Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторной работе

Построение гистограмм для сгенерированных выборок

По дисциплине «Теория вероятностей и математическая статистика»

Выполнил
Студент гр.3630201/80101 _______ В.Н. Сеннов

Руководитель
доцент к.ф.-м.н. _______ А.Н. Баженов

«___»__________ 202__г.

Санкт-Петербург
2020

Содержание

Список иллюстраций

1 Постановка задачи

Для заданных распределений нужно сгенерировать выборки размером 10, 100, 1000 элементов. Для каждой выборки нужно построить гистограмму и теоретическую функцию плотности вероятности на одном рисунке.

Заданные распределения:

1. Нормальное (гауссово) распределение с параметрами $\mu = 0$, $\sigma = 1$;
2. Распределение Коши с параметрами $\mu = 0$, $\lambda = 1$;
3. Распределение Лапласа с параметрами $\mu = 0$, $\lambda = \frac{1}{\sqrt{2}}$;
4. Распределение Пуассона с параметром $\mu = 10$;
5. Равномерное распределение с параметрами $a = -\sqrt{3}$, $b = \sqrt{3}$.

2 Математическое описание

2.1 Рассматриваемые распределения

2.1.1 Нормальное распределение

Рассматриваемое распределение имеет следующую функцию плотности вероятности:

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

Функция распределения имеет следующий вид:

$$F_n(x) = 0.5 + \Phi(x), \tag{2}$$

где $\Phi(x)$ — нормированная функция Лапласа [2].

Сгенерировать случайную величину, соответствующую данному нормальному распределению можно по следующей формуле:

$$x = \sum_{i=1}^{12} r_i - 6, \tag{3}$$

где r_i — случайная величина из интервала $(0, 1)$ [1].

2.1.2 Распределение Коши

Рассматриваемое распределение имеет следующую функцию плотности вероятности:

$$f_c(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} \tag{4}$$

Функция распределения имеет следующий вид:

$$F_c(x) = 0.5 + \frac{\operatorname{arctg} x}{\pi} \tag{5}$$

Сгенерировать случайную величину, соответствующую данному распределению можно по следующей формуле:

$$x = \operatorname{tg}(2\pi r_i), \tag{6}$$

где r_i — случайная величина из интервала $(0, 1)$ [1].

2.1.3 Распределение Лапласа

Рассматриваемое распределение имеет следующую функцию плотности вероятности:

$$f_L(x) = \frac{1}{2\sqrt{2}} e^{-\frac{|x|}{\sqrt{2}}} \tag{7}$$

Функция распределения задается следующим образом:

$$F_L(x) = \begin{cases} \frac{1}{2} e^{\frac{x}{\sqrt{2}}}, & x \geq 0 \\ 1 - \frac{1}{2} e^{-\frac{x}{\sqrt{2}}}, & x < 0 \end{cases} \tag{8}$$

Сгенерировать случайную величину, соответствующую данному распределению можно по следующей формуле:

$$x = \sqrt{2} \ln(r_{i+1}/r_i) \tag{9}$$

где r_i — случайная величина из интервала $(0, 1)$ [1].

2.1.4 Распределение Пуассона

Рассматриваемое распределение задается формулой:

$$p_k = \frac{10^k}{k!} e^{-10} \tag{10}$$

Случайная величина, соответствующая данному распределению может быть сгенерирована по алгоритму 1 [1].

Алгоритм 1 Генерирование случайной величины согласно распределению Пуассона

Вход: μ — параметр распределения, r_i — случайная величина из интервала $(0; 1)$
Выход: x — случайная величина

```
p ← e^{-μ}
x ← 0
r ← r_i − p
while r ≥ 0 do
   x ← x + 1
   p ← pμ/x
   r ← r − p
end while
return x
```

2.1.5 Равномерное распределение

Рассматриваемое равномерное распределение имеет следующую функцию плотности вероятности:

$$f_U(x) = \begin{cases} \frac{1}{2\sqrt{3}}, & -\sqrt{3} \le x \le \sqrt{3} \\ 0, & \text{иначе} \end{cases} \tag{11}$$

2.2 Гистограмма

Гистограмма в математической статистике — это функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него [3].

В процессе построения гистограммы числовой промежуток, покрывающий значения элементов выборки, разбивают на несколько интервалов. Эти интервалы называются бинами. Для каждого бина подсчитывается количество элементов выборки, попавших в него. Далее на графике строятся прямоугольники, основание которых совпадает с бинами, а высота пропорциональна количеству элементов, попавших в этот бин.

В рамках данной работы брался отрезок $[x_{min}; x_{max}]$, где x_{min} и x_{max} — минимальный и максимальный элементы выборки. Отрезок делился на бины равной ширины. Количество бинов n определялось по эвристике:

$$n = \left[\sqrt{N}\right], \tag{12}$$

где N — количество элементов в выборке.

