Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

ОТЧЕТ

Тема: ***Выявление выбросов***

Направление: 01.03.02 Прикладная математика и информатика

Выполнил студент гр. 33631/4 Камалетдинова Ю.

Преподаватель Баженов А.

Санкт-Петербург
2019

Содержание

Постановка задачи

Задачей данной работы является рассмотрение такого способа выявления выбросов как боксплот. Требуется сгенерировать выборки размерами $n = 20, 100$ и построить для них боксплот Тьюки. Для каждого распределения нужно определить процент выбросов экспериментально, сгенерировав выборку из распределения $N = 1000$ раз и вычислив средний процент выбросов, а затем сравнить с результатами, полученными теоретически. Рассматриваемые законы распределения приведены ниже

$$N(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \text{ – стандартное нормальное} \tag{1}$$

$$C(x, 0, 1) = \frac{1}{\pi(1 + x^2)} \text{ – Коши} \tag{2}$$

$$L(x, 0, \frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \text{ – Лаплас} \tag{3}$$

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}}, & |x| \le \sqrt{3} \\ 0, & |x| > \sqrt{3} \end{cases} \text{ – равномерное} \tag{4}$$

$$P(\lambda) = \frac{e^{-\lambda}}{k!} \lambda^k, \; \lambda = 7 \text{ – Пуассон} \tag{5}$$

Описание метода

Выброс – это некое наблюдение, которое нехарактерно далеко удалено от других значений в общей случайной выборке. В некотором смысле, это определение оставляет на усмотрение наблюдателя решение вопроса о том, что будет считаться выбросом.

Боксплот является удобным графическим способом описания поведения данных как в середине, так на концах распределения. Боксплот визуализирует положение медианы, нижнего и верхнего квартилей. Первый квартиль Q_1 определяется как медиана части выборки до медианного элемента всей выборки, третий квартиль Q_3 – медиана части выборки после медианного элемента всей выборки.

На графике присутствуют усы, границы которых определяются формулами

$$x_L = max(x_{(1)},\ Q_1 - 1.5 \cdot IQR) \text{ – нижняя граница уса} \tag{6}$$

$$x_U = min(x_{(n)},\ Q_3 + 1.5 \cdot IQR) \text{ – верхняя граница уса} \tag{7}$$

$$IQR = Q_3 - Q_1 \text{ – интерквартильная широта} \tag{8}$$

Будем считать элемент x_i выбросом, если $x_i \notin [x_L, x_U]$

Для сравнения теоретических и практических результатов вычислим квартили для непрерывных распределений. Квартили однозначно определяются уравнением

$$F_X(x_\alpha) = \alpha,\ \alpha = \left\{\frac{1}{4},\ \frac{3}{4}\right\} \tag{9}$$

Искомые квартили можно выразить, найдя обратную функцию к функции распределения.

Вычисление доли выбросов будем производить по формуле

$$p_{outliers} = \frac{\sum_i x_i}{\sum_{j=1}^{n} x_j},\ i : x_i \notin [x_L, x_U] \tag{10}$$

Реализация

Для выполнения поставленной задачи будем пользоваться библиотеками для языка Python: *numpy, scipy* – расчеты, законы распределения вероятностей; *matplotlib, seaborn* – визуализация результатов. Ход работы:

- Задаем распределение с заданными параметрами
- Генерируем случайные выборки из распределений размерами $n = 20, 100$
- Для каждого из распределений вычисляем доли выбросов $N = 1000$ раз
- Вычисляем теоретические квартили при помощи метода *.ppf(α)* (percent point function) - функция, обратная функции распределения, вычисляем доли выбросов с использованием данных квартилей
- Усредняем полученные суммы долей выбросов, разделив на $N = 1000$

Результат

Нормальное распределение с параметрами 0, 1

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.0222	0.0070
$n = 100$	0.0099	0.0070

