МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«МУРМАНСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра цифровых технологий, математики и экономики

Методическая разработка к выполнению расчетно-графической работы «Статистическая обработка экспериментальных данных»

по дисциплине: *«Специальные разделы высшей математики», часть* 2

для направления: 09.03.01 *«Информатика и вычислительная техника», бакалавриат, очная форма обучения*

Мурманск
2021

Составитель – Кацуба Валентина Сергеевна, канд. физ.-мат. наук, доцент кафедры цифровых технологий, математики и экономики МГТУ

Методическая разработка к выполнению расчетно-графической работы «Статистическая обработка экспериментальных данных» по дисциплине «Специальные разделы высшей математики, часть 2» рассмотрена и одобрена на заседании кафедры-разработчика цифровых технологий, математики и экономики МГТУ

21.06.2022г., протокол №12 .

дата

Рецензент – Романовская Юлия Владимировна, канд. физ.-мат. наук, доцент кафедры цифровых технологий, математики и экономики

Оглавление

1. Общие организационно-методические указания

Расчетно-графическая работа «Статистическая обработка экспериментальных данных» предусмотрена рабочей программой дисциплины «Специальные разделы высшей математики», во второй части которой изучаются основные разделы теории вероятностей и элементы математической статистики.

Целевая установка: при выполнении РГР студент должен освоить методы и сформировать практические навыки по простейшей обработке выборочных данных, представляющих один количественный признак или представляющих систему двух количественных признаков.

2. Задания РГР и планы их выполнения

РГР содержит 2 задачи на статистическую обработку одномерной и двумерной выборки. Ниже приведены условия обеих задач, планы их решения и необходимый числовой материал.

ЗАДАЧА №1

В результате эксперимента были получены значения количественного признака *X* генеральной совокупности (таблица 1). Требуется сформировать выборку 100 значений признака *X* из таблицы 1 (в соответствии с таблицей 2) и провести статистическую обработку результатов измерений по методу моментов в соответствии с приведенным ниже планом.

Цель обработки состоит в следующем:

1) привести обоснования статистической гипотезы о близости закона распределения случайной величины *X* к нормальному закону;
2) составить функцию плотности предполагаемого нормального распределения случайной величины *X*, используя точечные статистические оценки параметров этого распределения;
3) указать интервальные оценки с различными уровнями надёжности для математического ожидания предполагаемого нормального распределения;
4) проверить по критерию Пирсона согласование с выборочными данными выдвинутой гипотезы о нормальном распределении признака *X*.

План выполнения:

1этап. *Начальная обработка выборки*

Составить вариационный ряд; систематизировать выборку по 8 разрядам (левая граница каждого разряда включается в данный разряд, x_{max} включается в последний разряд) и составить группированный ряд распределения выборки. Составить статистическое распределение выборки, приняв середины разрядов за варианты x_i. Построить полигон частот n_i и полигон относительных частот p^*_i. Построить гистограмму частот и гистограмму относительных частот.

2 этап. *Эмпирическая функция распределения*

Найти эмпирическую функцию распределения $F^*(x)$, построить ее график.

3 этап. *Числовые характеристики выборки*

Найти числовые характеристики выборки: эмпирические начальные моменты $m_1^*, m_2^*, m_3^*, m_4^*$ и центральные моменты $\mu_1^*, \mu_2^*, \mu_3^*, \mu_4^*$, а также выборочную среднюю $\overline{x}_в$, выборочные СКО $\sigma_в$, асимметрию A_S, эксцесс E_S, моду M_0 и медиану M_e. Сделать вывод о возможной близости распределения признака X к нормальному распределению.

4 этап. *Гипотеза о теоретическом распределении признака X*

Сформулировать гипотезу о распределения признака X по нормальному закону. Записать вид плотности нормального распределения $f(x)$, используя точечные статистические оценки для математического ожидания $a \approx \overline{x}_в$, для среднего квадратического отклонения $\sigma \approx S = \sqrt{\frac{n}{n-1} D_в}$.

Построить график плотности $f(x)$, вычислив ее значения в граничных точках 8 разрядов; сравнить гистограмму относительных частот с графиком выравнивающей ее нормальной кривой $f(x)$. Вычислить теоретические частоты n_i' попадания СВ X в промежуток, совпадающий с i-тым разрядом группированного статистического ряда: $n_i' = n \cdot p_i$, где $p_i = P\{x_i < X < x_{i+1}\}$, n – объем выборки; составить сравнительную таблицу эмпирических частот n_i и теоретических частот n_i' и сравнить графики их полигонов.

5 этап. *Интервальные оценки параметра a*

Найти интервальные оценки для математического ожидания (для параметра a) нормально распределенного признака X генеральной совокупности, используя для этого следующие уровни надежности: а) γ = 0,95; б) γ = 0,99; в) γ = 0,999. Сформулировать связь между величиной надежности и длиной доверительного интервала.

6 этап. *Проверка статистической гипотезы о нормальном распределении признака X*

Проверить статистическую гипотезу о распределении признака X по нормальному закону, используя критерий Пирсона с заданным уровнем значимости $\alpha = 0{,}05$.

Рекомендуемая точность вычислений: 10^{-4}.

Таблица 1

Номера значений X	Выборочные значения признака X									
1 – 10	38	36	29	36	22	38	54	41	48	27
11 – 20	39	31	31	20	29	32	18	21	37	39
21 – 30	40	19	50	46	29	44	34	41	37	33
31 – 40	34	42	43	35	34	38	43	36	31	33
41 – 50	45	40	19	45	46	34	13	37	37	31
51 – 60	30	31	35	31	32	29	38	33	40	42
61 – 70	36	44	35	42	29	25	44	36	34	46
71 – 80	39	25	42	29	28	45	34	41	38	29
81 – 90	23	23	35	35	26	33	26	36	41	24
91 – 100	42	34	40	32	40	23	39	37	21	35
101 – 110	35	37	33	26	29	40	41	27	35	43
111 – 120	37	38	39	28	38	35	28	29	27	41
121 – 130	38	39	28	38	35	28	29	27	41	37
131 – 140	33	29	29	45	40	33	36	46	39	37
141 – 150	43	42	39	27	35	53	46	47	34	28

Таблица 2

Номер варианта	Номера значений X
1	1– 40, 51– 110
2	11 – 50, 61 – 120
3	21 – 60, 71 – 130
4	31 – 70, 81 – 140
5	41– 80, 91 – 150
6	1 – 100
7	11 – 110
8	21 –120
9	31 – 130
10	31-60, 71-140

Номер варианта	Номера значений X
11	41 – 140
12	51 – 150
13	1– 20, 31 – 110
14	11 – 30, 41 – 120
15	21 – 40, 51 – 130
16	31 – 50, 61 – 140
17	41 – 60, 71 – 150
18	1 – 30, 41 – 110
19	21 – 50, 61 – 130
20	41-70, 81-150

Задача №2

Собраны статистические данные о количестве уникальных посетителей некоторого сайта и количествах переходов по баннеру на главной странице сайта за сутки. В результате эксперимента было получено 100 измерений признаков X (количество уникальных посетителей сайта) и Y (количество переходов по баннеру). В корреляционной таблице представлены частоты значений пары (X, Y), которые наблюдались в этих измерениях.

Требуется провести статистическую обработку результатов измерений по методу моментов в соответствии с приведенным ниже планом.
Цель обработки состоит в следующем:

1) установить, являются ли корреляционно зависимыми СВ X и Y;
2) выявить степень близости корреляционной связи к линейной;
3) выполнить линейную аппроксимацию регрессии Y на x и оценить её точность;
4) получив интервальную оценку для углового коэффициента линейной корреляции, построить предельные положения прямой линейной регрессии.

План выполнения:

1 этап. *Выборочные распределения каждого из признаков X и Y*

Составить статистические ряды распределения и построить полигоны частот выборочных данных для каждого количественного признака X и Y.

2 этап. *Основные числовые характеристики выборки*

Найти числовые характеристики выборки: выборочные средние $\bar{x}$, $\bar{y}$, выборочные D_x, D_y и σ_x, σ_y, выборочную ковариацию $K_в$ и выборочный коэффициент корреляции $r_в$. Составить точечные оценки для числовых характеристик системы СВ (X,Y): $\tilde{m}_x$, $\tilde{m}_y$, $\tilde{\sigma}_x$, $\tilde{\sigma}_y$, K_{xy}, $\tilde{r}_{xy}$. Сделать вывод о наличии корреляционной связи между СВ X и Y и о близости этой связи к линейной.

3 этап. *Эмпирические линии регрессии*

Вычислить выборочные условные средние $\bar{y}_x$, $\bar{x}_y$ и построить эмпирические линии регрессии Y на x и X на y.

4 этап. *Линейная регрессия*

Найти уравнение линейной регрессии Y на x. Составить таблицу сравнения условных средних $\bar{y}_x$ и значений функции линейной регрессии $y_{регр}$. Построить эмпирическую ломаную и прямую линию регрессии на одном графике. Оценить точность линейной аппроксимации, вычислив величину $\frac{1}{n}\sum(\bar{y}_x - y_{регр})^2$.

5 этап. *Интервальная оценка для коэффициента линейной корреляции*

Найти интервальную оценку для углового коэффициента линейной корреляции k (для регрессии Y на x) с надежностью $\gamma = 0{,}99$. Построить предельные положения прямой линии регрессии.

Рекомендуемая точность вычислений: 10^{-4}.

В приложении А к данной методической разработке приведены 30 вариантов корреляционных таблиц с числовыми данными для задачи 2.

