

**Санкт-Петербургский Политехнический университет имени Петра
Великого**

Кафедра прикладной математики и информатики

ОТЧЁТ

по лабораторной работе №1

ГИСТОГРАММЫ И ПЛОТНОСТИ РАСПРЕДЕЛЕНИЙ

Выполнил:
студент группы 5030102/30003
Крутянский Роман Игоревич

Преподаватель: Баженов Александр
Николаевич

Санкт-Петербург
2025

Содержание

1.	Введение	3
2.	Лабораторная работа №1: Гистограммы и плотности распределений	4
2.1.	Постановка задачи	4
2.2.	Теоретическая часть	4
2.3.	Реализация	4
2.4.	Результаты	4
2.5.	Обсуждение	5
2.5.1.	Вопросы преподавателя:	6
2.6.	Выводы	6
3.	Лабораторная работа №2: Характеристики положения и рассеяния	7
3.1.	Постановка задачи	7
3.2.	Теоретическая часть	7
3.3.	Реализация	7
3.4.	Результаты	8
3.5.	Обсуждение	8
3.5.1.	Вопросы преподавателя	8
3.6.	Выводы	9
4.	Лабораторная работа №3: Боксплот Тьюки и анализ выбросов	10
4.1.	Постановка задачи	10
4.2.	Теоретическая часть	10
4.3.	Реализация	10
4.4.	Результаты	11
4.5.	Обсуждение	12
4.6.	Выводы	12
5.	Лабораторная работа №4: Эмпирическая функция распределения и ядерные оценки плотности	13
5.1.	Постановка задачи	13
5.2.	Теоретическая часть	13
5.3.	Реализация	13
5.4.	Результаты	14
5.5.	Обсуждение	16
5.6.	Выводы	17
6.	Общие выводы по блоку лабораторных работ	18
7.	Список литературы	19
8.	Приложение. Ссылка на репозиторий GitHub	20

1. Введение

Целью первого блока лабораторных работ является знакомство с основными методами описательной статистики:

- визуализацией распределений,
- вычислением характеристик положения и рассеяния,
- анализом выбросов,
- построением эмпирических оценок функций распределения и плотности.

Исследуются следующие распределения:

- 1) Нормальное $N(x; 0, 1)$
- 2) Коши $C(x; 0, 1)$
- 3) Лапласа $L\left(x; 0, \frac{1}{\sqrt{2}}\right)$
- 4) Пуассона $P(k; 5)$
- 5) Равномерное $U\left(x; -\sqrt{3}, \sqrt{3}\right)$

Их уравнения:

- 1) Нормальное: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- 2) Коши: $f(x) = \frac{1}{\pi(1+x^2)}$
- 3) Лапласа: $f(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|}$
- 4) Пуассона: $P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- 5) Равномерное: $f(x) = \begin{cases} \frac{1}{2\sqrt{3}} & x \in [-\sqrt{3}, \sqrt{3}] \\ 0 & x \notin [-\sqrt{3}, \sqrt{3}] \end{cases}$

2. Лабораторная работа №1: Гистограммы и плотности распределений

2.1. Постановка задачи

Цель: Освоить принцип группировки данных и построения гистограмм, исследовать влияние размера выборки на определение характера распределения.

Задача: Сгенерировать выборки объёмом $n = 10, 100, 1000$ для каждого из 5 распределений и построить на одном графике гистограмму выборки и теоретическую кривую плотности распределения.

2.2. Теоретическая часть

Гистограмма — это ступенчатая функция, аппроксимирующая плотность распределения вероятностей. Пусть имеется выборка x_1, x_2, \dots, x_n . Интервал значений разбивается на k бинов $[a_0, a_1), [a_1, a_2), \dots, [a_{k-1}, a_k]$.

Оценка плотности в j -м интервале:

$$\hat{f}(x) = \frac{n_j}{nh}, \quad x \in [a_{j-1}, a_j)]$$

где n_j — количество наблюдений в j -м интервале, $h = a_j - a_{j-1}$ — ширина интервала.

Выбор количества бинов:

- Слишком мало интервалов → потеря информации
- Слишком много интервалов → дырки в гистограмме, непонятки

2.3. Реализация

Язык программирования: Python 3.10

Библиотеки: NumPy, SciPy, Matplotlib

Основные методы:

- `scipy.stats` — генерация выборок из заданных распределений
- `matplotlib.pyplot.hist` с параметром `density=True` — построение нормированных гистограмм
- Аналитические функции плотности (`pdf`) — построение теоретических кривых

