

Санкт-Петербургский  
Политехнический университет Петра Великого

**Отчет по лабораторным работам №1-6  
по дисциплине  
"Математическая статистика"**

Студент:	Скворцов Владимир Сергеевич
Преподаватель:	Баженов Александр Николаевич
Группа:	5030102/10201

Санкт-Петербург  
2024

# Содержание

<b>1</b>	<b>Постановка задачи</b>	<b>3</b>
1.1	Описательная статистика . . . . .	3
1.2	Точечное оценивание характеристик положения и рассеяния . . . . .	3
<b>2</b>	<b>Теоретическое обоснование</b>	<b>3</b>
2.1	Функции распределения . . . . .	3
2.2	Характеристики положения и рассеяния . . . . .	4
<b>3</b>	<b>Описание работы</b>	<b>4</b>
<b>4</b>	<b>Результаты</b>	<b>5</b>
4.1	Гистограммы и графики плотности распределения . . . . .	5
4.2	Характеристики положения и рассеяния . . . . .	7
<b>5</b>	<b>Выводы</b>	<b>9</b>
<b>6</b>	<b>Постановка задачи</b>	<b>10</b>
6.1	Боксплот Тьюки . . . . .	10
6.2	Доверительные интервалы для параметров нормального распределения . . . . .	10
<b>7</b>	<b>Теоретическое обоснование</b>	<b>10</b>
7.1	Функции распределения . . . . .	10
7.2	Боксплот Тьюки . . . . .	11
7.3	Доверительные интервалы для параметров нормального распределения . . . . .	11
<b>8</b>	<b>Описание работы</b>	<b>11</b>
<b>9</b>	<b>Результаты</b>	<b>12</b>
9.1	Гистограммы и графики плотности распределения . . . . .	12
9.2	Доверительные интервалы для параметров распределений . . . . .	14
<b>10</b>	<b>Выводы</b>	<b>15</b>
<b>11</b>	<b>Постановка задачи</b>	<b>16</b>
11.1	Коэффициент корреляции . . . . .	16
11.2	Простая линейная регрессия . . . . .	16
<b>12</b>	<b>Теоретическое обоснование</b>	<b>16</b>
12.1	Двумерное нормальное распределение . . . . .	16
12.2	Корреляционный момент (ковариация) и коэффициент корреляции . . . . .	17
12.3	Выборочный коэффициент корреляции Пирсона . . . . .	17
12.4	Выборочный квадрантный коэффициент корреляции . . . . .	17
12.5	Выборочный коэффициент ранговой корреляции Спирмена . . . . .	17
12.6	Эллипсы рассеивания . . . . .	17
12.7	Метод наименьших квадратов . . . . .	18
12.8	Метод наименьших модулей . . . . .	18
<b>13</b>	<b>Описание работы</b>	<b>18</b>

<b>14 Результаты</b>	<b>18</b>
14.1 Коэффициент корреляции . . . . .	18
14.2 Простая линейная регрессия . . . . .	21
<b>15 Выводы</b>	<b>25</b>

# 1 Постановка задачи

## 1.1 Описательная статистика

Для 5 распределений:

- Нормальное распределение  $N(x, 0, 1)$
- распределение Коши  $C(x, 0, 1)$
- Распределение Стьюдента  $t(x, 0, 3)$  с тремя степенями свободы
- Распределение Пуассона  $P(k, 10)$
- Равномерное распределение  $U(x, -\sqrt{3}, \sqrt{3})$

Сгенерировать выборки размером 10, 50, 1000 элементов. Построить на одном рисунке гистограмму и график плотности распределения.

## 1.2 Точечное оценивание характеристик положения и рассеяния

Сгенерировать выборки размером 10, 50, 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных:  $\bar{x}$ ,  $med\ x$ ,  $z_Q$ ,  $z_R$ ,  $z_{tr}$ . Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:  $E(z) = \bar{z}$ . Вычислить оценку дисперсии по формуле  $D(z) = \overline{z^2} - \bar{z}^2$ .

# 2 Теоретическое обоснование

## 2.1 Функции распределения

- Нормальное распределение

$$N(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad (1)$$

- Распределение Коши

$$C(x, 0, 1) = \frac{1}{\pi} \frac{1}{x^2 + 1} \quad (2)$$

- Распределение Стьюдента  $t(x, 0, 3)$  с тремя степенями свободы

$$t(x, 0, 3) = \frac{6\sqrt{3}}{\pi(3 + t^2)^2} \quad (3)$$

- Распределение Пуассона

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \quad (4)$$

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при } |x| \leq \sqrt{3} \\ 0 & \text{при } |x| > \sqrt{3} \end{cases} \quad (5)$$

## 2.2 Характеристики положения и рассеяния

- Выборочное среднее

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (6)$$

- Выборочная медиана

$$\text{med } x = \begin{cases} x_{(l+1)} & \text{при } n = 2l + 1 \\ \frac{x_{(l)} + x_{(l+1)}}{2} & \text{при } n = 2l \end{cases} \quad (7)$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \quad (8)$$

- Полусумма квартилей

Выборочная квартиль  $z_p$  порядка  $p$  определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном} \\ x_{(np)} & \text{при } np \text{ целом} \end{cases} \quad (9)$$

Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \quad (10)$$

- Усечённое среднее

$$z_{tr} = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}, \quad r \approx \frac{n}{4} \quad (11)$$

- Среднее характеристики

$$E(z) = \bar{z} \quad (12)$$

- Оценка дисперсии

$$D(z) = \overline{z^2} - \bar{z}^2 \quad (13)$$

## 3 Описание работы

Лабораторные работы выполнены с использованием Python и его сторонних библиотек `numpy`, `pandas`, `matplotlib`, `seaborn` были построены гистограммы распределений и посчитаны характеристики положения.

Ссылка на GitHub репозиторий: <https://github.com/vladimir-skvortsov/spbstu-mathematical-statistics>

## 4 Результаты

### 4.1 Гистограммы и графики плотности распределения

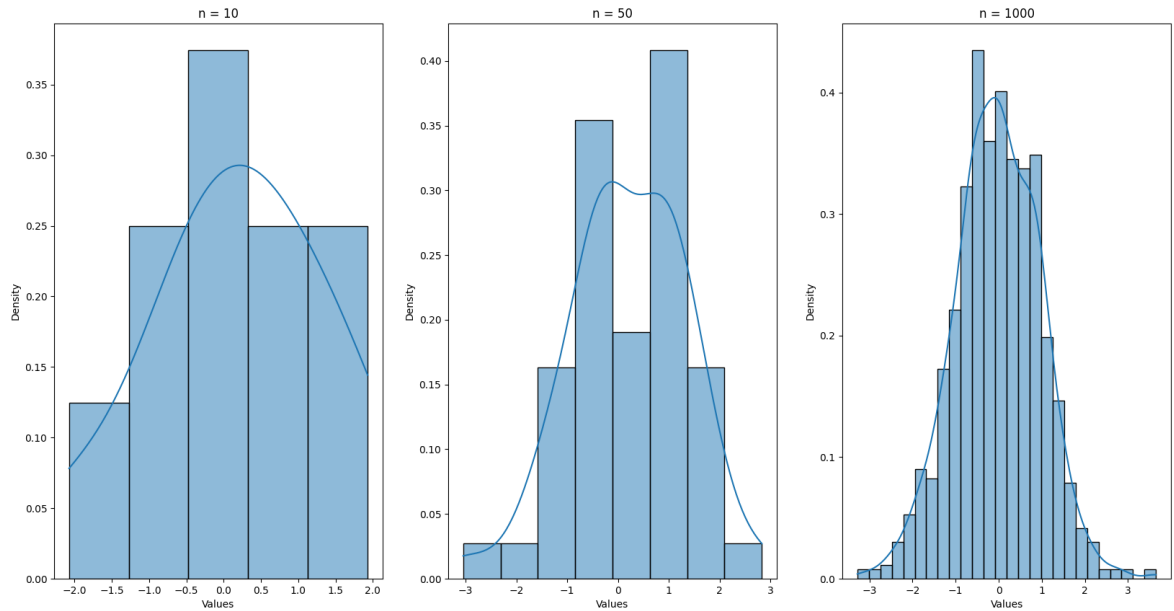


Рис. 1: Нормальное распределение (14)

