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

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: $\overline{x}$, $med\, x$, 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

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

- Распределение Коши

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

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

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

- Распределение Пуассона

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

- Равномерное распределение

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

2.2 Характеристики положения и рассеяния

- Выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

- Выборочная медиана

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

- Полусумма экстремальных выборочных элементов

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

- Полусумма квартилей
  Выборочная квартиль z_p порядка p определяется формулой

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

  Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{10}$$

- Усечённое среднее

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

- Среднее характеристики

$$E(z) = \overline{z} \tag{12}$$

- Оценка дисперсии

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

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

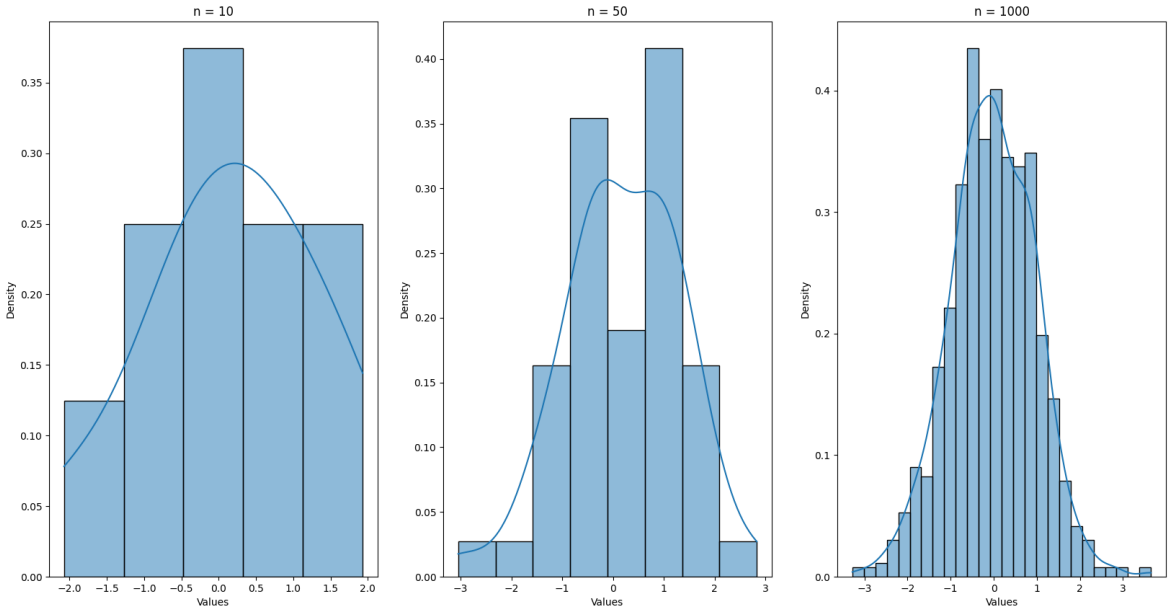

Рис. 1: Нормальное распределение (14)

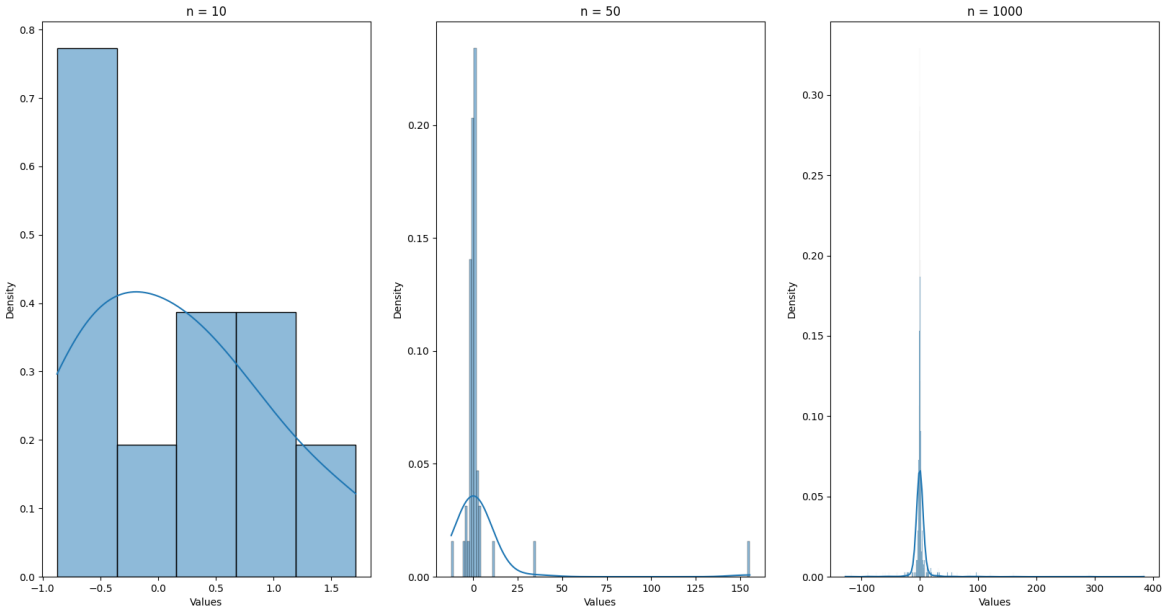

Рис. 2: Распределение Коши (15)