3. Список рекомендуемых учебных ресурсов

1. Конспект лекций «Теория вероятностей и элементы математической статистики» ведущего преподавателя дисциплины, в том числе в электронной форме.
2. Вентцель, Е. С. Теория вероятностей и ее инженерные приложения : учеб. пособие для вузов / Е. С. Вентцель, Л. А. Овчаров. - 2-е изд., стер. - Москва : Высш. шк., 2000. - 480 с. *(аб. 44).*
3. Гмурман, В. Е. Руководство к решению задач по теории вероятностей и математической статистике : учеб. пособие для бакалавров : [базовый курс] / В. Е. Гмурман. - 11-е изд., перераб. и доп. - Москва : Юрайт, 2013. – 403 с. *(аб. 5, чз. 2+предыдущие издания).*
4. Гмурман, В. Е. Теория вероятностей и математическая статистика : учеб. пособие / В. Е. Гмурман. - 12-е изд., перераб. - Москва : Юрайт : Высш. образование, 2009. - 478с. *(аб. 19, чз. 1+предыдущие издания).*
5. Данко, П. Е. Высшая математика в упражнениях и задачах. В 2 ч. Ч. 2 / П. Е. Данко, А. Г. Попов, Т. Я. Кожевникова. - 6-е изд. - Москва : Оникс 21 век : Мир и Образование, 2005, 2003. - 415 с. *(аб. 6, чз. 1+предыдущие издания).*

4.Образцы выполнения заданий РГР

Пример решения задачи **1** *для выборки объёмом* **50**

В результате эксперимента были получены 50 значений количественного признака X. Проведем статистическую обработку этой выборки по методу моментов в соответствии с предложенным планом. Цель обработки состоит в следующем:

1) привести обоснования статистической гипотезы о близости закона распределения случайной величины X к нормальному закону;

2) составить функцию плотности предполагаемого нормального распределения случайной величины X, используя точечные статистические оценки параметров этого распределения;

3) указать интервальные оценки с различными уровнями надёжности для математического ожидания предполагаемого нормального распределения;

4) проверить по критерию Пирсона согласование с выборочными данными выдвинутой гипотезы о нормальном распределении признака X.

Таблица 1. Выборочные значения признака X *(первичная статическая совокупность)*

40	19	50	46	29	44	34	41	37	33
34	42	43	35	34	38	43	36	31	33
45	40	19	45	46	34	13	37	37	31
36	44	35	42	29	25	44	36	34	46
39	25	42	29	28	45	34	41	38	29

1этап. Начальная обработка выборки

Вариационный ряд - это совокупность выборочных значений признака X в порядке их возрастания

Таблица 2. Вариационный ряд выборочных значений признака X

13	19	19	25	25	28	29	29	29	29
31	31	33	33	34	34	34	34	34	34
35	35	36	36	36	36	37	37	38	38
39	40	40	41	41	42	42	42	43	43
44	44	44	45	45	45	46	46	46	50

Чтобы упростить дальнейшую обработку выборки большого объема, систематизируем выборку по s разрядам и составим *группированный статистический ряд*. Для этого весь участок оси *OX*, на котором расположены наблюдаемые в выборке значения СВ X, разделим на промежутки [x_i ; x_{i+1}) или “разряды”. Длины разрядов выберем одинаковыми и целыми (можно четными), концы промежутков также назначим целыми числами (для удобства дальнейших вычислений).

Пусть число разрядов s =5. По обрабатываемой выборке фиксируем:

объем выборки	n=50
наименьшее выборочное значение признака X	x_{min} = 13
наибольшее выборочное значение признака X	x_{max} = 50
размах выборки X	$R = x_{max} - x_{min} = 37$

длину разряда h округлим до целого четного числа: $h = R / s = 7{,}4 \Rightarrow h = 8$.

Получаем 5 промежутков (разрядов):

[13 ; 21) , [21 ; 29) , [29 ; 37) , [37 ; 45) , [45 ; 53).

Группированный статический ряд включает строку получившихся разрядов [x; x_{i+1}), строку частот n_i и строку относительных частот $p_i^* = n_i / n$; при этом частоты n_i определяются как количество выборочных значений СВ X , попавших в i-тый разряд.

Таблица 3. Группированный статический ряд

i	1	2	3	4	5
$[x_i\,;\,x_{i+1})$	[13; 21)	[21;29)	[29;37)	[37;45)	[45;53)
n_i	3	3	19	18	7
p_i^*	0,06	0,06	0,38	0,36	0,14

$\sum\limits_i n_i = n$,

$\sum\limits_i p_i^* = 1$

Для построения *статистического распределения выборки* принимаем середины промежутков $[x_i\,;\,x_{i+1})$ за варианты x_i с частотами n_i и относительными частотами p_i^* .

Таблица 4. Статистический ряд распределения выборки

x_i	17	25	33	41	49	
n_i	3	3	19	18	7	$\sum\limits_i n_i = n$,
p_i^*	0,06	0,06	0,38	0,36	0,14	$\sum\limits_i p_i^* = 1$

Для построения *полигона частот* и *полигона относительных частот* на координатной плоскости изображают ломаную линию с вершинами в точках (x_i ; n_i) или (x_i ; p_i^*). Форма линии полигона характеризует частотность значений наблюдаемого признака X: чем выше очередная вершина полигона, тем чаще наблюдалось соответствующее ей значение варианты в рамках проведенного эксперимента (рис.1).

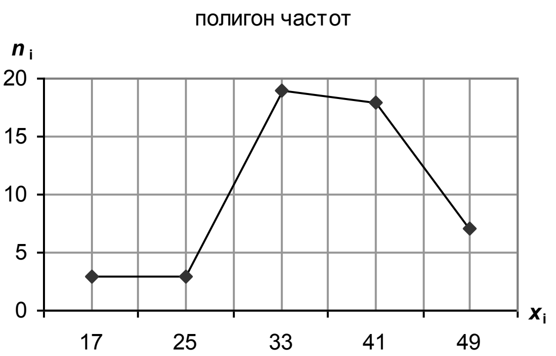

Рис.1

Гистограмма - это ступенчатая фигура, составленная из $s = 5$ прямоугольников. Основаниями прямоугольников служат промежутки $[x_i\,;\,x_{i+1})$, а высотами - отрезки плотностей частот $n_i\,/\,h$ или плотностей относительных частот $p_i^*\,/\,h$.

Для построения гистограммы удобно составить расширенный статистический ряд, включающий в себя группированный статистический ряд (таблица 3), а также плотности частот и плотности относительных частот.

Таблица 5. Расширенный статический ряд

$[x_i\,;\,x_{i+1})$	[13; 21)	[21;29)	[29;37)	[37;45)	[45;53)
n_i	3	3	19	18	7
p_i^*	0,06	0,06	0,38	0,36	0,14
$n_i/h = n_i/8$	0,375	0,375	2,375	2,25	0,875

$p_i^*/h = p_i^*/8$	0,0075	0,0075	0,0475	0,045	0,0175

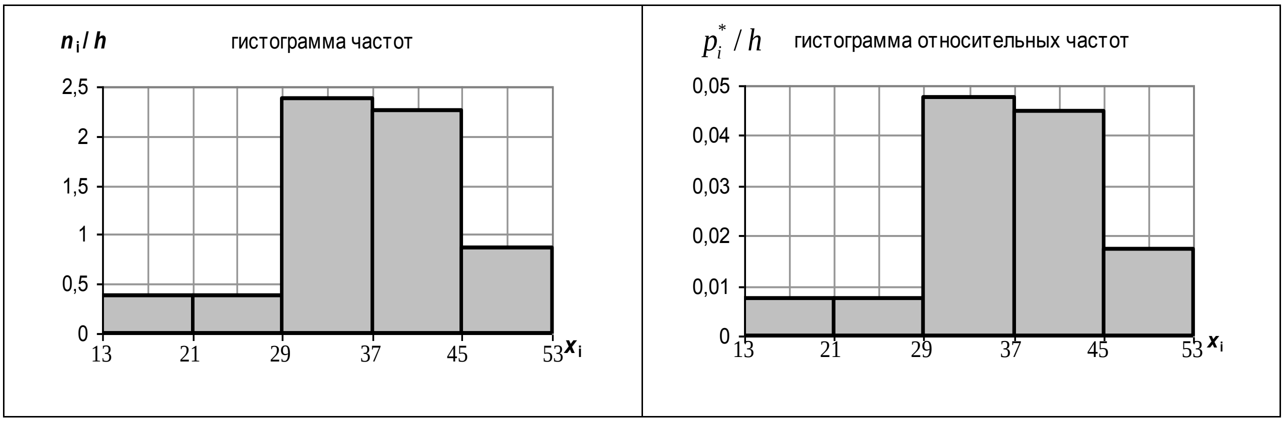

Рис.2

Площадь, ограниченная гистограммой частот, равна объему выборки *n*. Площадь, ограниченная гистограммой относительных частот, равна 1. Поэтому гистограмма относительных частот является статистическим аналогом графика плотности распределения вероятностей НСВ *X* в генеральной совокупности

2 этап. Построение эмпирической функции распределения

Эмпирическая функция распределения случайной величины *X* - это функция $F^*(x)$, равная относительной частоте события $X < x$: $F^*(x) = P^*\{X < x\} = \frac{n_x}{n}$,

n - объем выборки, n_x - количество выборочных вариант, значения которых меньше *x*.

