
n_x	8	13	15	7	7	50

Вариант № 6. Результаты исследования зависимости между средне-месячной выработкой продукции на одного рабочего Y (тыс. руб.) и стоимостью основных производственных средств X (млн. руб.) приведены в корреляционной таблице:

$Y \backslash X$	0,8	0,9	1,0	1,1	1,2	1,3	n_y
9,9	2						2
10	2	7				1	10
10,1		2	1				3
10,2		1	3	9			13
10,3			4	3	6		13
10,4					4	5	9
n_x	4	10	8	12	10	6	50

Вариант № 7. Распределение цилиндрических болванок по длине X (см) и по весу Y (кг) приведено в корреляционной таблице:

$Y \backslash X$	1,05	1,15	1,25	1,35	1,45	1,55	1,65	n_y
19,5	4	1						5
22,5		2	4	1				7
25,5			3	8	1			12
28,5				2	3	2		7
31,5					2	4		6
34,5						3	3	6
37,5							4	4
n_x	4	3	7	11	6	9	7	47

Вариант № 8. Результаты замера температуры X (°C) смазочного масла в двигателе и температуры Y (°C) масла в коробке передач автомобиля приведены в корреляционной таблице:

$Y \backslash X$	6	10	14	18	22	26	n_y
15	4	3					7
21	1	5	1				7
27		2	5				7
33			6	3			9
39				5	4		9
45					2	9	11
n_x	5	10	12	8	6	9	50

Вариант № 9. Результаты измерений сверл по твердости Y (HRC) и по стойкости X (час) приведены в корреляционной таблице:

$Y \backslash X$	20,5	25,5	30,5	35,5	40,5	45,5	n_y
25	2	3					5
35	1	7					8
43			2	4			6
51			1	9	4		14
59				4	1	1	6
67					3	8	11
n_x	3	10	3	17	8	9	50

Вариант № 10. Результаты измерений времени X (час) непрерывной работы и количества Y (шт.) полностью обработанных деталей приведены в корреляционной таблице:

$Y \backslash X$	6	10	14	18	22	26	n_y
0,95	5	5					10
1,85		4	2				6
2,75			6	8			14
3,65				1	6		7
4,55				5	3	5	13
n_x	5	9	8	14	9	5	50

Лабораторная работа №11 Кластерный анализ.

Цель лабораторной работы: освоить основы кластерного анализа и метод k ближайших соседей с различными метриками.

6.1. Общие понятия. Методика проверки.

Термин кластерный анализ, впервые был введен Трионом в 1939 году. Фактически, кластерный анализ является не столько обычным статистическим методом, сколько "набором" различных алгоритмов "распределения объектов по кластерам", включает в себя более 100 различных алгоритмов классификации.

Во многих отраслях науки перед исследователями постоянно встает вопрос о том, как организовать наблюдаемые данные в наглядные структуры, то есть развернуть таксономии (группы, классы).

Таксономия – это учение о принципах и практике классификации и систематизации, сложно организованных иерархически соотносящихся сущностей.

Так же таксономией называют **принцип организации объектов от низшего к высшему**. Часто его изображают в виде древовидной структуры или пирамиды.

Таким образом, в многомерном статистическом анализе кластерный анализ занимает особое место.

Кластерный анализ – это общее название множества вычислительных процедур, используемых при создании классификации. Главная цель кластерного анализа – нахождение групп схожих объектов в выборке данных. Эти группы называют кластерами.

Кластерный анализ – совокупность математических методов, предназначенных для формирования относительно "отдаленных" друг от друга групп "близких" между собой объектов по информации о расстояниях или связях (мерах близости) между ними.

Кластерный анализ объединяет различные процедуры, используемые для проведения классификации. В результате применения этих процедур

исходная совокупность объектов разделяется на **кластеры** или **группы** (классы) схожих между собой объектов.

Кластер – это группа объектов, обладающая свойством плотности (плотность объектов внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой (например, кластер может иметь очертания гиперсферы или эллипсоида), размером.

Данное определение не является строгим (строгого определения не существует вообще). Наиболее часто методы кластерного анализа используются в социологии, маркетинговых исследованиях, экономике, биологии, медицине, археологии.

Сложность задач кластерного анализа состоит в том, что реальные объекты являются многомерными, то есть описываются не одним, а несколькими параметрами (представьте, что объекты – это персональные компьютеры), и объединение объектов в группы проводится в пространстве многих измерений, что весьма нетривиально. Кроме того, данные могут носить нечисловой характер.

Применяя процедуры кластерного анализа, следует помнить, что:

1) Многие методы кластерного анализа – довольно простые процедуры, которые, как правило, не имеют достаточного статистического обоснования (то есть **большинство методов являются эвристическими**).

2) Методы кластерного анализа разрабатывались для многих дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.

3) Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.

4) Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, и эта структура может не совпадать с искомой «реальной».

Кластерный анализ, в отличие от других статистических методов, имеет несколько преимуществ:

1. С помощью кластерного анализа задачи распознавания образов решаются простыми и логичными методами, причем алгоритмы этих решений легко формализуются в компьютерные программы;

2. В кластерном анализе существуют методы комплексного изучения показателей различных типов данных (интервальных данных, частот, бинарных данных), при этом, не накладываются ограничения на представление исследуемых объектов;

3. Простота и доступность процедур кластерных методов распознавания образов позволяет сосредоточить внимание исследователя на содержании сложных многофакторных объектов;

4. Методы кластерного анализа позволяют накапливать знания с помощью информации, полученной в результате каждого эксперимента (измерения), выполненного в ходе использования кластерной модели. При этом характеристики

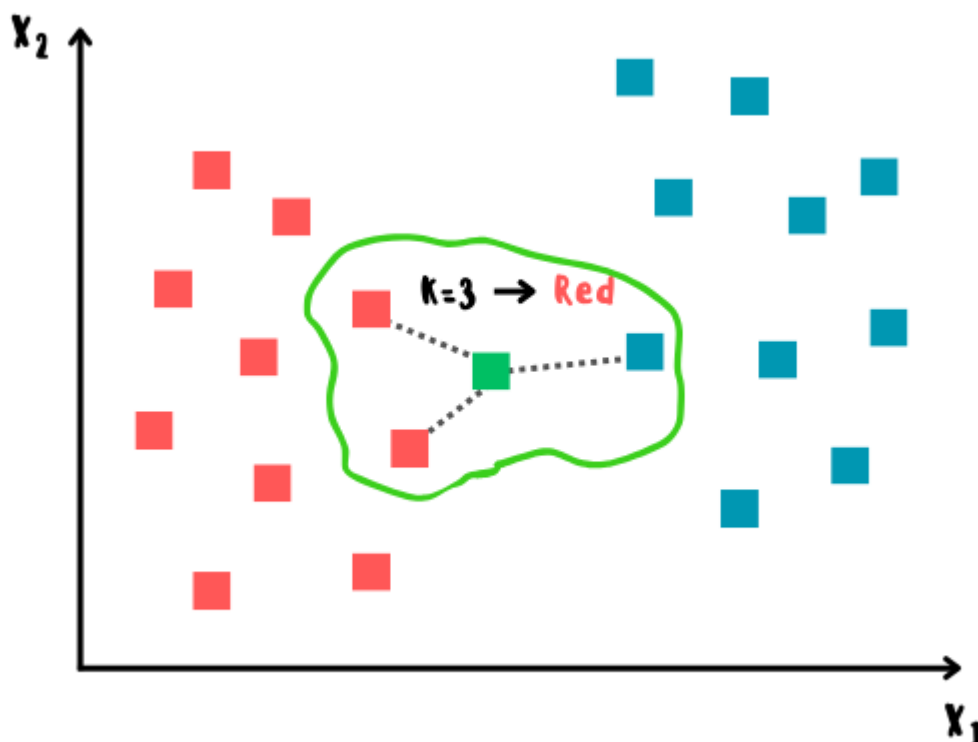
6.2. Метод k ближайших соседей

К-ближайших соседей (K-Nearest Neighbors или просто KNN) — алгоритм классификации и регрессии, основанный на гипотезе компактности, которая предполагает, что расположенные близко друг к другу объекты в пространстве признаков имеют схожие значения целевой переменной или принадлежат к одному классу.

Принцип работы KNN.

Алгоритм строится следующим образом:

- 1) сначала вычисляется расстояние между тестовым и всеми обучающими образцами;
- 2) далее из них выбирается k-ближайших образцов (соседей), где число k задаётся заранее;
- 3) итоговым прогнозом среди выбранных k-ближайших образцов будет мода в случае классификации и среднее арифметическое в случае регрессии;
- 4) предыдущие шаги повторяются для всех тестовых образцов.



