Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторной работе

Построение боксплотов и рассчет доли выбросов

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 таблиц

Список иллюстраций

1 Постановка задачи

Для заданных распределений нужно сгенерировать выборки размером 20, 100 элементов. Для каждой выборки нужно построить боксплот и рассчитать долю выбросов. Долю выбросов нужно посчитать 1000 раз и взять среднее значение. Также необходимо рассчитать теоретическую вероятность выбросов.

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 Построение боксплотов

Боксплотом называется диаграмма, компактно изображающая распределение одномерной величины. [1]

Диаграмма состоит из следующих частей:

1. Ящик. Границами ящика являются верхний и нижний квартили. Внутри ящика проводится линия — медиана.

2. Усы. Вне ящика изображают линии, напоминающие усы. Длина усов равна либо $3/2$ от межквартильного расстояния, либо разнице экстремального значения с квартилем. Тогда для выборки $x_1, \ldots, x_n$ концы усов будут иметь координаты X_1 и X_2:

$$X_1 = \max\left\{x_1,\ z_{1/4} - 3/2\left(z_{3/4} - z_{1/4}\right)\right\} \tag{1}$$

$$X_2 = \min\left\{x_n,\ z_{3/4} + 3/2\left(z_{3/4} - z_{1/4}\right)\right\} \tag{2}$$

3. Выбросы. Значения выборки, меньшие X_1 или большие X_2 отображаются на диаграмме кружочками.

[1]

2.2 Теоретическая вероятность выбросов

По формулам (1) и (2) можно вычислить теоретические значения X_1^T и X_2^T. Выбросом считается значение случайной величины, меньшее X_1^T или большее X_2^T. Тогда теоретическая вероятность выброса может быть рассчитана по формуле:

$$P_{sp} = P\left(x \in \left(-\infty; X_1^T\right) \cup \left(X_2^T; +\infty\right)\right) = F(X_1^T) + \left(1 - F(X_2^T)\right), \tag{3}$$

где $F(x)$ — функция распределения [2].

3 Особенности реализации

Программа для выполнения лабораторной была написана на языке Python 3.8.2. Для генерации выборок использовался модуль **distributions**, написанный для лабораторной №1. Для построения боксплотов использовалась библиотека Matplotlib.

Для вычисления доли выбросов был написан модуль **spikes**. В нем на основе формул (1) и (2) рассчитывается средняя доля выбросов.

В приложении А приведена ссылка на репозиторий с исходным кодом.

Теоритическая вероятность выбросов рассчитана вручную по формуле (3), математические выкладки приведены в приложении В.

4 Результаты работы программы

4.1 Боксплоты

На рис. 1 изображен боксплот для выборки, соответствующей нормальному распределению.

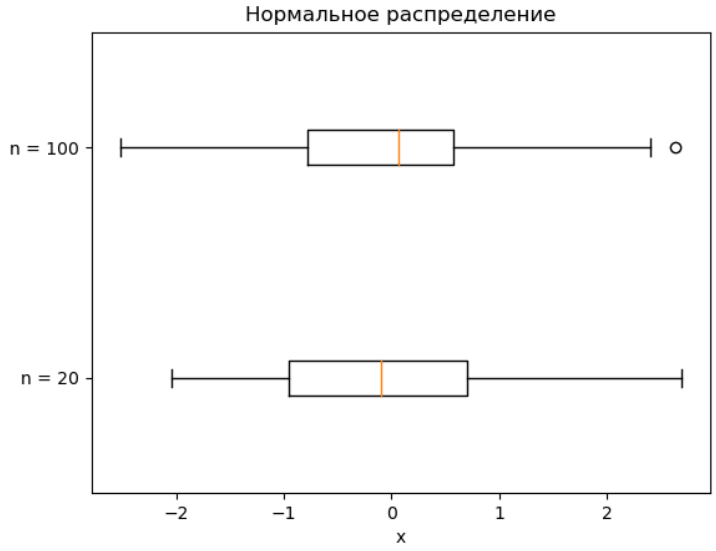

Рис. 1: Боксплот для выборки, соответствующей нормальному распределению

На рис. 2 изображен боксплот для выборки, соответствующей распределению Коши.