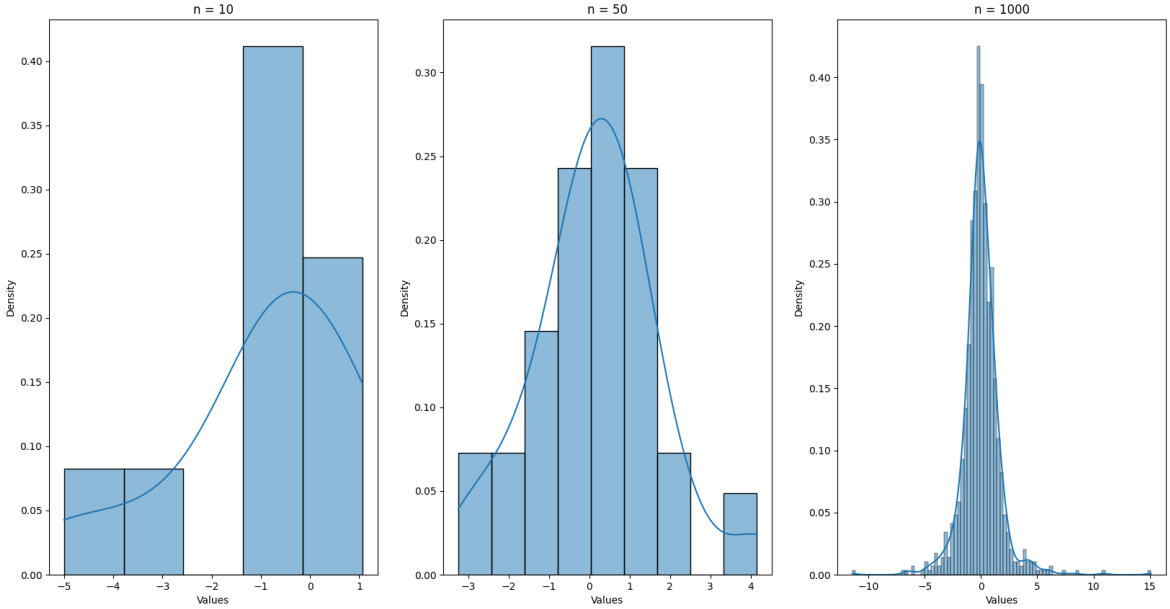

Рис. 3: Распределение Стьюдента (16)

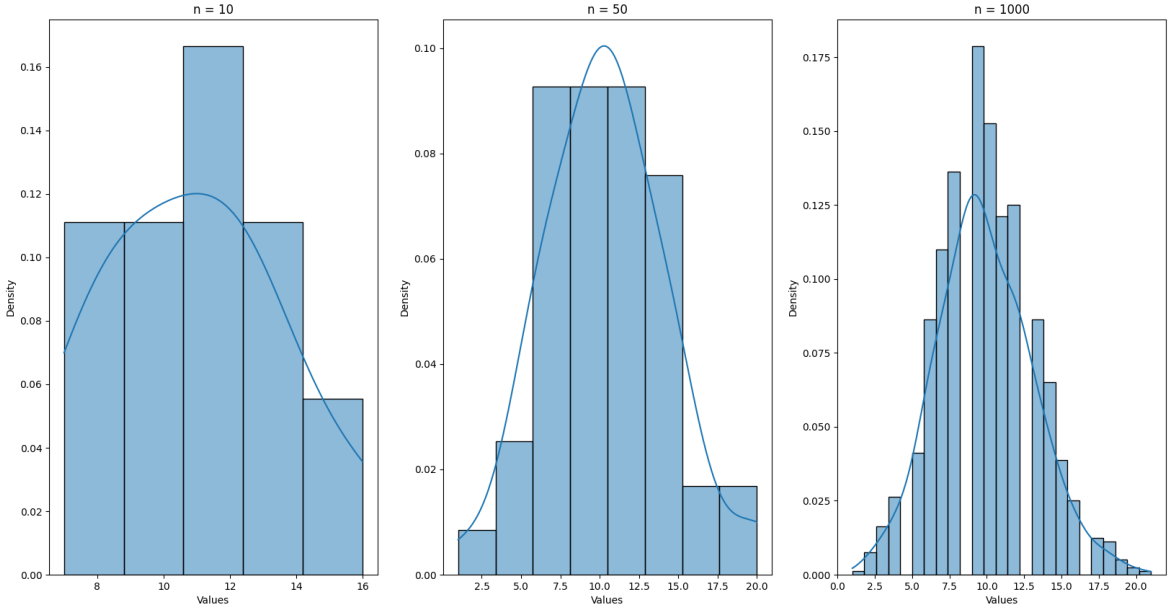

Рис. 4: Распределение Пуассона (17)

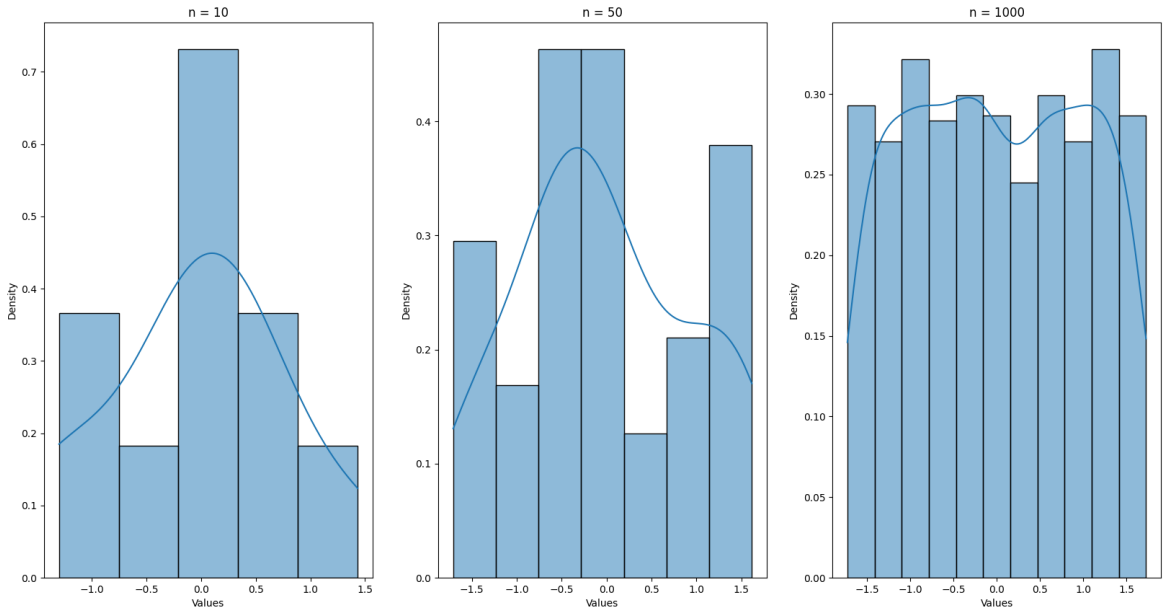

Рис. 5: Равномерное распределение (18)

4.2 Характеристики положения и рассеяния

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

Таблица 1: Нормальное распределение

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

Таблица 2: Распределение Коши

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

Таблица 3: Распределение Стьюдента

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

Таблица 4: Распределение Пуассона

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

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

- Распределение Коши

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

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

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

- Распределение Пуассона

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

- Равномерное распределение

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

7.2 Боксплот Тьюки

Боксплот (англ. box plot) — график, использующихся в описательной статистике, компактно изобрадающий 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 межквартальных расстояний и сумма третьего квартиля и полутора межквартальных расстояний. Формула имеет вид