Составить значения этой функции можно по дискретному статистическому ряду распределения выборки (таблица 4) или по группированному ряду распределения выборки (таблица 3). В первом случае $F^*(x)$ представляет собой разрывную ступенчатую функцию со скачками в точках x_i , равными относительным частотам p_i^* (рис.3).

$$F^*(x) = \begin{cases} 0, & x \le 17 \\ 0{,}06 \text{ при } & 17 < x \le 25 \\ 0{,}12 \text{ при } & 25 < x \le 33 \\ 0{,}5 \text{ при } & 33 < x \le 41 \\ 0{,}86 \text{ при } & 41 < x \le 49 \\ 1 \text{ при } & 49 < x < +\infty \end{cases}$$

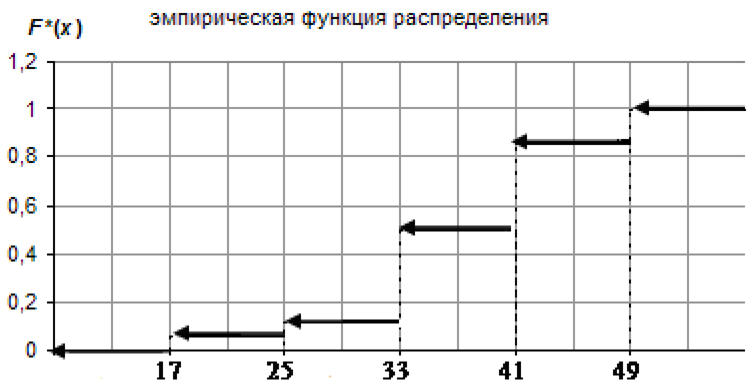

Рис.3

Во втором случае в качестве тех значений *x*, в которых вычисляются значения $F^*(x)$, нужно брать границы разрядов, и полученные точки соединить ломаной линией; в результате получится статистический аналог функции распределения НСВ *X* (рис.4).

x_i	$F^*(x_i)$
13	0
21	0,06
29	0,12
37	0,5
45	0,86
53	1

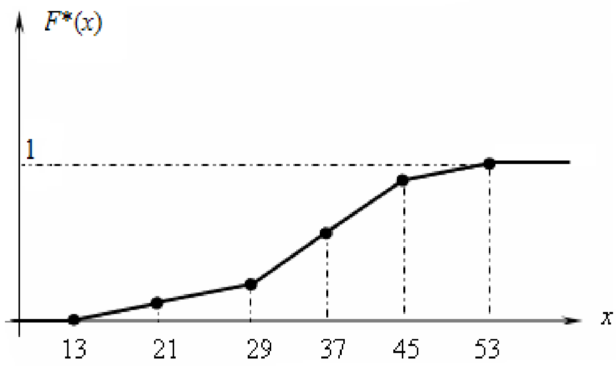

Рис.4

3этап.Числовые характеристики выборки

Вычисление числовых характеристик выборки проводится по укороченному статистическому ряду её распределения (таблица 4).

3.1. *Эмпирические начальные моменты k-того порядка* m_k^* и *эмпирические центральные моменты k-того порядка* μ_k^* вычисляются по формулам, аналогичным формулам для этих моментов дискретной СВ с заменой вероятностей на относительные частоты:

$$m_k^* = \sum_i (x_i)^k \cdot p_i^* = \frac{1}{n}\sum_{i=1}^{s}(x_i)^k \cdot n_i, \quad \mu_k^* = \sum_i (x_i - \overline{x}_в)^k \cdot p_i^* = \frac{1}{n}\sum_i (x_i - \overline{x}_в)^k \cdot n_i, \quad k = 1,2,3,4,...$$

$$\Rightarrow m_1^* = \frac{1}{n}\sum_{i=1}^{5} x_i \cdot n_i = 36,68; \qquad m_2^* = \frac{1}{n}\sum_{i=1}^{5}\left(x_i\right)^2 \cdot n_i = 1409,96;$$

$$m_3^* = \frac{1}{n}\sum_{i=1}^{5}\left(x_i\right)^3 \cdot n_i = 56170,76; \qquad m_4^* = \frac{1}{n}\sum_{i=1}^{5}\left(x_i\right)^4 \cdot n_i = 2303445.$$

Таблица 6. Вычисление эмпирических начальных моментов

x_i	17	25	33	41	49	m_k^*
n_i	3	3	19	18	7	
$x_i \cdot n_i$	51	75	627	738	343	m_1^* =36,68
$(x_i)^2 \cdot n_i$	867	1875	20691	30258	16807	m_2^* = 1409,96
$(x_i)^3 \cdot n_i$	14739	46875	682803	1240578	823543	m_3^* = 56170,76
$(x_i)^4 \cdot n_i$	250563	1171875	22532499	50863698	40353607	m_4^* = 2303445

В теории вероятностей известны выражения центральных моментов через начальные моменты, а также легко выводится значение первого центрального момента, равное 0 для любой СВ:

$\mu_1 = 0, \quad \mu_2 = m_2 - (m_1)^2, \quad \mu_3 = m_3 - 3m_1m_2 + 2(m_1)^3, \quad \mu_4 = m_4 - 4m_1m_3 + 6(m_1)^2 m_2 - 3(m_1)^4.$

Аналогичные формулы имеют место и в статистике. Используя их, определяем эмпирические центральные моменты обрабатываемой выборки:

$\mu_1^* = 0$; $\quad \mu_2^* = m_2^* - (m_1^*)^2 = 64,5376$; $\quad \mu_3^* = m_3^* - 3 \cdot m_1^* \cdot m_2^* + 2 \cdot \left(m_1^*\right)^3 = -281,051$;
$\mu_4^* = m_4^* - 4 \cdot m_1^* \cdot m_3^* + 6 \cdot (m_1^*)^2 \cdot m_2^* - 3(m_1^*)^4 = 13537,23$.

3.2. *Выборочная средняя* $\overline{x}_в$ (выборочное математическое ожидание) определяется как среднее взвешенное наблюдаемых значений признака *X*, когда каждое значение берется с учетом частоты его появлений:

$$\overline{x}_в = \frac{1}{n}\sum_i x_i n_i \quad \Rightarrow \quad \overline{x}_в = \frac{1}{50}\sum_{i=1}^{5} x_i n_i \quad \Rightarrow \quad \boxed{\overline{x}_в = 36,68}.$$

Выборочная дисперсия: $D_в = \sum_i x_i^2 \cdot p_i^* - (\overline{x}_в)^2 = m_2^* - (m_1^*)^2 = \mu_2 \quad \Rightarrow \quad \boxed{D_в = 64,5376}$.

Выборочное среднее квадратическое отклонение: $\sigma_в = \sqrt{D_в} \quad \Rightarrow \quad \boxed{\sigma_в = 8,0335}$.

Выборочные дисперсия и среднее квадратическое отклонение являются характеристиками рассеивания (разбросанности) наблюдаемых в выборке значений признака *X* около выборочной средней $\overline{x}_в$.

3.3. *Асимметрия As* и *эксцесс Es* для выборки вычисляются точно так же, как для случайной величины (СВ) в теории вероятностей:

$$A_s = \frac{\mu_3^*}{(\sigma_в)^3} = -0,5421 \ ; \qquad E_s = \frac{\mu_4^*}{(\sigma_в)^4} - 3 = 0,2502 .$$

Смысл этих характеристик описан для СВ в теории вероятностей.

Асимметрия характеризует "скошенность" распределения; если распределене симметрично относительно математического ожидания, то его все центральные моменты нечетного порядка равны 0, в частности $\mu_3 = 0$, поэтому $A_s = 0$.
Например, для нормального распределения $A_s = 0$.

Эксцесс используется для характеристики островершинности или плосковершинности графика плотности распределения НСВ. Для нормального распределения $E_s = 0$; кривые более островершинные, чем кривая Гаусса нормального распределения, обладают положительным эксцессом, более плосковершинные – отрицательным.

В обрабатываемой выборке наблюдается существенная асимметрия, т.е. скошенность распределения относительно выборочной средней $\overline{x}_в$; положительное значение эксцесса указывает на "островершинность" гистограммы относительных частот.

3.4. *Мода* СВ *X* - её наиболее вероятное значение (то значение, для которого вероятность p_i или плотность $f(x)$ распределения имеют наибольшее значение). Экспериментальные (статистические) аналоги моды: для дискретного распределения выборки - то значение, которое в данной серии опытов встречалось чаще всего; для группированного (интервального) распределения выборки - центр того разряда, для которого плотность частоты имеет наибольшее значение. Для обрабатываемой выборки моду определяем по таблицам 3 и 4: $\boxed{M_0 = 33.}$

Замечание. Наличие более чем одной моды обычно указывает на разнородность статистического материала.

Медиана СВ X - это такая характеристика, которая применяется, как правило, для НСВ и определяется как значение $x=Me$, для которого выполняется равенство:

$$P\{X < Me\}=P\{X > Me\}=\frac{1}{2},$$

то есть одинаково вероятно, окажется ли СВ X меньше значения $x=Me$ или больше этого значения.
В случае симметричного распределения (имеющего моду) значения моды, медианы и математического ожидания совпадают.
Для обрабатываемой выборки значение медианы определяем по группированному статистическому ряду (таблица 3):

$\boxed{Me=37}$, так как $P^*\{X<37\}=P^*\{x>37\}=0{,}5$.