2.4. Результаты

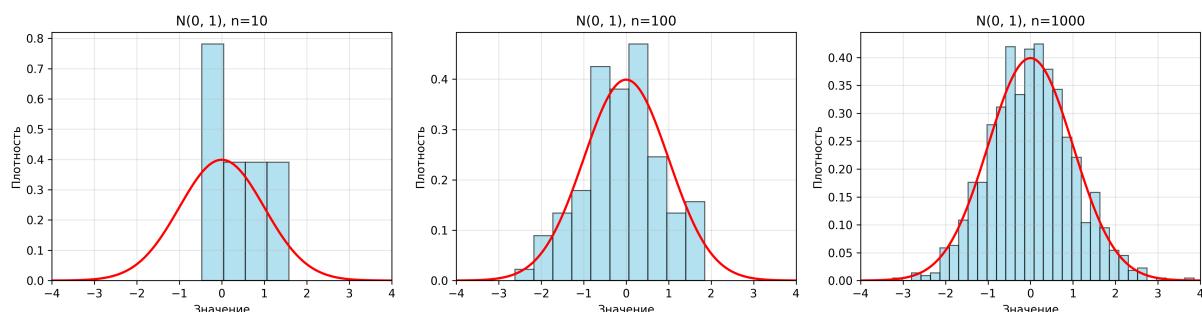


Рис. 1. Нормальное распределение: гистограммы для $n = 10, 100, 1000$

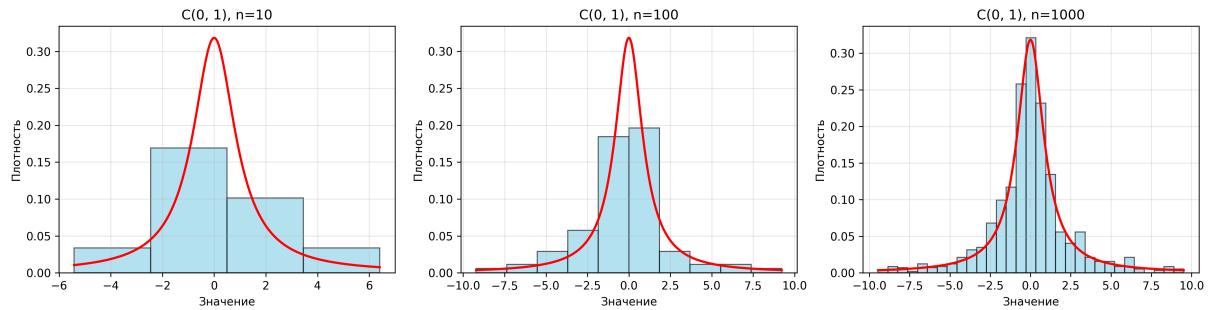


Рис. 2. Распределение Коши: гистограммы для $n = 10, 100, 1000$

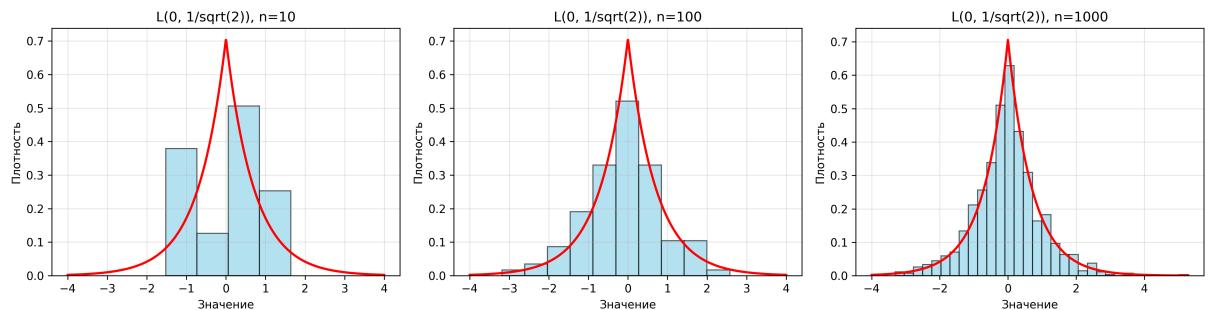


Рис. 3. Распределение Лапласа: гистограммы для $n = 10, 100, 1000$

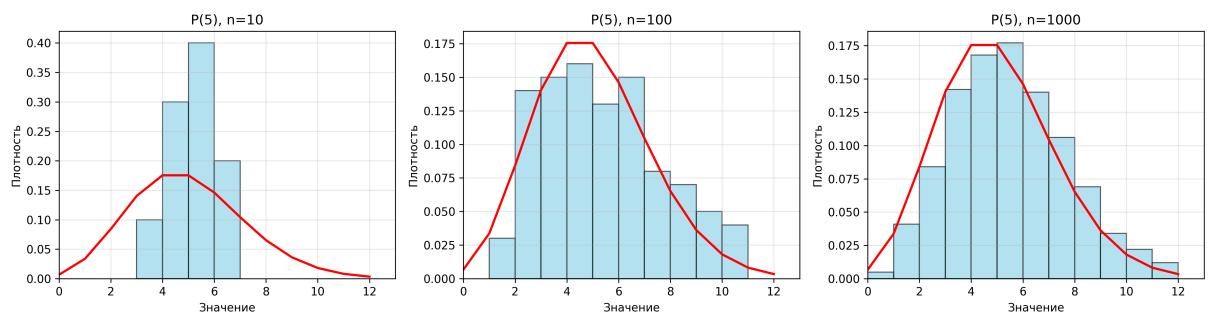


Рис. 4. Распределение Пуассона ($\lambda = 5$): гистограммы для $n = 10, 100, 1000$

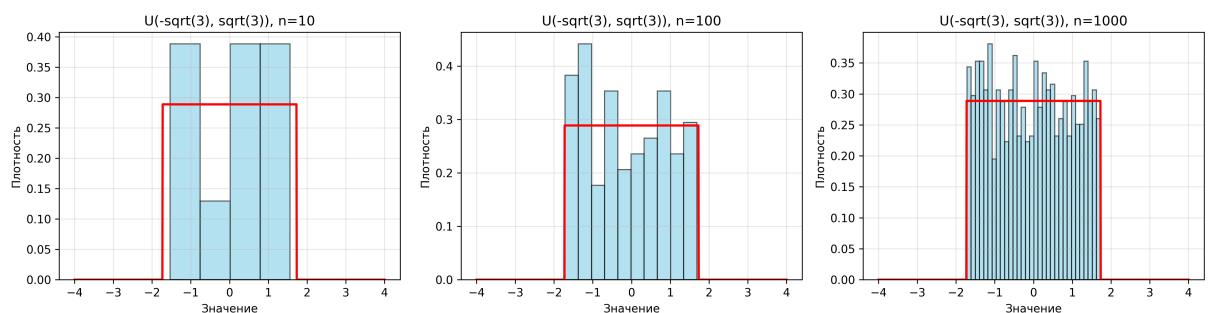


Рис. 5. Равномерное распределение: гистограммы для $n = 10, 100, 1000$

2.5. Обсуждение

Нормальное распределение: При $n = 10$ гистограмма плохо аппроксимирует теоретическую кривую из-за малого объёма выборки. При $n = 100$ форма становится узнаваемой, при $n = 1000$ наблюдается отличное совпадение.