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

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:

$$\left[\begin{array}{l} x < X_1^T \\ x > X_2^T \end{array} \right. \tag{20}$$

7.3 Доверительные интервалы для параметров нормального распределения

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

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

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 χ^2 с $n-1$ степенями свободы. Тогда

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

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

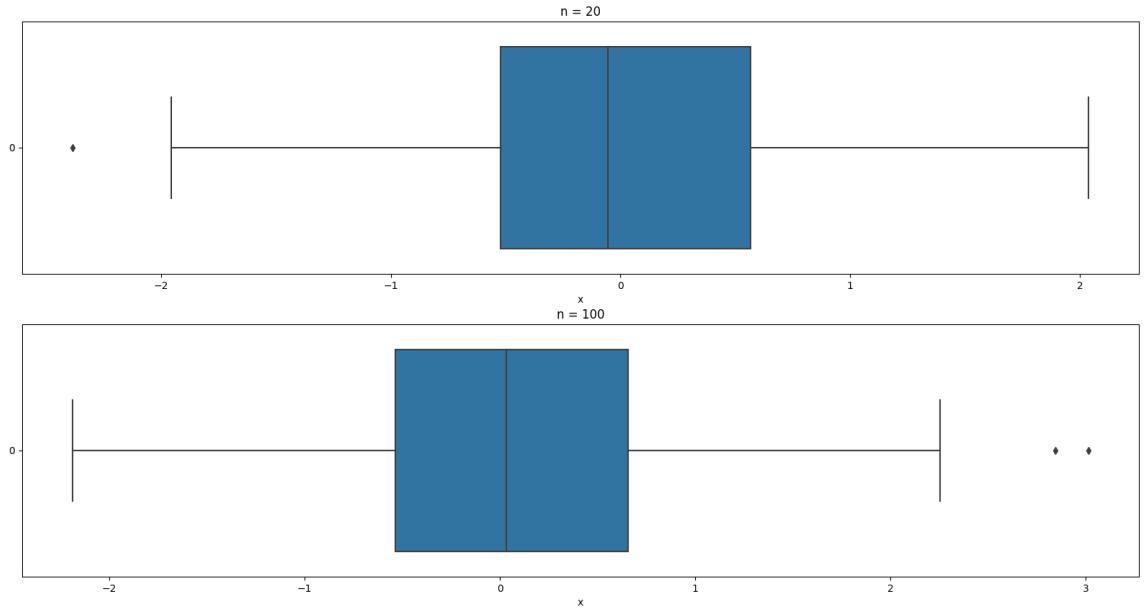

Рис. 6: Нормальное распределение (14)

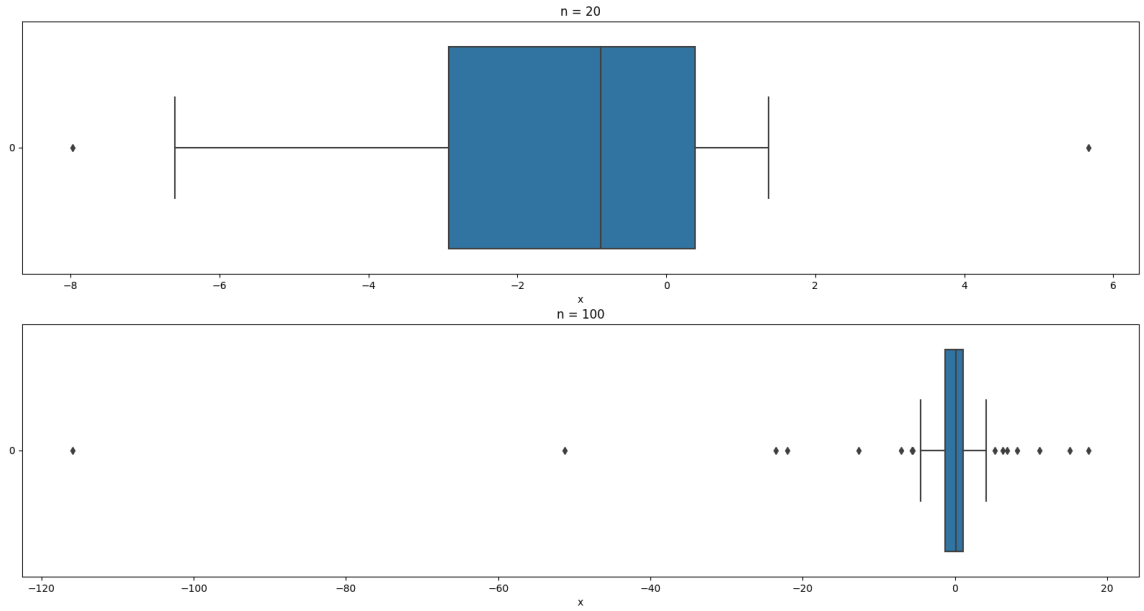

Рис. 7: Распределение Коши (15)

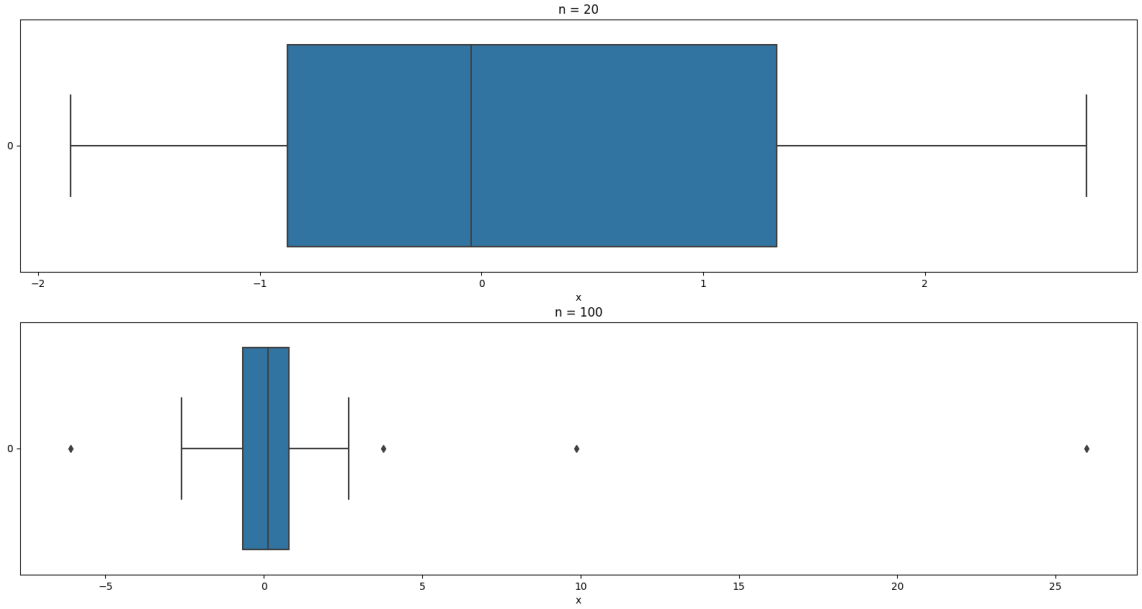

Рис. 8: Распределение Стьюдента (16)