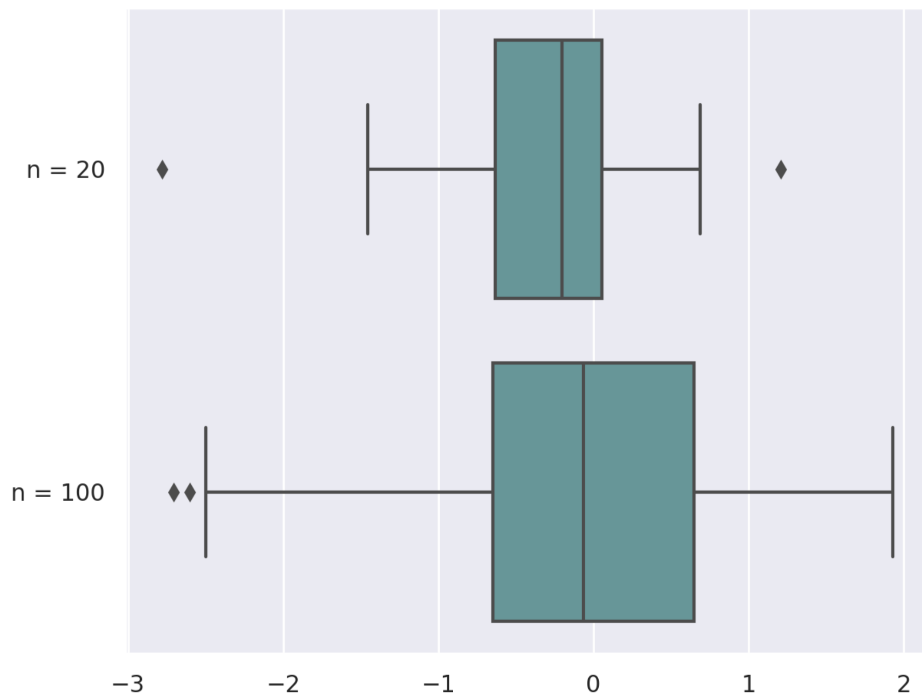

Рис. 1: Боксплот Тьюки для выборок из нормального распределения

Равномерное распределение на отрезке $[-\sqrt{3}, \sqrt{3}]$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.0027	0.0000
$n = 100$	0.0000	0.0000

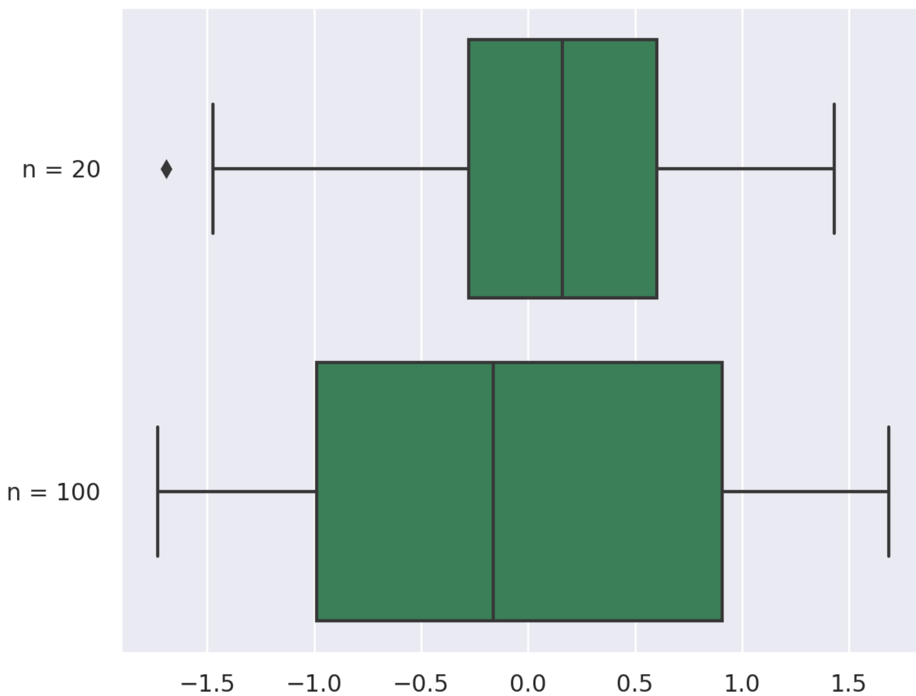

Рис. 2: Боксплот Тьюки для выборок из равномерного распределения

Распределение Коши с параметрами 0, 1

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.1525	0.1574
$n = 100$	0.1558	0.1562