Распределение Коши: Характеризуется тяжёлыми хвостами. При больших n появляются экстремальные выбросы, что требует ограничения диапазона визуализации.

Распределение Лапласа: Более острый пик и тяжёлые хвосты по сравнению с нормальным распределением.

Распределение Пуассона ($\lambda = 5$): Дискретное распределение, поэтому используются бины с целочисленными границами. При $\lambda = 5$ распределение асимметрично, но с ростом n приближается к нормальному.

Равномерное распределение: Гистограмма демонстрирует плоскую форму, которая становится более выраженной с ростом n .

2.5.1. Вопросы преподавателя:

- 1) Как вычислялось количество бинов в matplotlib? - раньше был `bins = auto` но из-за недетерминированности я поменял на `bins = sqrt(n)` теперь количество бинов $k = \sqrt{n}$, где n - количество значений в выборке.

2.6. Выводы

- 1) Размер выборки существенно влияет на качество аппроксимации: при $n \geq 1000$ гистограммы хорошо согласуются с теоретическими распределениями.
- 2) Для распределений с тяжёлыми хвостами (Коши) требуется особая обработка выбросов.

3. Лабораторная работа №2: Характеристики положения и рассеяния

3.1. Постановка задачи

Цель: Исследовать сходимость выборочных характеристик к теоретическим при росте n , оценить устойчивость различных оценок к выбросам.

Задача: Для выборок объёмом $n = 10, 100, 1000$ вычислить 5 характеристик положения:

- 1) Выборочное среднее \bar{x}
- 2) Медиану
- 3) Полусумму экстремальных элементов $z_R = \frac{x_{\min} + x_{\max}}{2}$
- 4) Полусумму квартилей $z_Q = \frac{Q_1 + Q_3}{2}$
- 5) Усечённое среднее z_{tr} (с отбрасыванием 10% наименьших и наибольших значений)

Повторить генерацию 1000 раз и найти:

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

Формат вывода: $x = E \pm \sqrt{D}$ с округлением до первого знака после запятой.

3.2. Теоретическая часть

Выборочное среднее:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Медиана: центральное значение в упорядоченной выборке.

Полусумма экстремумов (midrange): $z_R = \frac{x_{(1)} + x_{(n)}}{2}$

Полусумма квартилей (midhinge): $z_Q = \frac{Q_1 + Q_3}{2}$

Усечённое среднее: $z_{\text{tr}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}, \quad k = \lfloor 0.1n \rfloor$

3.3. Реализация

Язык программирования: Python 3.10

Библиотеки: NumPy, SciPy

Основные методы:

- `scipy.stats.rvs(size=n)` — генерация выборок
- `numpy.mean()`, `numpy.median()`, `numpy.percentile()` — вычисление характеристик
- `scipy.stats.trim_mean` — усечённое среднее
- Метод Монте-Карло (1000 повторений) для оценки $E(z)$ и $D(z)$

3.4. Результаты

n	Среднее	Медиана	Midrange	Midhinge	Trimmed
10	-0.0 ± 0.3	-0.0 ± 0.4	-0.0 ± 0.4	-0.0 ± 0.3	-0.0 ± 0.3
100	-0.0 ± 0.1	-0.0 ± 0.1	0.0 ± 0.3	-0.0 ± 0.1	-0.0 ± 0.1
1000	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.2	-0.0 ± 0.0	-0.0 ± 0.0

Таблица 1. Характеристики положения для нормального распределения

n	Среднее	Медиана	Midrange	Midhinge	Trimmed
10	1.2 ± 8.0	-0.0 ± 0.6	6.1 ± 17.9	-0.0 ± 0.9	-0.1 ± 1.4
100	0.3 ± 5.5	0.0 ± 0.2	14.0 ± 38.6	0.0 ± 0.2	0.0 ± 0.2
1000	-2.5 ± 10.5	0.0 ± 0.0	-1291.9 ± 234.2	0.0 ± 0.1	0.0 ± 0.1

Таблица 2. Характеристики положения для распределения Коши

n	Среднее	Медиана	Midrange	Midhinge	Trimmed
10	0.0 ± 0.3	0.0 ± 0.3	0.0 ± 0.6	0.0 ± 0.3	0.0 ± 0.3
100	0.0 ± 0.1	0.0 ± 0.1	-0.0 ± 0.6	0.0 ± 0.1	0.0 ± 0.1
1000	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.7	-0.0 ± 0.0	-0.0 ± 0.0

Таблица 3. Характеристики положения для распределения Лапласа

n	Среднее	Медиана	Midrange	Midhinge	Trimmed
10	5.0 ± 0.7	5.0 ± 1.0	5.3 ± 1.0	4.9 ± 0.8	4.9 ± 0.7
100	5.0 ± 0.2	5.0 ± 0.5	6.0 ± 0.7	4.9 ± 0.4	4.9 ± 0.2
1000	5.0 ± 0.1	5.0 ± 0.0	6.8 ± 0.6	4.7 ± 0.3	4.9 ± 0.1

Таблица 4. Характеристики положения для распределения Пуассона ($\lambda = 5$)

n	Среднее	Медиана	Midrange	Midhinge	Trimmed
10	-0.0 ± 0.3	-0.0 ± 0.5	-0.0 ± 0.2	-0.0 ± 0.4	-0.0 ± 0.4
100	-0.0 ± 0.1	-0.0 ± 0.2	-0.0 ± 0.0	-0.0 ± 0.1	-0.0 ± 0.1
1000	0.0 ± 0.0	0.0 ± 0.1	-0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Таблица 5. Характеристики положения для равномерного распределения