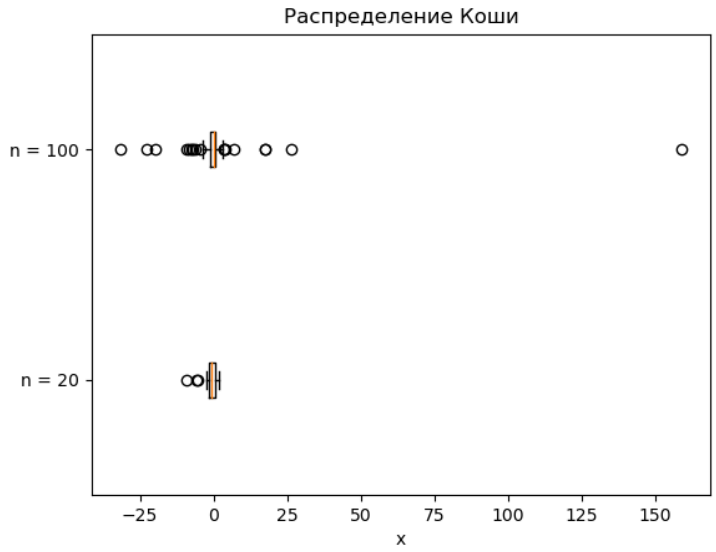

Рис. 2: Боксплот для выборки, соответствующей распределению Коши

На рис. 3 изображен боксплот для выборки, соответствующей распределению Лапласа.

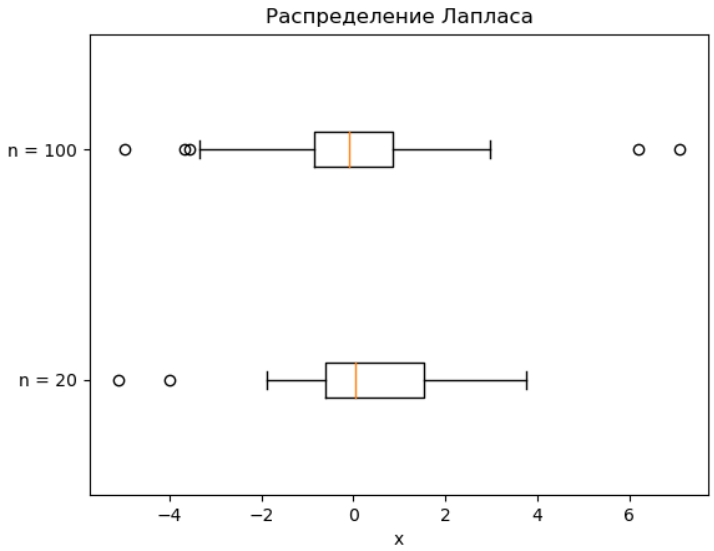

Рис. 3: Боксплот для выборки, соответствующей распределению Лапласа

На рис. 4 изображен боксплот для выборки, соответствующей распределению Пуассона.