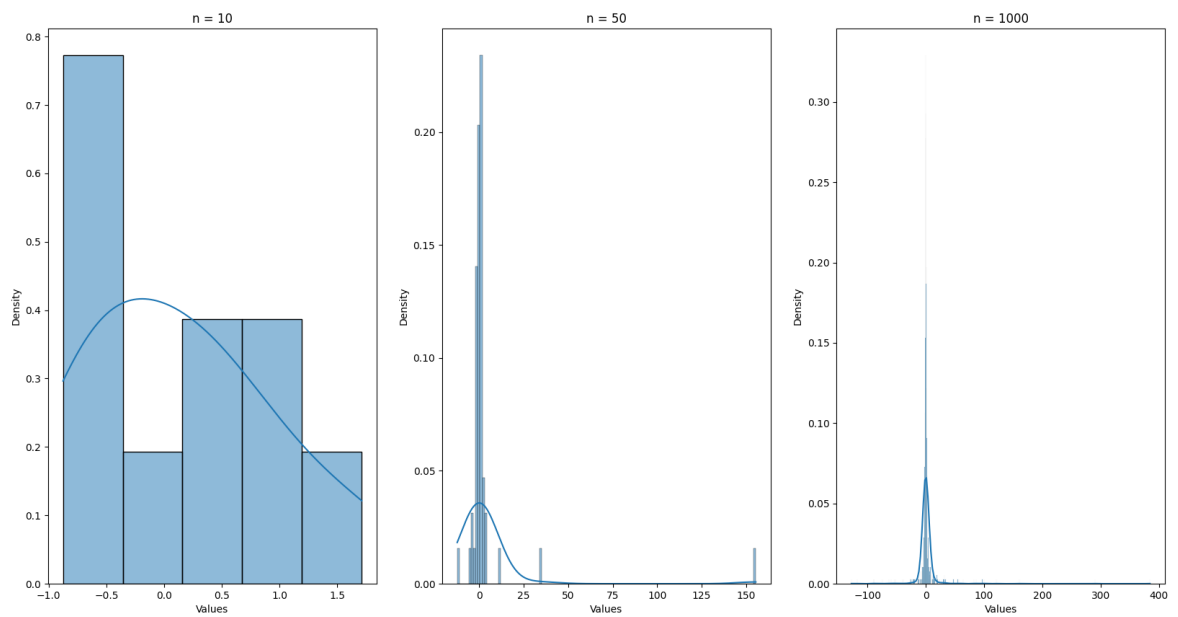


Рис. 2: Распределение Коши (15)

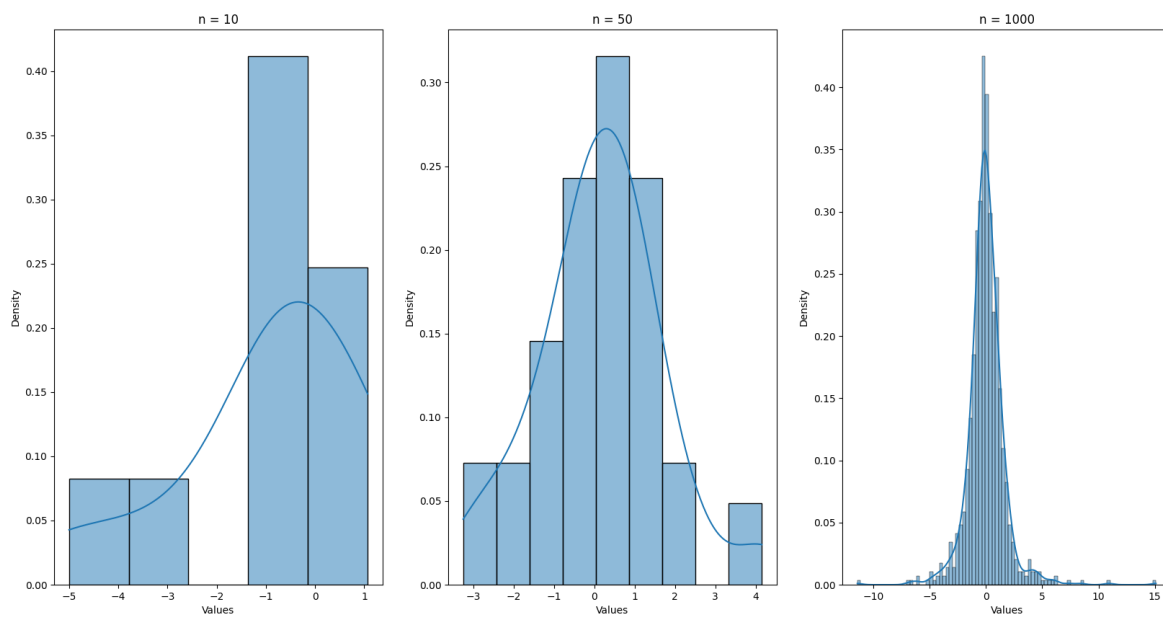


Рис. 3: Распределение Стьюдента (16)

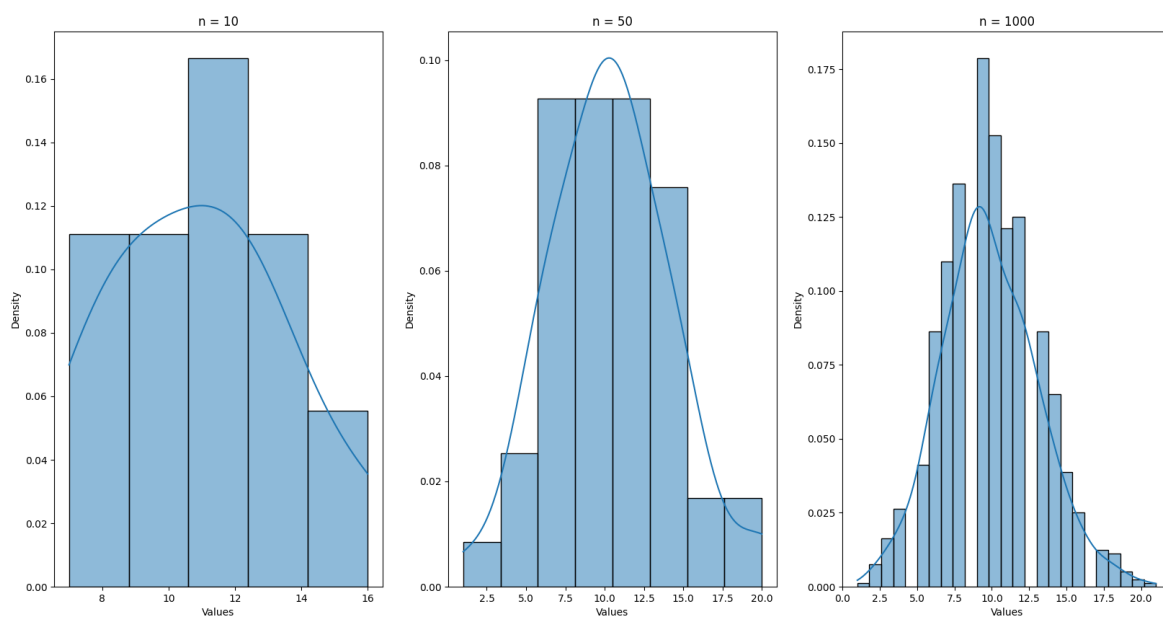


Рис. 4: Распределение Пуассона (17)

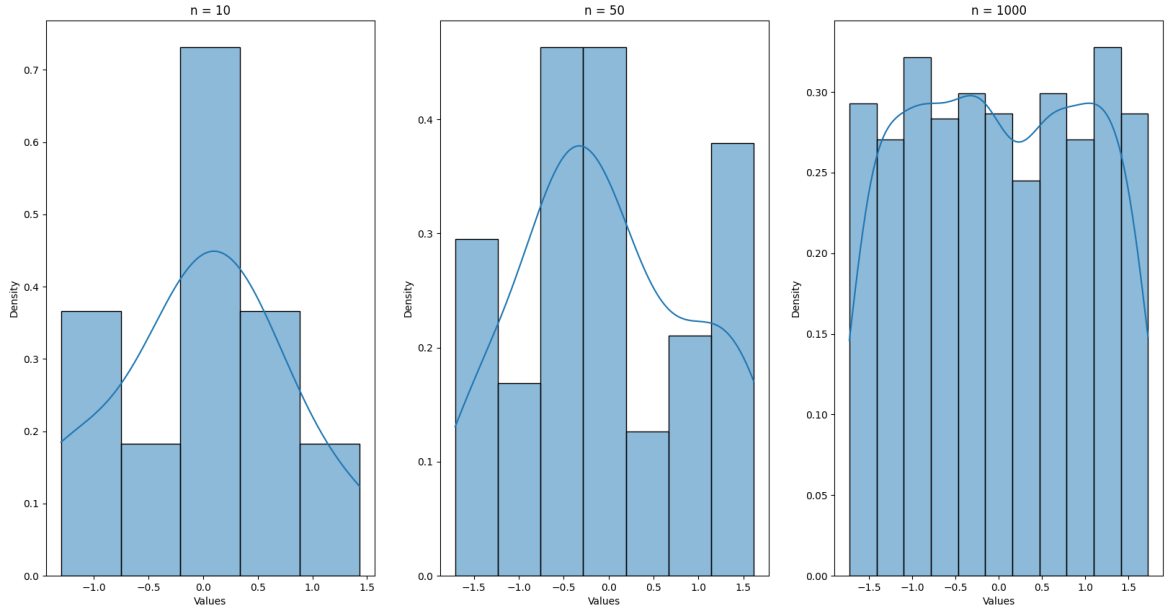


Рис. 5: Равномерное распределение (18)

## 4.2 Характеристики положения и рассеяния

n = 10					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-1.747 \times 10^{-2}$	$-1.928 \times 10^{-2}$	$-1.949 \times 10^{-2}$	$-1.449 \times 10^{-2}$	$-7.937 \times 10^{-3}$
$D(z)$ (13)	$1.009 \times 10^{-1}$	$1.427 \times 10^{-1}$	$1.878 \times 10^{-1}$	$1.154 \times 10^{-1}$	$1.608 \times 10^{-1}$
n = 50					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-7.937 \times 10^{-3}$	$1.009 \times 10^{-1}$	$1.427 \times 10^{-1}$	$1.878 \times 10^{-1}$	$1.154 \times 10^{-1}$
$D(z)$ (13)	$9.941 \times 10^{-3}$	$1.554 \times 10^{-2}$	$9.559 \times 10^{-2}$	$1.239 \times 10^{-2}$	$2.000 \times 10^{-2}$
n = 1000					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$3.800 \times 10^{-5}$	$-1.779 \times 10^{-3}$	$-2.971 \times 10^{-3}$	$1.002 \times 10^{-3}$	$-8.500 \times 10^{-5}$
$D(z)$ (13)	$9.850 \times 10^{-4}$	$1.682 \times 10^{-3}$	$6.138 \times 10^{-2}$	$1.243 \times 10^{-3}$	$1.939 \times 10^{-3}$