3.5. Обсуждение

- Полусумма экстремальных элементов (midrange) эффективна для распределений с ограниченным носителем (равномерное) и крайне неэффективна для распределений с тяжелыми хвостами (Коши, Лапласа)
- С увеличением объема выборки медиана, полусумма quartилей и усеченное среднее всегда сходятся к теоретическим значениям

3.5.1. Вопросы преподавателя

- Чему равно среднее и дисперсия у распределения Коши? Их нет, т.к. хвости слишком тяжелые и вообще интеграл не сходится

- Какие выводы можно сделать из картинок? Картинки были не говорящие и не имели надобности, потому я их убрал

3.6. Выводы

- 1) Среднее арифметическое не является рабастной оценкой. Для определения матожидания лучше использовать усеченное среднее или медиану
- 2) Полусумма экстремальных элементов эффективна для распределений с ограниченным носителем (равномерное) и крайне неэффективна для распределений с тяжелыми хвостами (Коши, Лапласа)
- 3) С увеличением объёма выборки медиана, полусумма квартилей и усеченное среднее всегда сходятся к теоретическим значениям

4. Лабораторная работа №3: Боксплот Тьюки и анализ выбросов

4.1. Постановка задачи

Цель: Научиться применять боксплот Тьюки для анализа одномерных распределений, исследовать влияние размера выборки на долю отсеиваемых аномальных значений.

Задача:

- 1) Сгенерировать выборки объёмом $n = 20, 100$
- 2) Построить боксплот Тьюки
- 3) Экспериментально определить долю выбросов (1000 повторений)

4.2. Теоретическая часть

Межквартильный размах (IQR): ($IQR = Q_3 - Q_1$) **Границы выбросов (при $k = 1.5$):**

$$\text{Lower} = Q_1 - 1.5 \cdot IQR$$

$$\text{Upper} = Q_3 + 1.5 \cdot IQR$$

Доля выбросов:

$$p_{\text{out}} = \frac{n_{\text{out}}}{n}$$

4.3. Реализация

Язык программирования: Python 3.10

Библиотеки: NumPy, SciPy, Matplotlib

Алгоритм:

- 1) Генерация 1000 выборок заданного объёма n
- 2) Вычисление Q_1, Q_3 и

$$IQR$$

для каждой выборки

- 3) Определение границ выбросов: $[Q_1 - 1.5 \cdot IQR; Q_3 + 1.5 \cdot IQR]$
- 4) Подсчёт доли выбросов $p_{\text{out}} = \frac{n_{\text{out}}}{n}$
- 5) Усреднение по 1000 повторениям

4.4. Результаты

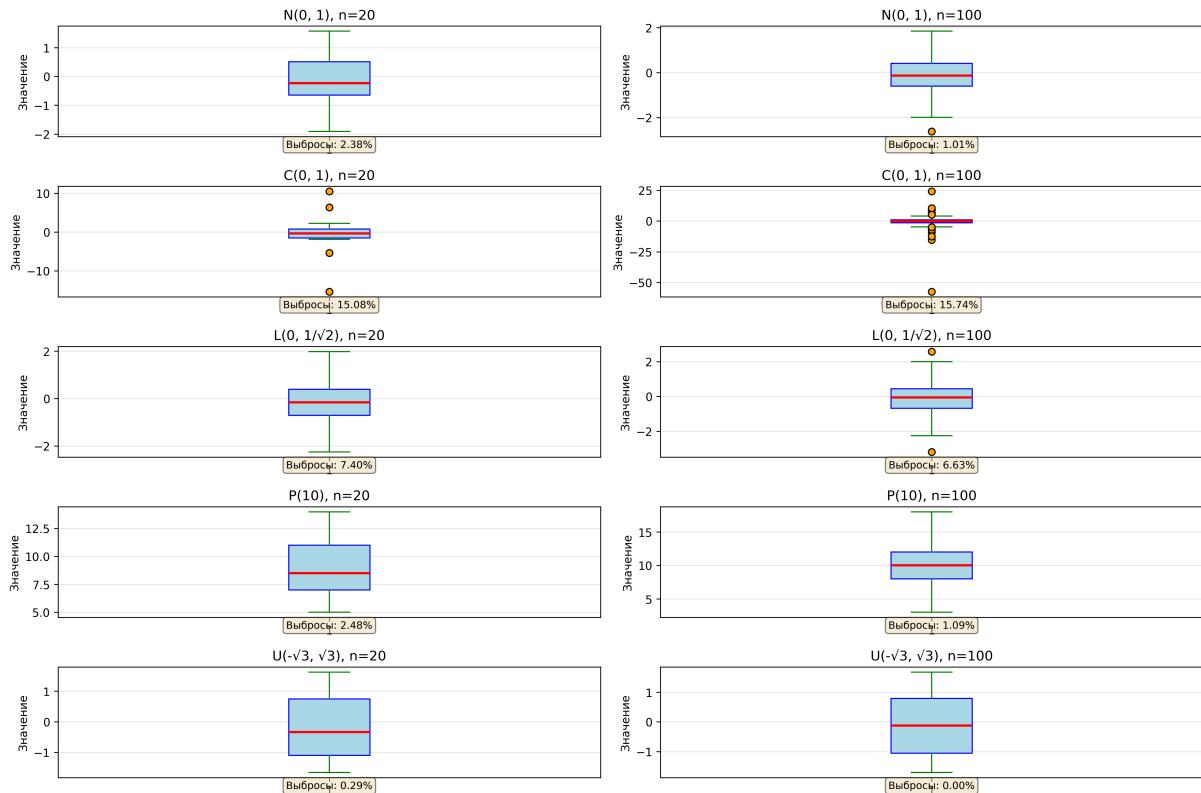


Рис. 6. Боксплоты Тьюки для всех распределений

Распределение	n	Средняя доля	Std
Нормальное	20	0.024	0.045
	100	0.010	0.014
Коши	20	0.155	0.072
	100	0.157	0.035
Лаплас	20	0.073	0.066
	100	0.066	0.029
Пуассон ($\lambda = 5$)	20	0.027	0.049
	100	0.015	0.016
Равномерное	20	0.002	0.015
	100	0.000	0.000