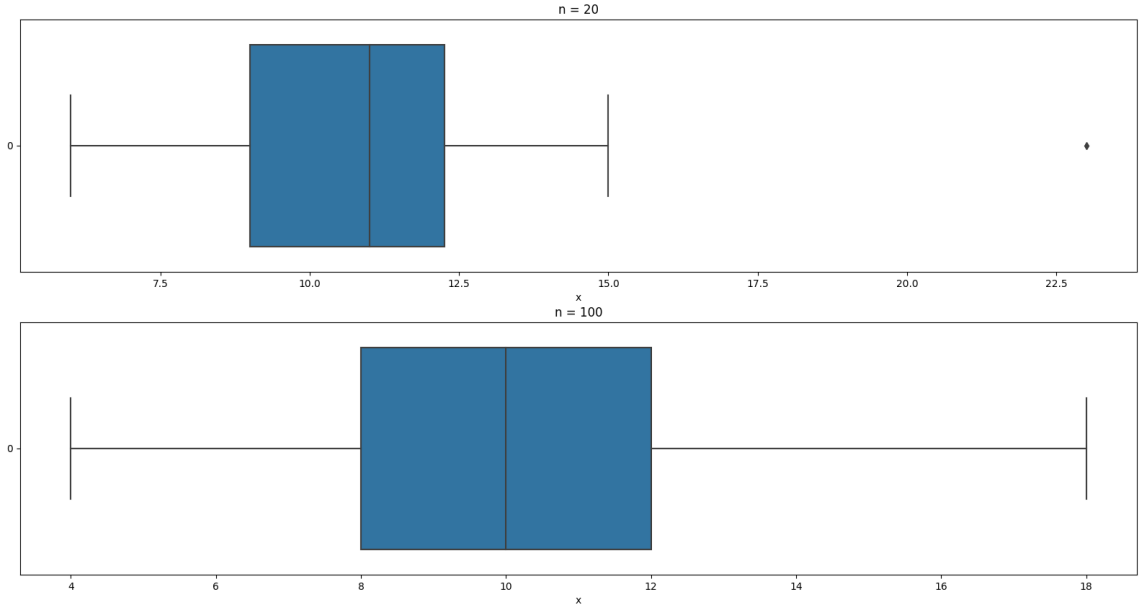

Рис. 9: Распределение Пуассона (17)

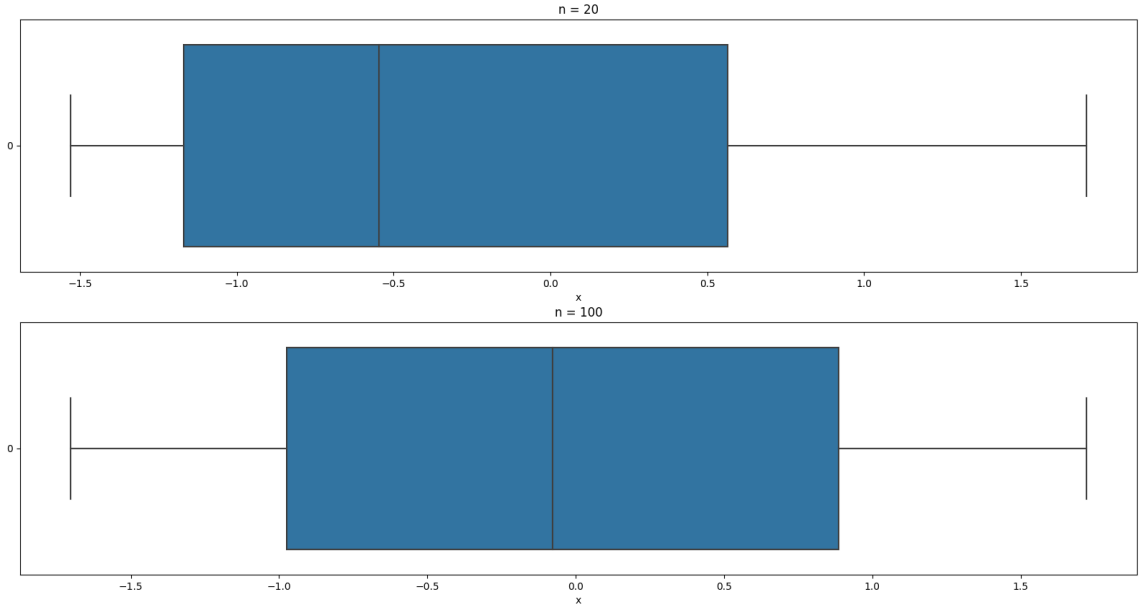

Рис. 10: Равномерное распределение (18)

