**Министерство науки и высшего образования Российской Федерации**

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО**

Статистика для анализа данных

Отчёт по лабораторной работе № 2

«Описательная статистика»

Выполнили работу:

**Семёнов Никита Викторович**

**Торши Ромдхан**

ИСУ: **467414**

ИСУ: **467746**

Академическая группа:

№ **J3113**

Санкт-Петербург

2025

#### **Цель работы**

Получить практические навыки в расчёте описательных статистик для различных типов распределений, построении соответствующих графиков и анализе устойчивости статистических характеристик к выбросам. Научиться визуально сравнивать эмпирические и теоретические распределения.

**Ход выполнения работы**

1. Были выбраны два распределения:

* **Дискретное**: Пуассоновское распределение с параметром λ = 3.
* **Непрерывное**: Нормальное распределение с параметрами μ = 0, σ = 1.

1. Для каждого распределения была сгенерирована выборка из 1000 значений.
2. Для каждой выборки рассчитаны основные описательные статистики:

* Квартили (Q1, Q2, Q3),
* Среднее, мода, медиана,
* Размах, интерквартильный размах, дисперсия, стандартное отклонение, коэффициент вариации, среднее абсолютное отклонение,
* Асимметрия и эксцесс,
* Первые 5 начальных и центральных моментов.

1. Построены графики:

* Эмпирическая функция распределения с наложением теоретической CDF.
* Гистограммы с наложением теоретической плотности PDF (для непрерывного).
* Многоугольник вероятностей (для дискретного).
* Boxplot (ящик с усами) с выделением выбросов.

1. Проведено исследование устойчивости характеристик при добавлении выбросов к выборке нормального распределения. Построены графики изменения мер вариабельности (стандартное отклонение, MAD и коэффициент вариации) при постепенном добавлении выбросов от 0% до 5%.

**Основная часть**

### **Краткие определения:**

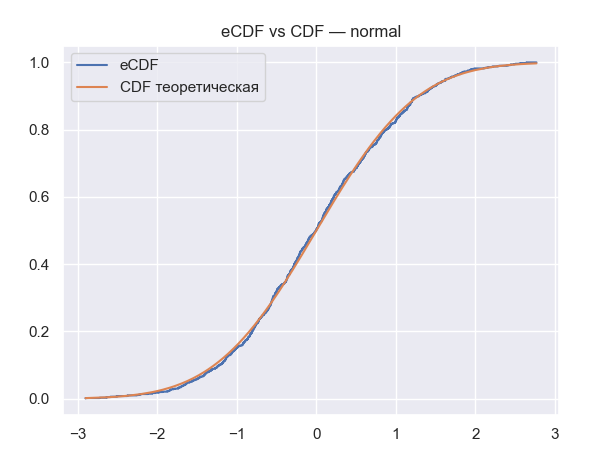
* **Квартили (Q1, Q2, Q3)** — значения, разделяющие выборку на четыре равные части по числу элементов.
* **Среднее** — сумма всех значений, делённая на их количество.
* **Мода** — наиболее часто встречающееся значение.
* **Медиана (Q2)** — значение в центре упорядоченной выборки.
* **Размах** — разница между максимальным и минимальным значением.
* **IQR (интерквартильный размах)** — Q3 − Q1, отражает "ядро" выборки.
* **Дисперсия** — средний квадрат отклонения от среднего. (Второй момент)
* **Стандартное отклонение (σ)** — корень из дисперсии, отражает "разброс".
* **Коэффициент вариации (CV)** — σ, делённое на среднее, мера относительного разброса.
* **MAD** — среднее абсолютное отклонение, альтернативная мера разброса.
* **Асимметрия** — показывает, насколько распределение "перекошено". Вправо или влево. Означает что справа или слева есть сильные выбросы.
* **Эксцесс** — степень "вытянутости" распределения. (Показывает насколько много выбросов – насколько стабильные данные. Описывает "остроконечность" и поведение хвостов)
* **Моменты** — числовые характеристики формы распределения. (Начальные моменты – отражают общую форму распределения относительно нуля. Центральные моменты – характеризуют форму распределения относительно среднего.)

