Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и
вычислительной физики, ИПММ

Направление подготовки
01.03.02 «Прикладная математика и информатика»

Отчет по лабораторным работам №1,2
по дисциплине «Математическая статистика»

Выполнил студент гр. 3630102/80201
Кирпиченко С. Р.
Руководитель
Баженов А. Н.

Санкт-Петербург
2021

Содержание

Страница

Список иллюстраций

Страница

Список таблиц

Страница

1 Постановка задачи

Для 5 распределений:

- Нормальное распределение $N(x, 0, 1)$
- Распределение Коши $C(x, 0, 1)$
- Распределение Лапласа $L(x, 0, \frac{1}{\sqrt{2}})$
- Распределение Пуассона $P(k, 10)$
- Равномерное распределение $U(x, -\sqrt{3}, \sqrt{3})$

Сгенерировать выборки размером 20 и 100 элементов. Построить для них боксплоты Тьюки. Для каждого распределения определить долю выбросов экспериментально (усредняя долю выбросов по 1000 выборок) и сравнить с результатами, полученными теоретически.

2 Теория

2.1 Определение

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей

2.2 Описание

Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

2.3 Построение

Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину «усов» определяют разность первого квартиля и полутора межквартильных расстояний и сумма третьего квартиля и полутора межквартильных расстояний. Формула имеет вид

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1) \tag{1}$$

где X_1 — нижняя граница уса, X_2 — верхняя граница уса, Q_1 — первый квартиль, Q_3 — третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков.

2.4 Теоретическая вероятность выбросов

Встроенными средствами языка программирования Python в среде разработки PyCharm можно вычислить теоретические первый и третий квартили распределений (Q_1^T и Q_3^T соответственно). По формуле (1) можно вычислить теоретические нижнюю и верхнюю границы уса (X_1^T и X_2^T соответственно). Выбросами считаются величины x, такие что:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix} \tag{2}$$

Теоретическая вероятность выбросов для непрерывных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T)) \tag{3}$$

где $F(X) = P(x \leq X)$ - функция распределения. Теоретическая вероятность выбросов для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > x_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T)) \tag{4}$$

где $F(X) = P(x \leq X)$ - функция распределения

3 Реализация

Лабораторная работа выполнена на языке Python 3.9 с использованием библиотек numpy, scipy, matplotlib, seaborn.