9.2 Доверительные интервалы для параметров распределений

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

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

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

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

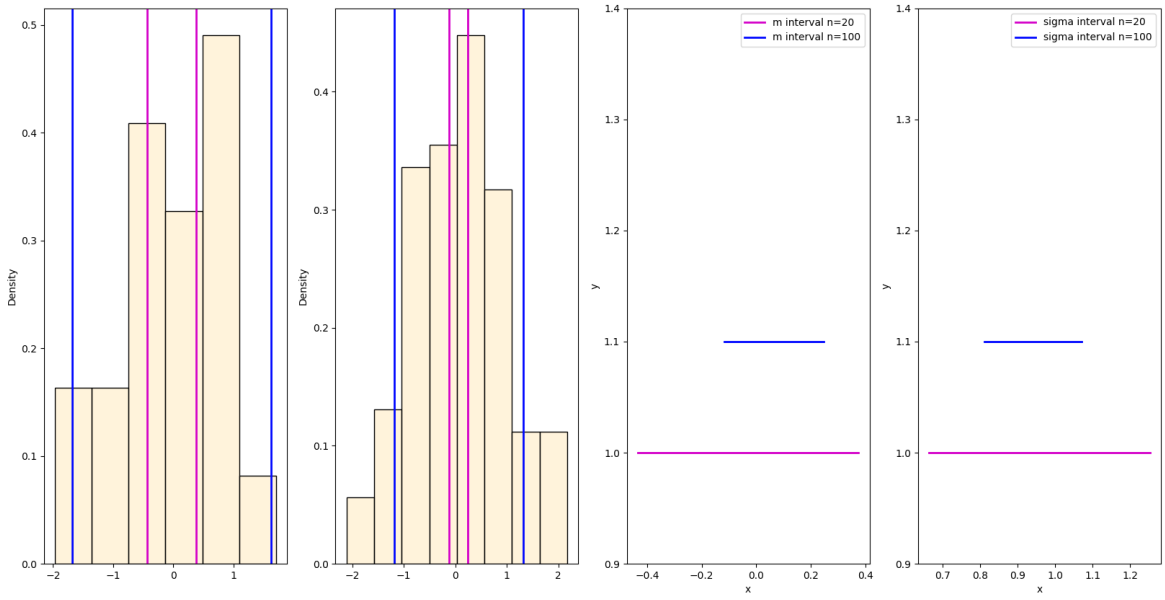

Рис. 11: Гистограммы и оценки для параметров нормального распределения

$$[[-0.434162, 0.374849], [0.663480, 1.252336]]$$

$$[[-0.117590, 0.248381], [0.810296, 1.070570]]$$

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, межквартальный 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 ρ взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреля- ции Пирсона, Спирмена и квадрантного 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. Проделать 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

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

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями x, y и средними квадратическими отклонениями σ_x, σ_y соответственно.

Параметр ρ называется коэффициентом корреляции.

12.2 Корреляционный момент (ковариация) и коэффициент корреляции

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

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

Коэффициент корреляции ρ двух случайных величин X и Y:

$$\rho = \frac{K}{\sigma_x \sigma_y} \tag{25}$$

12.3 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона:

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

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

12.4 Выборочный квадрантный коэффициент корреляции

Выборочный квадрантный коэффициент корреляции

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

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

Выборочный коэффициент ранговой корреляции Спирмена:

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

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

12.6 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость xOy:

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

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

$$\operatorname{tg} 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{30}$$

12.7 Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распрстранённых 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):

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

Задача минимизации квадратичного критерия (31) носит название задачи метода наименьших квадратов (МНК), а оценки β_0, β_1 параметров β_0, β_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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1}. \tag{32}$$

13 Описание работы

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

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

Таблица 8: Характеристики нормального двумерного распределения

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

Таблица 9: Характеристики смеси нормальных распределений

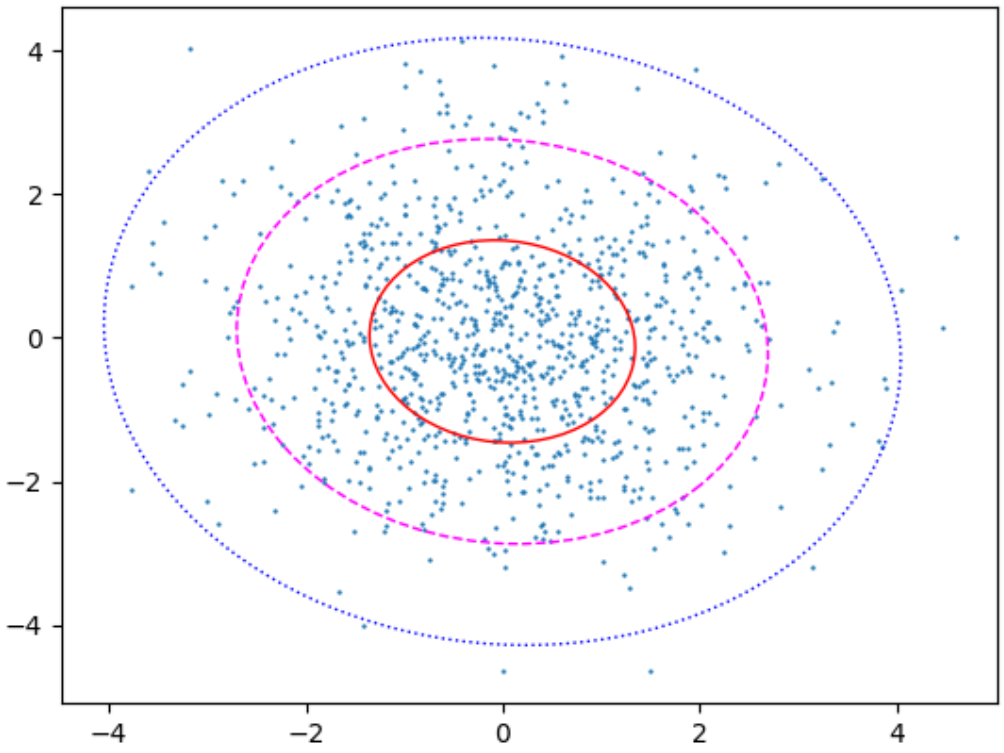

Рис. 12: Смесь нормальных распределений и эллипсы равновероятности ($\rho = 0$)

