Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №3

по дисциплине
«Математическая статистика»

Выполнил студент
группы 3630102/80401 Веденичев Дмитрий Александрович

Проверил
Доцент, к.ф.-м.н. Баженов Александр Николаевич

Санкт-Петербург
2021

Содержание

Список иллюстраций

1 Постановка задачи

Для 5 распределений:

1. $N(x, 0, 1)$ – нормальное распределение
2. $C(x, 0, 1)$ – распределение Коши
3. $L(x, 0, \frac{1}{\sqrt{2}})$ – распределение Лапласа
4. $P(k, 10)$ – распределение Пуассона
5. $U(x, -\sqrt{3}, \sqrt{3})$ – расномерное распределение

Сгенерировать выборки размером 20 и 100 элементов. Построить для них боксплот Тьюки. Для каждого распределения определить долю выбросов экспериментально (сгенерировав выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.

2 Теория

2.1 Боксплот Тьюки

2.1.1 Определение

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

2.1.2 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.

2.1.3 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.

2.2 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:

$$\left[\begin{array}{l} x < X_1^T \\ x > X_2^T \end{array}\right. \tag{2}$$

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 Программная реализация

Лабораторная работа выполнена на языке Python вресии 3.7 в среде разработки PyCharm. Использовались дополнительные библиотеки:

1. scipy
2. seaborn
3. matplotlib
4. math

В приложении находится ссылка на GitHub репозиторий с исходныи кодом.