Таблица 1: Нормальное распределение



n = 10					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	-4.724	$-1.599 \times 10^{-2}$	$-2.361 \times 10$	$-1.518 \times 10^{-2}$	-8.311
$D(z)$ (13)	$1.148 \times 10^4$	$3.371 \times 10^{-1}$	$2.865 \times 10^5$	1.164	$3.170 \times 10^4$
n = 50					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$7.817 \times 10^{-1}$	$1.222 \times 10^{-2}$	$3,703 \times 10$	$8.637 \times 10^{-3}$	$8.573 \times 10^{-1}$
$D(z)$ (13)	$4.319 \times 10^2$	$2.532 \times 10^{-2}$	$1.060 \times 10^6$	$5.501 \times 10^{-2}$	$1.677 \times 10^2$
n = 1000					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-3.361 \times 10^{-1}$	$-1.532 \times 10^{-3}$	$-1.290 \times 10^2$	$-1.540 \times 10^{-3}$	$-4.972 \times 10^{-2}$
$D(z)$ (13)	$2.406 \times 10^2$	$2.310 \times 10^{-3}$	$5.036 \times 10^7$	$4.735 \times 10^{-3}$	$1.743 \times 10^2$

Таблица 2: Распределение Коши

n = 10					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$1.626 \times 10^{-2}$	$4.667 \times 10^{-3}$	$4.092 \times 10^{-2}$	$1.432 \times 10^{-2}$	$7.500 \times 10^{-4}$
$D(z)$ (13)	$2.591 \times 10^{-1}$	$1.838 \times 10^{-1}$	1.659	$1.846 \times 10^{-1}$	$4.319 \times 10^{-1}$
n = 50					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-2.158 \times 10^{-3}$	$-1.389 \times 10^{-3}$	$2.124 \times 10^{-2}$	$3.592 \times 10^{-3}$	$-1.675 \times 10^{-2}$
$D(z)$ (13)	$2.691 \times 10^{-2}$	$1.905 \times 10^{-2}$	9.894	$1.848 \times 10^{-2}$	$5.278 \times 10^{-2}$
n = 1000					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$3.350 \times 10^{-4}$	$-2.380 \times 10^{-4}$	$-5.482 \times 10^{-2}$	$1.620 \times 10^{-4}$	$6.790 \times 10^{-4}$
$D(z)$ (13)	$2.898 \times 10^{-3}$	$1.903 \times 10^{-3}$	$3.253 \times 10$	$1.944 \times 10^{-3}$	$5.656 \times 10^{-3}$

Таблица 3: Распределение Стьюдента

n = 10					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$1.000 \times 10$	9.874	$1.029 \times 10$	9.918	9.937
$D(z)$ (13)	1.082	1.478	2.018	1.284	1.699
n = 50					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$1.001 \times 10$	9.856	$1.090 \times 10$	9.945	$1.001 \times 10$
$D(z)$ (13)	$9.575 \times 10^{-2}$	$1.974 \times 10^{-1}$	$9.572 \times 10^{-1}$	$1.398 \times 10^{-1}$	$2.048 \times 10^{-1}$
n = 1000					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$1.000 \times 10$	9.997	$1.163 \times 10$	9.994	$1.000 \times 10$
$D(z)$ (13)	$1.014 \times 10^{-2}$	$2.991 \times 10^{-3}$	$6.344 \times 10^{-1}$	$2.964 \times 10^{-3}$	$2.072 \times 10^{-2}$

Таблица 4: Распределение Пуассона

n = 10					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-5.450 \times 10^{-3}$	$-6.939 \times 10^{-3}$	$-5.412 \times 10^{-3}$	$-7.901 \times 10^{-3}$	$-1.561 \times 10^{-2}$
$D(z)$ (13)	$1.041 \times 10^{-1}$	$2.402 \times 10^{-1}$	$4.402 \times 10^{-2}$	$1.443 \times 10^{-1}$	$1.722 \times 10^{-1}$
n = 50					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$-1.915 \times 10^{-3}$	$-6.312 \times 10^{-3}$	$-1.349 \times 10^{-3}$	$1.960 \times 10^{-3}$	$-4.766 \times 10^{-3}$
$D(z)$ (13)	$1.002 \times 10^{-2}$	$2.972 \times 10^{-2}$	$5.990 \times 10^{-4}$	$1.428 \times 10^{-2}$	$1.894 \times 10^{-2}$
n = 1000					
	$\bar{x}$ (6)	$med\ x$ (7)	$z_R$ (8)	$z_Q$ (10)	$z_{tr}$ (11)
$E(z)$ (12)	$4.700 \times 10^{-4}$	$9.240 \times 10^{-4}$	$-1.330 \times 10^{-4}$	$-3.550 \times 10^{-4}$	$-3.870 \times 10^{-4}$
$D(z)$ (13)	$1.014 \times 10^{-3}$	$3.127 \times 10^{-3}$	$5.000 \times 10^{-6}$	$1.469 \times 10^{-3}$	$1.887 \times 10^{-3}$

Таблица 5: Равномерное распределение

## 5 Выводы

В процессе выполнения лабораторной работы был проведен анализ пяти уникальных распределений: нормальное, Коши, Стьюдента, Пуассона и равномерное. Были сгенерированы выборки разных объемов для каждого из них - 10, 50 и 1000 элементов. Были созданы гистограммы каждого распределения и нанесены на них графики плотности соответствующих распределений, что облегчило наглядное сопоставление формы распределения выборок с их теоретическими аналогами. Были также рассчитаны разные показатели положения и рассеяния для каждой выборки, включая выборочную среднюю величину, медиану, полусумму крайних элементов выборки, полусумму квартилей и усеченное среднее. Использовалась стандартная формула для оценки дисперсии.

На основании полученных данных были сделаны следующие выводы:

1. В случае нормального распределения, оценки показателей положения и рассеяния становятся ближе к их теоретическим значениям по мере увеличения размера выборки.
2. Для распределения Коши показатели положения и рассеяния менее стабильны и могут сильно отличаться от теоретических даже при больших размерах выборки.
3. Распределение Стьюдента при небольших размерах выборки также демонстрирует определенную нестабильность оценок, однако с увеличением размера выборки результаты становятся более точными.
4. Для распределения Пуассона и равномерного распределения, оценки показателей положения и рассеяния кажутся стабильными при любом объеме выборки.
5. В общем, выборочное среднее является наиболее чувствительным к экстремальным значениям по сравнению с медианой, особенно в меньших выборках. Однако с увеличением размера выборки, влияние этих экстремальных значений на среднее значение уменьшается. В то же время, медиана обычно более устойчива к выбросам и мало варьирует с изменением размера выборки.
6. Медиана является чувствительной к типу распределения: в нормальном и распределении Стьюдента медиана равна среднему, в распределении Коши она дает надежные, устойчивые к выбросам оценки, в Пуассоновском приближается к среднему, и в равномерном равна половине суммы минимального и максимального значений.

## 6 Постановка задачи

### 6.1 Боксплот Тьюки

Сгенерировать выборки размером 20 и 100 элементов. Построить для них боксплот Тьюки.

### 6.2 Доверительные интервалы для параметров нормального распределения

Сгенерировать выборки размером 20 и 100 элементов. Вычислить параметры положения и рассеяния:

- для нормального распределения,
- для произвольного распределения.

## 7 Теоретическое обоснование

### 7.1 Функции распределения

- Нормальное распределение

$$N(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad (14)$$

- Распределение Коши

$$C(x, 0, 1) = \frac{1}{\pi} \frac{1}{x^2 + 1} \quad (15)$$

- Распределение Стьюдента  $t(x, 0, 3)$  с тремя степенями свободы

$$t(x, 0, 3) = \frac{6\sqrt{3}}{\pi(3 + t^2)^2} \quad (16)$$

- Распределение Пуассона

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \quad (17)$$

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}}, & |x| \leq \sqrt{3} \\ 0, & |x| > \sqrt{3} \end{cases} \quad (18)$$

## 7.2 Боксплот Тьюки

Боксплот (англ. box plot) — график, использующихся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили и выбросы. Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выброса). Длину «усов» определяют разность первого квартиля и полутора межквартильных расстояний и сумма третьего квартиля и полутора межквартильных расстояний. Формула имеет вид

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), \quad X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1), \quad (19)$$

где  $X_1$  — нижняя граница уса,  $X_2$  — верхняя граница уса,  $Q_1$  — первый квартиль,  $Q_3$  — третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков. Выбросами считаются величины, такие что:

$$\begin{cases} x < X_1^T \\ x > X_2^T \end{cases} \quad (20)$$

## 7.3 Доверительные интервалы для параметров нормального распределения

Пусть  $F_T(x)$  — функция распределения Стюдента с  $n - 1$  степенями свободы. Полагая, что  $2F_T(x) - 1 = 1 - \alpha$ , где  $\alpha$  — выбранный уровень значимости. Тогда  $F_T(x) = 1 - \alpha/2$ . Пусть  $st_{1-\alpha/2}(n - 1)$  — квантиль распределения Стюдента с  $n - 1$  степенями свободы и порядка  $1 - \alpha/2$ . Тогда получаем

$$P\left(\bar{x} - \frac{st_{1-\alpha/2}(n - 1)}{\sqrt{n - 1}} < m < \bar{x} + \frac{st_{1-\alpha/2}(n - 1)}{\sqrt{n - 1}}\right) = 1 - \alpha, \quad (21)$$

что и даст доверительный интервал для  $m$  с доверительной вероятностью  $\gamma = 1 - \alpha$  для нормального распределения.