Таблица 6. Средняя доля выбросов по методу Тьюки (1000 повторений)

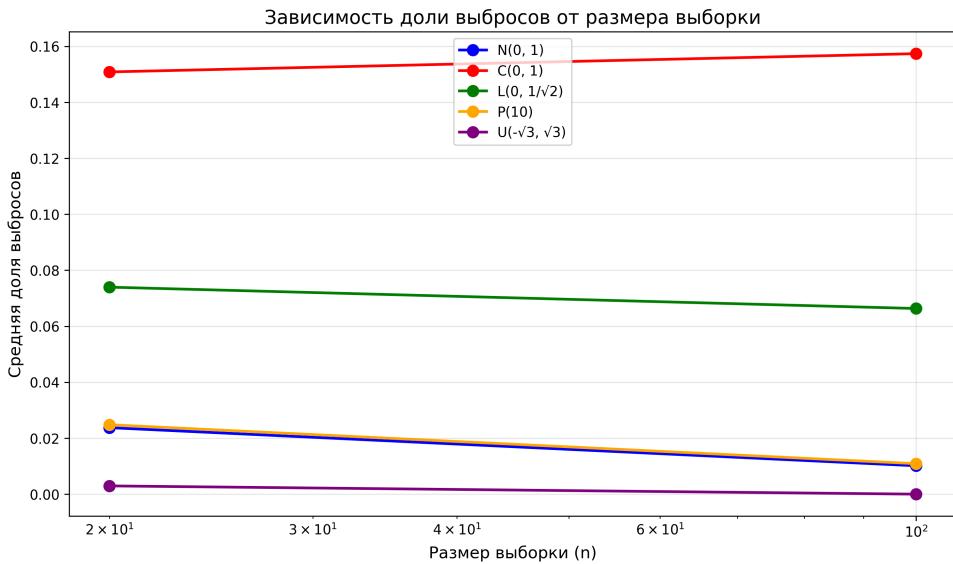


Рис. 7. Средняя доля выбросов в зависимости от размера выборки

4.5. Обсуждение

Нормальное распределение: Экспериментальная доля выбросов (2.4% при $n = 20$, 1.0% при $n = 100$) близка к теоретическому значению $\approx 0.7\%$. Метод Тьюки работает корректно для распределений с лёгкими хвостами.

Распределение Коши: Доля выбросов стабильно высокая ($\approx 15.6\%$) — это особенность распределения с чрезвычайно тяжёлыми хвостами: точки, маркируемые как аномалии, являются типичными для закона Коши.

Распределение Лапласа: Доля выбросов ($6.6\text{--}7.3\%$) занимает промежуточное положение, что согласуется с теорией: хвосты Лапласа тяжелее нормальных, но легче хвостов Коши.

Распределение Пуассона ($\lambda = 5$): Доля выбросов ($1.5\text{--}2.7\%$) превышает теоретическое значение 0.7% из-за правосторонней асимметрии: метод Тьюки маркирует как выбросы значения из длинного правого хвоста.

Равномерное распределение: Теоретическая доля выбросов равна нулю. Эксперимент подтверждает: 0.000% при $n = 100$.

4.6. Выводы

- 1) Метод Тьюки эффективен для обнаружения аномалий в распределениях с лёгкими хвостами.
- 2) Для распределений с тяжёлыми хвостами (Коши) метод даёт много «ложных» выбросов.
- 3) При интерпретации результатов необходимо учитывать природу распределения.
- 4) С ростом объёма выборки оценка доли выбросов становится более устойчивой.

5. Лабораторная работа №4: Эмпирическая функция распределения и ядерные оценки плотности

5.1. Постановка задачи

Цель: Исследовать сходимость эмпирической функции распределения к теоретической, сравнить ядерные оценки плотности с гистограммами.

Задача: Для выборок объёмом $n = 20, 60, 100$ построить:

- 1) Эмпирическую функцию распределения (ЭФР) и теоретическую функцию распределения
- 2) Ядерную оценку плотности (гауссово ядро) и теоретическую плотность

Диапазоны: $[-4, 4]$ для непрерывных распределений, $[6, 14]$ для Пуассона.

5.2. Теоретическая часть

Эмпирическая функция распределения:

$$F_{n(x)} = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

где $I(\cdot)$ — индикаторная функция.