4 Результаты

4.1 Боксплот Тьюки

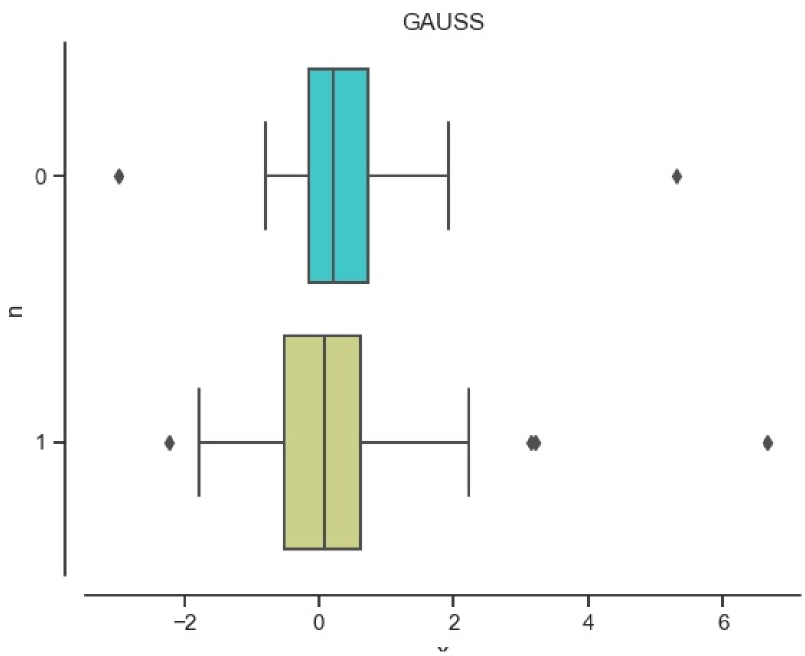

Рис. 1: Нормальное распределение

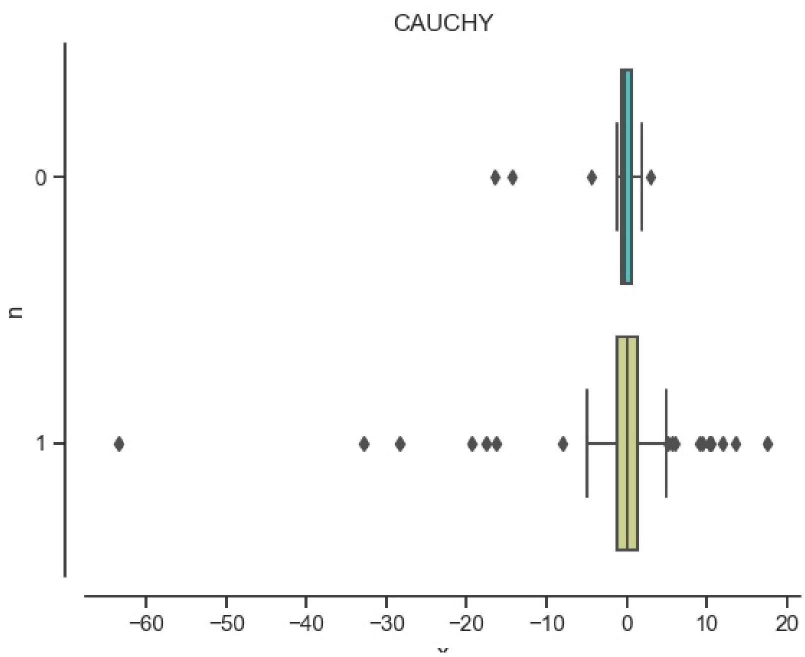

Рис. 2: распределение Коши

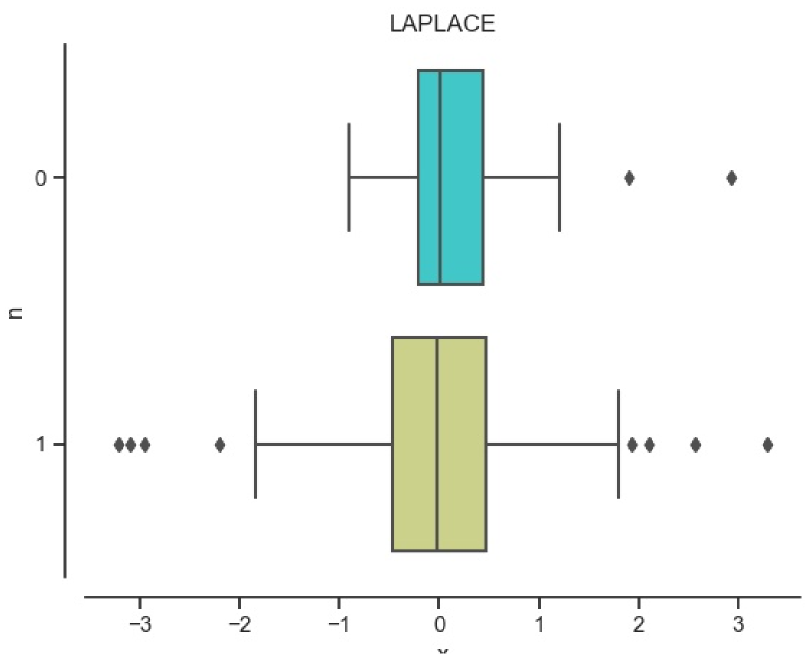

Рис. 3: распределение Лапласа

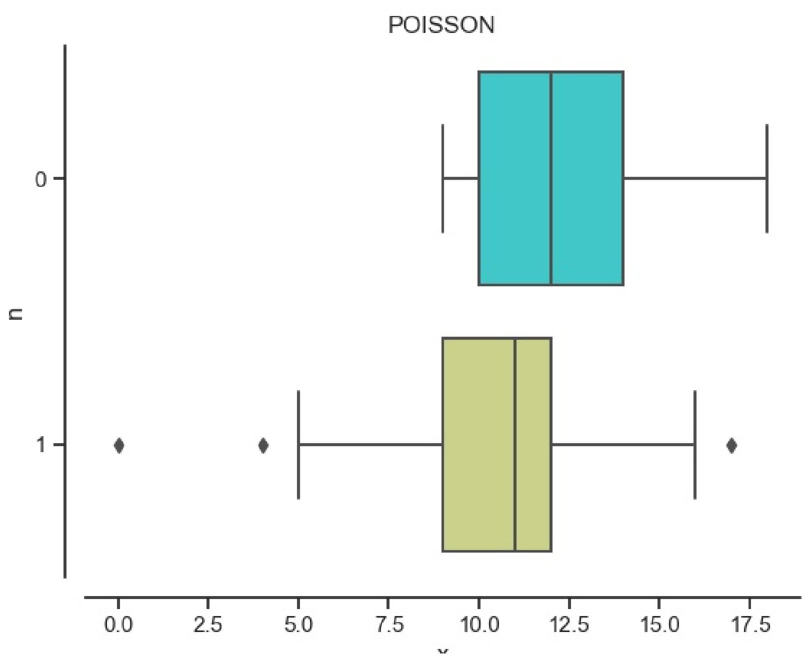

Рис. 4: распределение Пуассона

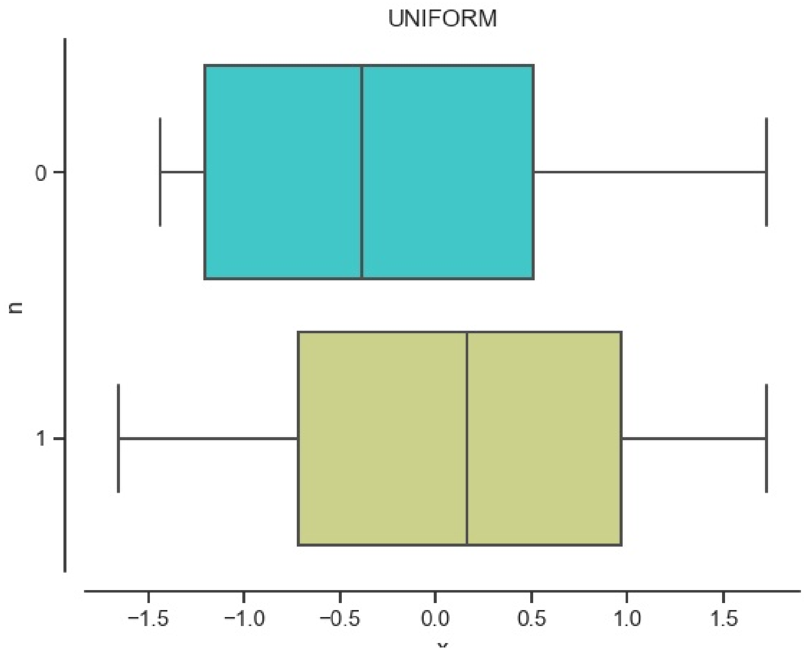

Рис. 5: равномерное распределение

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.

Выборка	Доля выбросов	P_B^T
Normal n=20	0.058	0.007
Normal n=100	0.057	0.007
Cauchy n=20	0.149	0.156
Cauchy n=100	0.184	0.156
Laplace n=20	0.078	0.063
Laplace n=100	0.081	0.063
Poisson n=20	0.023	0.008
Poisson n=100	0.015	0.008
Uniform n=20	0.002	0
Uniform n=100	0.0	0

Таблица 1: Доля выбросов

4.3 Теоретическая вероятность выбросов

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

Таблица 2: Теоретическая вероятность выбросов

5 Обсуждение

По данным, приведенным в таблице, можно сказать, что чем больше выборка, тем ближе доля выбросов будет к теоретической оценке. Однако в каждом правиле есть исключение, у нас им в очередной раз является распределение Коши. Доля выбросов у него не только не уменьшилась, но возросла. Хотя справедливо будет заметить, что тоже самое случилось с Лапласом. Тем не менее, Коши продолжает выделятся за счет повышенной доли выбросов по сравнению с другими распределениями.
Равномерное распределение не показало никаких выбросов. Для равномерного, Лапласа и Пуассона погрешность при выборке n=100 составила не более 2 %.

Боксплоты Тьюки в удобной форме показывает многие важные характеристики выборки, такие как медиана, первый и третий квартили и другие. Так, исходя из полученных рисунков, наглядно видно то, что мы довольно трудоёмко анализировали в предыдущих частях.

6 Приложение

Код программы GitHub URL:
https://github.com/PopeyeTheSailorsCat/math_stat_2021/blob/main/lab3/src/lab3.py