Случайная величина  $n \frac{s^2}{\sigma^2}$  распределена по закону  $\chi^2$  с  $n - 1$  степенями свободы. Тогда

$$P\left(\bar{x} - \frac{st_{1-\alpha/2}(n - 1)}{\sqrt{n - 1}} < m < \bar{x} + \frac{st_{1-\alpha/2}(n - 1)}{\sqrt{n - 1}}\right) = 1 - \alpha, \quad (22)$$

## 8 Описание работы

Лабораторные работы выполнены с использованием Python и его сторонних библиотек: `numpy`, `pandas`, `matplotlib`, `seaborn`.

Ссылка на GitHub репозиторий: <https://github.com/vladimir-skvortsov/spbstu-mathematical-statistics>

# 9 Результаты

## 9.1 Гистограммы и графики плотности распределения

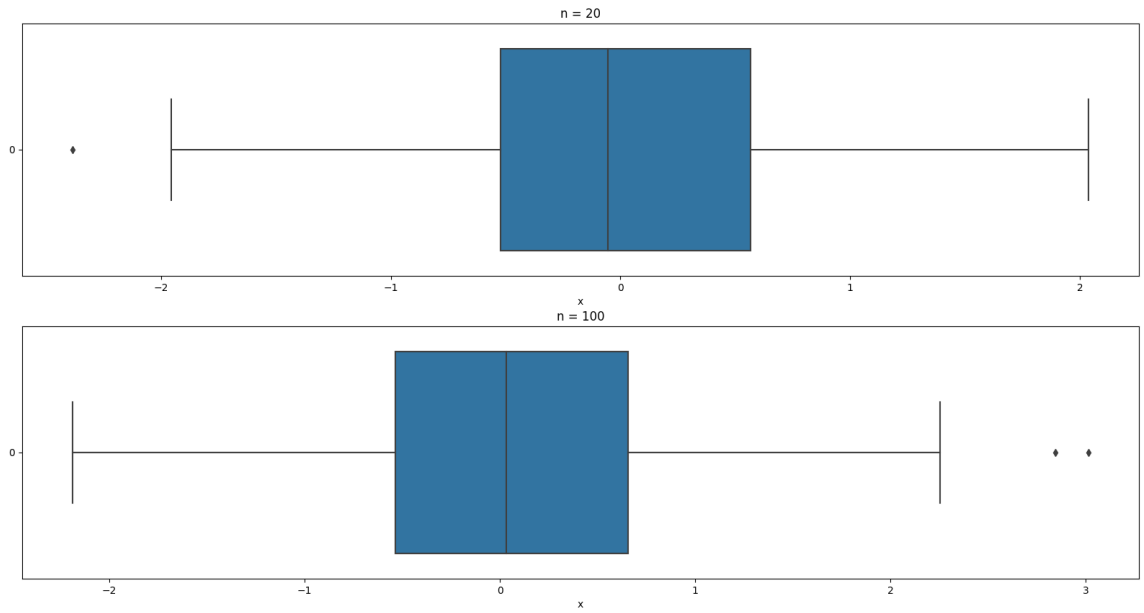


Рис. 6: Нормальное распределение (14)

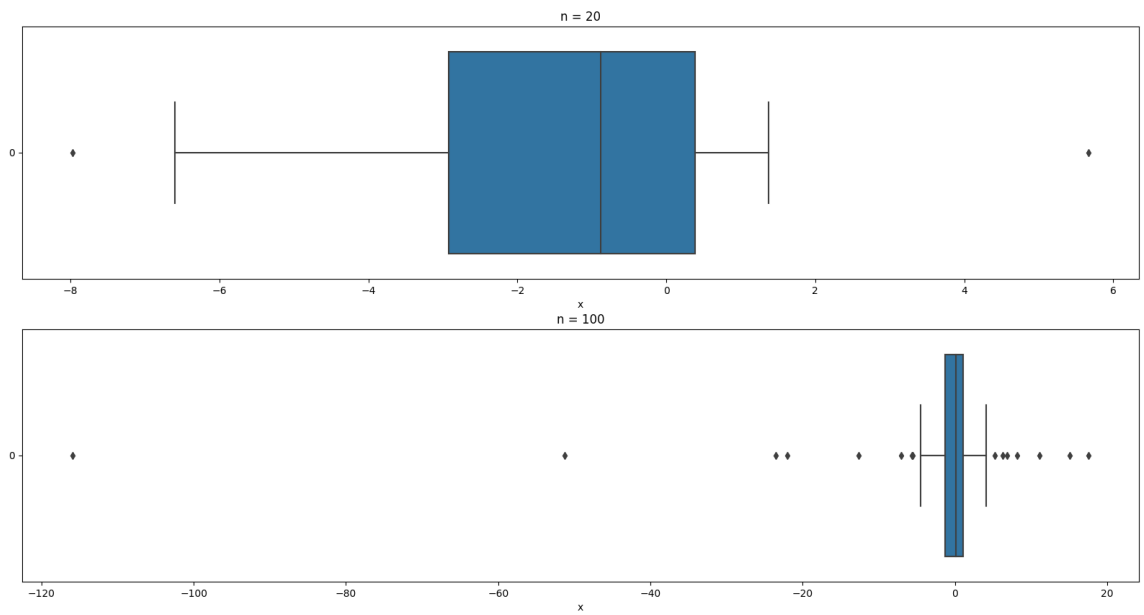


Рис. 7: Распределение Коши (15)

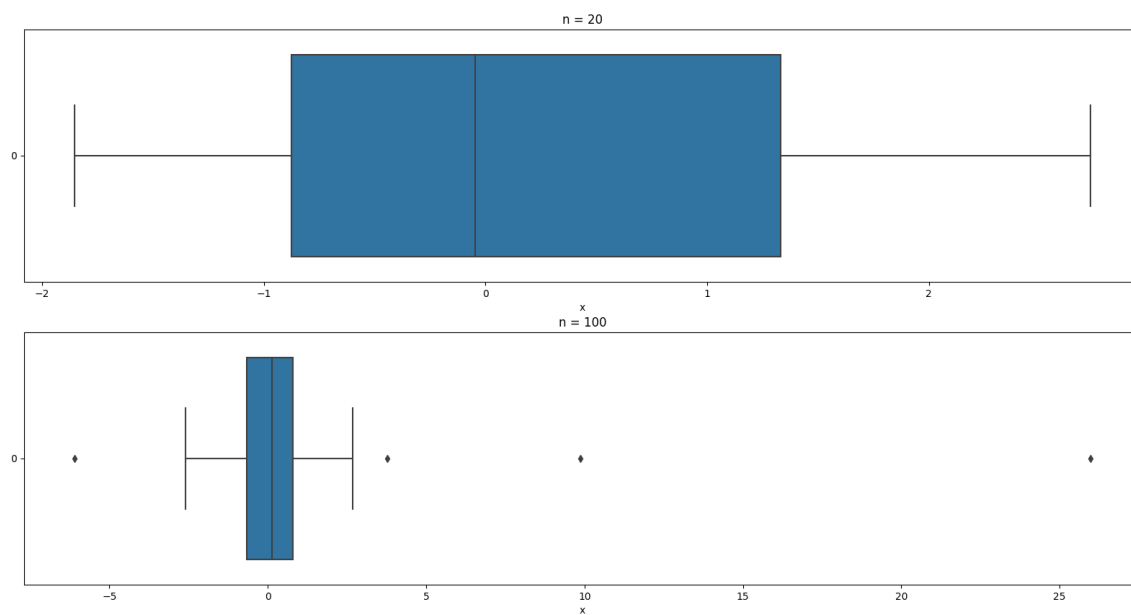


Рис. 8: Распределение Стьюдента (16)

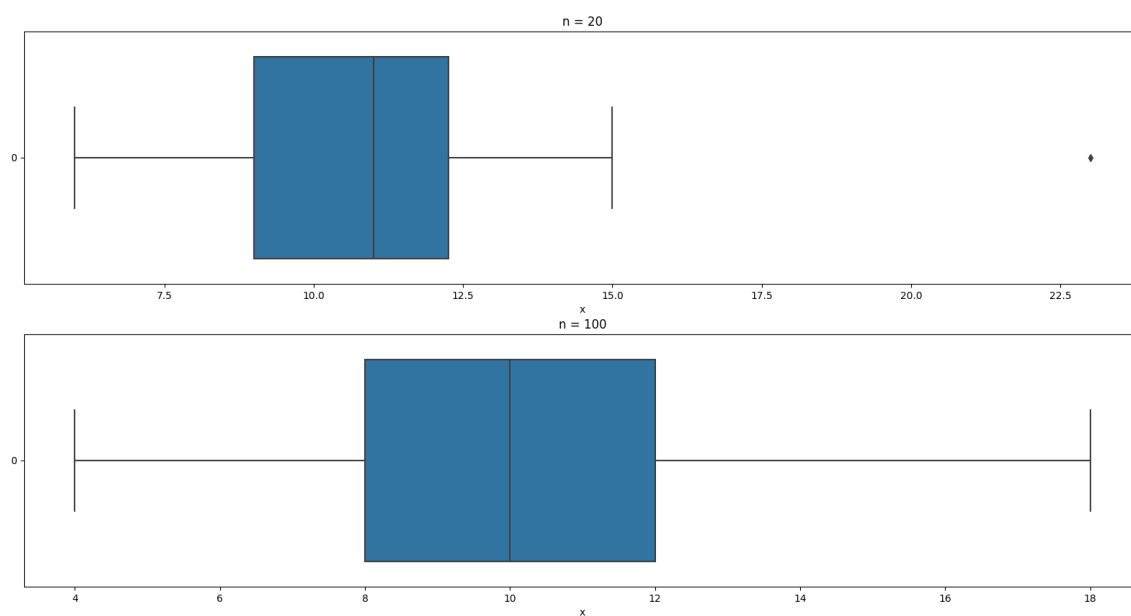


Рис. 9: Распределение Пуассона (17)