6.3 Метрики

Существует множество метрик для вычисления расстояния между объектами.

Исходными данными для анализа могут быть собственно объекты и их параметры. Данные для анализа могут быть также представлены матрицей расстояний между объектами, в которой на пересечении строки с номером i и столбца с номером j записано расстояние между i -м и j -м объектом.

Если расстояния не даны сразу, то агломеративные алгоритмы начинаются с вычисления расстояний между объектами.

Переход от объектов к расстояниям между объектами – важный момент.

Расстояние между объектами – одна из мер сходства.

Интуитивно понятно, что, чем меньше расстояние между объектами, тем они более схожи. Но как выбрать естественную метрику, то есть, как естественно для данной задачи измерить расстояние между объектами?

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания.

Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько.

Наиболее распространенный способ – вычисление **евклидова расстояния** между двумя точками i и j на плоскости, когда известны их координаты X и Y :

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

Манхэттенское расстояние, или «расстояние городских кварталов». В этом случае просто берутся абсолютные значения по координатным расстояниям и суммируются. Свое название эта метрика получила из-за того, что моделирует расстояние, пройденное человеком в городе, когда перемещаться можно только по улицам, и нельзя, например, пересечь квартал по диагонали. Аналогия в декартовой плоскости приводит к перемещениям только по линиям, параллельным осям координат, и соответственно, к манхэттенскому расстоянию, которое вычисляют по формуле

$$(x,y) = \sum |x_i - y_i|$$

Метрика Чебышева. Максимальная метрика, также известная как расстояние Чебышева, рассматривает наибольшую разницу среди всех разниц по осям координат.

Расстояние, с использованием данной метрики определяется

$$(x,y) = \max |x_i - y_i|.$$

Расстояние Минковского (метрика Минковского) – параметрическая метрика для измерения расстояния между двумя точками в нормированном векторном пространстве. Обобщает евклидово расстояние и расстояние городских кварталов, введя параметр p , который определяет чувствительность к различиям по измерениям. Названа в честь немецкого математика Германа Минковского, впервые систематически изучившего это семейство функций расстояния

Расстояние Минковского порядка p между двумя точками принадлежащих пространству R^n определяется

$$\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

где: p – положительное вещественное число; $|x_i - y_i|$ — абсолютная разница между i -й координатами двух точек.

Специальные случаи:

$p = 1$ – расстояние Манхэттена (расстояние городских кварталов);

$p = 2$ – евклидово расстояние;

$p \rightarrow \infty$ – расстояние Чебышёва (максимальная абсолютная разность между координатами).

Свойства

Для $p \geq 1$ расстояние Минковского – метрика вследствие неравенства Минковского.

Для $p < 1$ расстояние не является метрикой, поскольку нарушается неравенство треугольника.

По мере увеличения p расстояние Минковского приближается к расстоянию Чебышёва, что означает, что только наибольшая разница координат доминирует в общем измерении расстояния.

Расстояние Махаланобиса между двумя точками — мера расстояния между двумя случайными точками U и V , одна из которых может (или обе могут) принадлежать некоторому классу C с матрицей ковариаций COV :

$$d_M(U, V, COV^{-1}) = \sqrt{(U - V) \cdot COV^{-1} \cdot (U - V)^T}$$

Символ T означает операцию транспонирования, а под COV^{-1} подразумевается матрица, обратная ковариационной.

Если матрица ковариаций является единичной матрицей, то расстояние Махаланобиса становится равным расстоянию Евклида. Иначе говоря, если класс представляет собой упорядоченное множество нормированных (дисперсии равны 1) независимых (ковариации равны 0) точек, то расстояние Махаланобиса равно расстоянию Евклида.

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.

Пусть даны две выборки $x^m = (x_1, \dots, x_m)$, $y^m = (y_1, \dots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}} = \frac{cov(x, y)}{\sqrt{s_x^2 s_y^2}},$$

где $\bar{x}, \bar{y}$ – выборочные средние x^m и y^m , s_x^2, s_y^2 – выборочные дисперсии, $r_{xy} \in [-1, 1]$.

Коэффициент корреляции Пирсона называют также теснотой линейной связи:

- $|r_{xy}| = 1 \Rightarrow x, y$ линейно зависимы,
- $r_{xy} = 0 \Rightarrow x, y$ линейно независимы.

6.4. Пример выполнения кластеризации.

Осуществим кластеризацию набора данных «Ирисы Фишера».

Выполнение кластеризации будем осуществлять с помощью языка программирования python. Для выполнения кластеризации и использования методов визуализации, нам потребуется следующие пакеты:

- Numpy;
- Pandas;
- Matplotlib.pyplot;
- Seaborn;
- Scikit-learn;
- Scikit-learn-extra;
- Scipy.

Пример команды для установки необходимых библиотек:

```
pip install numpy pandas matplotlib seaborn scikit-learn scikit-learn-extra scipy
```

Программный код с пояснениями представлен в файле «Код программы.ру», набор данных представлен в файле «iris.data.txt».

Выполнив данный код, мы получим следующий результат:



Рисунок 1. Результат выполнения программы.

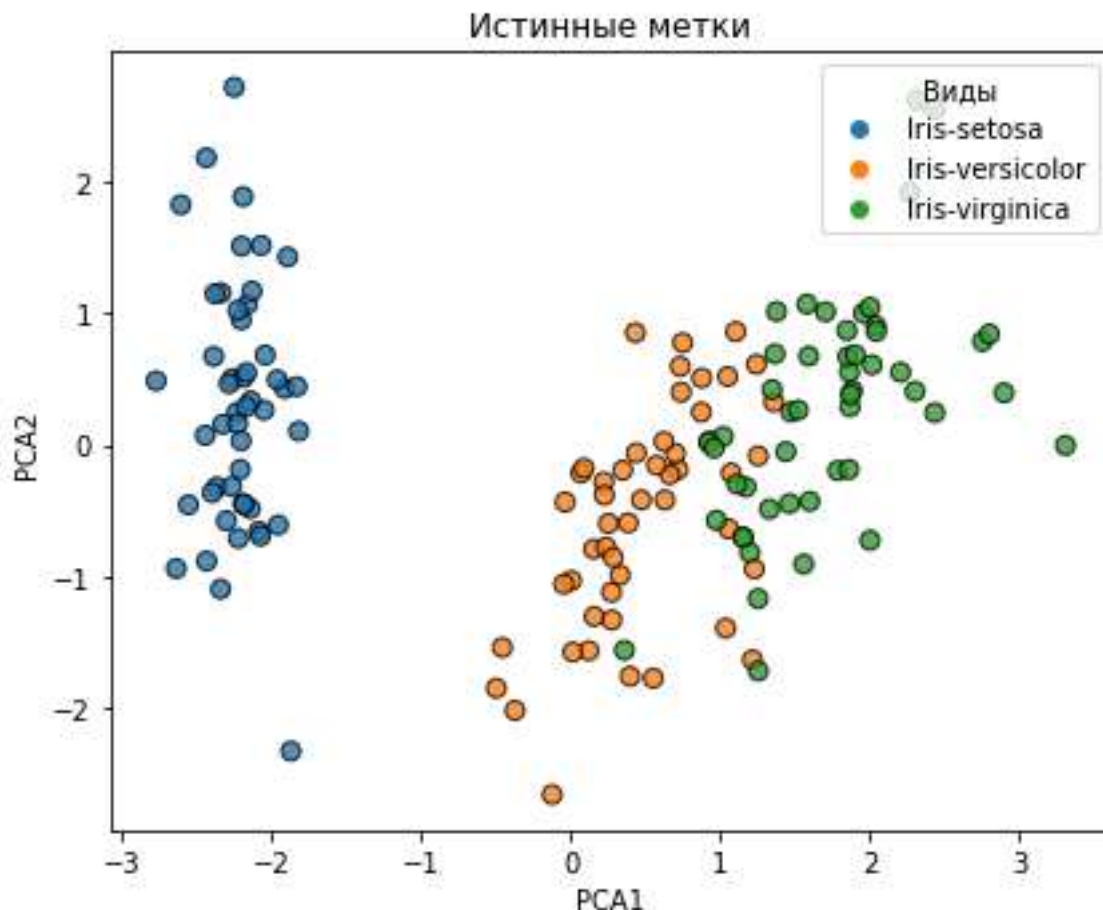


Рисунок 2. Истинная принадлежность.