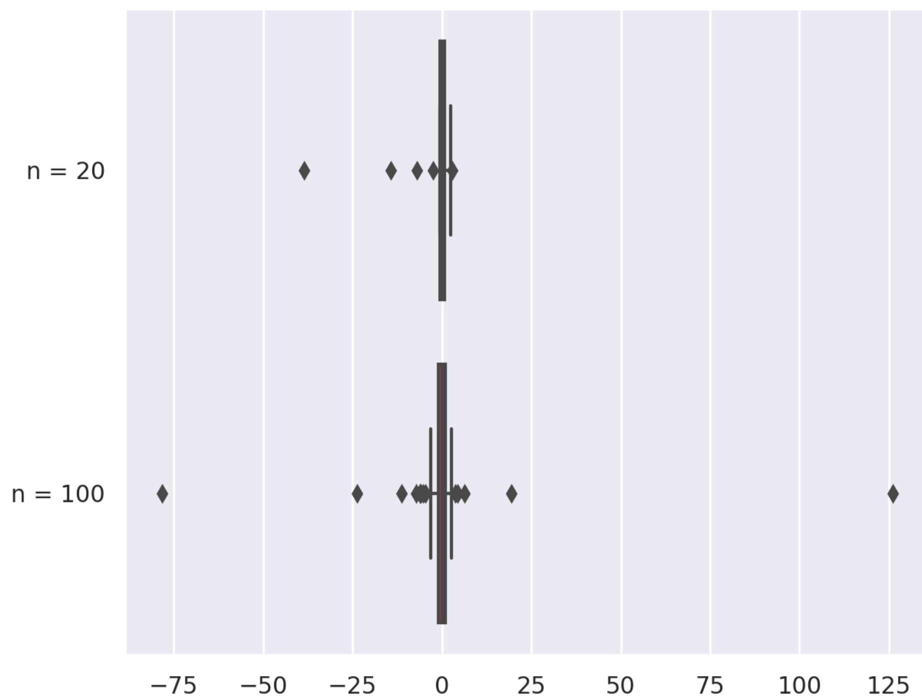

Рис. 3: Боксплот Тьюки для выборок из распределения Коши

Распределение Лапласа с параметрами 0, $\frac{1}{\sqrt{2}}$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.0749	0.0608
$n = 100$	0.0642	0.0610