Теорема Гливенко–Кантелли:

$$\sup_x |F_{n(x)} - F(x)| \xrightarrow{n \rightarrow \infty} 0$$

Ядерная оценка плотности (KDE):

$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

Для гауссова ядра:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

5.3. Реализация

Язык программирования: Python 3.10

Библиотеки: NumPy, SciPy, Matplotlib, scikit-learn

Основные методы:

- `numpy.searchsorted` — эффективное вычисление ЭФР
- `sklearn.neighbors.KernelDensity` — ядерная оценка плотности с гауссовым ядром
- Автоматический выбор ширины окна по правилу Сильвермана: $h = 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}$

5.4. Результаты

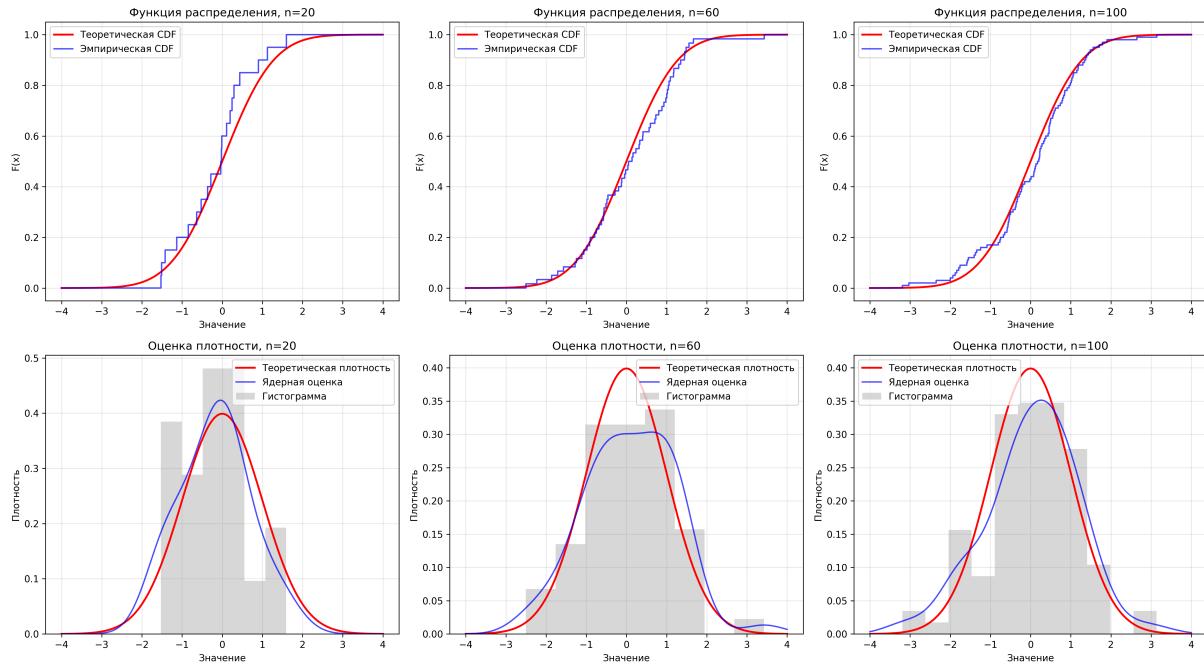


Рис. 8. Нормальное распределение: ЭФР и KDE

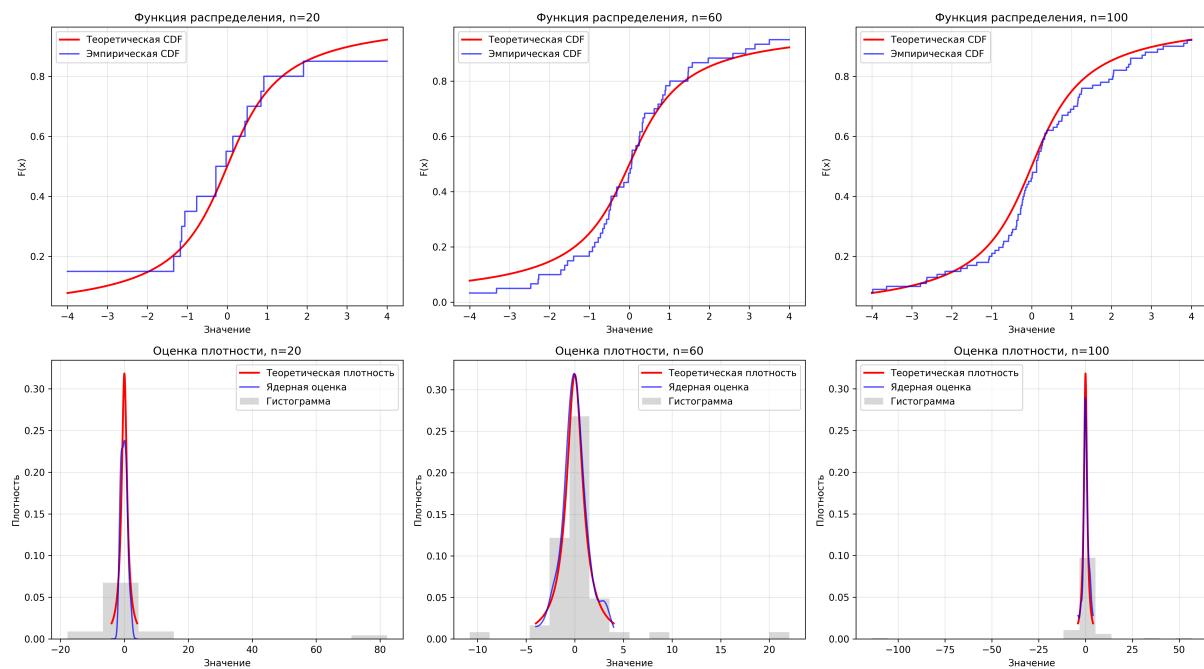


Рис. 9. Распределение Коши: ЭФР и KDE

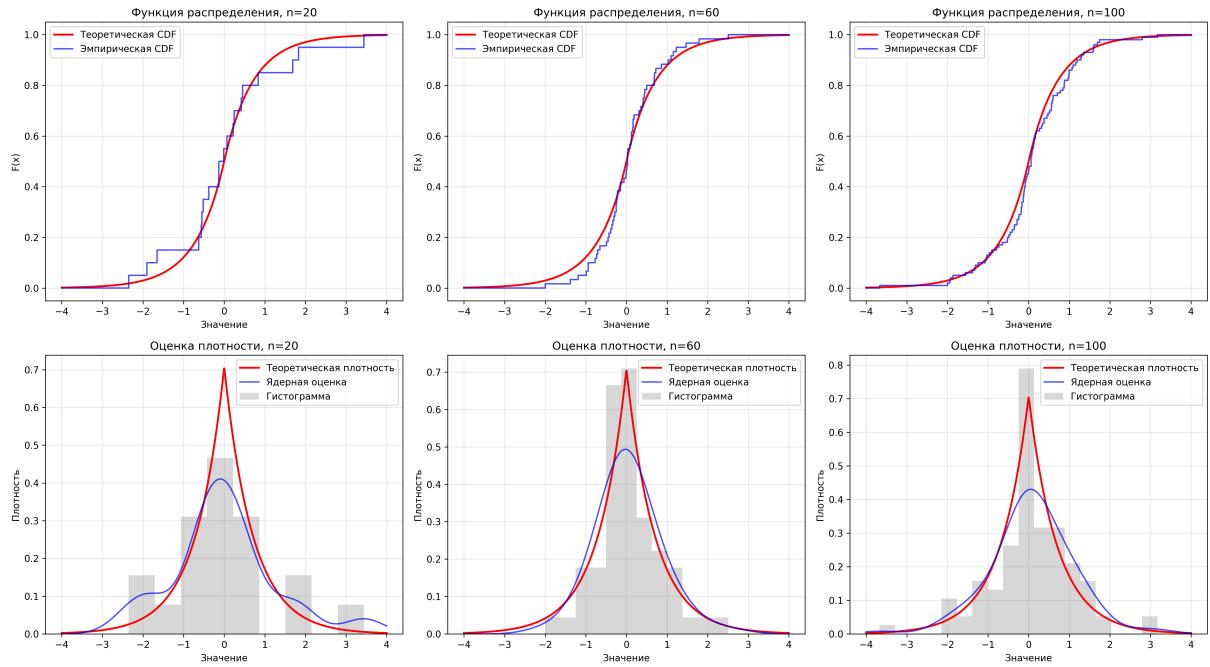


Рис. 10. Распределение Лапласа: ЭФР и KDE

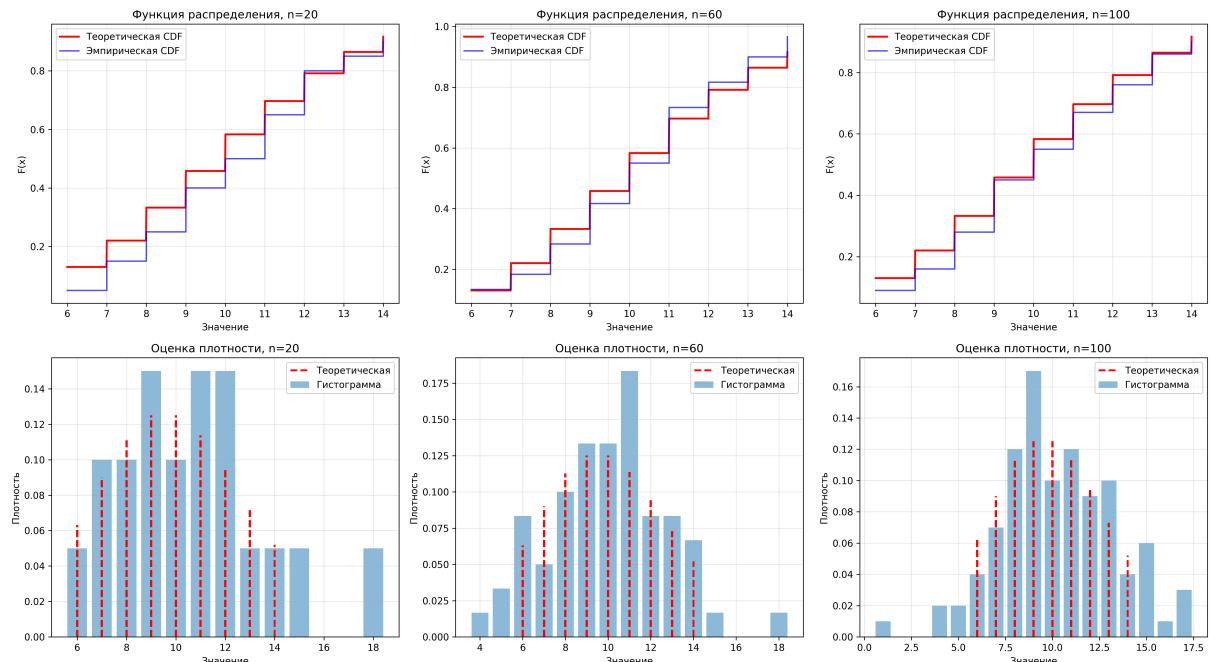


Рис. 11. Распределение Пуассона ($\lambda = 5$): ЭФР и KDE

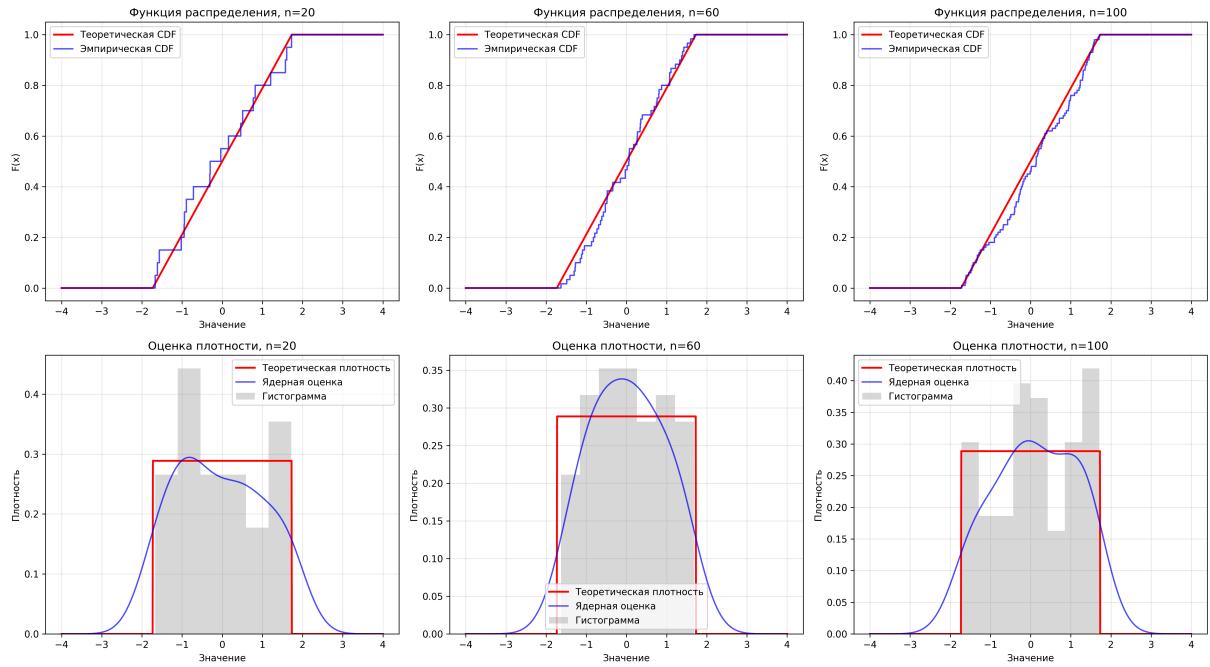


Рис. 12. Равномерное распределение: ЭФР и KDE

Распределение	n	MSE (CDF)	MSE (PDF)	Std (PDF)
Нормальное	20	0.0036	0.0025	0.0020
	60	0.0012	0.0010	0.0008
	100	0.0007	0.0007	0.0005
Коши	20	0.0073	0.0026	0.0017
	60	0.0024	0.0013	0.0006
	100	0.0014	0.0011	0.0004
Лаплас	20	0.0034	0.0046	0.0027
	60	0.0011	0.0025	0.0013
	100	0.0006	0.0019	0.0009
Пуассон	20	0.0043	—	—
	60	0.0014	—	—
	100	0.0008	—	—
Равномерное	20	0.0033	0.0045	0.0019