x̄ - среднее, k – номер момента

**Графики и Анализ**

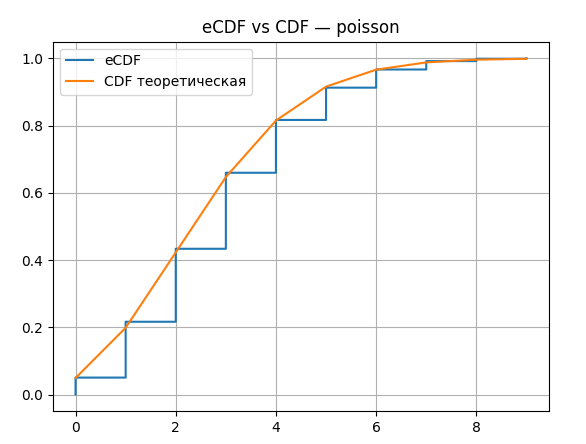
| **Показатель** | **Выборка 1 (дискретная)** | **Выборка 2 (нормальное распределение)** |
| --- | --- | --- |
| Q1 | 2.0 | –0.657 |
| Медиана (Q2) | 3.0 | –0.001 |
| Q3 | 4.0 | 0.709 |
| Среднее | 2.96 | 0.014 |
| Мода | 3 | –2.9 |
| Размах | 9 | 5.67 |
| IQR | 2.0 | 1.37 |
| Дисперсия | 3.00 | 0.96 |
| Ст. отклонение (σ) | 1.73 | 0.98 |
| Коэффициент вариации | 0.586 | 69.52 |
| MAD | 1.36 | 0.79 |
| Асимметрия | 0.53 (вправо) | 0.06 (почти симметрия) |
| Эксцесс | –0.03 | –0.19 |
| Момент 1 | 0.0 | 0.0 |
| Момент 2 | 3.00 | 0.96 |
| Момент 3 | 2.76 | 0.057 |
| Момент 4 | 26.78 | 2.60 |
| Момент 5 | 67.14 | 0.17 |

#### **1. Эмпирическая и теоретическая CDF — нормальное распределение**



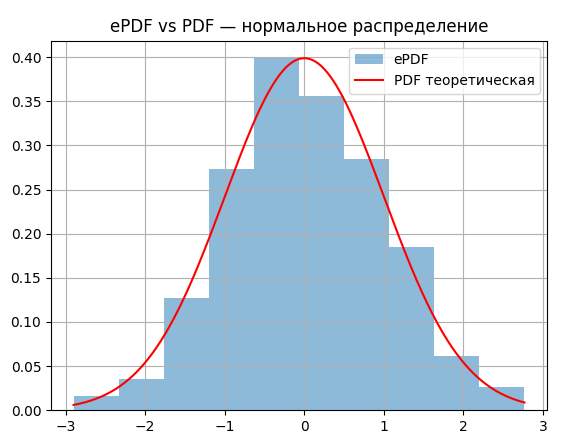
Линии эмпирической и теоретической функции распределения почти совпадают, что говорит о хорошем соответствии выборки нормальному распределению.

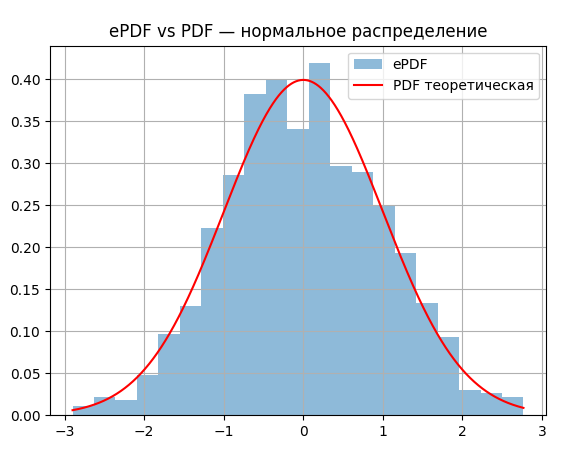
**2. Эмпирическая и теоретическая CDF — Пуассоновское распределение**



Эмпирическая CDF ступенчатая, как и ожидается для дискретных данных. Совпадает с теоретической функцией распределения Пуассона.

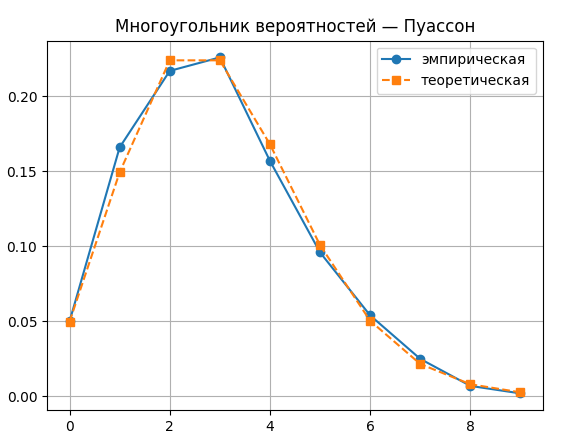
#### **3. PDF (нормальное распределение)**