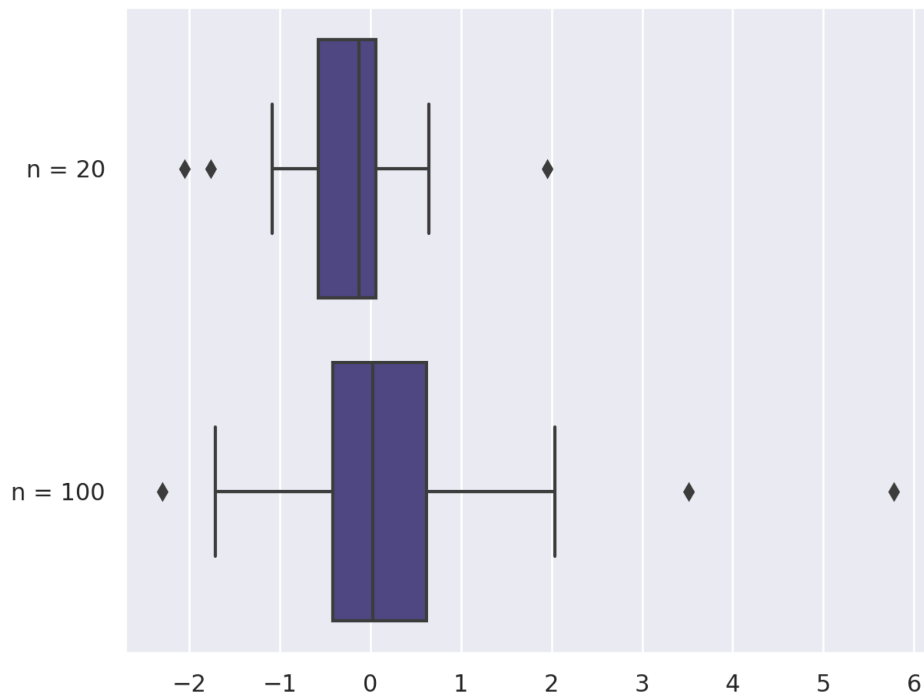

Рис. 4: Боксплот Тьюки для выборок из распределения Лапласа

Распределение Пуассона с параметром $\lambda = 7$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.0278	0.0027
$n = 100$	0.0119	0.0024

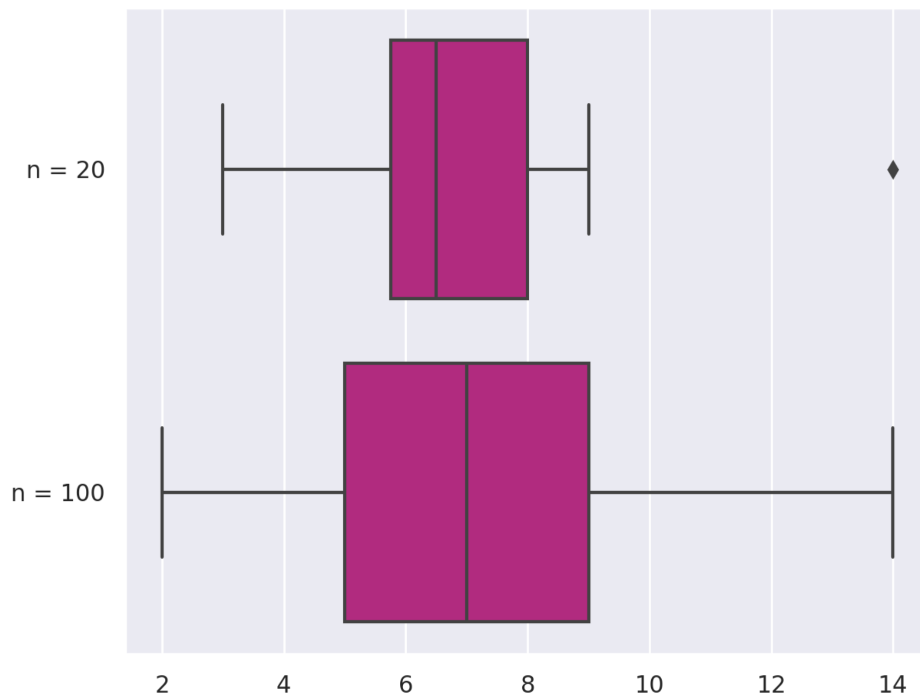

Рис. 5: Боксплот Тьюки для выборок из распределения Пуассона

Взглянув на полученные результаты, можно сделать вывод о том, что равномерное распределение не имеет выбросов. Ненулевое значение в малой выборке при расчете квантилей с использованием значений выборки есть показатель того, что нельзя судить о характере распределения по выборке размером $n = 20$ окончательно. Отсутствие выбросов объясняется тем, что функция плотности такого респределения постоянна, и элемент выборки никогда будет вне отрезка, ограниченного (6), (7) Наибольший процент выбросов установлен для выборов из распределения Коши. Также увеличение выборки не влияет на изменение результата, математического ожидания нет. Обратимся к результатам предыдущей лабораторной работы. Дисперсия размаха распределения на выборке $n = 100$ уже принимает значения порядка 10^8. Это объясняет тот факт, что доля выбросов достаточно высока. Распределение Коши имеет тяжелые хвосты, и доля элементов, принимающих далекие от центра распределения значения, на порядок выше по сравнению с остальными распределениями, такими как нормальное, Лапласа, Пуассона.

Список литературы

[1] *Chrisman, L.* How the strange Cauchy distribution proved useful. *Lumina Decision Systems* (2018). URL: http://www.lumina.com/blog/how-the-strange-cauchy-distribution-proved-useful