Гистограмма задавалась следующей функцией:

$$h(x) = \begin{cases} \frac{n_k}{N \cdot b}, & x \in [x_{min}; x_{max}] \\ 0, & \text{иначе} \end{cases} \tag{13}$$

где b — ширина одного бина, N — количество элементов в выборке, x_{min} и x_{max} — минимальный и максимальный элементы выборки, $k = \left[\frac{x-x_{min}}{b}\right]$, n_k — количество элементов выборки в k-том бине.

3 Особенности реализации

Программа для выполнения лабораторной была написана на языке Python 3.8.2. Для отображения графиков использовалась библиотека Mathplotlib.

Программа состоит из 3 модулей:

1. Модуль **distributions**. Содержит классы, реализующие необходимые распределения. Все классы унаследованы от класса **AbstractDistribution**. Он имеет 2 поля:

    - **name: str** — название распределения, используется для подписи на на графиках;
    - **parameters: dict** — словарь с параметрами распределения.

    Класс имеет следующие методы:

    - **Конструктор** — инициализирует поля класса.
    - **x()->float** — возвращает случайную величину, соответствующую распределению.
    - **f(x:float)->float** — возвращает значение функции плотности вероятности в точке **x**.
    - **F(x:float)->float** — возвращает значение функции распределения в точке **x**.
    - **discrete()** — возвращает True, если распределение дискретное, иначе False. По умолчанию возвращает False.

    От этого класса унаследованы классы **Normal**, **Cauchy**, **Laplace**, **Poisson**, **Uniform**. Они реализуют формулы для вычислений, приведенные в разделе 2.1. Интерес представляет только метод вычисления функции распределения для нормального распределения (2). Значение функции вычисляется по таблице, сохраненной в отдельном файле. Генерация случайной величины производилась согласно формулам (3), (6), (9) и алгоритму 1. Также стоит указать, что диапазон возможных значений случайной величины, соответсвующей распределению Коши был насильно ограничен до интервала $(-1591; 1591)$, чтобы не возникало ошибок области определения математических фукнций.

2. Модуль **histogram**. Содержит класс **Histogram**, реализующий преобразование выборки в гистограмму. Количество бинов вычисляется при помощи эвристики (12). Высота столбца в гистограмме вычисляется по формуле (13).

3. Модуль **plot**. Модуль верхнего уровня, в нем инициализируются заданные распределения, генерируются выборки и отображаются графики.

В приложении А размещена ссылка на исходный код.

4 Результаты работы программы

В данном разделе представлены построенные графики. На каждом рисунке представлены графики функции плотности вероятности $f(x)$ и функции распределения $F(x)$ для выборок разного размера.

На рис. 1 представлены графики для нормального распределения. Теоретические значения вычислены по формулам 1 и 2.

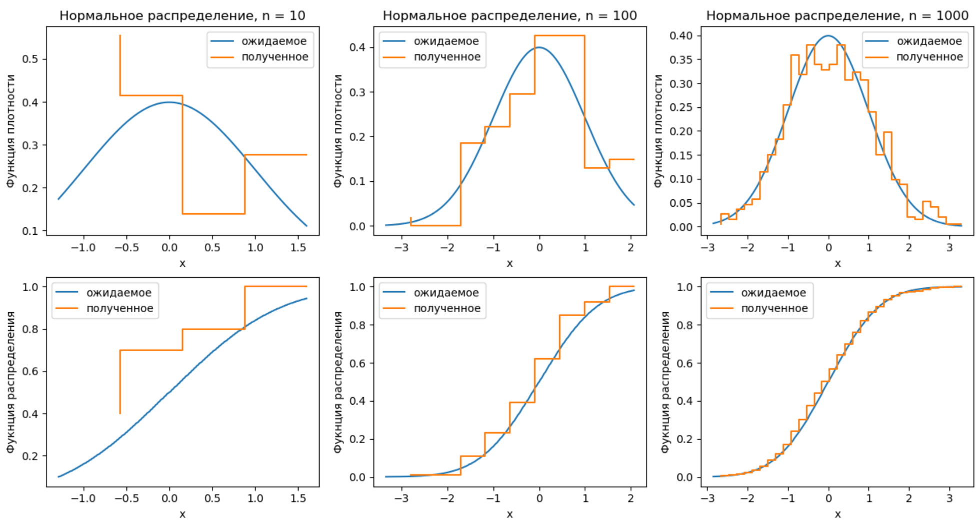

Рис. 1: Графики для нормального распределения

На рис. 2 представлены графики для распределения Коши. Теоретические значения вычислены по формулам (4) и (5).