Как мы можем наблюдать, при использовании всех признаков мы получили достаточно хорошую точность при использовании Манхэттановской метрики (88%). Однако, мы можем улучшить данный результат. Для этого нам необходимо использовать корреляционный анализ или методы дисперсионного анализа.

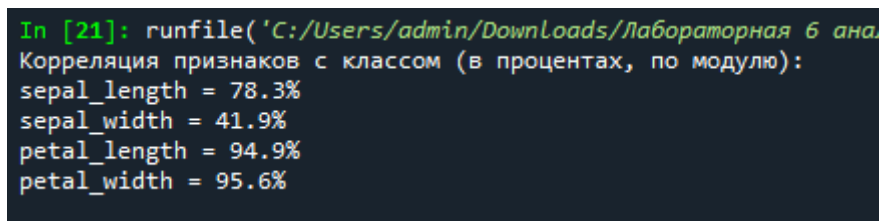
Пример программы, которая производит оценку корреляции методом Пирсона:

```
import pandas as pd
import numpy as np
from scipy.stats import pearsonr
# 1. Загрузка данных из файла
columns = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
df = pd.read_csv('C:/Users/admin/Downloads/iris.data.txt', names=columns)
# 2. Преобразование классов в числовые метки
class_labels, class_names = pd.factorize(df['class']) # class_labels: 0,1,2; class_names:
названия
```

```

df['class_numeric'] = class_labels
# 3. Имена признаков
feature_columns = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']
# 4. Вычисление корреляции для каждого признака с классом
print("Корреляция признаков с классом (в процентах, по модулю):")
for feature in feature_columns:
    # Коэффициент корреляции Пирсона между признаком и числовой меткой класса
    corr, _ = pearsonr(df[feature], df['class_numeric'])
    # Берём модуль и переводим в проценты
    corr_percent = abs(corr) * 100
    print(f"{feature} = {corr_percent:.1f}%")

```



```

In [21]: runfile('C:/Users/admin/Downloads/Лабораторная 6 ана
Корреляция признаков с классом (в процентах, по модулю):
sepal_length = 78.3%
sepal_width = 41.9%
petal_length = 94.9%
petal_width = 95.6%

```

Рисунок 3. Результат поиска корреляции признаков.

Из полученного результата очевидно, что повысить качество кластеризации возможно используя только два признака – длина и ширина лепестка ириса.

Однако, для подтверждения результата, воспользуемся η^2 – долей дисперсии признака, объясняемой принадлежностью к классу, что интерпретируется как мера силы связи между количественным признаком и категориальной переменной (классом).

Код программы:

```

import pandas as pd
import numpy as np

# 1. Загрузка данных
columns = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
df = pd.read_csv('C:/Users/admin/Downloads/iris.data.txt', names=columns)

# 2. Имена признаков
feature_columns = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']

```

3. Общая сумма квадратов (Total Sum of Squares)

```
def eta_squared(feature_series, class_series):
```

```
    """
```

Вычисляет η^2 для одного признака относительно категориальной переменной.

$$\eta^2 = \text{SSB} / \text{SST},$$

где:

$$\text{SST} = \text{total sum of squares} = \sum (x_i - \text{grand_mean})^2$$

$$\text{SSB} = \text{between-group sum of squares} = \sum n_j * (\text{mean}_j - \text{grand_mean})^2$$

```
    """
```

```
    grand_mean = feature_series.mean()
```

```
    SST = ((feature_series - grand_mean) ** 2).sum()
```

```
    SSB = 0
```

```
    for class_label in class_series.unique():
```

```
        group = feature_series[class_series == class_label]
```

```
        n_j = len(group)
```

```
        mean_j = group.mean()
```

```
        SSB += n_j * (mean_j - grand_mean) ** 2
```

```
    if SST == 0:
```

```
        return 0.0
```

```
    return SSB / SST
```

4. Вычисление и вывод η^2 для каждого признака

```
print("Сила связи признаков с классом ( $\eta^2$  в процентах):")
```

```
for feature in feature_columns:
```

```
    eta2 = eta_squared(df[feature], df['class'])
```

```
    eta2_percent = eta2 * 100
```

```
    print(f'{feature} = {eta2_percent:.1f}%')
```

```
In [22]: runfile('C:/Users/admin/Downloads/Лабораторная 6 анализ данных/Код 1.py')
Сила связи признаков с классом ( $\eta^2$  в процентах):
sepal_length = 61.9%
sepal_width = 39.2%
petal_length = 94.1%
petal_width = 92.9%
```

Рисунок 4. Результат выполнения программы.

Таким образом, мы подтвердили гипотезу о том, что некоторые признаки имеют тесную связь с классами (сортами ириса, в примере). Проведем повторную кластеризацию с использованием только указанных признаков.

```
In [23]: runfile('C:/Users/admin/Downloads/Лабораторная 6 анализ данных/Код 1.py')
Данные успешно загружены
Обнаружены классы: ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']

Доступные признаки:
[0] sepal_length
[1] sepal_width
[2] petal_length
[3] petal_width

Введите номера признаков через запятую (например: 2,3): 2,3

Введите количество кластеров (рекомендуется 2-5): 3

Выбраны признаки: petal_length, petal_width
Число кластеров: 3
Кластеризация с метрикой: euclidean...
Кластеризация с метрикой: manhattan (Minkowski p=1)...
Кластеризация с метрикой: chebyshev...
Кластеризация с метрикой: minkowski (p=3)...
Кластеризация с метрикой: mahalanobis...
Кластеризация с метрикой: pearson correlation...

=== Результаты кластеризации ===
      Metric  Accuracy
manhattan  euclidean      0.96
manhattan (Minkowski p=1)      0.96
      chebyshev      0.97
      minkowski (p=3)      0.96
      mahalanobis      0.85
      pearson correlation      0.33
```

Рисунок 5. Результат выполнения программы.

Как видно из рисунка, удалось повысить точность на 9 процентов.

Задания для выполнения лабораторной работы

Требуется выполнить кластеризацию данных, согласно варианту. В качестве метрик необходимо использовать расстояние Махаланобиса, Евклидово расстояние, Коэффициент корреляции Пирсона, расстояние Минковского, метрику Чебышева. Получить оценку точности кластеризации.

Проранжировать и определить наилучшую метрику.

Провести поиск корреляции между признаками и целевыми классами. При наличии физического смысла в используемых признаках – пояснить их связь с классом.

Повторно произвести кластеризацию данных с использованием того же набора метрик, проранжировать их и определить лучшую метрику.

При выполнении работы обратите внимание на тип файла с данными, наличие или отсутствие заголовков и дополнительной информации в наборе данных.

Вариант 1.

В файле (Вариант 1.xlsx) представлены наблюдения за параметрами фиников. В качестве признаков использованы размер, ширина и высота. Данные размечены по двум классам: сорт Ajwa и сорт Medjool.

Дополнительную информацию о наборе данных можно получить по ссылке: <http://archive.ics.uci.edu/dataset/879/ajwa+or+medjool>

Вариант 2.

В файле (Вариант 2) представлен набор данных дистанционного зондирования высокого разрешения. В качестве признаков представлены средняя текстура GLCM, среднее значение зеленого канала, ближнего инфракрасного и стандартное отклонение SD_Pan. Данные размечены по двум классам: большие деревья (w) и остальной растительный покров (n).

Дополнительную информацию о наборе данных можно получить по ссылке: <http://archive.ics.uci.edu/dataset/285/wilt>

Вариант 3.

В файле (Вариант 3) представлен набор данных о состоянии знаний студентов по предмету «Электрические машины постоянного тока». В

качестве признаков использованы: процент времени изучения материалов; процент количества повторений пользователем материалов; процент времени изучения пользователем материала связанного с дисциплиной, результаты экзамена пользователя по смежным дисциплинам, результаты экзамена пользователя по дисциплине «Электрические машины постоянного тока». Данные размечены по следующим классам: очень низкий, низкий, средний, высокий коэффициент освоения дисциплины.

Дополнительную информацию о наборе данных можно получить по ссылке: <http://archive.ics.uci.edu/dataset/257/user+knowledge+modeling>

Вариант 4.

В файле (Вариант 4) представлен набор данных оценки эффективности преподавания. В качестве признаков использованы: является ли преподаватель носителем английского языка (1 = носитель английского языка, 2 = не носитель английского языка); преподаватель курса (25 категорий); направление обучения (26 категорий); семестр; размер группы. Данные размечены по следующим классам (последний столбец): низкий (1), средний (2), высокий (3).