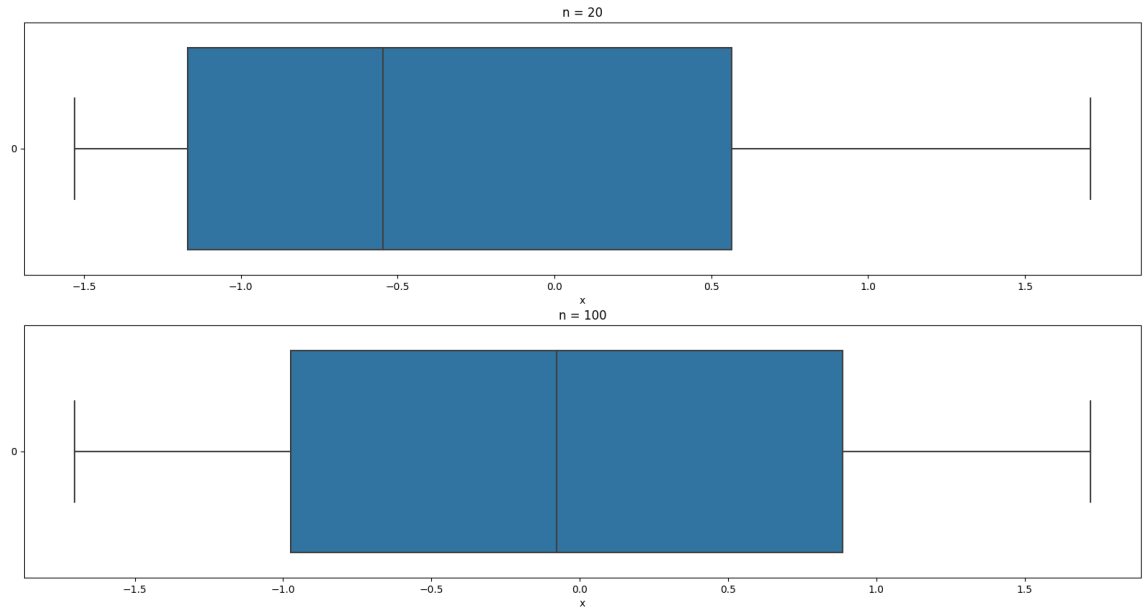


Рис. 10: Равномерное распределение (18)

## 9.2 Доверительные интервалы для параметров распределений

n = 20	m	$\sigma$
	$-0.43 < m < 0.37$	$0.66 < \sigma < 1.25$
n = 100	m	$\sigma$
	$-0.12 < m < 0.24$	$0.81 < \sigma < 1.07$

Таблица 6: Доверительные интервалы для параметров нормального распределения (14)

n = 20	$m$	$\sigma$
	$0.11 < m < 0.97$	$0.29 < \sigma < 0.33$
n = 100	$m$	$\sigma$
	$0.30 < m < 0.67$	$0.28 < \sigma < 0.33$

Таблица 7: Доверительные интервалы для параметров произвольного распределения. Асимптотический подход

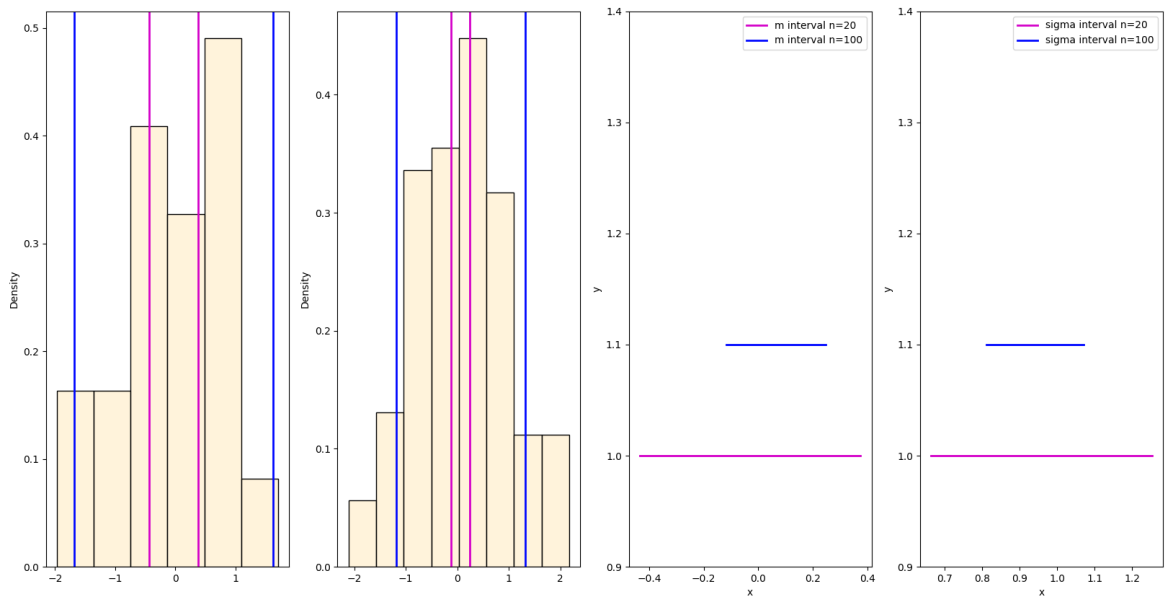


Рис. 11: Гистограммы и оценки для параметров нормального распределения

$(0.663480, -0.434162, 0.374849, 1.252336)$

$(-0.117590, 0.248381, 0.810296, 1.070570)$

## 10 Выводы

По результатам выполнения лабораторной работы были сгенерированы выборки размером 20 и 100 элементов и построены для них боксплоты Тьюки.

Боксплот позволяет наглядно представить основные характеристики выборки - медиану, квартили, межквартильный размах и выбросы. На основе построенных графиков можно увидеть разницу в распределении данных для двух выборок. Для выборки размером в 100 элементов представленные метрики имеют более проработанный вид, ведь с увеличением размера выборки улучшается точность оценок параметров распределения, но при этом количество выбросов растет.

Также в ходе выполнения лабораторной работы были сгенерированы две выборки размерами 20 и 100 элементов для нормального и произвольного распределения. Затем для каждой из них были вычислены параметры распределения: среднее значение и дисперсия.



Результаты, представленные графически, демонстрируют, что количество элементов в выборке влияет на точность оценок параметров. Более большое количество наблюдений (т.е. 100 элементов) приводит к более точным и стабильным оценкам среднего и дисперсии, как для нормального, так и для произвольного распределения. Для выборки с меньшим количеством элементов (20 элементов) оценки могут сильно варьироваться в зависимости от конкретной выборки, что также наглядно отображено на графиках.

Лабораторная работа иллюстрирует важнейший статистический принцип: точность статистической оценки увеличивается с ростом объема выборки. Результаты этого исследования подчеркивают значимость использования достаточно больших выборок для надежного анализа данных.

## 11 Постановка задачи

### 11.1 Коэффициент корреляции

Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения  $N(x, y, 0, 0, 1, 1, \rho)$ . Коэффициент корреляции  $\rho$  взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадратного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

$$f(x, y) = 0.9N(x, y, 0, 0, 1, 1, 0.9) + 0.1N(x, y, 0, 0, 10, 10, -0.9).$$

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

### 11.2 Простая линейная регрессия

Найти оценки коэффициентов линейной регрессии  $y_i = a + bx_i + e_i$ , используя 20 точек на отрезке  $[-1.8; 2]$  с равномерным шагом равным 0.2. Ошибку  $e_i$  считать нормально распределённой с параметрами  $(0, 1)$ . В качестве эталонной зависимости взять  $y_i = 2 + 2x_i + e_i$ . При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Прodelать то же самое для выборки, у которой в значения  $y_1$  и  $y_{20}$  вносятся возмущения 10 и -10.

## 12 Теоретическое обоснование

### 12.1 Двумерное нормальное распределение

Двумерная случайная величина  $(X, Y)$  называется распределённой нормально (или просто нормальной), если её плотность вероятности определена формулой

$$N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho \frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} \right] \right\} \quad (23)$$

Компоненты  $X, Y$  двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями  $\bar{x}, \bar{y}$  и средними квадратическими отклонениями  $\sigma_x, \sigma_y$  соответственно.

Параметр  $\rho$  называется коэффициентом корреляции.