4 Результаты

4.1 Боксплот Тьюки

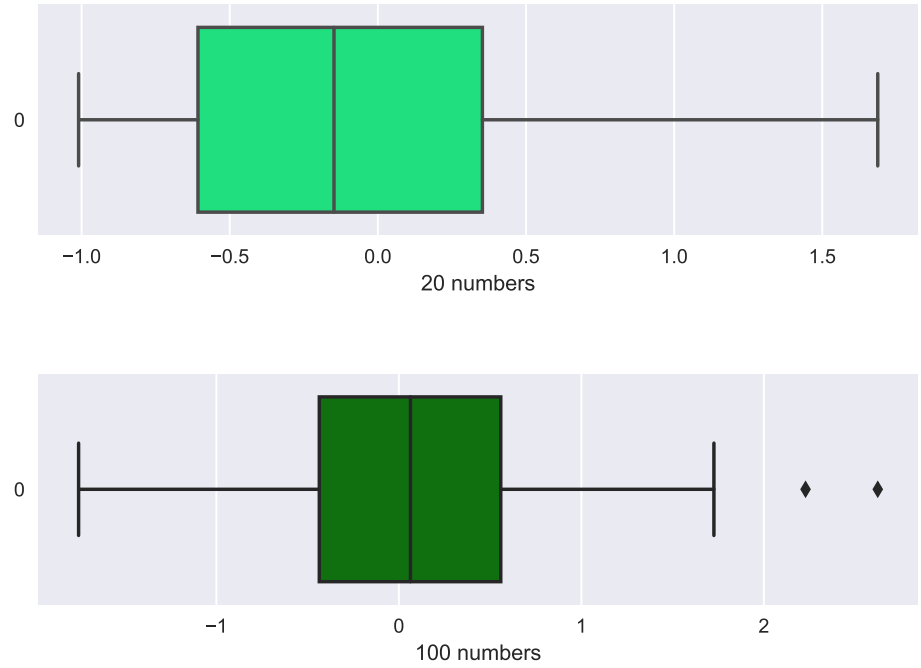

Рис. 1: Нормальное распределение

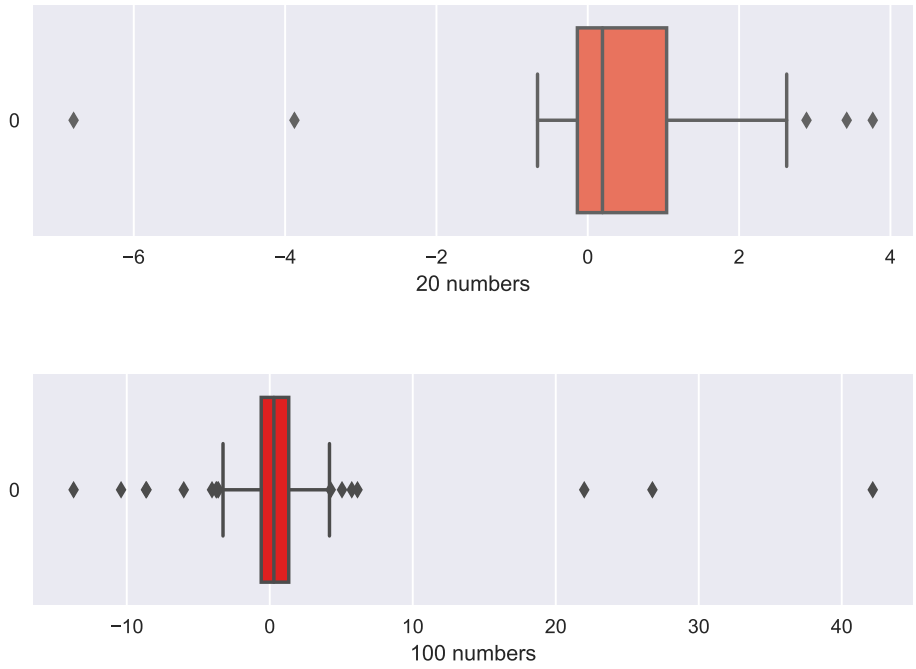

Рис. 2: Распределение Коши

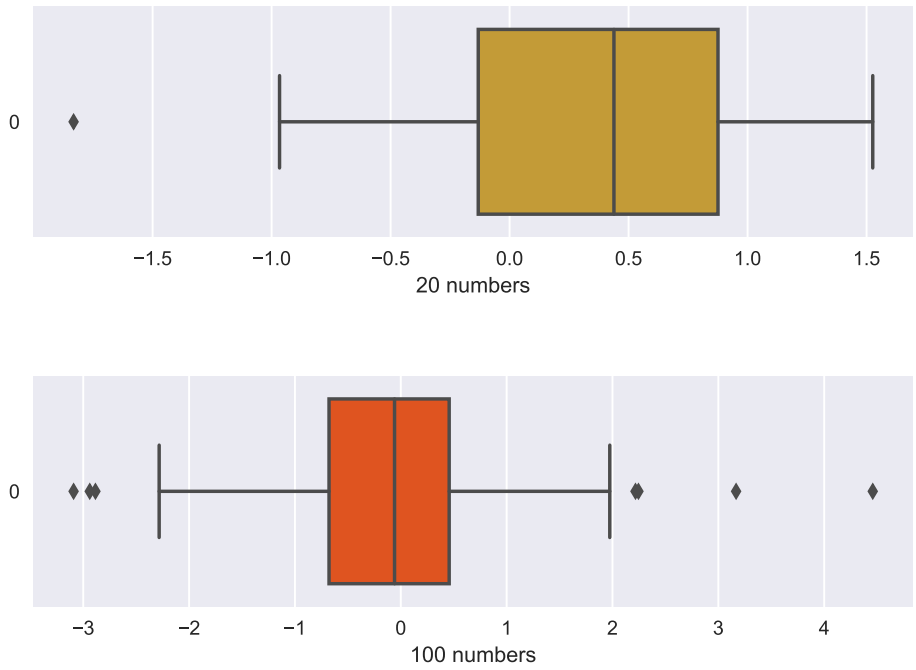

Рис. 3: Распределение Лапласа

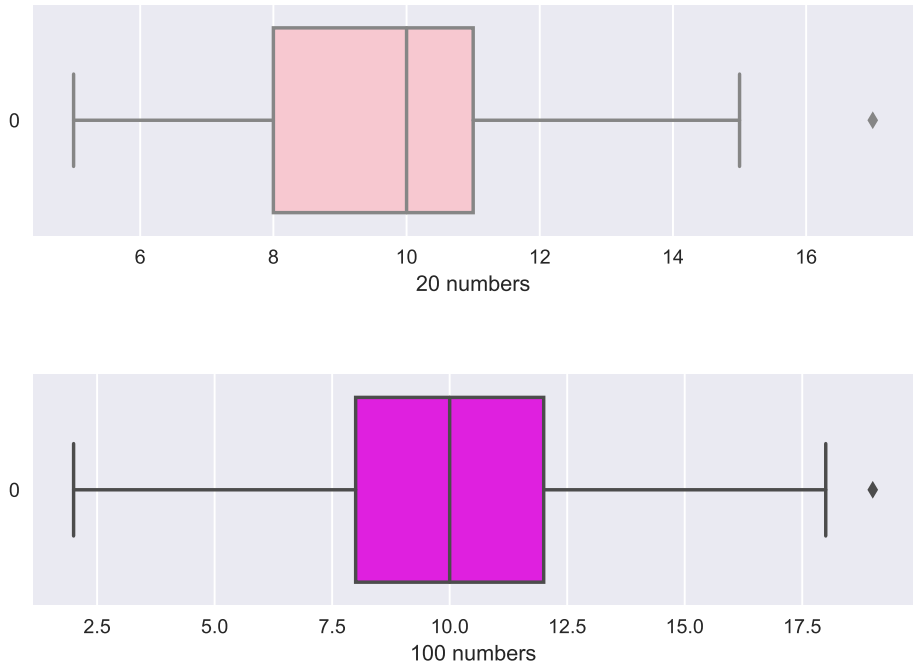

Рис. 4: Распределение Пуассона

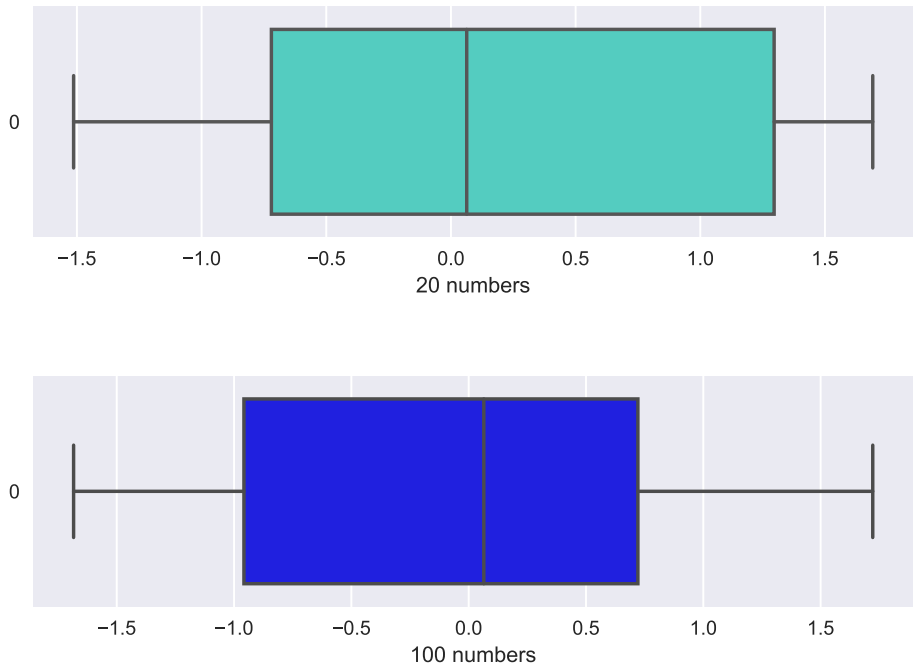

Рис. 5: Равномерное распределение

4.2 Доля выбросов

Округление доли выбросов:

Выборка случайна, поэтому в качестве оценки рассеяния можно взять дисперсию пуассоновского потока: $D_n \approx \sqrt{n}$

Доля $p_n = \frac{D_n}{n} = \frac{1}{\sqrt{n}}$

Доля $n = 20 : p_n = \frac{1}{\sqrt{20}}$ - примерно 0.2 или 20%

Для $n = 100 : p_n = \frac{1}{\sqrt{100}}$ - примерно 0.1 или 10%

Исходя из этого можно решить, сколько знаков оставлять в доле выброса.

Распределение	Размер выборки	Доля выбросов
Нормальное	20	0.0198
Нормальное	100	0.00919
Коши	20	0.1488
Коши	100	0.1521
Лапласа	20	0.0652
Лапласа	100	0.06295
Пуассона	20	0.0265
Пуассона	100	0.01527
Равномерное	20	0
Равномерное	100	0

Таблица 1: Теоретическая вероятность выбросов

4.3 Теоретическая вероятность выбросов

Распределение	Q_1^T	Q_3^T	X_1^T	X_2^T	P_B^T
Нормальное	-0.674	0.674	-2.698	2.698	0.007
Коши	-1	1	-4	4	0.156
Лапласа	-0.490	0.490	-1.961	1.961	0.063
Пуассона	8	12	2	18	0.008
Равномерное	-0.866	0.866	-3.464	3.464	0

Таблица 2: Доля выбросов

5 Обсуждение

5.1 Доля и теоретическая вероятность выбросов

По данным, приведенных в таблицах,можно сделать вывод, что увеличение выборки ведет к приближению доли выбросов к теоретической оценке. Доля выбросов для распределения Коши значительно больше, чем для остальных распределений. В равномерном распределении выбросы отсутствуют.
Боксплоты Тьюки веьма наглядно визуализируют характеристики выборок, проводить анализ по ним намного проще, чем по табличным данным.

Примечание

С исходным кодом работы и данного отчета можно ознакомиться в репозитории `https://github.com/Stasychbr/MatStat`