Дополнительную информацию о наборе данных можно получить по ссылке: <https://archive.ics.uci.edu/dataset/100/teaching+assistant+evaluation>

Вариант 5.

В файле (Вариант 5) представлен набор данных рисовых зерен двух видов. В качестве признаков использованы: область пикселей изображения риса; периметр риса; наибольшая длина главной оси; наименьшая длина главной оси; эксцентриситет; выпуклость зерна; отношение площади, образованной рисовым зерном, к ограничивающему прямоугольнику. Данные размечены по следующим классам: сорт Cammeo, сорт Osmancik.

Дополнительную информацию о наборе данных можно получить по ссылке: <http://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik>

Теоретические вопросы:

- 1. Что такое кластерный анализ? Приведите пример.**
- 2. Приведите примеры метрик, используемых в кластерном анализе.**
- 3. Как связаны между собой метрики Минковского, Манхэттена, Евклида и Чебышёва? Почему при $p \rightarrow \infty$ расстояние Минковского стремится к расстоянию Чебышёва?**
- 4. Как интерпретировать значение η^2 (эта-квадрат) в контексте связи количественного признака с категориальной переменной (классом)? Какие пороговые значения указывают на слабую, среднюю и сильную связь?**
- 5. Могут ли разные метрики приводить к принципиально разным результатам кластеризации на одних и тех же данных? Обоснуйте ответ с примером**

Приложения

Приложение 1

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	39986	3984	3982	3980	3977	3973
0,1	0,3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	0,3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	0,3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	0,3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	0,3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	0,3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	0,3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	0,2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	0,2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	0,2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	0,1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	0,1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	0,1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	0,1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	0,1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0,0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0,0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0,0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0,0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0,0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0,0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0,0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0,0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0,0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0,0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0,0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0,0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0,0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0,0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0,0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0,0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0,0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0,0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0,0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0,0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0,0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

$$\text{Значения функции } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x^2}{2}} dx$$

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0,00	0,0000	0,52	0,1985	1,04	0,3508	1,56	0,4406	2,16	0,4846
0,01	0,0040	0,53	0,2019	1,05	0,3531	1,57	0,4418	2,18	0,4854
0,02	0,0080	0,54	0,2054	1,06	0,3554	1,58	0,4429	2,20	0,4861
0,03	0,0120	0,55	0,2088	1,07	0,3577	1,59	0,4441	2,22	0,4868
0,04	0,0160	0,56	0,2123	1,08	0,3599	1,60	0,4452	2,24	0,4875
0,05	0,0199	0,57	0,2157	1,09	0,3621	1,61	0,4463	2,26	0,4881
0,06	0,0239	0,58	0,2190	1,10	0,3643	1,62	0,4474	2,28	0,4887
0,07	0,0279	0,59	0,2224	1,11	0,3665	1,63	0,4484	2,30	0,4893
0,08	0,0319	0,60	0,2257	1,12	0,3686	1,64	0,4495	2,32	0,4898
0,09	0,0359	0,61	0,2291	1,13	0,3708	1,65	0,4505	2,34	0,4904
0,10	0,0398	0,62	0,2324	1,14	0,3729	1,66	0,4515	2,36	0,4908
0,11	0,0438	0,63	0,2357	1,15	0,3749	1,67	0,4525	2,38	0,4913
0,12	0,0478	0,64	0,2389	1,16	0,3770	1,68	0,4535	2,40	0,4918
0,13	0,0517	0,65	0,2422	1,17	0,3790	1,69	0,4545	2,42	0,4922
0,14	0,0557	0,66	0,2454	1,18	0,3810	1,70	0,4554	2,44	0,4927
0,15	0,0596	0,67	0,2486	1,19	0,3830	1,71	0,4564	2,46	0,4931
0,16	0,0636	0,68	0,2517	1,20	0,3849	1,72	0,4573	2,48	0,4934
0,17	0,0675	0,69	0,2549	1,21	0,3869	1,73	0,4582	2,50	0,4938
0,18	0,0714	0,70	0,2580	1,22	0,3888	1,74	0,4591	2,52	0,4941
0,19	0,0753	0,71	0,2611	1,23	0,3907	1,75	0,4599	2,54	0,4945
0,20	0,0793	0,72	0,2642	1,24	0,3925	1,76	0,4608	2,56	0,4948
0,21	0,0832	0,73	0,2673	1,25	0,3914	1,77	0,4616	2,58	0,4951
0,22	0,0871	0,74	0,2703	1,26	0,3962	1,78	0,4625	2,60	0,4953
0,23	0,0910	0,75	0,2734	1,27	0,3980	1,79	0,4633	2,62	0,4956
0,24	0,0948	0,76	0,2764	1,28	0,3997	1,80	0,4641	2,64	0,4959
0,25	0,0987	0,77	0,2794	1,29	0,4015	1,81	0,4649	2,66	0,4961
0,26	0,1026	0,78	0,2823	1,30	0,4032	1,82	0,4656	2,68	0,4963
0,27	0,1064	0,79	0,2852	1,31	0,4049	1,83	0,4664	2,70	0,4965
0,28	0,1103	0,80	0,2881	1,32	0,4066	1,84	0,4671	2,72	0,4967
0,29	0,1141	0,81	0,2910	1,33	0,4082	1,85	0,4678	2,74	0,4969
0,30	0,1179	0,82	0,2939	1,34	0,4099	1,86	0,4686	2,76	0,4971
0,31	0,1217	0,83	0,2967	1,35	0,4115	1,87	0,4693	2,78	0,4973
0,32	0,1255	0,84	0,2995	1,36	0,4131	1,88	0,4699	2,80	0,4974
0,33	0,1293	0,85	0,3023	1,37	0,4147	1,89	0,4706	2,82	0,4976
0,34	0,1331	0,86	0,3051	1,38	0,4162	1,90	0,4713	2,84	0,4977
0,35	0,1368	0,87	0,3078	1,39	0,4177	1,91	0,4719	2,86	0,4979
0,36	0,1406	0,88	0,3106	1,40	0,4192	1,92	0,4726	2,88	0,4980
0,37	0,1443	0,89	0,3133	1,41	0,4207	1,93	0,4732	2,90	0,4981
0,38	0,1480	0,90	0,3159	1,42	0,4222	1,94	0,4738	2,92	0,4982
0,39	0,1517	0,91	0,3186	1,43	0,4236	1,95	0,4744	2,94	0,4984
0,40	0,1554	0,92	0,3212	1,44	0,4251	1,96	0,4750	2,96	0,4985
0,41	0,1591	0,93	0,3238	1,45	0,4265	1,97	0,4756	2,98	0,4986
0,42	0,1628	0,94	0,3264	1,46	0,4279	1,98	0,4761	3,00	0,49865
0,43	0,1664	0,95	0,3289	1,47	0,4292	1,99	0,4767	3,20	0,49931
0,44	0,1700	0,96	0,3315	1,48	0,4306	2,00	0,4772	3,40	0,49966
0,45	0,1736	0,97	0,3340	1,49	0,4319	2,02	0,4783	3,60	0,499841
0,46	0,1772	0,98	0,3365	1,50	0,4332	2,04	0,4793	3,80	0,499928
0,47	0,1808	0,99	0,3389	1,51	0,4345	2,06	0,4803	4,00	0,499968
0,48	0,1844	1,00	0,3413	1,52	0,4357	2,08	0,4812	4,50	0,499997
0,49	0,1879	1,01	0,3438	1,53	0,4370	2,10	0,4821	5,00	0,4999997
0,50	0,1915	1,02	0,3461	1,54	0,4382	2,12	0,4830		
0,51	0,1950	1,03	0,3485	1,55	0,4394	2,14	0,4838		

Приложение 3

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

$n \backslash \gamma$	0,95	0,99	0,999	$n \backslash \gamma$	0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,361	3,883
6	2,57	4,03	6,86	25	2,064	2,297	3,745
7	2,45	3,71	5,96	30	2,045	2,256	3,659
8	2,37	3,50	5,41	35	2,032	2,220	3,600
9	2,31	3,36	5,04	40	2,023	2,208	3,558
10	2,26	3,25	4,78	45	2,016	2,192	3,527
11	2,23	3,17	4,59	50	2,009	2,179	3,502
12	2,20	3,11	4,44	60	2,001	2,166	3,464
13	2,18	3,06	4,32	70	1,996	2,149	3,439
14	2,16	3,01	4,22	80	1,991	2,140	3,418
15	2,15	2,98	4,14	90	1,987	2,133	3,403
16	2,13	2,95	4,07	100	1,984	2,127	3,392
17	2,12	2,92	4,02	120	1,980	2,117	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Приложение 4

