Санкт-Петербургский Политехнический
университет имени Петра Великого

Отчёт по лабораторной работе №1

Гистограммы и плотности распределений

Студент: Расторгуев Михаил Павлович
Группа: 5030102/30201

Санкт-Петербург
2026

1 Постановка задачи

В рамках данной лабораторной работы было необходимо:

- Освоить принцип группировки данных и построения гистограмм.
- Сделать вывод о том, как влияет размер выборки на определение характера распределения величины с помощью гистограммы.

2 Теоретическая часть

В данной работе требовалось изучить следующие распределения случайных величин:

1. Нормальное распределение $N(x; 0, 1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

2. Распределение Коши $C(x; 0, 1)$

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}$$

3. Распределение Лапласа $L\left(x; 0, \frac{1}{\sqrt{2}}\right)$

$$f(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|}, \quad x \in \mathbb{R}$$

4. Распределение Пуассона $P(k; 5)$

$$P(X = k) = \frac{5^k e^{-5}}{k!}, \quad k = 0, 1, 2, \ldots$$

5. Равномерное распределение $U(x; -\sqrt{3}, \sqrt{3})$

$$f(x) = \frac{1}{2\sqrt{3}}, \quad x \in [-\sqrt{3}, \sqrt{3}]$$

3 Реализация

Практическая часть работы была реализована на языке программирования `Python` с использованием библиотек: `scipy,seaborn,numpy,pandas`.
Код писался в интерактивном блокноте `Jupyter Notebook`.

Были реализованы следующие функции:

`generate_and_draw(n,a,d,**param)` - функция, генерирующая набор из случайных данных и затем отображающая его на графике. На вход подаётся `n` - размер генерируемых данных; `a` - ось matplotlib, на которую следует отобразить график; `d` - распределение, по которому генерируются данные; `**param` - параметры распределения.

`show_different_n(d,name,**parameters)` - функция - обёртка над `generate_and_draw`.

4 Результаты

Ниже представлены гистограммы распределений сгенерированных выборок (синие столбики) и наложенные на них теоретические кривые плотностей распределений (оранжевые линии).