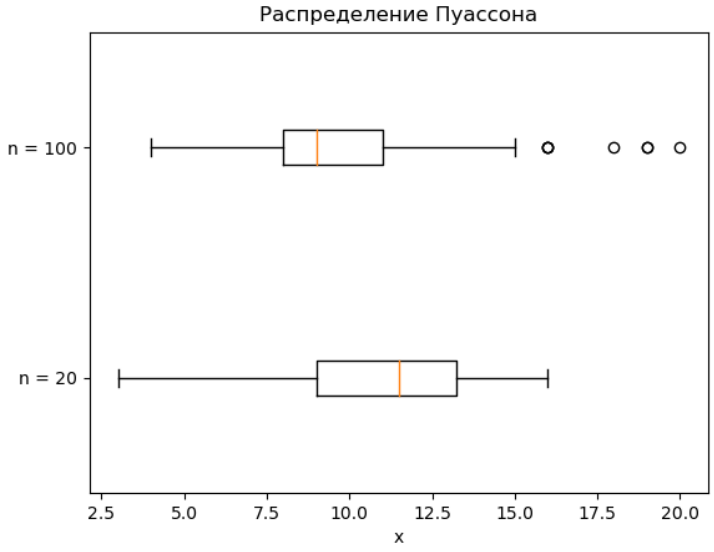

Рис. 4: Боксплот для выборки, соответствующей распределению Пуассона

На рис. 5 изображен боксплот для выборки, соответствующей равномерному распределению.

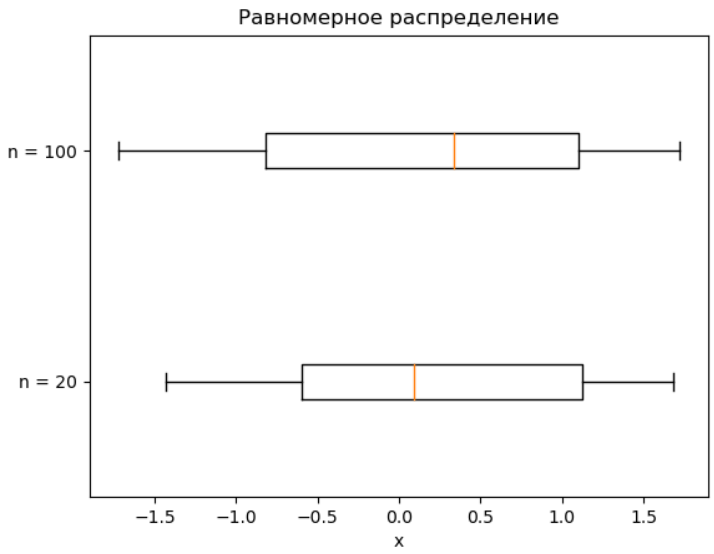

Рис. 5: Боксплот для выборки, соответствующей равномерному распределению

4.2 Доля выбросов

В таблице 1 представлены полученные экспериментально доли выбросов и теоретические значения. Экспериментальные значения приведены с округлением, погрешность рассчитана по следующей формуле:

$$\Delta_z = \sqrt{\frac{1}{n}\sum z_i^2 - \frac{1}{n}\left(\sum z_i\right)^2}$$

Распределение	n	Доля выбросов	P_{sp}
Нормальное	20 100	0.02 0.007	0.00694
Коши	20 100	0.14 0.15	0.156
Лапласа	20 100	0.06 0.06	0.0625
Пуассона	20 100	0.02 0.01	0.00996
Равномерное	20 100	0.001 0.0	0

Таблица 1: Доли выбросов и теоретические вероятности выбросов P_{sp}

Заключение

В рамках лабораторной работы были построены боксплоты для заданных распределений, были вычислены доли выбросов и теоретические вероятности выбросов.

По построенным боксплотам можно легко отличить на вид распределение Лапласа и Коши, а вот нормальное распределение и равномерное распределение оказываются очень похожи.

Заметно, что если выборка больше, то доля выбросов ближе к теоретической. Также заметно, что разные распределения весьма заметно отличаются долями выбросов.

Программа для лабораторной была написана языке Python 3.8.2, для построения графиков использовалась библиотека Matplotlib.

Список литературы

[1] Box plot. // Wikipedia, the free encyclopedia. — URL: https://en.wikipedia.org/wiki/Box_plot. — (дата обращения: 09.11.2020)

[2] 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.