Таблица значений $q_\gamma = q(\gamma, n)$

$n \backslash \gamma$	0,95	0,99	0,999	$n \backslash \gamma$	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение 5

Критические точки распределения χ^2

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,98
1	6,6	5,024	3,841	0,0039	0,00098	0,00016
2	9,2	7,378	5,991	0,103	0,051	0,020
3	11,3	9,348	7,815	0,352	0,216	0,115
4	13,3	11,143	9,488	0,711	0,484	0,297
5	15,1	12,832	11,070	1,15	0,831	0,554
6	16,8	14,449	12,592	1,64	1,24	0,872
7	18,5	16,013	14,067	2,17	1,69	1,24
8	20,1	17,535	15,507	2,73	2,18	1,65
9	21,7	19,023	16,919	3,33	2,70	2,09
10	23,2	20,483	18,307	3,94	3,25	2,56
11	24,7	21,920	19,676	4,57	3,82	3,05
12	26,2	23,336	21,026	5,23	4,40	3,57
13	27,7	24,736	22,362	5,89	5,01	4,11
14	29,1	26,129	23,685	6,57	5,63	4,66
15	30,6	27,488	24,996	7,26	6,26	5,23
16	32,0	28,845	26,296	7,96	6,91	5,81
17	33,4	30,191	27,587	8,67	7,56	6,41
18	34,8	31,536	28,869	9,39	8,23	7,01
19	36,2	32,852	30,144	10,1	8,91	7,63
20	37,6	34,170	31,410	10,9	9,59	8,26
21	38,9	35,479	32,671	11,6	10,3	8,90
22	40,3	36,781	33,924	12,3	11,0	9,54
23	41,6	38,076	35,172	13,1	11,7	10,2
24	43,0	39,364	36,415	13,8	12,4	10,9
25	44,3	40,646	37,652	14,6	13,1	11,5
26	45,6	41,923	38,885	15,4	13,8	12,2
27	47,0	43,194	40,113	16,2	14,6	12,9
28	48,3	44,461	41,337	16,9	15,3	13,6
29	49,6	45,722	42,557	17,7	16,0	14,3
30	50,9	46,979	43,773	18,5	16,8	15,0

Критические точки распределения Фишера — Снедекора

 $(k_1$ — число степеней свободы большей дисперсии, k_2 — число степеней свободы меньшей дисперсии)

Уровень значимости $\alpha = 0,01$												
$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	90,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40	99,41	99,42
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,86	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45
Уровень значимости $\alpha = 0,05$												
$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

Приложение 8

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,42	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
по организации самостоятельной работы студентов по
дисциплине «Анализ данных»

для студентов направления подготовки
09.03.01 Информатика и вычислительная техника

(ЭЛЕКТРОННЫЙ ДОКУМЕНТ)

Ставрополь

Методические указания по организации самостоятельной работы студентов составлено в соответствии с требованиями программы дисциплины «Анализ данных» для бакалавров направления подготовки 09.03.01 Информатика и вычислительная техника. Учебно-методическое пособие посвящено планируемой учебной работе студентов, выполняемой во внеаудиторное (аудиторное) время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия, при частичном, непосредственном участии преподавателя.

СОДЕРЖАНИЕ

1. Самостоятельная работа как важнейшая форма учебного процесса
2. Цели и основные задачи СРС
3. Виды самостоятельной работы
4. Организация СРС
5. Общие рекомендации по организации самостоятельной работы
6. Самостоятельная работа студента - необходимое звено становления исследователя
7. Методические рекомендации для студентов по отдельным формам самостоятельной работы
8. Учебно-методическое и информационное обеспечение дисциплины

1. Самостоятельная работа как важнейшая форма учебного процесса

Самостоятельная работа - планируемая учебная, учебно-исследовательская, научно-исследовательская работа студентов, выполняемая во внеаудиторное (аудиторное) время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия (при частичном непосредственном участии преподавателя, оставляющем ведущую роль за работой студентов).

Самостоятельная работа студентов в ВУЗе является важным видом учебной и научной деятельности студента. Самостоятельная работа студентов играет значительную роль в рейтинговой технологии обучения. Обучение в ВУЗе включает в себя две, практически одинаковые по объему и взаимовлиянию части – процесса обучения и процесса самообучения. Поэтому СРС должна стать эффективной и целенаправленной работой студента.

Концепцией модернизации российского образования определены основные задачи профессионального образования - "подготовка квалифицированного работника соответствующего уровня и профиля, конкурентоспособного на рынке труда, компетентного, ответственного, свободно владеющего своей профессией и ориентированного в смежных областях деятельности, способного к эффективной работе по специальности на уровне мировых стандартов, готового к постоянному профессиональному росту, социальной и профессиональной мобильности".

Решение этих задач невозможно без повышения роли самостоятельной работы студентов над учебным материалом, усиления ответственности преподавателей за развитие навыков самостоятельной работы, за стимулирование профессионального роста студентов, воспитание творческой активности и инициативы.

К современному специалисту общество предъявляет достаточно широкий перечень требований, среди которых немаловажное значение имеет наличие у выпускников определенных способностей и умения самостоятельно добывать знания из различных источников, систематизировать полученную информацию, давать оценку конкретной финансовой ситуации. Формирование такого умения происходит в течение всего периода обучения через участие студентов в лабораторных занятиях, выполнение контрольных заданий и тестов, написание курсовых и выпускных квалификационных работ. При этом самостоятельная работа студентов играет решающую роль в ходе всего учебного процесса.

Формы самостоятельной работы студентов разнообразны. По данной дисциплине «Анализ данных» предусмотрены следующие:

Использование материала учебно-методического комплекса дисциплины

На первом этапе необходимо ознакомиться с обязательным минимумом содержания основной образовательной программы по изучаемой дисциплине. Далее необходимо изучить рабочую программу дисциплины, в которой рассмотрено содержание тем дисциплины лекционного курса, взаимосвязь тем лекций с лабораторных занятий, темы и виды самостоятельной работы. По каждому виду самостоятельной работы предусмотрены определённые формы отчетности, представленные в учебной программе по дисциплине «Основы теории управления».

Работа с литературой

Для успешного освоения дисциплины, необходимо самостоятельно детально изучить представленные темы по рекомендуемым источникам информации, которые представлены в учебной программе.

Самостоятельная работа приобщает студентов к научному творчеству, поиску и решению актуальных современных проблем.

2. Цели и основные задачи СРС

Ведущая цель организации и осуществления СРС должна совпадать с целью обучения студента – подготовкой бакалавра с высшим образованием. При организации СРС важным и необходимым условием становятся формирование умения самостоятельной работы для приобретения знаний, навыков и возможности организации учебной и научной деятельности.

Целью самостоятельной работы студентов является овладение фундаментальными знаниями, профессиональными умениями и навыками деятельности по профилю, опытом творческой, исследовательской деятельности. Самостоятельная работа студентов способствует развитию самостоятельности, ответственности и организованности, творческого подхода к решению проблем учебного и профессионального уровня.

Задачами СРС являются:

- систематизация и закрепление полученных теоретических знаний и практических умений студентов;
- углубление и расширение теоретических знаний;
- формирование умений использовать нормативную, правовую, справочную документацию и специальную литературу;

- развитие познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности;
- формирование самостоятельности мышления, способностей к саморазвитию, самосовершенствованию и самореализации;
- развитие исследовательских умений;
- использование материала, собранного и полученного в ходе самостоятельных занятий на семинарах, на лабораторных занятиях, при написании курсовых и выпускной квалификационной работ, для эффективной подготовки к итоговым зачетам и экзаменам.

3. Виды самостоятельной работы

В образовательном процессе высшего профессионального образовательного учреждения выделяется два вида самостоятельной работы – аудиторная, под руководством преподавателя, и внеаудиторная. Тесная взаимосвязь этих видов работ предусматривает дифференциацию и эффективность результатов ее выполнения и зависит от организации, содержания, логики учебного процесса (межпредметных связей, перспективных знаний и др.):

Аудиторная самостоятельная работа по дисциплине выполняется на учебных занятиях под непосредственным руководством преподавателя и по его заданию.

Внеаудиторная самостоятельная работа выполняется студентом по заданию преподавателя, но без его непосредственного участия.