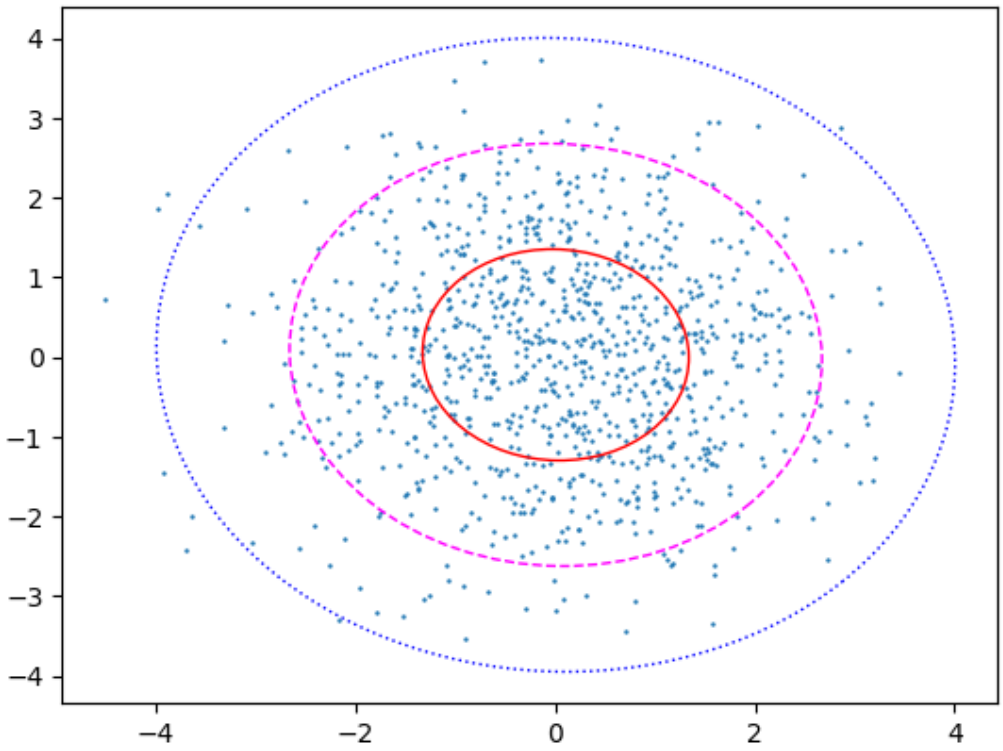

Рис. 13: Смесь нормальных распределений и эллипсы равновероятности ($\rho = 0.5$)

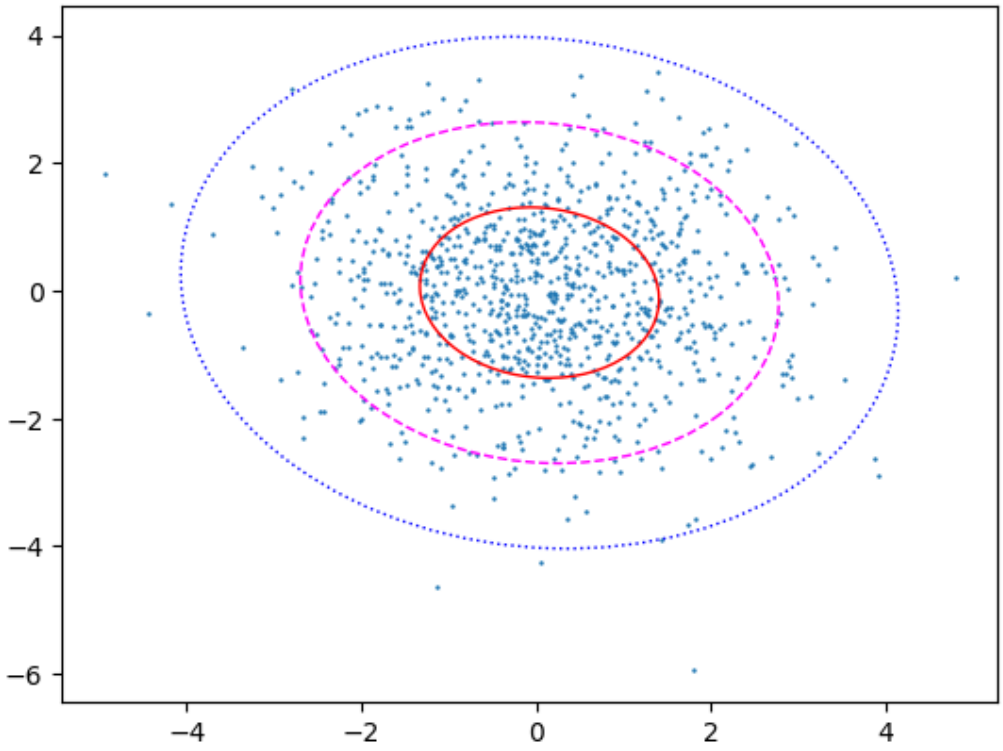

Рис. 14: Смесь нормальных распределений и эллипсы равновероятности ($\rho = 0.9$)