	60	0.0012	0.0030	0.0007
	100	0.0007	0.0025	0.0005

Таблица 7. MSE для ЭФР и ядерной оценки плотности (500 повторений)

5.5. Обсуждение

Эмпирическая функция распределения:

- При $n = 20$ наблюдаются заметные отклонения, особенно в хвостах
- При $n = 100$ ЭФР практически совпадает с теоретической функцией
- Подтверждается теорема Гливенко–Кантелли

Ядерная оценка плотности:

- KDE даёт гладкую оценку в отличие от ступенчатой гистограммы

- При малых n возможно избыточное сглаживание
- Для распределения Коши качество оценки ниже из-за тяжёлых хвостов

Сравнение методов:

- **Гистограмма:** проста, но зависит от выбора бинов, ступенчатая
- **KDE:** гладкая оценка, не требует выбора бинов, вычислительно сложнее
- При больших n оба метода дают схожие результаты

5.6. Выводы

- 1) Эмпирическая функция распределения сходится к теоретической с ростом объёма выборки.
- 2) Ядерные оценки обеспечивают гладкую аппроксимацию плотности, превосходящую гистограммы.
- 3) Для распределений с тяжёлыми хвостами требуется больший объём выборки для качественного приближения.
- 4) KDE предпочтительнее гистограмм для визуализации непрерывных распределений.

6. Общие выводы по блоку лабораторных работ

В ходе выполнения блока лабораторных работ №1–4 были исследованы основные методы описательной статистики:

- 1) **Визуализация распределений:** Гистограммы позволяют оценить форму распределения, но качество аппроксимации сильно зависит от объёма выборки и количества бинов.