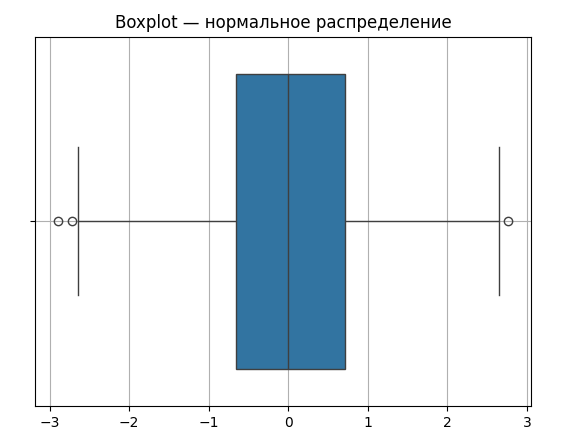
Чем больше количество бинов, тем точнее гистограмма повторяет форму колокола нормального распределения. При слишком малом числе бинов (например, 10) теряются детали.

**4. Многоугольник вероятностей — Пуассон**



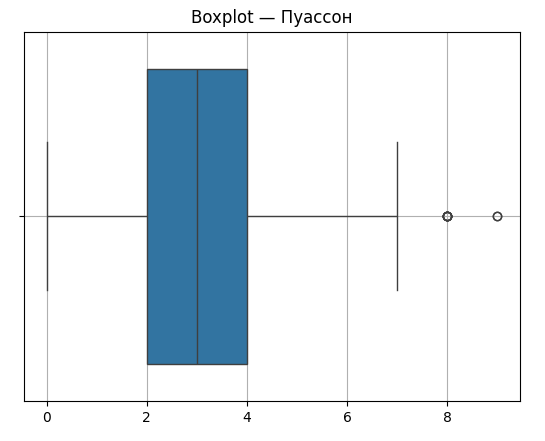
Эмпирическая кривая вероятностей схожа с теоретической. Это подтверждает, что распределение действительно похоже на Пуассоновское с λ = 3.

#### **5. Boxplot — нормальное распределение**



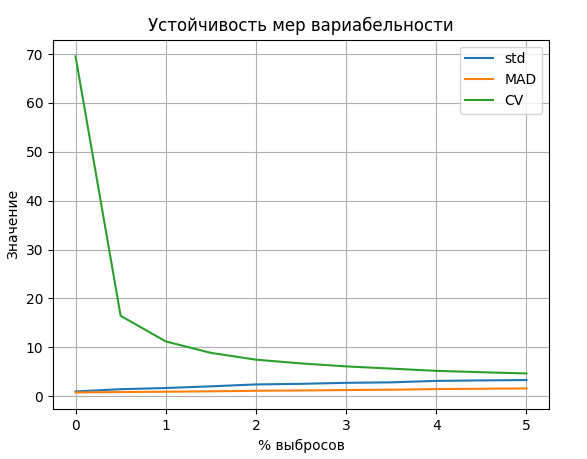
Большинство значений лежат в пределах «ящика» (между Q1 и Q3). Выбросов немного, как и ожидается для нормального распределения.

#### **6. Boxplot — Пуассоновское распределение**



Распределение смещено влево (асимметрия положительная), присутствуют выбросы на больших значениях. Медиана около 3.

**7. Графики устойчивости статистик при добавлении выбросов**



**MAD (оранжевая линия)** — **самая устойчивая**. Почти не меняется при добавлении выбросов. Это подтверждает, что **MAD — хорошая мера разброса в выборках с шумом**.

**Стандартное отклонение (синяя линия)** — **чувствительно к выбросам**, но изменение умеренное. Оно растёт с увеличением количества выбросов.

**Коэффициент вариации (зелёная линия)** — **самый нестабильный**. Он резко изменяется при добавлении даже 1% выбросов. Это потому, что в его формуле есть деление на среднее, которое само искажается при наличии экстремальных значений. (В начале среднее близко к нулю, поэтому такая высота)

**Вывод**

В этой лабораторной работе мы научились рассчитывать и сравнивать описательные статистики для разных распределений. Мы увидели, как выглядят выборки на графиках и как они соотносятся с теоретическими функциями.

Также мы увидели, как выбросы влияют на статистики: коэффициент вариации сильно меняется, стандартное отклонение и MAD более устойчивы. Это показывает, что не все характеристики одинаково надёжны при наличии выбросов.

**Ссылка на гит хаб:**

[user6778899/Statistic\_labs](https://github.com/user6778899/Statistic_labs)