Основными видами самостоятельной работы студентов без участия преподавателя по данной дисциплине являются:

- формирование и усвоение содержания конспекта лекций на базе рекомендованной лектором учебной литературы, включая информационные образовательные ресурсы (электронные учебники, электронные библиотеки и др.);
- подготовка к лабораторным работам, их оформление.

Основными видами самостоятельной работы студентов с участием преподавателей являются:

- текущие консультации;
- прием и защита лабораторных работ (во время проведения л/р).

4. Организация СРС

Методика организации самостоятельной работы студентов зависит от структуры, характера и особенностей дисциплины «Основы теории управления», объема часов на ее изучение, вида заданий для самостоятельной работы

студентов, индивидуальных качеств студентов и условий учебной деятельности.

Процесс организации самостоятельной работы студентов включает в себя следующие этапы:

- подготовительный (определение целей, составление программы, подготовка методического обеспечения, подготовка оборудования);
- основной (реализация программы, использование приемов поиска информации, усвоения, переработки, применения, передачи знаний, фиксирование результатов, самоорганизация процесса работы);
- заключительный (оценка значимости и анализ результатов, их систематизация, оценка эффективности программы и приемов работы, выводы о направлениях оптимизации труда).

Деятельность студентов по формированию и развитию навыков учебной самостоятельной работы

В процессе самостоятельной работы студент приобретает навыки самоорганизации, самоконтроля, самоуправления, саморефлексии и становится активным самостоятельным субъектом учебной деятельности.

Выполняя самостоятельную работу под контролем преподавателя студент должен:

- освоить минимум содержания, выносимый на самостоятельную работу студентов и предложенный преподавателем в соответствии с образовательными стандартами высшего образования;
- планировать самостоятельную работу в соответствии с графиком самостоятельной работы, предложенным преподавателем;
- самостоятельную работу студент должен осуществлять в организационных формах, предусмотренных учебным планом и рабочей программой преподавателя;
- выполнять самостоятельную работу и отчитываться по ее результатам в соответствии с графиком представления результатов, видами и сроками отчетности по самостоятельной работе студентов.

студент может:

сверх предложенного преподавателем (при обосновании и согласовании с ним) и минимума обязательного содержания, определяемого программой по данной дисциплине:

- самостоятельно определять уровень (глубину) проработки содержания материала;

- предлагать темы и вопросы для самостоятельной проработки;

- в рамках общего графика выполнения самостоятельной работы предлагать обоснованный индивидуальный график выполнения и отчетности по результатам самостоятельной работы;

- предлагать свои варианты организационных форм самостоятельной работы;

- использовать для самостоятельной работы методические пособия, учебные пособия, разработки сверх предложенного преподавателем перечня;

- использовать не только контроль, но и самоконтроль результатов самостоятельной работы в соответствии с методами самоконтроля, предложенными преподавателем или выбранными самостоятельно.

Самостоятельная работа студентов должна оказывать важное влияние на формирование личности будущего бакалавра, она планируется студентом самостоятельно. Каждый студент самостоятельно определяет режим своей работы и меру труда, затрачиваемого на овладение учебным содержанием по каждой дисциплине. Он выполняет внеаудиторную работу по личному индивидуальному плану, в зависимости от его подготовки, времени и других условий.

5. Общие рекомендации по организации самостоятельной работы

Основной формой самостоятельной работы студента является изучение конспекта лекций, их дополнение, рекомендованной литературы, активное участие на лабораторных занятиях. Но для успешной учебной деятельности, ее интенсификации, необходимо учитывать следующие субъективные факторы:

1. Знание школьного программного материала, наличие прочной системы знаний, необходимой для усвоения основных вузовских курсов. Это особенно важно для математических дисциплин. Необходимо отличать пробелы в знаниях, затрудняющие усвоение нового материала, от малых способностей. Затратив силы на преодоление этих пробелов, студент обеспечит себе нормальную успеваемость и поверит в свои способности.

2. Наличие умений, навыков умственного труда:

- а) умение конспектировать на лекции и при работе с книгой;

- б) владение логическими операциями: сравнение, анализ, синтез, обобщение, определение понятий, правила систематизации и классификации.

3. Специфика познавательных психических процессов: внимание, память, речь, наблюдательность, интеллект и мышление. Слабое развитие каждого из них становится серьезным препятствием в учебе.

4. Хорошая работоспособность, которая обеспечивается нормальным физическим состоянием. Ведь серьезное учение - это большой многосторонний и разнообразный труд. Результат обучения оценивается не количеством сообщаемой информации, а качеством ее усвоения, умением ее использовать и развитием у себя способности к дальнейшему самостоятельному образованию.

5. Соответствие избранной деятельности, профессии индивидуальным способностям. Необходимо выработать у себя умение саморегулировать свое эмоциональное состояние и устранять обстоятельства, нарушающие деловой настрой, мешающие намеченной работе.

6. Овладение оптимальным стилем работы, обеспечивающим успех в деятельности. Чередование труда и пауз в работе, периоды отдыха, индивидуально обоснованная норма продолжительности сна, предпочтение вечерних или утренних занятий, стрессоустойчивость на экзаменах и особенности подготовки к ним,

7. Уровень требований к себе, определяемый сложившейся самооценкой.

Адекватная оценка знаний, достоинств, недостатков - важная составляющая самоорганизации человека, без нее невозможна успешная работа по управлению своим поведением, деятельностью.

Одна из основных особенностей обучения в высшей школе заключается в том, что постоянный внешний контроль заменяется самоконтролем, активная роль в обучении принадлежит уже не столько преподавателю, сколько студенту.

Зная основные методы научной организации умственного труда, можно при наименьших затратах времени, средств и трудовых усилий достичь наилучших результатов.

Эффективность усвоения поступающей информации зависит от работоспособности человека в тот или иной момент его деятельности.

Работоспособность - способность человека к труду с высокой степенью напряженности в течение определенного времени. Различают внутренние и внешние факторы работоспособности.

К внутренним факторам работоспособности относятся интеллектуальные особенности, воля, состояние здоровья.

К внешним:

- организация рабочего места, режим труда и отдыха;
- уровень организации труда - умение получить справку и пользоваться информацией;

- величина умственной нагрузки.

Выдающийся русский физиолог Н. Е. Введенский выделил следующие условия продуктивности умственной деятельности:

- во всякий труд нужно входить постепенно;
- мерность и ритм работы. Разным людям присущ более или менее разный темп работы;
- привычная последовательность и систематичность деятельности;
- правильное чередование труда и отдыха.

Отдых не предполагает обязательного полного бездействия со стороны человека, он может быть достигнут простой переменой дела. В течение дня работоспособность изменяется. Наиболее плодотворным является *утреннее время* (с 8 до 14 часов), причем максимальная работоспособность приходится на период с 10 до 13 часов, затем *послеобеденное* - (с 16 до 19 часов) и *вечернее* (с 20 до 24 часов). Очень трудный для понимания материал лучше изучать в начале каждого отрезка времени (лучше всего утреннего) после хорошего отдыха. Через 1-1,5 часа нужны перерывы по 10 - 15 мин, через 3 - 4 часа работы отдых должен быть продолжительным - около часа.

Составной частью научной организации умственного труда является овладение техникой умственного труда.

Время, которым располагает студент для выполнения учебного плана, складывается из двух составляющих: одна из них - это аудиторная работа в вузе по расписанию занятий, другая - внеаудиторная самостоятельная работа. Задания и материалы для самостоятельной работы выдаются во время учебных занятий по расписанию, на этих же занятиях преподаватель осуществляет контроль за самостоятельной работой, а также оказывает помощь студентам по правильной организации работы.

Самостоятельные занятия потребуют интенсивного умственного труда, который необходимо не только правильно организовать, но и стимулировать. При этом очень важно уметь поддерживать устойчивое внимание к изучаемому материалу. Выработка внимания требует значительных волевых усилий. Именно поэтому, если студент замечает, что он часто отвлекается во время самостоятельных занятий, ему надо заставить себя сосредоточиться. Подобную процедуру необходимо проделывать постоянно, так как это является тренировкой внимания. Устойчивое внимание появляется тогда, когда человек относится к делу с интересом.

Следует правильно организовать свои занятия по времени: 50 минут - работа, 5-10 минут - перерыв; после 3 часов работы перерыв - 20-25 минут. Иначе нарастающее утомление повлечет неустойчивость внимания. Очень существен-

ным фактором, влияющим на повышение умственной работоспособности, являются систематические занятия физической культурой. Организация активного отдыха предусматривает чередование умственной и физической деятельности, что полностью восстанавливает работоспособность человека.