- 2) **Характеристики положения:** Выборочное среднее не всегда является оптимальной оценкой центра. Для распределений с тяжёлыми хвостами робастные оценки (медиана, усечённое среднее) существенно превосходят среднее по устойчивости.
- 3) **Анализ выбросов:** Метод Тьюки на основе IQR эффективен для распределений с лёгкими хвостами, но даёт много «ложных» выбросов для распределений с тяжёлыми хвостами (Коши).
- 4) **Эмпирические оценки:** ЭФР и KDE обеспечивают состоятельные оценки функций распределения и плотности. KDE предпочтительнее гистограмм благодаря гладкости и независимости от выбора бинов.

Практическая значимость: Полученные результаты демонстрируют важность выбора статистических методов в зависимости от природы данных. Для «нормальных» данных подходят классические методы, но при наличии тяжёлых хвостов или выбросов необходимо использовать робастные подходы.

7. Список литературы

- 1) Гмурман В.Е. *Теория вероятностей и математическая статистика.* — 2003.
- 2) Кремер Н.Ш. *Теория вероятностей и математическая статистика.* — 2012.
- 3) Елисеева И.И., Юзбашев М.М. *Общая теория статистики.* — 2002.
- 4) Большев Л.Н., Смирнов Н.В. *Таблицы математической статистики.* — 1983.
- 5) Крамер Г. *Математические методы статистики.* — 1975.

8. Приложение. Ссылка на репозиторий GitHub

Исходный код программ, разработанных в ходе выполнения лабораторных работ, размещён в репозитории GitHub:

<https://github.com/Fromant/matstat>