14.2 Простая линейная регрессия

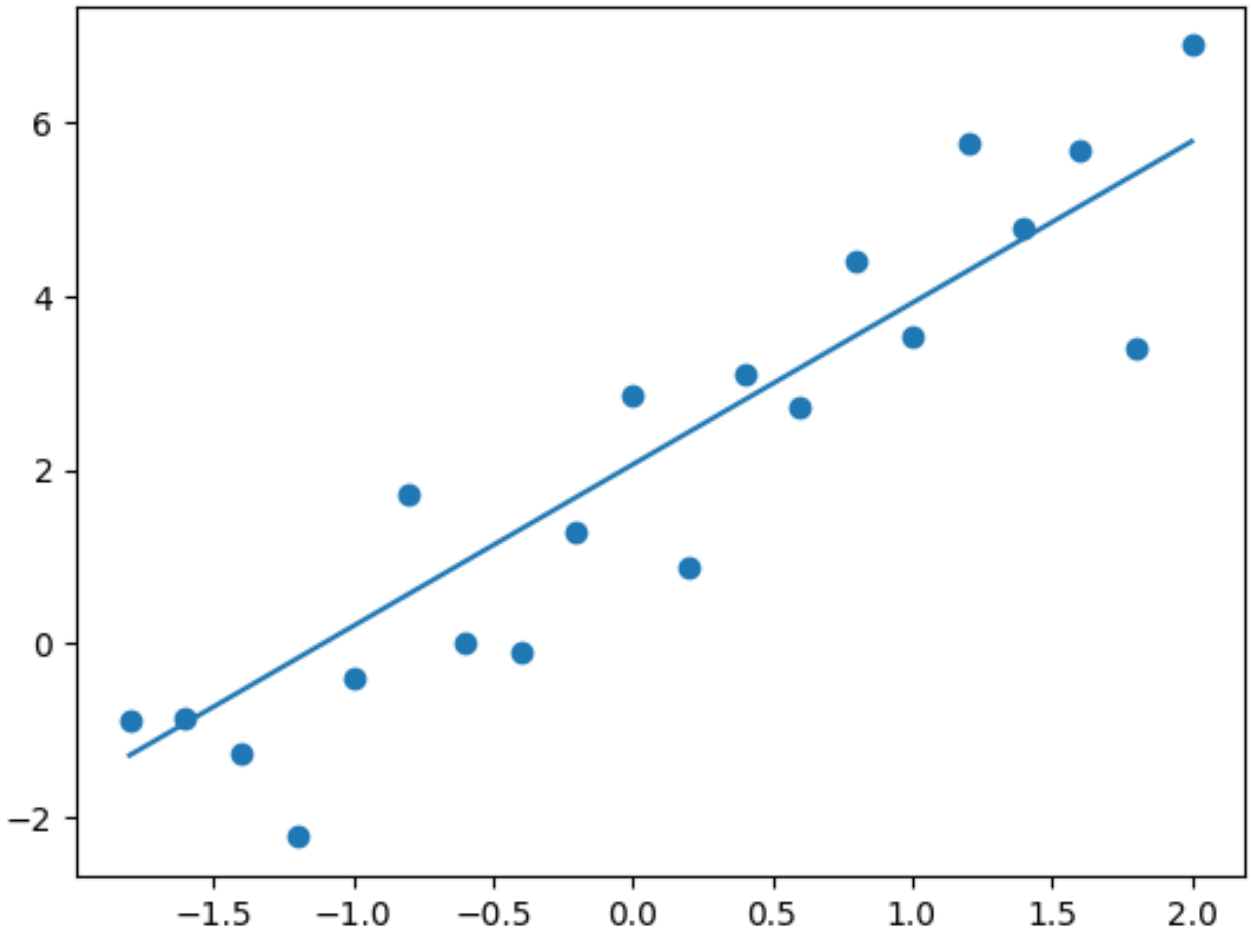

Рис. 15: Метод наименьших квадратов $(1.861, 2.061)$

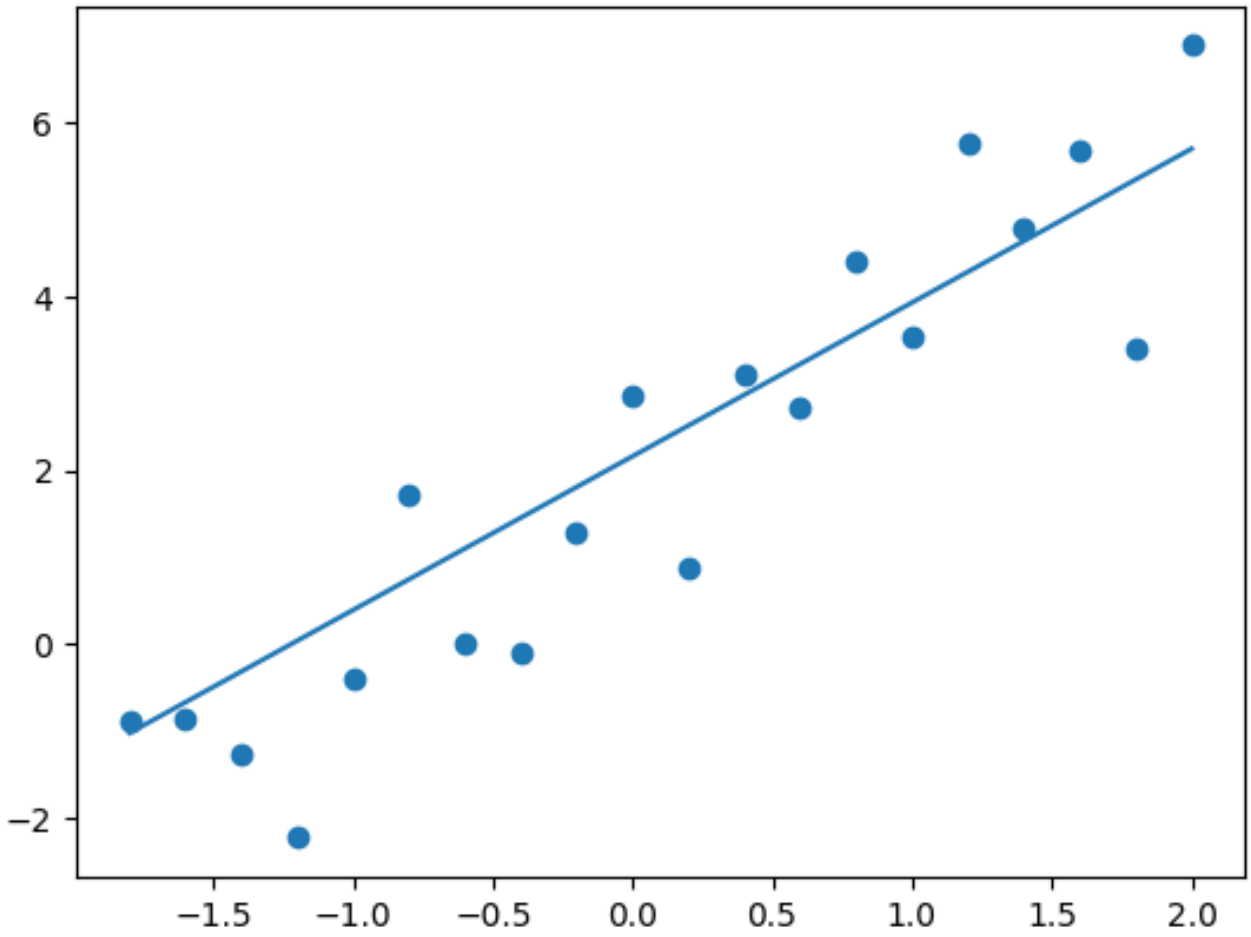

Рис. 16: Метод наименьших модулей $(1.7691, 2.162)$

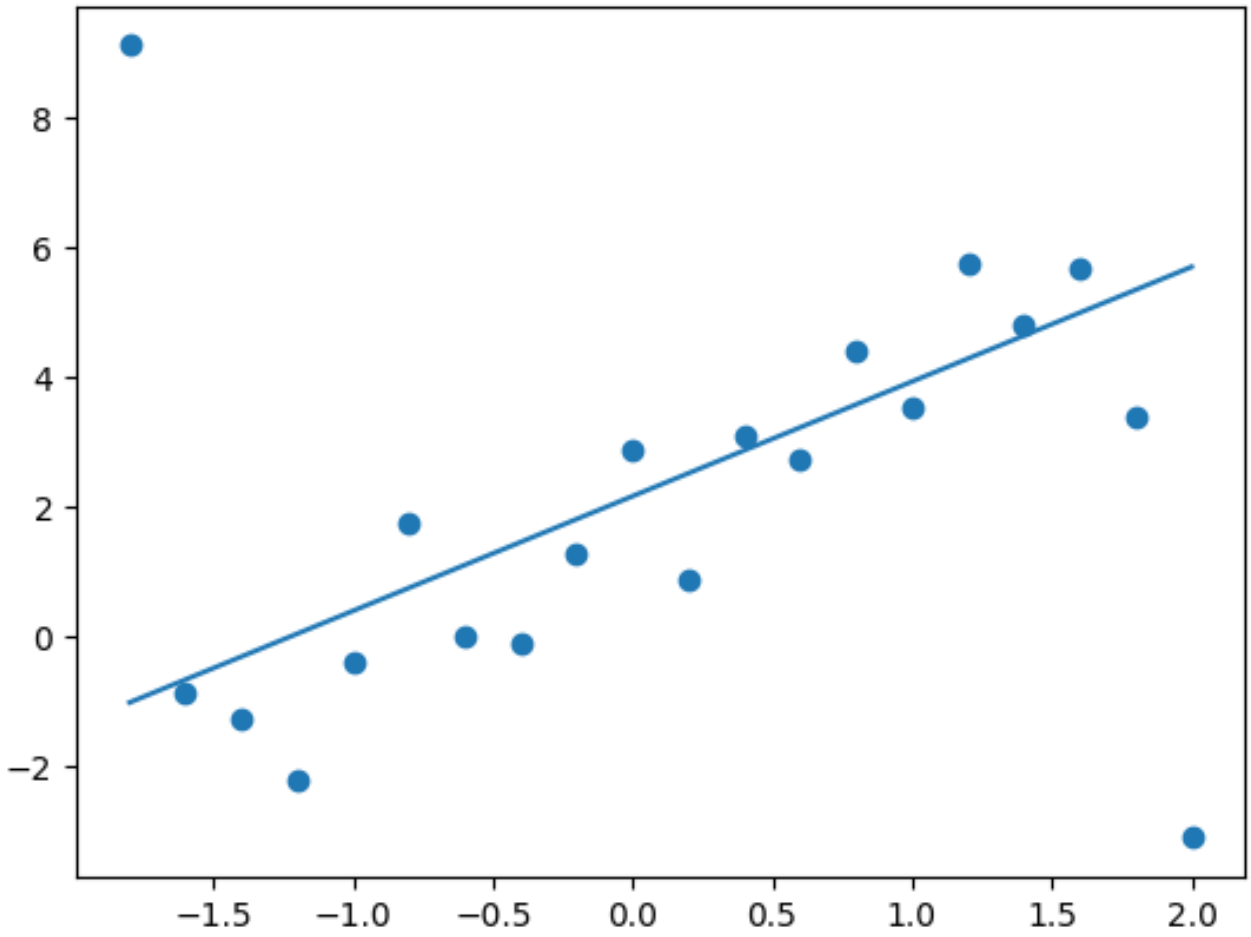

Рис. 17: Метод наименьших квадратов с возмущениями $(1.668, 1.990)$