## 12.2 Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционный момент, иначе ковариация, двух случайных величин  $X$  и  $Y$ :

$$K = \mathbf{cov}(X, Y) = \mathbf{M}[(X - \bar{x})(Y - \bar{y})] \quad (24)$$

Коэффициент корреляции  $\rho$  двух случайных величин  $X$  и  $Y$ :

$$\rho = \frac{K}{\sigma_x \sigma_y} \quad (25)$$

## 12.3 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона:

$$r = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2 \frac{1}{n} \sum (y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}, \quad (26)$$

где  $K$ ,  $s_X^2$ ,  $s_Y^2$  — выборочные ковариации и дисперсии случайных величин  $X$  и  $Y$ .

## 12.4 Выборочный квадрантный коэффициент корреляции

Выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \quad (27)$$

где  $n_1, n_2, n_3, n_4$  — количество точек с координатами  $(x_i, y_i)$ , попавшими, соответственно, в I, II, III, IV квадранты декартовой системы с осями  $x' = x - \mathbf{med}x$ ,  $y' = y - \mathbf{med}y$ .

## 12.5 Выборочный коэффициент ранговой корреляции Спирмена

Обозначим ранги, соответствующие значениям переменной  $X$ , через  $u$ , а ранги, соответствующие значениям переменной  $Y$ , — через  $v$ .

Выборочный коэффициент ранговой корреляции Спирмена:

$$r_S = \frac{\frac{1}{n} \sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n} \sum (u_i - \bar{u})^2 \frac{1}{n} \sum (v_i - \bar{v})^2}}, \quad (28)$$

где  $\bar{u} = \bar{v} = \frac{1+2+\dots+n}{n} = \frac{n+1}{2}$  — среднее значение рангов.

## 12.6 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость  $xOy$ :

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - 2\rho \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \text{const.} \quad (29)$$

Центр эллипса [30](#) находится в точке с координатами  $(x, y)$ ; оси симметрии эллипса составляют с осью  $Ox$  углы, определяемые уравнением

$$\text{tg } 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \quad (30)$$

## 12.7 Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распространённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \min_{\beta_0, \beta_1}. \quad (31)$$

Задача минимизации квадратичного критерия (31) носит название задачи метода наименьших квадратов (МНК), а оценки  $\beta_0$ ,  $\beta_1$  параметров  $\beta_0$ ,  $\beta_1$ , реализующие минимум критерия (31), называют МНК-оценками.

## 12.8 Метод наименьших модулей

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является использование метода наименьших модулей вместо метода наименьших квадратов:

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \rightarrow \min_{\beta_0, \beta_1}. \quad (32)$$

## 13 Описание работы

Лабораторные работы выполнены с использованием Python и его сторонних библиотек `numpy`, `pandas`, `matplotlib`, `seaborn` были построены гистограммы распределений и посчитаны характеристики положения.

Ссылка на GitHub репозиторий: <https://github.com/vladimir-skvortsov/spbstu-mathematical-statistics>

## 14 Результаты

### 14.1 Коэффициент корреляции

$n = 20, \rho = 0$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$8.051 \times 10^{-3}$	$8.633 \times 10^{-3}$	$1.216 \times 10^{-2}$
Среднее квадратов	$5.501 \times 10^{-2}$	$5.418 \times 10^{-2}$	$1.033 \times 10^{-1}$
Дисперсия	$5.495 \times 10^{-2}$	$5.410 \times 10^{-2}$	$1.031 \times 10^{-1}$
$n = 20, \rho = 0.5$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$4.933 \times 10^{-1}$	$4.674 \times 10^{-1}$	$4.644 \times 10^{-1}$
Среднее квадратов	$2.743 \times 10^{-1}$	$2.534 \times 10^{-1}$	$3.139 \times 10^{-1}$
Дисперсия	$3.093 \times 10^{-2}$	$3.496 \times 10^{-2}$	$9.823 \times 10^{-2}$
$n = 20, \rho = 0.9$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$8.938 \times 10^{-1}$	$8.646 \times 10^{-1}$	$9.837 \times 10^{-1}$
Среднее квадратов	$8.014 \times 10^{-1}$	$7.527 \times 10^{-1}$	1.026
Дисперсия	$2.454 \times 10^{-3}$	$5.209 \times 10^{-3}$	$5.804 \times 10^{-2}$
$n = 60, \rho = 0$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$8.143 \times 10^{-3}$	$8.747 \times 10^{-3}$	$8.485 \times 10^{-3}$
Среднее квадратов	$1.709 \times 10^{-2}$	$1.689 \times 10^{-2}$	$3.111 \times 10^{-2}$
Дисперсия	$1.703 \times 10^{-2}$	$1.682 \times 10^{-2}$	$3.104 \times 10^{-2}$
$n = 60, \rho = 0.5$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$4.985 \times 10^{-1}$	$4.757 \times 10^{-1}$	$4.668 \times 10^{-1}$
Среднее квадратов	$2.585 \times 10^{-1}$	$2.373 \times 10^{-1}$	$2.504 \times 10^{-1}$
Дисперсия	$1.000 \times 10^{-2}$	$1.094 \times 10^{-2}$	$3.256 \times 10^{-2}$
$n = 60, \rho = 0.9$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$8.979 \times 10^{-1}$	$8.810 \times 10^{-1}$	$9.937 \times 10^{-1}$
Среднее квадратов	$8.069 \times 10^{-1}$	$7.774 \times 10^{-1}$	1.004
Дисперсия	$7.297 \times 10^{-4}$	$1.202 \times 10^{-3}$	$1.700 \times 10^{-2}$
$n = 100, \rho = 0$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$1.396 \times 10^{-3}$	$8.326 \times 10^{-5}$	$1.584 \times 10^{-3}$
Среднее квадратов	$9.856 \times 10^{-3}$	$9.848 \times 10^{-3}$	$1.972 \times 10^{-2}$
Дисперсия	$9.854 \times 10^{-3}$	$9.848 \times 10^{-3}$	$1.972 \times 10^{-2}$
$n = 100, \rho = 0.5$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$5.013 \times 10^{-1}$	$4.812 \times 10^{-1}$	$4.723 \times 10^{-1}$
Среднее квадратов	$2.568 \times 10^{-1}$	$2.375 \times 10^{-1}$	$2.407 \times 10^{-1}$
Дисперсия	$5.481 \times 10^{-3}$	$6.013 \times 10^{-3}$	$1.762 \times 10^{-2}$
$n = 100, \rho = 0.9$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$8.999 \times 10^{-1}$	$8.866 \times 10^{-1}$	1.003
Среднее квадратов	$8.103 \times 10^{-1}$	$7.868 \times 10^{-1}$	1.017
Дисперсия	$4.017 \times 10^{-4}$	$6.665 \times 10^{-4}$	$1.049 \times 10^{-2}$

Таблица 8: Характеристики нормального двумерного распределения

$n = 20$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$-7.987 \times 10^{-2}$	$-7.020 \times 10^{-2}$	$-6.336 \times 10^{-2}$
Среднее квадратов	$5.968 \times 10^{-2}$	$5.944 \times 10^{-2}$	$1.112 \times 10^{-1}$
Дисперсия	$5.330 \times 10^{-2}$	$5.451 \times 10^{-2}$	$1.072 \times 10^{-1}$
$n = 60$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$9.290 \times 10^{-2}$	$-8.988 \times 10^{-2}$	$-8.730 \times 10^{-2}$
Среднее квадратов	$2.606 \times 10^{-2}$	$2.553 \times 10^{-2}$	$4.290 \times 10^{-2}$
Дисперсия	$1.743 \times 10^{-2}$	$1.745 \times 10^{-2}$	$3.528 \times 10^{-2}$
$n = 100$			
	$r$ (26)	$r_S$ (28)	$r_Q$ (27)
Среднее	$-1.013 \times 10^{-1}$	$-9.639 \times 10^{-2}$	$-9.011 \times 10^{-2}$
Среднее квадратов	$2.047 \times 10^{-2}$	$1.984 \times 10^{-2}$	$2.968 \times 10^{-2}$
Дисперсия	$1.021 \times 10^{-2}$	$1.054 \times 10^{-2}$	$2.156 \times 10^{-2}$