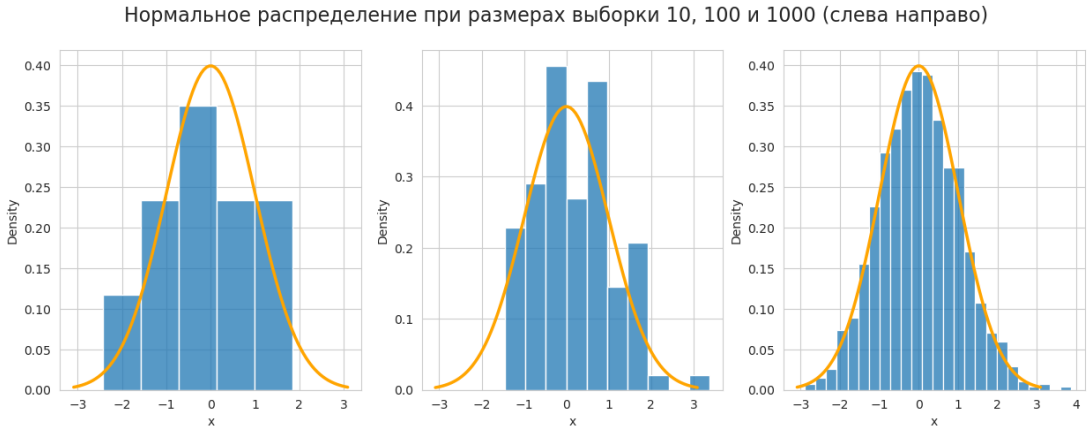

Рис. 1: $N(0,1)$

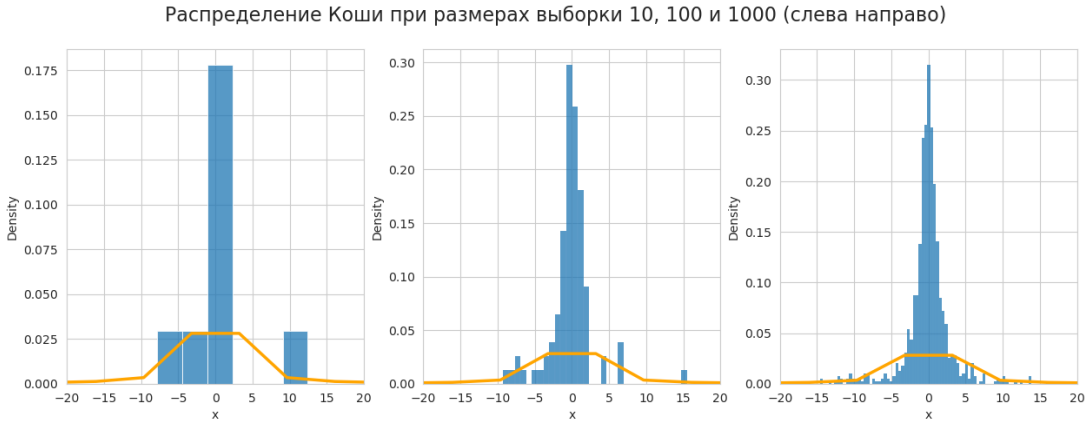

Рис. 2: $C(0, 1)$

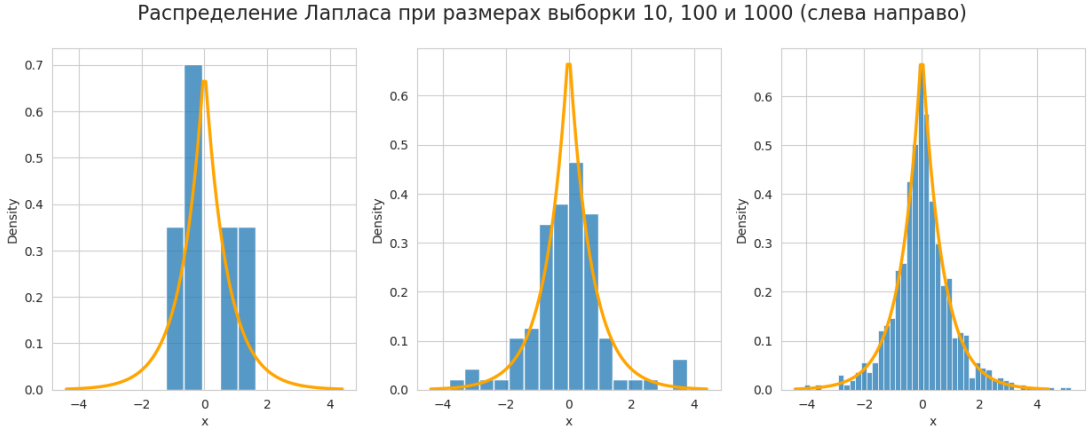

Рис. 3: $L(0, \frac{1}{\sqrt{2}})$

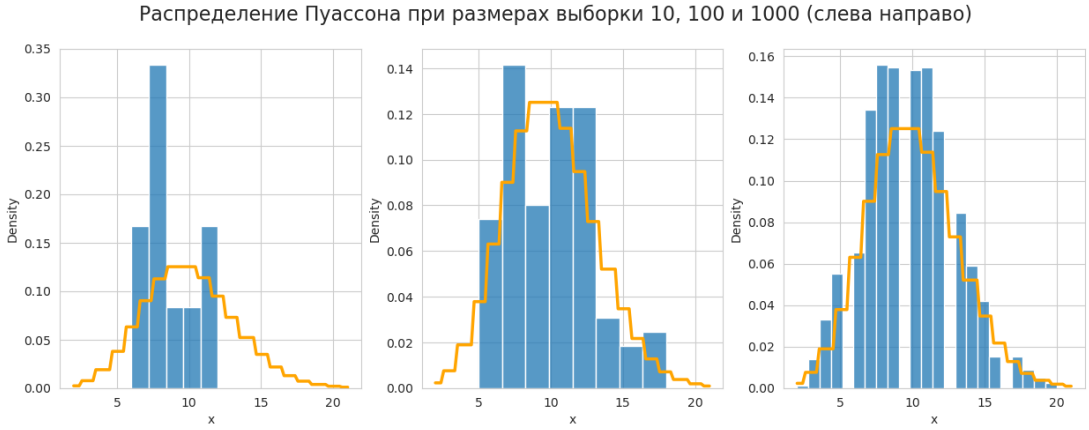

Рис. 4: $P(5)$

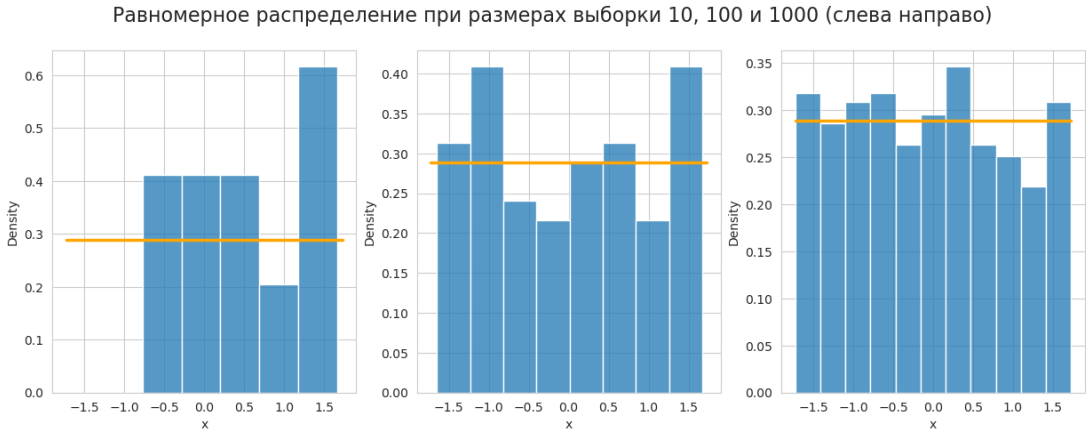

Рис. 5: $U(-\sqrt{3}, \sqrt{3})$

5 Обсуждение

Из всех графиков (кроме рис.2) отчётливо видно, что с увеличением размеров данных, их распределения стремятся к теоретическим.

График, представленный на Рис.2 трудно назвать информативным.
Дело в том, что в данной программной реализации подбор бинов осуществляется автоматически: либо по формуле $k = log_2 n + 1$ (для нормальных данных), либо $w = 2 \cdot IQR \cdot n^{-\frac{1}{3}}$, где IQR = межквартильный размах (75-й перцентиль минус 25-й).

На Рис.2 было автоматически выбрано второе правило. Из-за тяжёлых хвостов распределения Коши, IQR не может иметь значение, подходящее для определения ширины бинов. Таким образом, мы видим, как столбики гистограммы возвышаются над теоретической кривой.

6 Выводы

1. С увеличением размера выборки, распределение данных стремится к теоретическому.
2. От выбора ширины/количества бинов гистограммы принципиально зависит информативность графика. Не всегда можно положиться на инструменты автоматического подбора этого параметра.

7 Список литературы

1. Документация `scipy` - https://docs.scipy.org/doc/scipy/

2. Документация `seaborn` - https://seaborn.pydata.org/

8 Приложение

https://github.com/haskell-md2/MatStat/blob/main/lab1-4/lab1-4.ipynb - Параграф "Лабораторная 1"