3.5. Для того, чтобы сделать предположение о возможной близости распределения признака X к нормальному распределению, нужно провести следующие сравнения:

1) форма гистограммы относительных частот обрабатываемой выборки имеет аналогию с формой графика плотности нормального распределения (с гауссовой кривой);

2) график эмпирической функции распределения $F^*(x)$ имеет аналогично с графиком функции нормального распределения $F(x) = 0{,}5 + \Phi(x)$, где $\Phi(x)$ - функция Лапласа;

3) асимметрия и эксцесс для выборки имеют небольшие значения - для нормального распределения эти значения равны 0;

4) для нормального распределения мода и медиана совпадают с математическим ожиданием - для выборки получились близкие друг к другу значения:
$M_o = 33$, $Me = 37$, $\overline{x}_в = 36{,}68$.

На основании этих сравнений можно сделать предположение о том, что закон распределения случайной величины X в генеральной совокупности является близким к одному из специальных законов непрерывных СВ, который называется нормальным законом.

4 этап. Гипотеза о теоретическом распределении признака *X*

Выдвигаем гипотезу H_0: количественный признак X в генеральной совокупности распределен по нормальному закону с параметрами a и σ. Следовательно, *теоретическая плотность распределения* признака X имеет следующий вид:

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

a – это математическое ожидание, σ - это с.к.о. признака X.

Используем точечные статистические оценки для математического ожидания и с.к.о.:

$$a \approx \overline{x}_в, \ \sigma \approx \sqrt{\frac{n}{n-1}} \cdot \sigma_в \quad \Rightarrow \quad \boxed{a \approx 36{,}68}; \qquad \sigma \approx \sqrt{\frac{50}{49}} \cdot 8{,}0335 \approx 8{,}1151 \quad \Rightarrow \quad \boxed{\sigma = 8{,}1151}.$$

Для построения *графика плотности распределения* $f(x)$ можно вычислить ее значения в граничных точках 5 разрядов (таблица 3); для вычислений удобно использовать известную таблицу значений функции

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad \text{тогда} \quad f(x_i) = \frac{1}{\sigma}\varphi(z_i), \ \text{где} \ z_i = \frac{x_i - a}{\sigma} = \frac{x_i - 36{,}68}{8{,}1151}.$$

Чтобы сопоставить график теоретической плотности распределения $f(x)$ с построенной ранее гистограммой относительных частот, следует сравнивать значения $f(x_i)$ и p_i^*/h. Поэтому полезно эти значения поместить в одну таблицу (таблица 7).

Таблица 7. Значения плотности $f(x)$ распределения СВ X: $N(a, \sigma)$

x_i	13	21	29	37	45	53
$z_i = \frac{x_i - a}{\sigma}$	-2,92	-1,93	-0,95	0,04	1,03	2,01
$\varphi(z_i)$	0,0056	0,0617	0,2549	0,3986	0,2359	0,0528
$f(x_i)$	0,0007	0,0074	0,0314	0,0496	0,0290	0,0063
p_i^*/h	0,0075	0,0075	0,0475	0,045	0,0175	0

x	$f(x)$
13	0,0006
15	0,0013
17	0,0025
19	0,0044
21	0,0074
23	0,0117
25	0,0173
27	0,0240
29	0,0314
31	0,0387
33	0,0447
35	0,0486
37	0,0496
39	0,0476
41	0,0430
43	0,0364
45	0,0290
47	0,0218
49	0,0153
51	0,0101
53	0,0063
55	0,0037
57	0,0020
59	0,0010
61	0,0005
63	0,0002

Рис.5

Эта таблица значений функции $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, в которой $a=36,68$ и $\sigma=8,1151$, и её график построены в программе MS Excel. Промежуток значений аргумента x выбран симметричным относительно математического ожидания a.

Сравнение графика $f(x)$ с гистограммой относительных частот

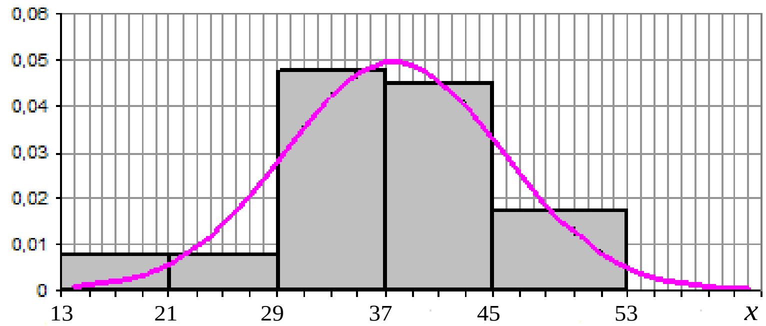

Рис.6

На основании предполагаемого нормального закона распределения признака X найдем *теоретические частоты* n_i' попадания НСВ X в промежуток, совпадающий с i-тым разрядом группированного ряда распределения выборки: $\boxed{n_i' = n\cdot p_i}$,

где $p_i = P\{x_i < X < x_{i+1}\}$, n -объем выборки.

Вероятности p_i нужно находить, используя нормальный закон $N(a,\ \sigma)$ распределения признака X в генеральной совокупности. Для нормального закона известны формулы вычисления вероятности попадания СВ в заданный промежуток :

$P\{\alpha < X < \beta\} = \Phi(\frac{\beta - a}{\sigma}) - \Phi(\frac{\alpha - a}{\sigma})$, где $\quad \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{\frac{z^2}{2}} dz$ - функция Лапласса =>

=> $P\{x_i < X < x_{i+1}\} = \Phi(\frac{x_{i+1} - a}{\sigma}) - \Phi(\frac{x_i - a}{\sigma})$.

Если заменить $z_i = (x_i - a)/\sigma$, где a=36,68 и σ=8,1151, то получим расчетную формулу для вероятностей p_i : $\boxed{p_i = \Phi(z_{i+1}) - \Phi(z_i)}$

Таблица 8. Вычисление значений теоретических частот $n_i^{'}$

i	x_i	x_{i+1}	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	p_i	$n_i = p_i \cdot n$
1	13	21	-2,9476	-1,9518	-0,4984	-0,4745	0,0239	1,1939
2	21	29	-1,9518	-0,9560	-0,4745	-0,3305	0,1441	7,2029
3	29	37	-0,9560	0,0398	-0,3305	0,0159	0,3463	17,3175
4	37	45	0,0398	1,0357	0,0159	0,3498	0,3339	16,6966
5	45	53	1,0357	2,0315	0,3498	0,4789	0,1291	6,4539

Контроль: $\sum_i n_i = 48,8648$ - близко к n=50; $\quad \sum_i p_i^* = 0,9773$ - близко к 1.

Сравнение эмпирических частот n_i и теоретических частот $n_i^{'}$ можно сделать для вариант x_i (середины промежутков $[x_i; x_{i+1})$), а также с помощью полигонов этих частот (таблица 9 и рис.7).

Таблица 9. Сравнение эмпирических и теоретических частот

x_i	17	25	33	41	49
n_i	3	3	19	18	7
$n_i^{'}$	1,1939	7,2029	17,3175	16,6966	6,4539

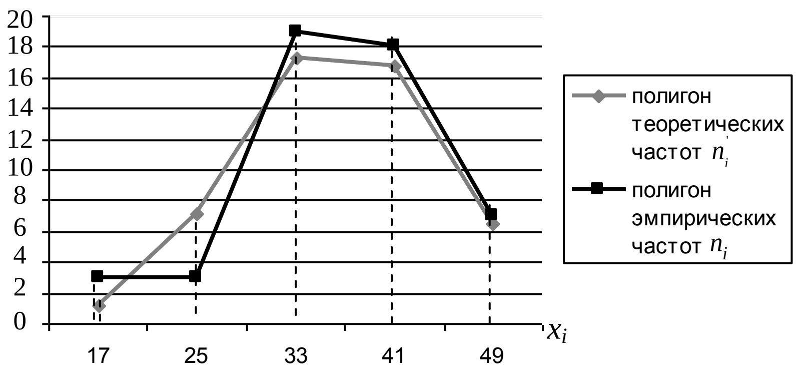

Рис.7

Графики на рис.7 показывают сходство формы полигона теоретических частот, которые получены на основе предполагаемого нормального распределения, с формой полигона эмпирических частот, рассчитанных по выборке. Это сходство на качественном уровне подтверждает правомерность гипотезы о нормальном законе распределения НСВ X.

5 этап. Интервальные оценки параметра a

Интервальной оценкой параметра $a = M(X)$ является доверительный интервал, который определяется заданным уровнем надежности γ точечной оценки параметра a. Известно из теории, что качественной точечной оценкой параметра a является выборочная средняя $\overline{x_в}$, то есть $a \approx \overline{x_в}$.

По определению уровня надежности γ имеем, что

$$P\{|a - \overline{x_в}| < \varepsilon\} = \gamma, \text{ где } \gamma = 0{,}95 \text{ или } \gamma = 0{,}99 \text{ или } \gamma = 0{,}999.$$

Если из этого определения по фиксированному значению γ вычислить число ε и раскрыть неравенство:

$$|a - \overline{x_в}| < \varepsilon \Leftrightarrow \overline{x_в} - \varepsilon < a < \overline{x_в} + \varepsilon \Leftrightarrow a \in (\overline{x_в} - \varepsilon; \overline{x_в} + \varepsilon),$$

то получим доверительный интервал $(\overline{x_в} - \varepsilon; \overline{x_в} + \varepsilon)$ для точечной статистической оценки $\overline{x_в}$ или, другими словами, интервальную оценку параметра a. Смысл интервальной оценки любого параметра состоит в том, что найденный доверительный интервал с наперёд заданной надежностью γ накрывает значение оцениваемого параметра. Длина доверительного интервала определяется числом ε, которое должно быть вычислено по заданной надежности γ.