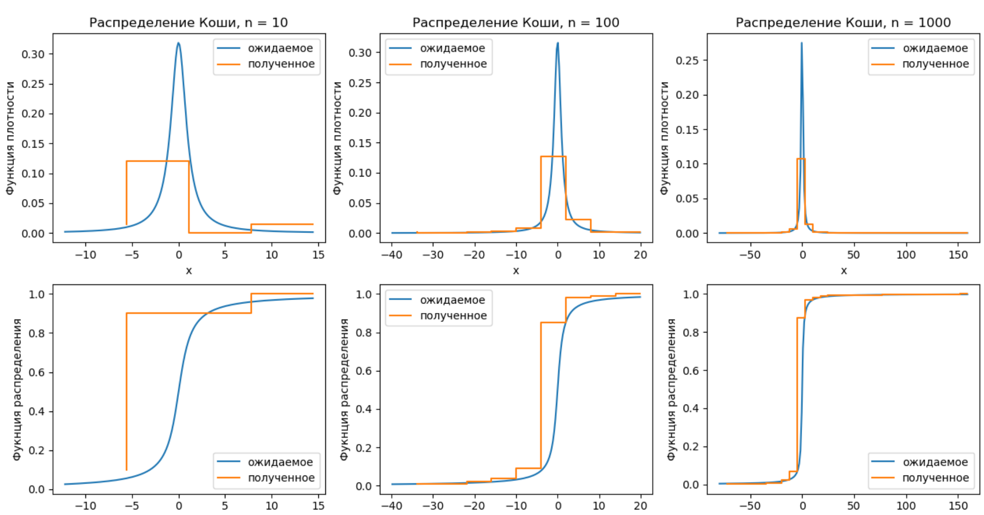

Рис. 2: Графики для распределения Коши

На рис. 3 представлены графики для распределения Лапласа. Теоретические значения вычислены по формулам (7) и (8).

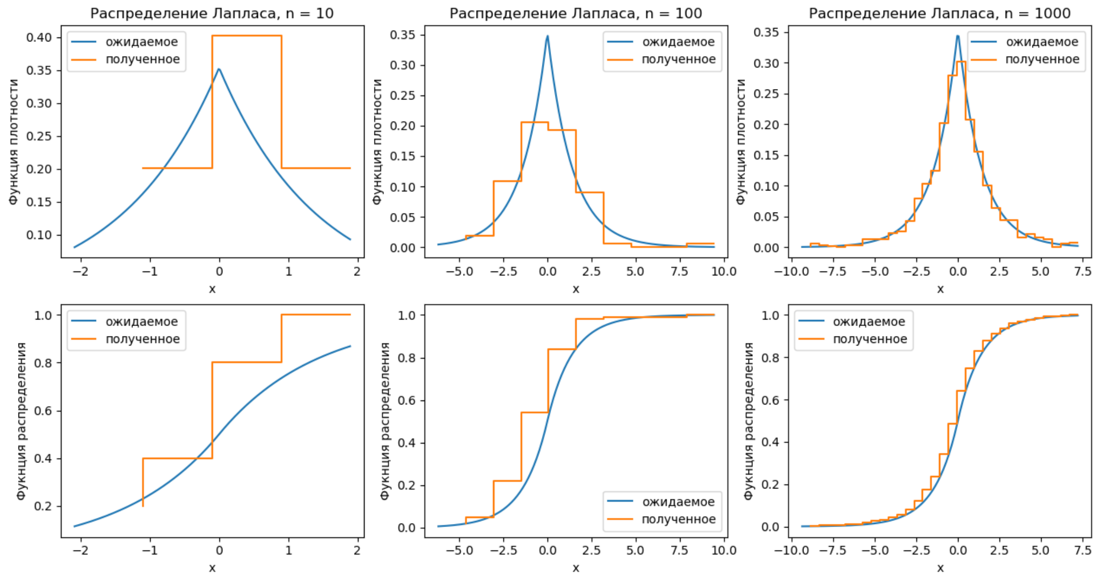

Рис. 3: Графики для распределения Лапласа

На рис. 4 представлены графики для распределения Пуассона. Теоретические значения вычислены по формуле (10).