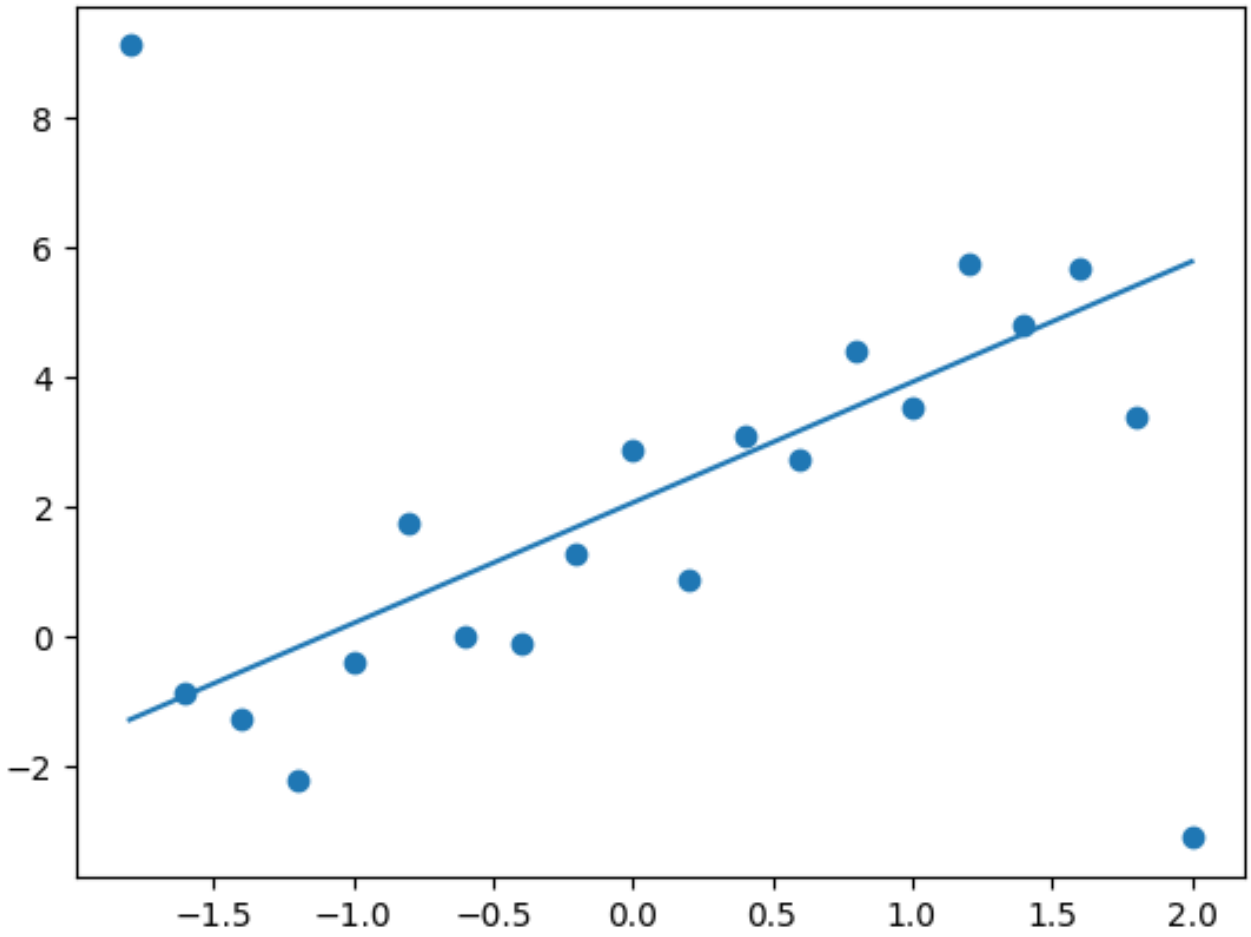

Рис. 18: Метод наименьших модулей с возмущениями $(2.004, 0.632)$

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 ρ, средние значения коэффициентов Пирсона, Спирмена и квадратичного коэффициента корреляции тоже увеличиваются. Это указывает на прямую связь между ρ 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.