В случае оценивания математического ожидания a нормально распределенного признака X нужно провести следующие рассуждения:

1) выборочную среднюю $\overline{x_в}$ рассматриваем как одну из реализаций случайной величины $\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$, где X_i - это случайная величина, равная варианте, попавшей под номером i в первоначальную статистическую совокупность; все СВ X_i независимы друг от друга и имеют одинаковые распределения с исследуемым признаком X, поэтому в данной задаче являются нормально распределенными с параметрами $a = M(X)$ и $\sigma = \sigma(X)$;

2) легко вычисляются математическое ожидание и дисперсия СВ $\overline{X}$, если использовать свойства этих характеристик:

$$M(\overline{X}) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n}\cdot a \cdot n = a,$$

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\cdot \sigma^2 \cdot n = \frac{\sigma^2}{n} \Rightarrow \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}};$$

3) для выборки большого объёма n естественно предположить, что СВ $\overline{X}$ имеет также нормальное распределение, параметрами которого являются числа $(a, \frac{\sigma}{\sqrt{n}})$; тогда можно использовать известную формулу для вероятности отклонения нормально распределенной СВ от её математического ожидания: $P\{|X - a| < \varepsilon\} = 2\Phi(\frac{\varepsilon}{\sigma})$;

при условии, что $X := \overline{x_в}$ и $\sigma := \frac{\sigma}{\sqrt{n}}$, получаем следующую формулу для вероятности отклонения выборочной средней $\overline{x_в}$ от её математического ожидания a менее чем на число ε: $P\{|\overline{x_в} - a| < \varepsilon\} = 2\Phi(\frac{\varepsilon\sqrt{n}}{\sigma})$ **;**

эта вероятность должна быть равна заданной надежности γ;

4) теперь имеем два равенства, из которых можно найти число ε:

$$\begin{cases} P\{|\overline{x}_в - a| < \varepsilon\} = \gamma \\ P\{|\overline{x}_в - a| < \varepsilon\} = 2\Phi(\dfrac{\varepsilon\sqrt{n}}{\sigma}) \end{cases} \Rightarrow \Phi(t_\gamma) = \frac{\gamma}{2}, \text{ где } t_\gamma = \frac{\varepsilon\sqrt{n}}{\sigma};$$

значение t_γ находится по известному значению γ с помощью таблицы значений функции Лапласа $\Phi(x)$, используя обратную интерполяцию; по найденному значению t_γ определяем число ε: $\varepsilon = \dfrac{t_\gamma \cdot \sigma}{\sqrt{n}}$ и далее записываем доверительный интервал $(\overline{x}_в - \varepsilon; \overline{x}_в + \varepsilon)$ для оцениваемого параметра a.

В обрабатываемой выборке имеем:
$\overline{x}_в = 36,68$ и $\sigma = 8,1151,\ n = 50$, поэтому $\varepsilon = t_\gamma \cdot \dfrac{\sigma}{\sqrt{n}} = t_\gamma \cdot \dfrac{8,1151}{\sqrt{50}} \approx 1,148 \cdot t_\gamma$, доверительный интервал $(36,68 - \varepsilon; 36,68 + \varepsilon)$ или $36,68 - \varepsilon < a < 36,68 + \varepsilon$, его длина равна 2ε. Каждый такой интервал, рассчитанный по заданной надёжности γ с вероятностью, равной числу γ, накрывает оцениваемый параметр a. В таблице 10 представлены несколько интервальных оценок для параметра a, полученных при различных значениях надёжности γ.

Таблица 10. Интервальные оценки параметра a

γ	$\frac{\gamma}{2}$	t_γ	ε	доверительный интервал	длина доверительного интервала
0,95	0,475	1,96	2,25	34,43< a <38,93	4,5
0,99	0,495	2,58	2,96	33,72< a <39,61	5,92
0,999	0,4995	3,29	3,78	32,9< a <40,46	7,56

Сделанные расчеты показывают, что с увеличением надежности γ длина доверительного интервала также увеличивается.

6 этап. Проверка статистической гипотезы о нормальном распределении признака X

Справедливость выдвинутой гипотезы H_0 о законе распределения признака X в генеральной совокупности чаще всего проверяют по критерию согласия Пирсона, в соответствии с которым статистика критерия имеет распределение «хи-квадрат»:

$$\chi^2 = \sum_{i=1}^{s} \frac{\left(n_i - np_i\right)^2}{np_i} = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'}$$

здесь n_i – это эмпирические частоты вариант выборки,

$n_i' = n \cdot p_i$ – это теоретические частоты попадания СВ X в промежуток $[x_i; x_{i+1})$, вычисленные с помощью плотности $f(x)$ распределения, предполагаемого по гипотезе H_0.

Распределение χ^2 зависит ещё от параметра r, называемого «числом степеней свободы», значение которого находится по формуле $r=s-l$,
где s – число разрядов $\left[x_i; x_{i+1}\right)$, l – число независимых условий (связей), накладываемых на относительные частоты p_i^*; для нормального распределения количество этих связей равно 3 и их содержание определяется следующими равенствами:

$$\sum_i p_i^* = 1, \qquad \sum_i x_i p_i^* = m_x, \qquad \sum_i (x_i - \bar{x}_в)^2 \cdot p_i^* = D_x.$$

Для распределения χ^2 составлена таблица, в которой входами являются вероятности p и число степеней свободы r; числа, стоящие в таблице, представляют собой соответствующие значения функции χ^2. Использовать таблицу нужно следующим образом:

1) если вычисленному значению величины χ^2 соответствует очень маленькая вероятность p ($p \le \alpha$, где α - уровень значимости), то это означает, что значение статистики χ^2 попало в критическую область $V_к$. В этом случае гипотезу H_0 о законе распределения нужно признать неверной, так как в результате лишь одного испытания (т.е. обработки одной выборки) наблюдаем событие $\chi^2 \in V_к$, вероятность которого очень мала: $P\{\chi^2 \in V_к \middle| H_0\} \le \alpha$, т.е. наблюдаем событие, которое считается практически невозможным;

2) если значение вероятности p не маленькое ($p > \alpha$), то гипотезу H_0 нужно признать правдоподобной, так как она не противоречит выборочным данным.

Для вычисления значения χ^2 в обрабатываемой выборке нужно использовать таблицу 9 эмпирических n_i и теоретических $n_i^{'}$ частот, дополненную вспомогательной строкой.

Таблица 11. Вычисление значения χ^2

x_i	17	25	33	41	49
n_i	3	3	19	18	7
$n_i^{'}$	1,1939	7,2029	17,3175	16,6966	6,4539
$\dfrac{(n_i - n_i^{'})^2}{n_i^{'}}$	2,7320	2,4524	0,1635	0,1017	0,0462

Значение $\chi^2 = 5,4958$.
Вычисляем число степеней свободы для обрабатываемой выборки: $r = s - l = 5 - 3 = 2$.
По таблице значений χ^2 обратной интерполяцией находим, что при $r=2$ значение $\chi^2=5,4958$ соответствует $p < 0,05$, так как при $p=0,05$ в таблице есть $\chi^2=5,99$.

Делаем вывод: при уровне значимости $\alpha = 0,05$ вычисленное значение χ^2 не попадает в критическую область (т.е. $p > \alpha$), следовательно, гипотеза о нормальном распределении согласуется с обрабатываемой выборкой.

***Ответ по задаче* 1:**

Проведена статистическая обработка по методу моментов одномерной выборки объемом n=50 значений количественного признака X в соответствии с заданным планом. В результате этой обработки получены следующие результаты:

1) приведены обоснования гипотезы о нормальном законе распределения признака X в генеральной совокупности его значений:

- форма гистограммы относительных частот обрабатываемой выборки имеет аналогию с формой графика плотности нормального распределения (с гауссовой кривой);
- график эмпирической функции распределения $F^*(x)$ имеет аналогично с графиком функции нормального распределения $F(x) = 0,5 + \Phi(x)$, где $\Phi(x)$ - функция Лапласа;
- асимметрия и эксцесс для выборки имеют небольшие значения - для нормального распределения эти значения равны 0;
- для нормального распределения мода и медиана совпадают с математическим ожиданием - для выборки получились близкие друг к другу значения:

$M_o = 33$, $Me = 37$, $\overline{x}_в = 36,68$;

2) составлена функция плотности предполагаемого нормального закона распределения с использованием точечных статистических оценок для ее параметров - математического ожидания и среднего квадратичного отклонения:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad a \approx \overline{x}_в = 36,68, \quad \sigma \approx \sqrt{\frac{n}{n-1}} \cdot \sigma_в = 8,1151;$$

3) найдены интервальные оценки параметра a для нескольких уровней надежности: γ = 0,95; γ = 0,99; γ =0,999; выявлено, что длина доверительного интервала увеличивается с повышением надежности γ (таблица 10);

4) проведена проверка статистической гипотезы о нормальном распределении признака X на основании критерия Пирсона с заданным уровнем значимости $\alpha = 0,05$; полученное значение статистики критерия «хи-квадрат»: $\chi^2 = 5,4958$ указывает на то, что на заданном уровне значимости гипотеза о нормальном распределении исследуемого признака X согласуется с обработанной выборкой.