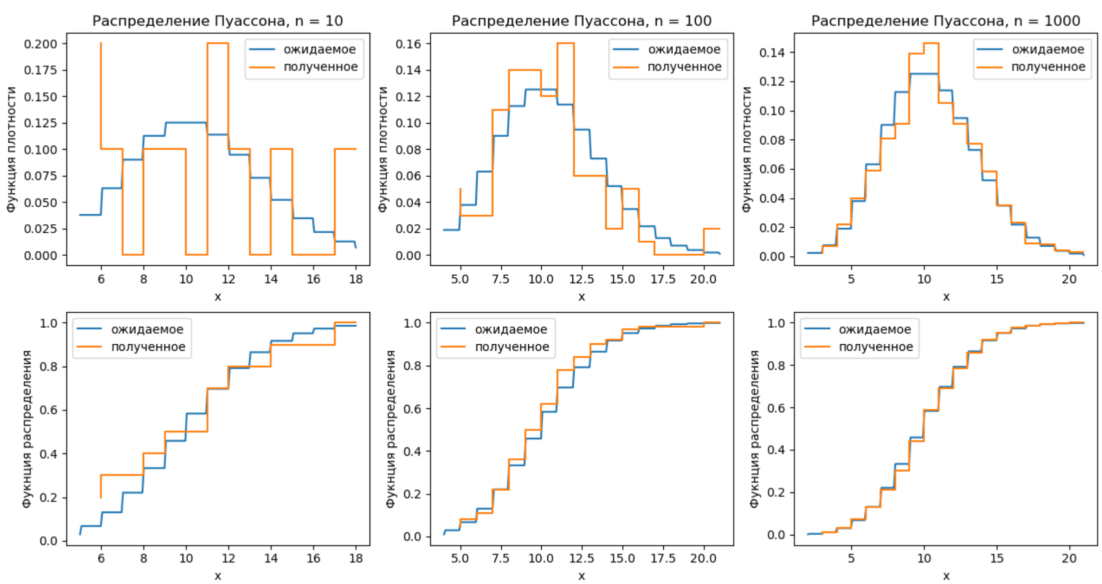

Рис. 4: Графики для распределения Пуассона

На рис. 5 представлены графики для равномерного распределения. Теоретические значения вычислены по формуле (11).

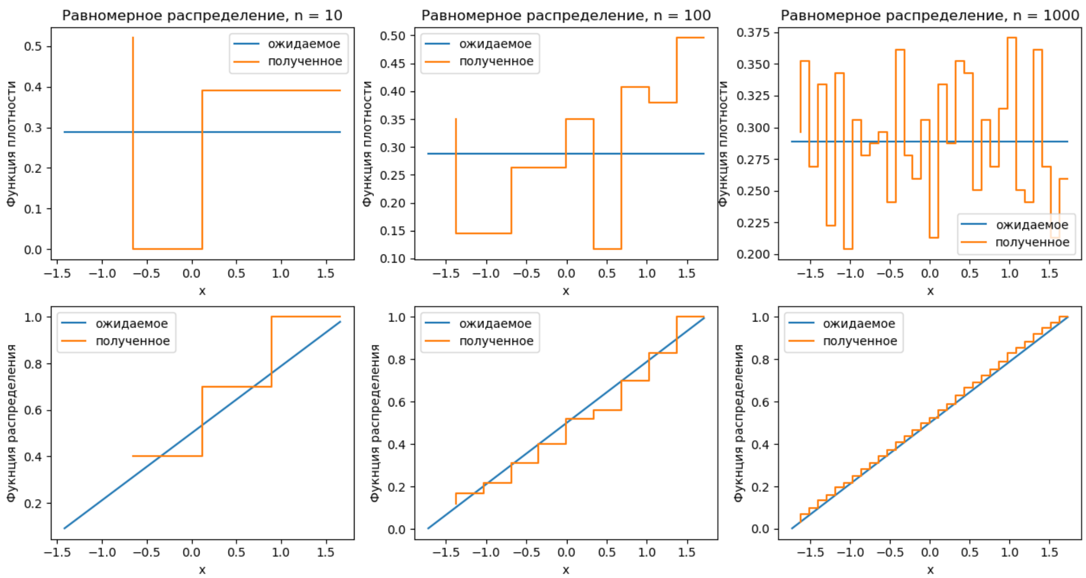

Рис. 5: Графики для равномерного распределения

Заключение

В рамках лабораторной работы были сгенерированы выборки трех разных размеров для пяти различных распределений. Были построены графики гистограмм полученных выборок и теоретических функций.

Заметно, что при малой выборке гистограммы лишь отдаленно напоминает теоретические функции, но при большей выборке функции распределения становятся очень похожи.

Программа для лабораторной была написана языке Python 3.8.2. Для отображения графиков использовалась библиотека Mathplotlib.

Список литературы

[1] Вадзинский Р. Н. Справочник по вероятностным распределениям. — СПб.: Наука, 2001. — 295 с., ил. 116.

[2] Максимов Ю. Д. Математика. Теория вероятностей и случайных процессов; Учебное пособие / Под ред. В. И. Антонова. — СПб.: Изд-во Политехн. ун-та, 2008. — 384 с.

[3] Теоритическое приложение к лабораторным работам №1-4 по дисциплине «Математическая статистика». — Спб.: Сантк-Петербургский политехнический университет, 2020. — 12 с.

A Репозиторий с исходным кодом

Исходный код программы для данной лабораторной размещен на сервисе GitHub.

Ссылка на репозиторий: https://github.com/Vovan-S/TV-Lab1.