6. Самостоятельная работа студента - необходимое звено становления исследователя

Прогресс науки и техники, информационных технологий приводит к значительному увеличению научной информации, что предъявляет более высокие требования не только к моральным, нравственным свойствам человека, но и в особенности, постоянно возрастающие требования в области образования – обновление, модернизация общих и профессиональных знаний, умений специалиста.

Всякое образование должно выступать как динамический процесс, присущий человеку и продолжающийся всю его жизнь. Овладение научной мыслью и языком науки является необходимой составляющей в самоорганизации будущего специалиста исследователя. Под этим понимается не столько накопление знаний, сколько овладение научно обоснованными способами их приобретения. В этом, вообще говоря, состоит основная задача вуза.

Специфика вузовского учебного процесса, в организации которого самостоятельной работе студента отводятся все больше места, состоит в том, что он является как будто бы последним и самым адекватным звеном для реализации этой задачи. Ибо во время учебы в вузе происходит выработка стиля, навыков учебной (познавательной) деятельности, рациональный характер которых будет способствовать постоянному обновлению знаний высококвалифицированного выпускника вуза.

Однако до этом пути существуют определенные трудности, в частности, переход студента от синтетического процесса обучения в средней школе, к аналитическому в высшей. Это связано как с новым содержанием обучения (расширение общего образования и углубление профессиональной подготовки), так и с новыми, неизвестными до сих пор формами: обучения (лекции, семинары, лабораторные занятия и т.д.). Студент получает не только знания, предусмотренные программой и учебными пособиями, но он также должен познакомиться со способами приобретения знаний так, чтобы суметь оценить, что мы знаем, откуда мы это знаем и как этого знания мы достигли. Ко всему этому приходят через собственную самостоятельную работу.

Это и потому, что самостоятельно приобретенные знания являются более оперативными, они становятся личной собственностью, а также мотивом поведения, развивают интеллектуальные черты, внимание, наблюдательность, критичность, умение оценивать. Роль преподавателя в основном заключается в руководстве накопления знаний (по отношению к первокурсникам), а в последующие годы учебы, на старших курсах, в совместном установлении проблем и заботе о самостоятельных поисках студента, а также контролирования за их деятельностью. Отметим, что нельзя ограничиваться только приобретением знаний предусмотренных программой изучаемой дисциплины, надо постоянно углублять полученные знания, сосредотачивая их на какой-нибудь узкой определенной области, соответствующей интересам студента. Углубленное изучение всех предметов, предусмотренных программой, на практике является возможным, и хорошая организация работы позволяет экономить время, что создает условия для глубокого, систематического, заинтересованного изучения самостоятельно выбранной студентом темы.

Конечно, все советы, примеры, рекомендации в этой области, даваемые преподавателем, или определенными публикациями, или другими источниками, не гарантируют никакого успеха без проявления собственной активности в этом деле, т. е. они не дают готовых рецептов, а должны способствовать анализу собственной работы, ее целей, организации в соответствии с индивидуальными особенностями. Учитывая личные возможности, существующие условия жизни и работы, навыки, на основе этих рекомендаций, возможно, выработать индивидуально обоснованную совокупность методов, способов, найти свой стиль или усовершенствовать его, чтобы изучив определенный материал, иметь время оценить его значимость, пригодность и возможности его применения, чтобы, в конечном счете, обеспечить успешность своей учебы с будущей профессиональной деятельности

7. Методические рекомендации для студентов по отдельным формам самостоятельной работы

Система вузовского обучения подразумевает значительно большую самостоятельность студентов в планировании и организации своей деятельности.

Работа с литературой

При работе с литературой необходимо подобрать литературу, научиться правильно ее читать, вести записи. Для подбора литературы в библиотеке используются алфавитный и систематический каталоги.

Важно помнить, что рациональные навыки работы с книгой - это всегда большая экономия времени и сил.

Правильный подбор учебников рекомендуется преподавателем, читающим лекционный курс. Необходимая литература может быть также указана в методических разработках по данному курсу.

Изучая материал по учебнику, следует переходить к следующему вопросу только после правильного уяснения предыдущего, описывая на бумаге все выкладки и вычисления (в том числе те, которые в учебнике опущены или на лекции даны для самостоятельного вывода).

При изучении любой дисциплины большую и важную роль играет самостоятельная индивидуальная работа.

Особое внимание следует обратить на определение основных понятий курса. Студент должен подробно разбирать примеры, которые поясняют такие определения, и уметь строить аналогичные примеры самостоятельно. Нужно добиваться точного представления о том, что изучаешь. Полезно составлять опорные конспекты. При изучении материала по учебнику полезно в тетради (на специально отведенных полях) дополнять конспект лекций. Там же следует отмечать вопросы, выделенные студентом для консультации с преподавателем.

Выводы, полученные в результате изучения, рекомендуется в конспекте выделять, чтобы они при перечитывании записей лучше запоминались.

Опыт показывает, что многим студентам помогает составление листа опорных сигналов, содержащего важнейшие и наиболее часто употребляемые формулы и понятия. Такой лист помогает запомнить формулы, основные положения лекции, а также может служить постоянным справочником для студента.

Различают два вида чтения; первичное и вторичное. *Первичное* - это внимательное, неторопливое чтение, при котором можно остановиться на трудных местах. После него не должно остаться ни одного непонятого слова. Содержание не всегда может быть понятно после первичного чтения.

Задача *вторичного* чтения полное усвоение смысла целого (по счету это чтение может быть и не вторым, а третьим или четвертым).

Чтение научного текста является частью познавательной деятельности. Ее цель – извлечение из текста необходимой информации. От того на сколько осознанна читающим собственная внутренняя установка при обращении к печатному слову (найти нужные сведения, усвоить информацию полностью или частично, критически проанализировать материал и т.п.) во многом зависит эффективность осуществляемого действия.

Выделяют **четыре основные установки в чтении научного текста**:

1. информационно-поисковый (задача – найти, выделить искомую информацию)

2. усваивающая (усилия читателя направлены на то, чтобы как можно полнее осознать и запомнить как сами сведения излагаемые автором, так и всю логику его рассуждений)

3. аналитико-критическая (читатель стремится критически осмыслить материал, проанализировав его, определив свое отношение к нему)

4. творческая (создает у читателя готовность в том или ином виде – как отправной пункт для своих рассуждений, как образ для действия по аналогии и т.п. – использовать суждения автора, ход его мыслей, результат наблюдения, разработанную методику, дополнить их, подвергнуть новой проверке).

С наличием различных установок обращения к научному тексту связано существование и нескольких **видов чтения**:

1. библиографическое – просматривание карточек каталога, рекомендательных списков, сводных списков журналов и статей за год и т.п.;

2. просмотрное – используется для поиска материалов, содержащих нужную информацию, обычно к нему прибегают сразу после работы со списками литературы и каталогами, в результате такого просмотра читатель устанавливает, какие из источников будут использованы в дальнейшей работе;

3. ознакомительное – подразумевает сплошное, достаточно подробное прочтение отобранных статей, глав, отдельных страниц, цель – познакомиться с характером информации, узнать, какие вопросы вынесены автором на рассмотрение, провести сортировку материала;

4. изучающее – предполагает доскональное освоение материала; в ходе такого чтения проявляется доверие читателя к автору, готовность принять изложенную информацию, реализуется установка на предельно полное понимание материала;

5. аналитико-критическое и творческое чтение – два вида чтения близкие между собой тем, что участвуют в решении исследовательских задач. Первый из них предполагает направленный критический анализ, как самой информации, так и способов ее получения и подачи автором; второе – поиск тех суждений, фактов, по которым или в связи с которыми, читатель считает нужным высказать собственные мысли.

Из всех рассмотренных видов чтения основным для студентов является изучающее – именно оно позволяет в работе с учебной литературой накапливать знания в различных областях. Вот почему именно этот вид чтения в рамках учебной деятельности должен быть освоен в первую очередь. Кроме того, при овладении данным видом чтения формируются основные приемы, повышающие эффективность работы с научным текстом.

Основные виды систематизированной записи прочитанного:

1. Аннотирование – предельно краткое связное описание просмотренной или прочитанной книги (статьи), ее содержания, источников, характера и назначения;
2. Планирование – краткая логическая организация текста, раскрывающая содержание и структуру изучаемого материала;
3. Тезирование – лаконичное воспроизведение основных утверждений автора без привлечения фактического материала;
4. Цитирование – дословное выписывание из текста выдержек, извлечений, наиболее существенно отражающих ту или иную мысль автора;
5. Конспектирование – краткое и последовательное изложение содержания прочитанного.