Таблица 9: Характеристики смеси нормальных распределений

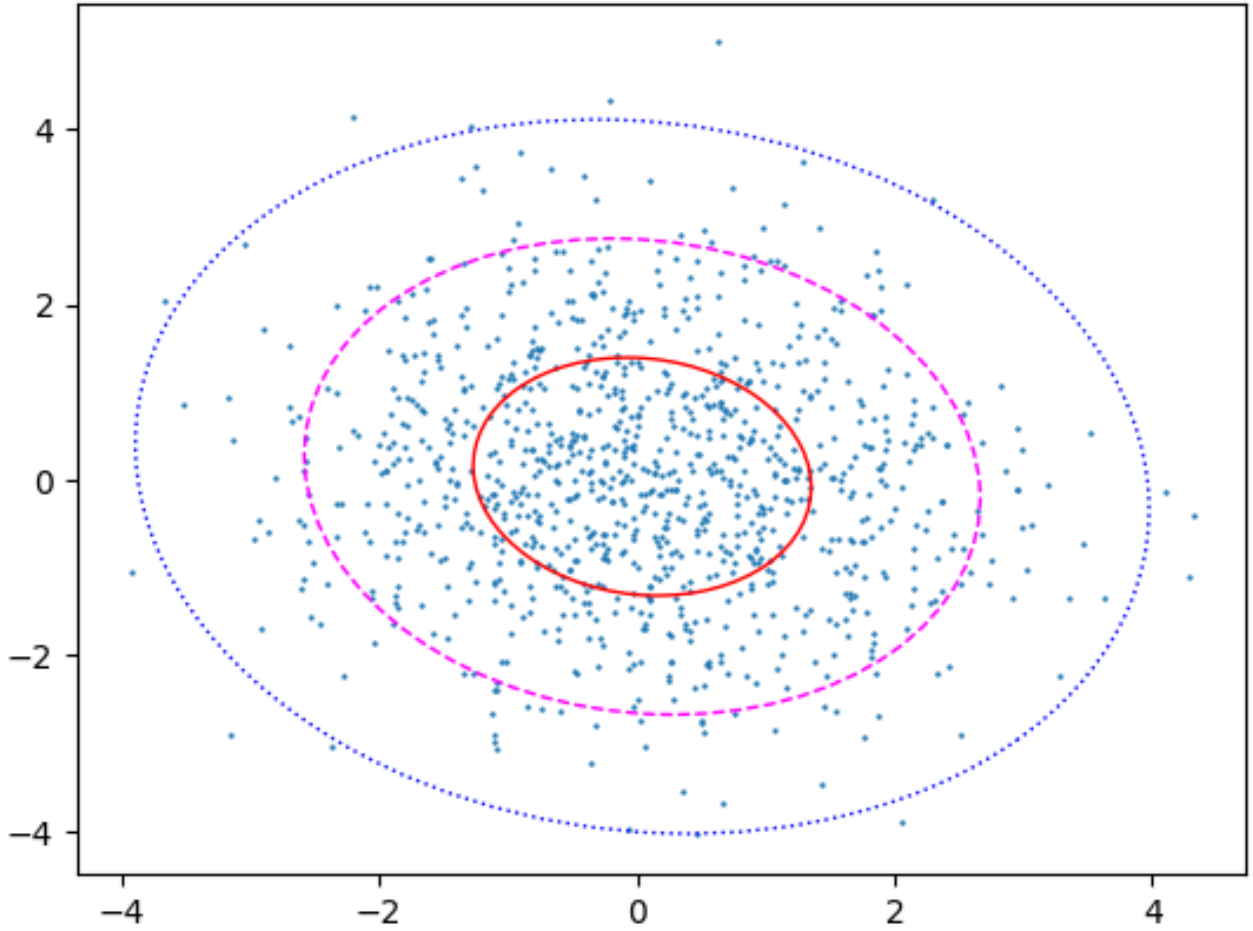


Рис. 12: Смесь нормальных распределений и эллипсы равновероятности

## 14.2 Простая линейная регрессия

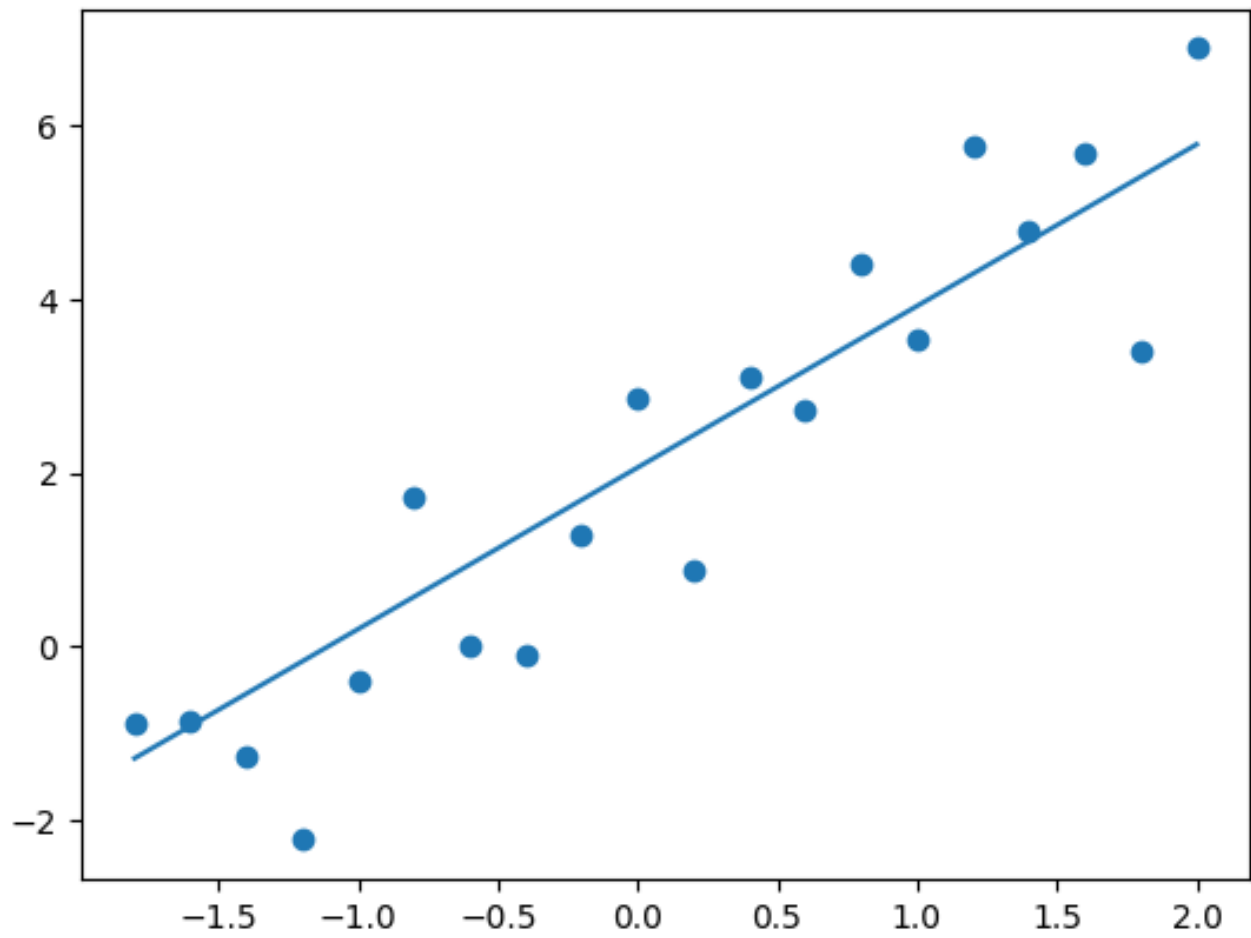


Рис. 13: Метод наименьших квадратов (1.861, 2.061)

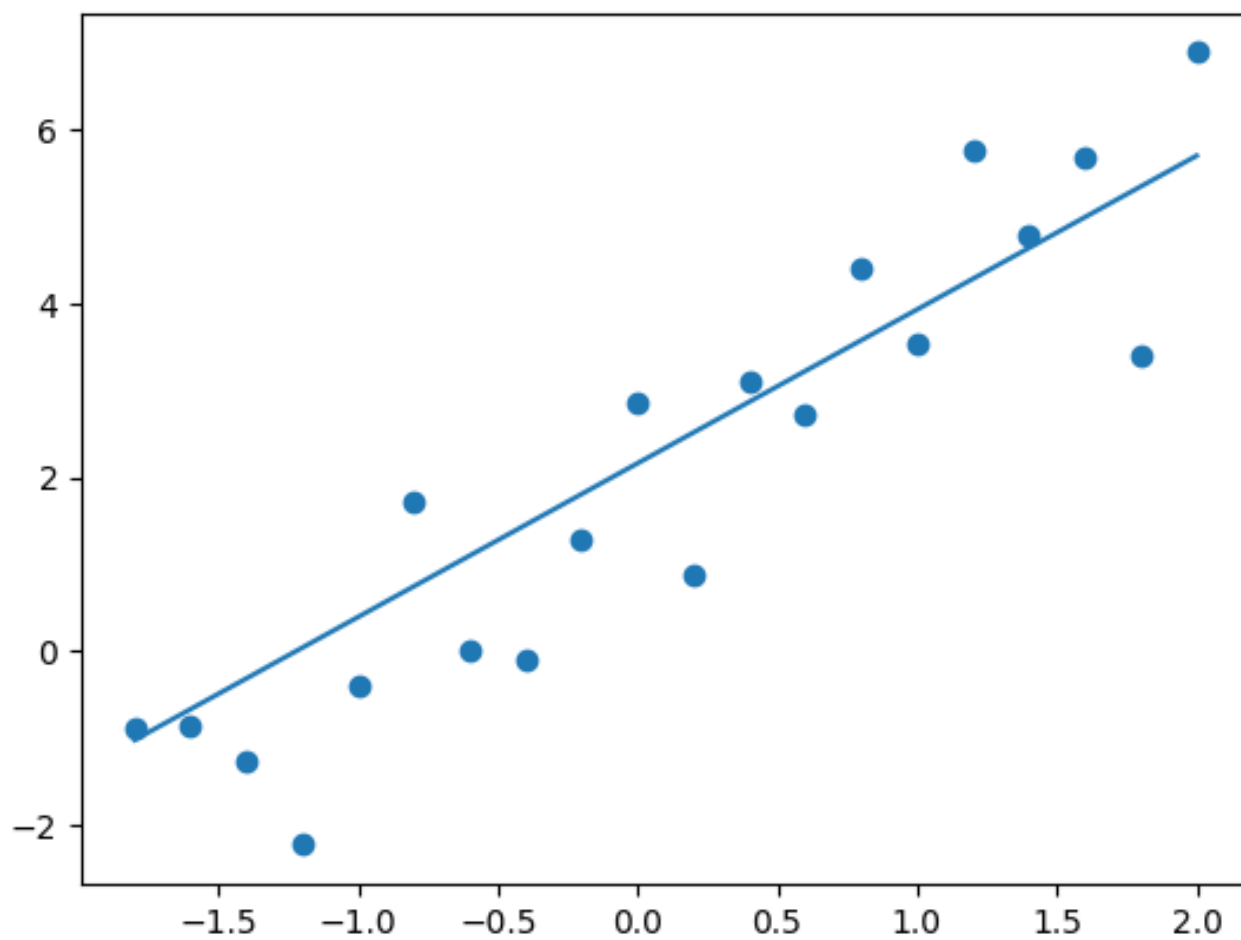


Рис. 14: Метод наименьших модулей (1.7691, 2.162)