Пример решения задачи **№2** *для выборки объёмом* **40**

Собраны статистические данные о количестве уникальных посетителей некоторого сайта и количествах переходов по баннеру на главной странице сайта за сутки. В результате эксперимента было получено 40 измерений признаков X (количество уникальных посетителей сайта) и Y (количество переходов по баннеру). В корреляционной таблице представлены частоты значений пары (X, Y), которые наблюдались в этих измерениях.
Проведем статистическую обработку результатов измерений по методу моментов в соответствии с данным планом.

Цель обработки состоит в следующем:

1) установить, являются ли корреляционно зависимыми СВ X и Y;
2) выявить степень близости корреляционной связи к линейной;
3) выполнить линейную аппроксимацию регрессии Y на x и оценить её точность;
4) получив интервальную оценку для углового коэффициента линейной корреляции, построить предельные положения прямой линейной регрессии.

В результате эксперимента было получено 40 измерений признаков *X* и *Y*, частоты которых представлены в табл. 1.

Таблица 1. Корреляционная таблица наблюдаемых частот признаков *X* и *Y*

x_i \ y_j	2	3	5	6	8
1	1	2	1	0	0
2	0	8	4	1	0
4	1	7	5	1	0
7	0	0	4	4	1

число измерений: $n = 40$

n_{ij} - частота пары $(x_i; y_j)$ в выборке, $\sum\limits_{i,j} n_{ij} = 40$

Проведем статистическую обработку результатов, используя метод моментов. В процессе решения будем пополнять корреляционную таблицу эмпирических частот вспомогательными строками и столбцами.

1 этап. Выборочные распределения каждого из признаков *X* и *Y*.

Вычислим частоты и относительные частоты значений $x_i (i = \overline{1,4})$ и $y_j (j = \overline{1,5})$ каждого из признаков *X* и *Y* в отдельности, составляя для этого формулы, аналогичные формулам, которые использовались при работе с корреляционной матрицей системы двух дискретных случайных величин:

$\boxed{n_{x_i} = \sum\limits_{j=1}^{5} n_{ij}, \quad p^*_{x_i} = \frac{n_{x_i}}{n}, \quad i = \overline{1,4}}$, $\boxed{n_{y_j} = \sum\limits_{j=1}^{4} n_{ij}, \quad p^*_{y_j} = \frac{n_{y_j}}{n}, \quad j = \overline{1,5}}$, $\quad n = 40;$

Вспомогательная таблица 1.1.

x_i \ y_j	2	3	5	6	8	n_{x_i}
1	1	2	1	0	0	4
2	0	8	4	1	0	13
4	1	7	5	1	0	14
7	0	0	4	4	1	9
n_{y_j}	2	17	14	6	1	**40**

Составим статистические ряды распределения выборочных данных каждого из признаков *X* и *Y* (таблицы 2 и 3) и построим полигоны частот этих признаков:

Рисунок 1.

Таблица 2. Статистический ряд распределения признака *X*.

x_i	1	2	4	7
n_{x_i}	4	13	14	9
$p^*_{x_i}$	0,01	0,325	0,35	0,225

$$\sum_i n_{x_i} = 40$$

$$\sum_i p^*_{x_i} = 1$$

Таблица 3. Статистический ряд распределения признака *Y*.

y_j	2	3	5	6	8
n_{y_j}	2	17	14	6	1
$p^*_{y_j}$	0,005	0,425	0,125	0,15	0,00375

$$\sum_i n_{y_j} = 40$$

$$\sum_i p^*_{x_i} = 1$$

2 этап. Найдем основные числовые характеристики выборки.

Выборочные числовые характеристики каждой СВ X и Y находим, используя их статистические ряды распределения (таблицы 2 и 3) и формулы, известные из решения задачи 1.

Вспомогательная таблица 1.2.

$X \backslash Y$	2	3	5	6	8	n_{x_i}	$x_i \cdot n_{x_i}$	x_i^2	$x_i^2 \cdot n_{x_i}$
1	1	2	1	0	0	4	4	1	4
2	0	8	4	1	0	13	26	4	52
4	1	7	5	1	0	14	56	16	224
7	0	0	4	4	1	9	63	49	441
n_{y_j}	2	17	14	6	1		**3,725**		**18,025**
$y_j \cdot n_{y_j}$	4	51	70	36	8	**4,225**			
y_j^2	4	9	25	36	64				
$y_j^2 \cdot n_{y_j}$	8	153	350	216	64	**19,775**			

Средние выборочные: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i \cdot n_{x_i} = 3{,}725$; $\bar{y} = \frac{1}{n} \cdot \sum_{j=1}^{4} y_j \cdot n_{y_j} = 4{,}225$;

средние по квадратам: $\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i^2 \cdot n_{x_i} = 18{,}025,$ $\overline{y^2} = \frac{1}{n} \cdot \sum_{j=1}^{4} y_j^2 \cdot n_{y_j} = 19{,}775$;

выборочные дисперсии: $D_x = \overline{x^2} - \left(\bar{x}\right)^2 = 18{,}025 - (3{,}725)^2 \approx 4{,}149;$

$$D_y = \overline{y^2} - \left(\bar{y}\right)^2 = 19{,}775 - (4{,}225)^2 \approx 1{,}924;$$

выборочные с.к.о.: $\sigma_x = \sqrt{D_x} \approx 2{,}037$, $\sigma_y = \sqrt{D_y} \approx 1{,}387$.

Выборочная ковариация $K_в$ и выборочный коэффициент корреляции $r_в$ вычисляются по аналогии с этими характеристиками в теории вероятностей:

$$K_{xy} = M(X \cdot Y) - m_x \cdot m_y, \text{ где } M(X \cdot Y) = \sum_i \sum_j x_i \cdot y_j \cdot p_{ij}$$

$$\Rightarrow \boxed{K_в = \overline{x \cdot y} - \bar{x} \cdot \bar{y}}, \text{ где } \quad \overline{x \cdot y} = \frac{1}{n} \cdot \sum_i \sum_j x_i \cdot y_j \cdot n_{ij};$$

$$r_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y} \Rightarrow \qquad \boxed{r_в = \frac{K_в}{\sqrt{D_x} \cdot \sqrt{D_y}}};$$

для вычисления $\overline{x \cdot y}$ удобно провести следующие преобразования:

$$\overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{4} \sum_{j=1}^{5} x_i \cdot y_j \cdot n_{ij} = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i \sum_{j=1}^{5} y_j \cdot n_{ij} \Rightarrow \overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i \cdot s_i, \quad \text{где } s_i = \sum_{j=1}^{5} y_j \cdot n_{ij}$$

и далее промежуточные результаты вычислений оформить вспомогательной таблицей 1.3.

Вспомогательная таблица 1.3.

x_i \ y_j	2	3	5	6	8	s_i	$x_i \cdot s_i$
1	1	2	1	0	0	13	13
2	0	8	4	1	0	50	100
4	1	7	5	1	0	54	216
7	0	0	4	4	1	52	364
							693

$$\Rightarrow \overline{x \cdot y} = \frac{693}{40} \approx 17{,}325; \quad K_в = 17{,}325 - 3{,}725 \cdot 4{,}225 \approx 1{,}587; \quad r_в \approx 0{,}562 \Rightarrow$$

$$\boxed{K_в = 1{,}587}, \qquad \boxed{r_в = 0{,}562}.$$

Составляем точечные оценки для числовых характеристик системы СВ (X,Y):

$$\tilde{m}_x = \overline{x_в} = 3{,}725, \qquad \tilde{m}_y = \overline{y_в} = 4{,}225;$$

$$\tilde{\sigma}_x = \sqrt{D_x \cdot \frac{n}{n-1}} = \sqrt{4{,}1494 \cdot \frac{40}{39}} \approx 2{,}063; \qquad \tilde{\sigma}_y = \sqrt{D_y \cdot \frac{n}{n-1}} = \sqrt{1{,}9244 \cdot \frac{40}{39}} \approx 1{,}405;$$

$$\tilde{K}_{xy} = K_в \cdot \frac{n}{n-1} = 1{,}5869 \cdot \frac{40}{39} \approx 1{,}628; \qquad \tilde{r}_{xy} = r_в = 0{,}5616 \approx 0{,}562.$$

При вычислении этих точечных оценок учтено, что

1) статистическая оценка для математического ожидания исследуемой СВ совпадает с ее выборочной средней;

2) для получения статистических оценок дисперсий и ковариации нужно вводить поправочный коэффициент $\dfrac{n}{n-1}$;

3) статистическая оценка для коэффициента корреляции совпадает с выборочным коэффициентом корреляции $r_в$, так как

$$\tilde{r}_{xy} = \frac{\tilde{K}_{xy}}{\tilde{\sigma}_x \cdot \tilde{\sigma}_y} = \frac{K_в \cdot \dfrac{n}{n-1}}{\sqrt{D_x \cdot \dfrac{n}{n-1}} \cdot \sqrt{D_y \cdot \dfrac{n}{n-1}}} = \frac{K_в}{\sqrt{D_x} \cdot \sqrt{D_y}} = r_в.$$

По значению $\tilde{K}_{xy} \neq 0$ и $\tilde{K}_{xy} > 0$ заключаем, что между признаками X и Y есть корреляционная связь, причем положительная, которая характеризуется тем, что с увеличением значений одного из признаков X или Y значения другого имеют тенденцию также увеличиваться. По значению $\tilde{r}_{xy}$, близкому к числу 0,5, делаем вывод, что линейная связь между X и Y проявляется как умеренная.