B Вычисление теоретической вероятности выбросов

Нормальное распределение

Заданное распределение имеет параметры $\mu = 0$, $\sigma = 1$, то есть является стандартным. Значение верхнего и нижнего квартилей найдем приблизительно по таблице значений функции Лапласа, поскольку $\Phi(z_{1/4}) = 0.25$, $\Phi(z_{3/4}) = 0.75$.

По таблице находим значения: $z_{1/4} = -0.675$, $z_{3/4} = 0.675$.

Тогда $X_1^T = -2.700$, $X_2^T = 2.700$. Тогда

$$P_{sp} = \Phi(-2.7) + 1 - \Phi(2.7) = 1 - 2\Phi_0(2.7) = 1 - 2 \cdot 0.49653 = 0.00694$$

Распределение Коши

Заданное распределение имеет функцию распределения $F_C(x) = 0.5 + \frac{\operatorname{arctg} x}{\pi}$. Найдем $z_{1/4}$:

$$0.25 = 0.5 + \frac{\operatorname{arctg} z_{1/4}}{\pi}$$

$$\operatorname{arctg} z_{1/4} = -\frac{\pi}{4}$$

$$z_{1/4} = -1$$

Очевидно, что $z_{3/4} = -z_{1/4} = 1$. Тогда $X_1^T = -4$, $X_1^T = 4$. Тогда:

$$P_{sp} = F_C(-4) + 1 - F_C(4) = 1 - 2\frac{\operatorname{arctg} 4}{\pi} \approx 0.156$$

Распределение Лапласа

Заданное распределение имеет следующую функцию распределения:

$$F_L(x) = \begin{cases} \frac{1}{2}e^{\frac{x}{\sqrt{2}}}, & x \le 0 \\ 1 - \frac{1}{2}e^{-\frac{x}{\sqrt{2}}}, & x > 0 \end{cases}$$

Найдем $z_{1/4}$:

$$0.25 = 0.5e^{\frac{z_{1/4}}{\sqrt{2}}}$$

$$z_{1/4} = -\sqrt{2} \cdot \ln 2$$

Очевидно, что $z_{3/4} = -z_{1/4} = \sqrt{2} \cdot \ln 2$. Тогда $X_1^T = -4\sqrt{2}\ln 2$, $X_2^T = 4\sqrt{2}\ln 2$. Вычислим P_{sp}:

$$P_{sp} = F_L(X_1^T) + 1 - F_L(X_2^T) = 0.5e^{\frac{-4\sqrt{2}\ln 2}{\sqrt{2}}} + 0.5e^{\frac{-4\sqrt{2}\ln 2}{\sqrt{2}}} = 2^{-4} = 1/16 = 0.0625$$

Распределение Пуассона

При помощи написанной программы посчитаем значения функции распределения, меньшие 0.9:

k	1	2	3	4	5	6	7
F(k)	0.00005	0.00049	0.00277	0.01034	0.02925	0.06709	0.13014
k	8	9	10	11	12	13	14
F(k)	0.22022	0.33281	0.45793	0.58304	0.69678	0.79156	0.86446

Возьмем $z_{1/4} = 8.5$, $z_{3/4} = 12.5$. Тогда $X_1^T = 2.5$, $X_2^T = 18.5$. Тогда:

$$P_{sp} = F(3) + 1 - F(19) = 0.00277 + 1 - 0.99281 = 0.00996$$

Равномерное распределение

Теоретическая вероятность выбросов равна 0. Действительно, $z_{1/4} = a + \frac{b-a}{4}$, $z_{3/4} = a + 3\frac{b-a}{4}$. Значит:

$$X_1^T = a + \frac{b-a}{4} - \frac{3}{2} \cdot \frac{b-a}{2} < a,$$

$$X_2^T = a + 3\frac{b-a}{4} + \frac{3}{2} \cdot \frac{b-a}{2} > b.$$

То есть случайная величина не может оказаться вне $[X_1^T; X_2^T]$.