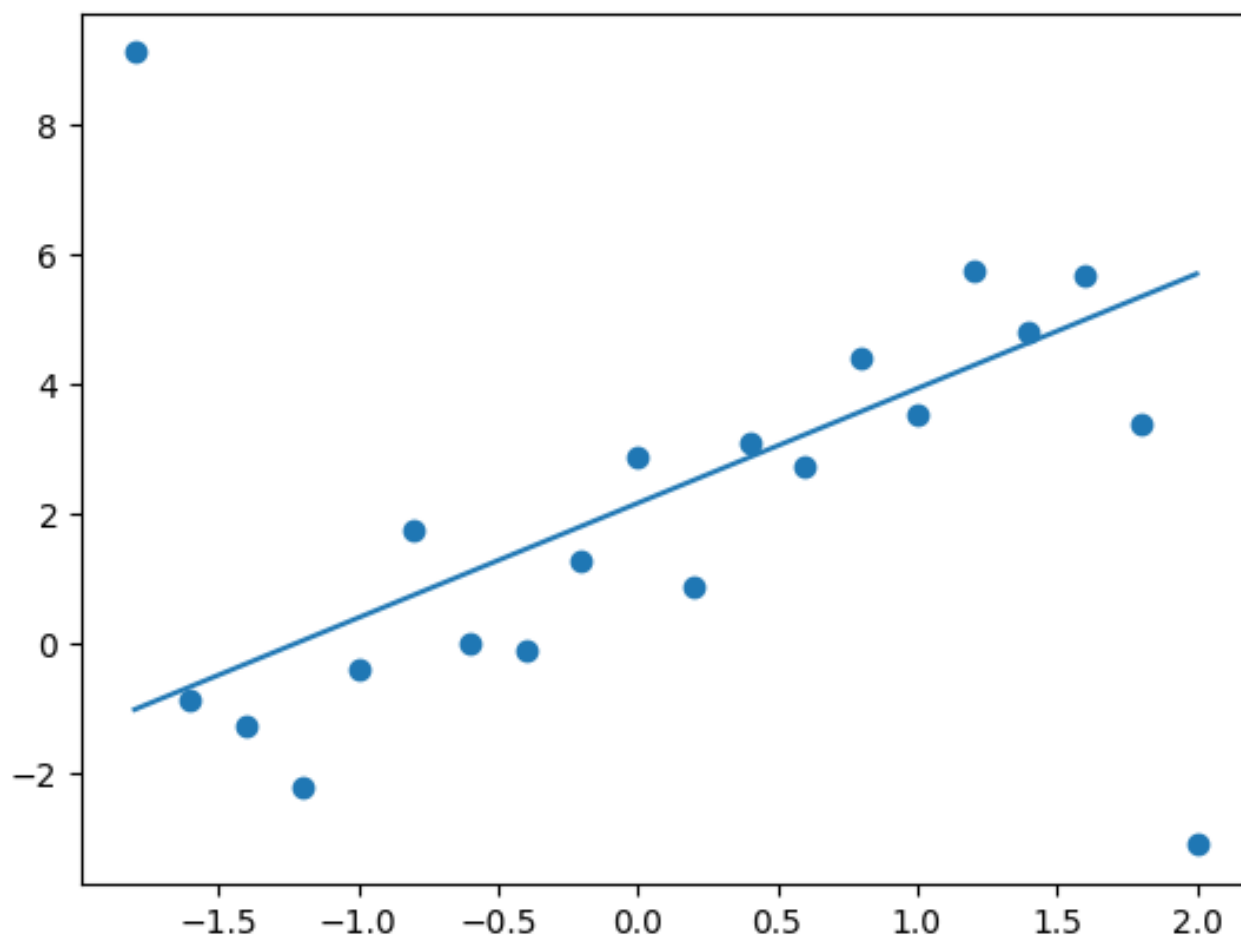


Рис. 15: Метод наименьших квадратов с возмущениями (1.668, 1.990)



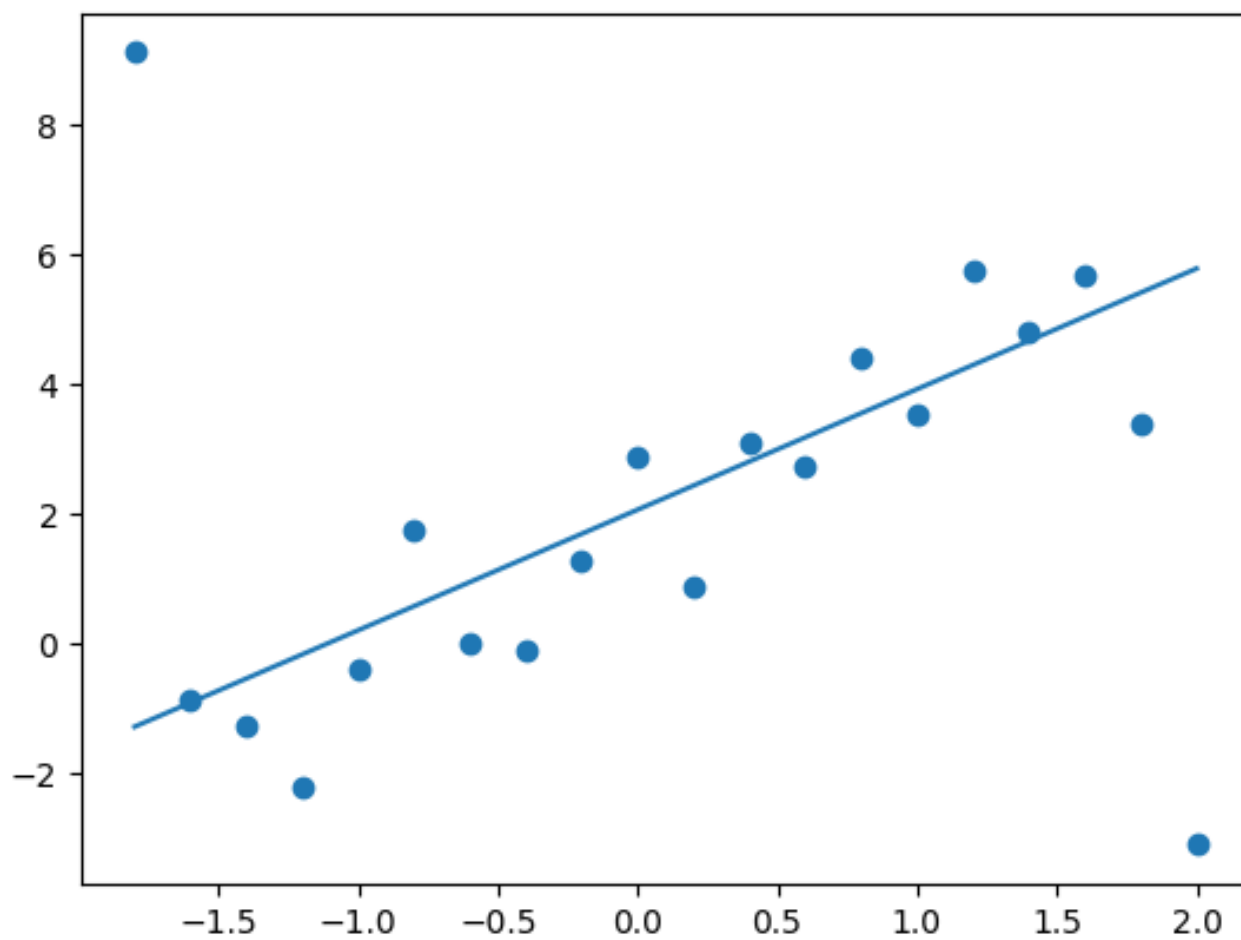


Рис. 16: Метод наименьших модулей с возмущениями (2.004, 0.632)

## 15 Выводы

На основе полученных характеристик (включая среднее значение, среднее значение квадрата и дисперсию) для различных коэффициентов корреляции и размеров выборки, можно сделать следующие наблюдения:

1. При увеличении размера выборки повышается точность оценок, что видно по уменьшению дисперсий коэффициентов корреляции. Это соответствует принципам центральной предельной теоремы и закона больших чисел.
2. При увеличении коэффициента корреляции  $\rho$ , средние значения коэффициентов Пирсона, Спирмена и квадратичного коэффициента корреляции тоже увеличиваются. Это указывает на прямую связь между  $\rho$  и другими коэффициентами корреляции.

Из результатов оценок коэффициентов линейной регрессии при использовании двух критериев (критерий наименьших квадратов и критерий наименьших модулей) можно сделать следующие выводы:

1. Метод наименьших квадратов показал себя эффективно в случае, когда нет значительных выбросов в данных, в то время как метод наименьших модулей проявил себя лучше в присутствии значительных возмущений.
2. Важно выбирать метод, исходя из особенностей данных. Если в данных присутствуют выбросы, метод наименьших модулей будет предпочтительнее из-за его устойчивости к выбросам.