3 этап. Эмпирические линии регрессии

Условные средние $\overline{y_{x_i}}$ вычисляются по статистическим рядам распределения частот признака *Y* при фиксированных значениях признака *X* (аналогично системе дискретных СВ в теории вероятностей):

$$P^*\{Y = y_j \mid X = x_i\} = \frac{p^*_{ij}}{p^*_{x_i}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{x_i}}{n}} = \frac{n_{ij}}{n_{x_i}} \Rightarrow \overline{y_{x_i}} = \sum_{j=1}^{5} y_i \cdot P^*\{Y = y_j \mid X = x_i\} = \sum_{j=1}^{5} y_j \cdot \frac{n_{ij}}{n_{x_i}} \Rightarrow$$

$\overline{y_x} = \frac{1}{n_{x_i}} \cdot s_i$, где $s_i = \sum_{j=1}^{5} y_j \cdot n_{ij}$.

Для этих вычислений используем таблицу 2 и числа s_i из таблицы 1.3, результаты вычислений в таблице 4.

Эмпирическая линия регрессии *Y* на *x* иллюстрирует зависимость условных средних $\overline{y_{x_i}}$ от значений x_i:

Таблица 4. Условные средние значения $\overline{y_{x_i}}$

x_i	1	2	4	7
n_{x_i}	4	13	14	9
s_i	13	50	54	52
$\overline{y_{x_i}}$	3,25	3,846	3,857	5,778

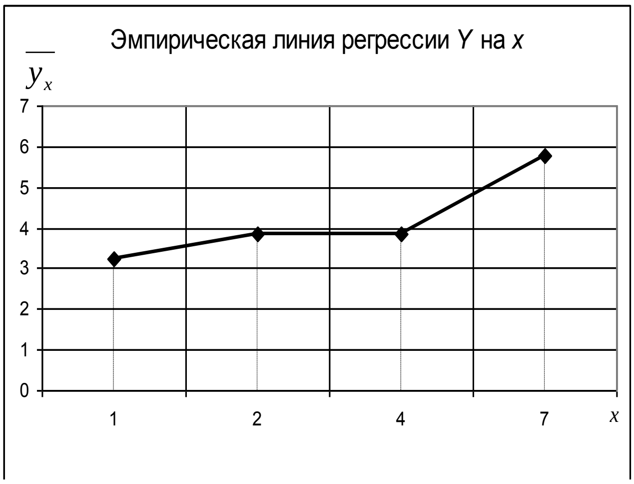

Рисунок 2.

Аналогично находим условные средние $\overline{x_{y_j}}$:

$$P^*\{X = x_i \mid Y = y_j\} = \frac{p^*_{ij}}{p^*_{y_j}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{y_j}}{n}} = \frac{n_{ij}}{n_{y_j}} \Rightarrow \overline{x_{y_j}} = \sum_{j=1}^{5} x_i \cdot P^*\{X = x_i \mid Y = y_j\} = \sum_{i=1}^{5} x_i \cdot \frac{n_{ij}}{n_{y_j}} \Rightarrow$$

$\overline{x_{y_j}} = \frac{1}{n_{y_j}} \cdot s'_j$, где $s'_j = \sum_{i=1}^{4} x_j \cdot n_{ij}$.

Для счета используем таблицу 3 и числа s_i вычисляем дополнительно по таблице 1; результаты помещаем в таблицу 5.

Эмпирическая линия регрессии *X* на *y* иллюстрирует зависимость условных средних $\overline{x_y}$ от значений *y* .

Таблица 5. Условные средние значения $\overline{x_{y_j}}$

y_j	2	3	5	6	8

n_{y_j}	2	17	14	6	1
s_i	5	46	57	34	7
$\overline{x_{y_j}}$	2,5	2,706	4,071	5,667	7

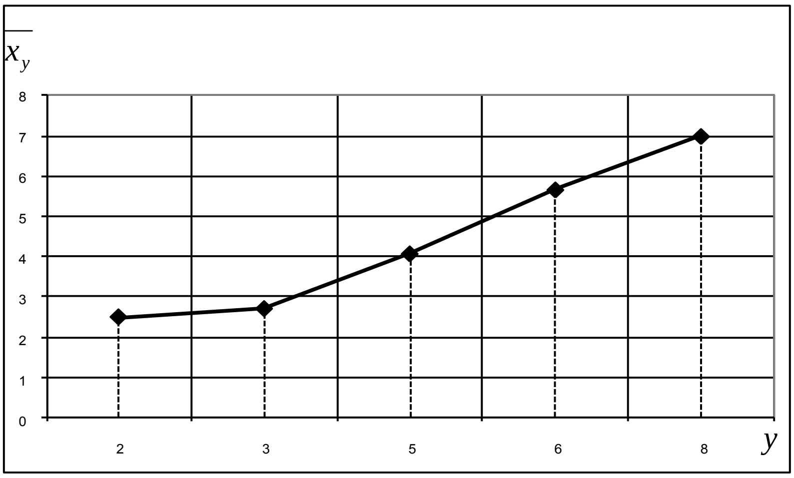

Рисунок 3.

4 этап. Линейная регрессия

Уравнение линейной регрессии Y на x (другое обозначение $Y(x)$) имеет такой же вид, как в теории вероятности для системы двух СВ (X,Y):

$y - m_y = \dfrac{r_{xy} \cdot \sigma_y}{\sigma_x} \cdot (x - m_x) \Rightarrow$ заменяем $m_y = \overline{y},\ \ m_x = \overline{x},\ r_{xy} = r_в$ и получаем выборочное уравнение линейной регрессии:

$$\boxed{y - \overline{y} = r_в \cdot \frac{\sigma_y}{\sigma_x} \cdot (x - \overline{x})};$$

для построения этой прямой в осях XOY её уравнение можно записать в виде $y = kx + b$,

где $k = r_в \cdot \dfrac{\sigma_y}{\sigma_x}, \quad b = \overline{y} - k \cdot \overline{x}.$

Найдем коэффициенты уравнения линейной регрессии Y на x, записанного в виде $y = kx + b$:

$k = r_в \cdot \dfrac{\sigma_y}{\sigma_x} = 0{,}382437 \approx 0{,}382, \quad b = \overline{y} - k \cdot \overline{x} = 2{,}800422 \approx 2{,}800.$

Следовательно, уравнение линейной регрессии в решаемой задаче имеет следующий вид:

$$y_{регр} = 0{,}382x + 2{,}800\,.$$

Таблица 6. Сравнение условных средних $\overline{y_x}$ и значений $y_{регр}$

Y / x_i	$\overline{y_{x_i}}$	$y_{регр}$	$y_{регр} - \overline{y_{x_i}}$	$(y_{регр} - \overline{y_{x_i}})^2$
1	3,25	3,183	-0,067	0,0045
2	3,846	3,565	-0,281	0,07896
4	3,857	4,33	0,473	0,22373
7	5,778	5,478	-0,3	0,09000

	0,3973

Построим график линейной регрессии $y_{регр}$, наложив его на эмпирическую ломаную регрессии с узлами $(x_i, \overline{y_{x_i}})$:

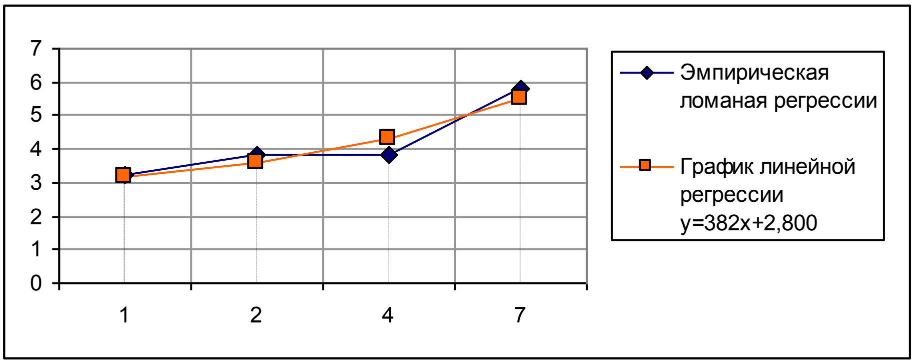

Рисунок 4.

Оценим точность линейной аппроксимации выборочных данных:

$$\frac{1}{n} \cdot \sum_{i=1}^{4} \left(\left(y_{регр} \right)_i - y_{x_i} \right)^2 = \frac{0{,}3973}{40} \approx 0{,}010.$$

5 этап. Интервальная оценка для углового коэффициента линейной корреляции с надежностью $\gamma = 0{,}99$

Интервальная оценка углового коэффициента линейной корреляции определяется промежутком : $(k - \varepsilon; k + \varepsilon)$, где $k = r_в \cdot \dfrac{\sigma_y}{\sigma_x} = 0{,}382.$

Число ε находим из условия:

$\varepsilon = t_\gamma \cdot \dfrac{1 - r_в^2}{\sqrt{n}} \cdot \dfrac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$, где t_γ - это такое значение аргумента функции Лапласа, при котором $\Phi(x) = \dfrac{\gamma}{2}$.

Для $\gamma = 0{,}99$ получаем $t_\gamma = 2{,}575829 \approx 2{,}576 \Rightarrow \varepsilon = 2{,}576 \cdot \dfrac{1 - (0{,}562)^2}{\sqrt{40}} \cdot \dfrac{1{,}974}{2{,}063} \approx 0{,}267.$

Таким образом, искомая интервальная оценка получилась следующей:

$$k_{Y/x} \in (0{,}193\ ;\ 0{,}572).$$

Предельные положения прямой регрессии определяются уравнением $y - \bar{y} = k \cdot (x - \bar{x}) \Leftrightarrow y = kx + b$, в котором берутся предельные значения коэффициента k, фиксируемые концами доверительного интервала:

если $k = 0{,}193$, то $b = \bar{y} - k \cdot \bar{x} \approx 3{,}508 \Rightarrow y_1 = 0{,}193x + 3{,}508;$

если $k = 0{,}572$, то $b = \bar{y} - k \cdot \bar{x} \approx 2{,}093 \Rightarrow y_2 = 0{,}572x + 2{,}093.$

Все три прямые проходят через точку $(\bar{x}; \bar{y}) = (3{,}725\ ;\ 4{,}225)$ и строятся по крайним значениям отрезков этих прямых.