Конспект – сложный способ изложения содержания книги или статьи в логической последовательности. Конспект аккумулирует в себе предыдущие виды записи, позволяет всесторонне охватить содержание книги, статьи. Поэтому умение составлять план, тезисы, делать выписки и другие записи определяет и технологию составления конспекта.

Методические рекомендации по составлению конспекта:

1. Внимательно прочитайте текст. Уточните в справочной литературе непонятные слова. При записи не забудьте вынести справочные данные на поля конспекта.
2. Выделите главное, составьте план.
3. Кратко сформулируйте основные положения текста, отметьте аргументацию автора.
4. Законспектируйте материал, четко следуя пунктам плана. При конспектировании старайтесь выразить мысль своими словами. Записи следует вести четко, ясно.
5. Грамотно записывайте цитаты. Цитируя, учитывайте лаконичность, значимость мысли.

В тексте конспекта желательно приводить не только тезисные положения, но и их доказательства. При оформлении конспекта необходимо стремиться к емкости каждого предложения. Мысли автора книги следует излагать кратко, заботясь о стиле и выразительности написанного. Число дополнительных элементов конспекта должно быть логически обоснованным, записи должны распределяться в определенной последовательности, отвечающей логической структуре произведения. Для уточнения и дополнения необходимо оставлять поля.

Овладение навыками конспектирования требует от студента целеустремленности, повседневной самостоятельной работы.

Лабораторные занятия

Следует подчеркнуть, что только после усвоения лекционного материала с определенной точки зрения (а именно с той, с которой он излагается на лекциях) он будет закрепляться на лабораторных занятиях как в результате обсуждения и анализа лекционного материала, так и с помощью выполнения заданий лабораторной работы, поставленных задач. При этих условиях студент не только хорошо усвоит материал, но и научится применять его на практике, а также получит дополнительный стимул (и это очень важно) для активной проработки лекции.

При самостоятельном выполнении заданий нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если студент видит несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи). Решение задач или примеров следует излагать подробно, вычисления располагать в строгом порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что выполнение каждого учебного задания должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом. Выполнение заданий данного типа нужно продолжать до приобретения твердых навыков в их решении.

Самопроверка

После изучения определенной темы по записям в конспекте и учебнику, а также решения достаточного количества соответствующих заданий на лабораторных занятиях и самостоятельно студенту рекомендуется, используя лист опорных материалов, воспроизвести по памяти определения, выводы формул, формулировки основных положений и доказательств.

В случае необходимости нужно еще раз внимательно разобраться в материале.

Иногда недостаточность усвоения того или иного вопроса выясняется только при изучении дальнейшего материала. В этом случае надо вернуться назад и повторить плохо усвоенный материал. Важный критерий усвоения теоретического материала - умение решать задачи или пройти тестирование по пройденному материалу. Однако следует помнить, что правильное решение за-

дачи может получиться в результате применения механически заученных формул без понимания сущности теоретических положений.

Консультации

Если в процессе самостоятельной работы над изучением теоретического материала или при выполнении заданий у студента возникают вопросы, разрешить которые самостоятельно не удастся, необходимо обратиться к преподавателю для получения у него разъяснений или указаний. В своих вопросах студент должен четко выразить, в чем он испытывает затруднения, характер этого затруднения. За консультацией следует обращаться и в случае, если возникнут сомнения в правильности ответов на вопросы самопроверки.

Подготовка к экзамену по дисциплине «Анализ данных».

Изучение многих общепрофессиональных и специальных дисциплин завершается экзаменом. Подготовка к экзамену способствует закреплению, углублению и обобщению знаний, получаемых, в процессе обучения, а также применению их к решению практических задач. Готовясь к экзамену, студент ликвидирует имеющиеся пробелы в знаниях, углубляет, систематизирует и упорядочивает свои знания.

Экзамен - форма промежуточной аттестации студентов по результатам освоения теоретических знаний, приобретения практических навыков и компетенций, целью которой является контроль результатов освоения студентами учебного материала по программе конкретной дисциплины, проверка и оценка знаний, полученных за семестр (курс), их прочности, развития творческого мышления, приобретения навыков самостоятельной работы, умения применять теоретические знания при решении практических задач. Экзамен может быть проведен в устной форме, в устно-письменной форме, в форме письменной работы или тестирования.

Число вопросов, включаемых в экзаменационный билет, должно быть не менее двух и не более пяти, при этом вопросы могут носить как теоретический, так и прикладной характер. На экзамен могут выноситься типовые задачи, проработанные в течение семестра на аудиторных занятиях и в процессе самостоятельной работы. Содержание вопросов и задач, включаемых в экзаменационный билет, должно соответствовать учебной программе дисциплины и фондам оценочных средств. Перечень примерных вопросов, заданий и критерии оценки уровня обученности и уровня сформированности компетенций доводятся до сведения обучающихся в начале семестра изучения дисциплины.

Систематическая подготовка к занятиям в течение семестра позволит использовать время экзаменационной сессии для систематизации знаний.

Текущая аттестация студентов проводится преподавателями, ведущими лабораторные занятия по дисциплине в следующих формах: собеседование при защите лабораторной работы с предоставлением письменного отчета. Допуск к лабораторным работам происходит при наличии у студентов печатного вари-

анта отчета. Защита отчета проходит в форме собеседования. Максимальное количество баллов студент получает, если оформление отчета соответствует установленным требованиям, а отчет полностью раскрывает суть работы. Основанием для снижения оценки являются:

- не качественно выполненная работа на персональном компьютере,
- письменный отчет выполнен в небрежной форме и с ошибками,
- студент не дает исчерпывающий ответ на вопрос преподавателя,
- путается или затрудняется в демонстрации работы на компьютере.

Отчет может быть отправлен на доработку в следующих случаях:

- небрежное выполнение и с ошибками,
- сокращенное содержание отчета.

Критерии оценивания собеседования приведены в Фонде оценочных средств по дисциплине Анализ данных.

8. Учебно-методическое и информационное обеспечение

Дятлов, А. В.; Анализ данных в социологии Электронный ресурс : Учебник / А. В. Дятлов, Д. А. Гугуева. - Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2018. - 226 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-9275-2690-1

Карпенко, О. М.; Высшее образование в странах мира. Анализ данных образовательной статистики и глобальных рейтингов в сфере образования Электронный ресурс / О. М. Карпенко, М. Д. Бершадская. - Высшее образование в странах мира. Анализ данных образовательной статистики и глобальных рейтингов в сфере образования, 2019-09-20. - Москва : Современная гуманитарная академия, 2009. - 244 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-8323-0580-6

Крутиков, В. И.; Анализ данных Электронный ресурс / Крутиков В. И., Мешечкин В. В. - Кемерово : КемГУ, 2014. - 138 с. - ISBN 978-5-8353-1770-7

Низаметдинов, Ш. У.; Анализ данных Электронный ресурс / Низаметдинов Ш. У., Румянцев В. П. : учебное пособие для вузов. - Москва : НИЯУ МИФИ, 2012. - 288 с. - Рекомендовано УМО «Ядерные физика и технологии» в качестве учебного пособия для студентов высших учебных заведений. - ISBN 978-5-7262-1687-4

Фаткуллина, Р. Р.; Анализ технологических данных с использованием Microsoft Excel Электронный ресурс : Учебное пособие / Р. Р. Фаткуллина. - Казань : Казанский национальный исследовательский технологический университет, 2014. - 80 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-7882-1555-6

Фендель, Т. В.; Анализ и статистическая обработка исследовательских данных Электронный ресурс / Фендель Т. В. : учебно-методическое пособия. - Чайковский : ЧГИФК, 2017. - 25 с.

Перечень основной литературы:

1 Коробко, В. И. Теория управления Электронный ресурс : Учебное пособие для студентов вузов, обучающихся по специальности «Государственное и муниципальное управление», «Менеджмент организации» / В. И. Коробко. – Теория управления, 2022-03-26. - Москва : ЮНИТИ-ДАНА, 2017. - 383 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-238-01483-8

2 Цветкова, О. Л. Теория автоматического управления : учебник / О.Л. Цветкова. - Москва|Берлин : Директ-Медиа, 2016. - 207 с. : ил., схем., табл. - <http://biblioclub.ru/>. - Библиогр. в кн. - ISBN 978-5-4475-8334-7