Таблица 7. Координаты крайних точек на прямых линиях регрессии.

x	y_1	$y_{регр}$	y_2
1	3,7	3,183	2,665
3,725	4,225	4,225	4,225
7	4,856	5,477	6,099

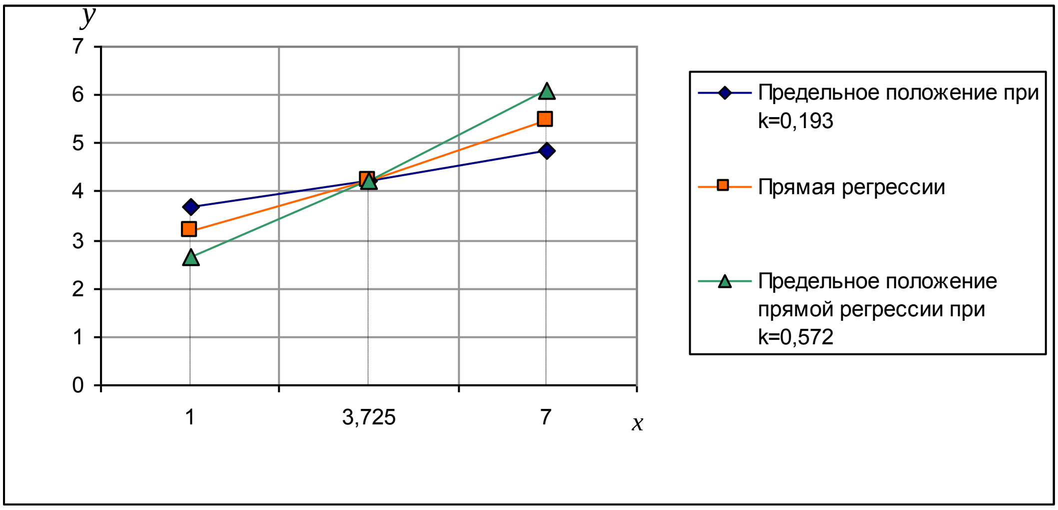

Рисунок 5.

Задача решена полностью.

Ответ по задаче 2:

Проведена статистическая обработка по методу моментов двумерной выборки (объем выборки n=40 значений) пары количественных признаков (X, Y), где X – это количество уникальных посетителей некоторого сайта и Y – это количество переходов по баннеру на главной странице сайта за сутки. Получены следующие результаты обработки:

1) по значению ковариации $K_{xy} \neq 0$ и $K_{xy} > 0$ сделано заключение, что между признаками X и Y есть корреляционная связь, причем корреляция является положительной, которая характеризуется тем, что с увеличением значений одного из признаков X или Y значения другого признака имеют тенденцию также увеличиваться;

2) по значению коэффициента корреляции r_{xy}, близкому к числу 0,5 сделан вывод, что линейная составляющая связи между признаками X и Y проявляется как умеренная;

3) составлено уравнение линейной среднеквадратической регрессии Y на x:

$$y_{регр} = 0{,}382x + 2{,}800\,;$$

точность линейной аппроксимации выборочных данных определена числом 0,010;

4) найдена интервальная оценка для углового коэффициента линейной корреляции k (для регрессии Y на x) с надежностью $\gamma = 0{,}99$; построены предельные положения прямой линейной регрессии (рисунок 5).

Приложение А. Варианты корреляционных таблиц для задачи 2

Вариант 1

X \ Y	23	45	67	89	111	133
18	2	3	2			
33		1	5	6	1	
48		2	49	4	1	
63			4	2	4	1
78			1	3	5	4

Вариант 2

X \ Y	7	14	21	28	35	42
16	1	1	3			
30			3	3		
44			3	38	14	5
58				3	10	3
72				2	4	7

Вариант 3

X \ Y	25	48	71	94	117	140
21	2	4	1			
36			4	2		
51			10	59	3	
66				4	7	
81					2	2

Вариант 4

X \ Y	10	17	24	31	38	45
21	3	5				
35		6	7	4	5	
49			2	32	15	7
63				4	6	
77					3	1

Вариант 5

X \ Y	51	66	81	96	111	126
43	3	4				
56		5	5	4		
69			3	52	1	
82			1	10	5	
95				1	4	2

Вариант 6

X \ Y	20	25	30	35	40	45
24	4	5	1			
35			6	5		
46			2	41	12	
57				7	13	2
68						2

Вариант 7

X \ Y	41	52	63	74	85	96
35	3	4	1			
43			7	1	5	
51			4	45	1	
59			3	7	8	2
67				4	4	1

Вариант 8

X \ Y	35	39	43	47	51	55
79	1	3	2			
90			7	8	3	
101			4	51	3	
112				5	4	2
123					5	2

Вариант 9

X \ Y	13	22	31	40	49	58
22	1	4	3			
37			14	8	1	
52			39	7	6	
67			1	4		
82				5	6	1

Вариант 10

X \ Y	19	24	29	34	39	44
48	9	1	1			
58	2	14				
68		3	42	2		
78				15		
80					4	7

Вариант 11

X \ Y	21	27	33	39	45	51
15	3	4				
18			5	7		

Вариант 12

X \ Y	25	30	35	40	45	50
50	2	3				
60		7	3			

21	1	6	48	5		
24			2	4	4	
27		1	1	2	3	4

70			2	50	2	
80			1	10	6	
90				4	7	3

Вариант 13

X \ *Y*	34	50	64	80	114	115
30	1	1	2			
35		3	4	5		
39			9	59	2	
44				1	3	
49				7	1	2

Вариант 14

X \ *Y*	54	59	62	67	73	76
74	4	3				
96		2	5	6		
118			6	30	14	6
140			2	5	9	5
162					1	2

Вариант 15

X \ *Y*	23	45	67	89	111	133
18	2	1	4			
33		1	5	6	1	
48		1	40	5	10	
63			4	2	3	2
78			1	1	7	4

Вариант 16

X \ *Y*	7	14	21	28	35	42
16	1	1	3			
30			4	2		
44			6	35	14	5
58				8	5	3
72				2	3	8

Вариант 17

X \ *Y*	25	48	71	94	117	140
21	3	2	1			
36			5	1		
51			20	49	3	
66				5	6	
81					1	3

Вариант 18

X \ *Y*	10	17	24	31	38	45
21	3	5				
35		5	8	6	3	
49			12	22	15	7
63				4	6	
77					2	2

Вариант 19

X \ *Y*	51	66	81	96	111	126
43	3	4				
56		8	2	4		
69			23	32	1	
82			1	10	5	
95				2	3	2

Вариант 20

X \ *Y*	20	25	30	35	40	45
24	3	6	1			
35			6	5		
46			22	21	12	
57				7	13	3
68						1

Вариант 21

X \ *Y*	41	52	63	74	85	96
35	3	4	1			
43			8	1	4	
51			4	35	11	
59			3	7	9	1
67				4	4	1

Вариант 22

X \ *Y*	35	39	43	47	51	55
79	1	3	2			
90			5	10	3	
101			4	51	3	
112				4	5	2
123					3	5

Вариант 23

X \ *Y*	13	22	31	40	49	58
22	1	4	3			
37			8	14	1	

Вариант 24

X \ *Y*	19	24	29	34	39	44
48	7	3	1			
58	5	11				

52			39	6	7	
67			1	4		
82				3	8	1

68		3	42	2		
78				15		
80					3	8

Вариант 25

X \ *Y*	21	27	33	39	45	51
15	3	4				
18			5	7		
21	1	6	38	15		
24			2	2	6	
27		1	1	3	2	4

Вариант 26

X \ *Y*	25	30	35	40	45	50
50	2	3				
60		8	2			
70			2	50	2	
80			1	12	4	
90				4	5	5

Вариант 27

X \ *Y*	34	50	64	80	114	115
30	1	1	2			
35		1	6	5		
39			19	49	2	
44				1	3	
49				7	1	2

Вариант 28

X \ *Y*	54	59	62	67	73	76
74	3	4				
96		1	6	6		
118			16	20	14	6
140			2	5	9	5
162					1	2

Вариант 29

X \ *Y*	51	66	81	96	111	126
43	3	4				
56		4	6	4		
69			13	32	11	
82			1	5	10	
95				1	3	3

Вариант 30

X \ *Y*	20	25	30	35	40	45
24	3	4	3			
35			2	9		
46			22	21	12	
57				5	15	3
68						1

Приложение Б. Образец оформления титульного листа

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ФГАОУ ВО «МУРМАНСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра цифровых технологий,
математики и экономики

Расчетно-графическая работа
«Статистическая обработка экспериментальных данных»

по дисциплине «Специальные разделы высшей математики», часть 2

Вариант 10

Выполнил: Петров Н.К.,
студент группы ИВТ-20о

Проверил: Кацуба В.С.,
доцент кафедры ЦТМиЭ

Оценка: ____________________

Дата: ______________________

Мурманск, 2022