Санкт-Петербургский Политехнический
Университет им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

Отчёт по лабораторным работам №1 – 4
Исследование распределений

3 курс, группа 33631/2

Студент Д. А. Плаксин

Преподаватель Баженов А. Н.

Санкт-Петербург
2019 г.

Содержание

1 Список иллюстраций

2 Список таблиц

3 Постановка задачи

В данных работах надо было сгенерировать выборки разных мощностей для пяти распределений: Распределения [13]:

1. Стандартное нормальное распределение
2. Стандартное распределение Коши
3. Распределение Лапласа с коэффициентом масштаба $\sqrt{2}$ и нулевым коэффициентом сдвига.
4. Равномерное распределение на отрезке $\left[-\sqrt{3}, \sqrt{3}\right]$
5. Распределение Пуассона со значением матожидания равным двум.

Для этих выборок надо было:

1. Построить гистограммы и графики плотности распределения.
2. Вычислить характеристики положения. Вычисления необходимо повторить 1000 раз для каждой выборки, найти среднее характеристик положения и их квадратов, вычислить оценку дисперсии, полученные данные представить в виде таблиц.
3. Построить боксплот Тьюки, затем для каждого распределения определить процент выбросов (вычислив средний процент выбросов из 1000 экспериментов) и сравнить полученные результаты с теоретическими.
4. Построить эмпирические функции распределения и ядерные оценки плотности распределения на отрезке $[-4; 4]$.

4 Теория

4.1 Плотности распределения вероятностей

Плотностью распределения вероятностей непрерывной случайной величины называется первая производная от функции распределения. Смысл плотности распределения в том, что она показывает, как часто появляется случайная величина в некоторой окрестности точки при повторении опытов.

Для рассматриваемых в этой работе пяти распределений известны плотности их вероятности:

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

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

$$\text{Распределение Лапласа } L\left(x, 0, \frac{1}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \tag{3}$$

$$\text{Распределение Пуассона } P(\lambda, k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{4}$$

$$\text{Равномерное распределение } M(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & |x| \leqslant \sqrt{3} \\ 0 & |x| > \sqrt{3} \end{cases} \tag{5}$$

4.2 Характеристики положения

Характеристики положения дают информацию о расположении значений выборки на числовой прямой и характеризуют этот признак с точки зрения некоторого среднего значения.

Для рассматриваемых в этой работе характеристик положения данных известны формулы для их вычисления:

1. Выборочное среднее [3]:
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

2. Выборочная медиана [8]:
$$med\ x = \begin{cases} x_{k+1}, & n = 2k+1 \\ \frac{1}{2}\left(x_k + x_{k+1}\right), & n = 2k \end{cases} \tag{7}$$

3. Полусумма экстремальных значений [5]:
$$Z_R = \frac{1}{2}\left(x_1 + x_n\right) \tag{8}$$

4. Полусумма квартилей [6]:
$$Z_Q = \frac{1}{2}\left(Z_{\frac{1}{4}} + Z_{\frac{3}{4}}\right) \tag{9}$$

5. Усечённое среднее [7]:
$$Z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_i \tag{10}$$

4.3 Боксплот Тьюки

Боксплот Тьюки - график, использующийся в описательной статистике, изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.

Выбросом в статистике называют результат измерения, выделяющийся из общей выборки.

4.4 Эмпирические функции и ядерные оценки

Эмпирическая функция распределения [14], построенная по выборке $X = (X_1, \ldots, X_n)$ есть случайная функция $F_n(y)$, определённая на $\mathbb{R}$:

$$F_n(y) = \sum_{i=1}^{n} I\left(X_i < y\right) \quad \text{где } I\left(X_i < y\right) = \begin{cases} 1, & X_i < y \\ 0, & \text{иначе} \end{cases} \tag{11}$$

$X = (X_1, \ldots, X_n)$ есть одномерная выборка одинаково распределённых элементов, с плотностью распределения f.

Ядерная оценка плотности [15]:

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{12}$$

где K является ядром, а $h > 0$ является сглаживающим параметром, и называется шириной полосы.

В данной работе в качестве ядра была выбрана плотность вероятности стандартного нормального распределения [16]:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{13}$$

5 Реализация

Для генерации выборки был использован *Python* 3.7: модуль *random* библиотеки *numpy* [1] для генерации случайных чисел с различными распределениями.

Обработка функций распределений была сделана с помощью модуля *scipy* [12].

После вычисления характеристик положения 1000 раз среднее значение и дисперсия находились по следующей формуле:

$$D(z) = E\left(z^2\right) - E^2(z), \text{ где } E(z) = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{14}$$

Боксплот Тьюки был построен средствами библиотеки matplotlib [11].

Правая и левые границы: $R - LQ - l(UQ - LQ)$, $L = UQ - k(UQ - LQ)$, где k обычно полагают равным 1.5 [10]

Число выбросов определялось таким образом: если значение из выборки находится вне установленных левой и правых границ, то оно является выбросом.

6 Результаты

6.1 Плотности распределения вероятностей

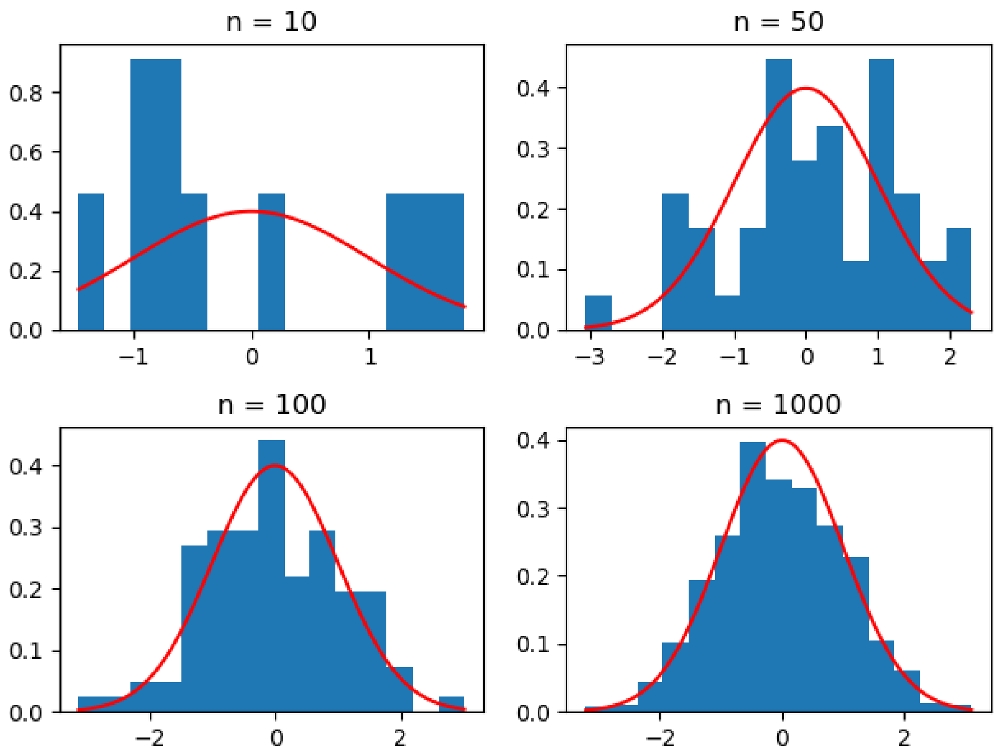


Рис. 1: Нормальное распределение (1)

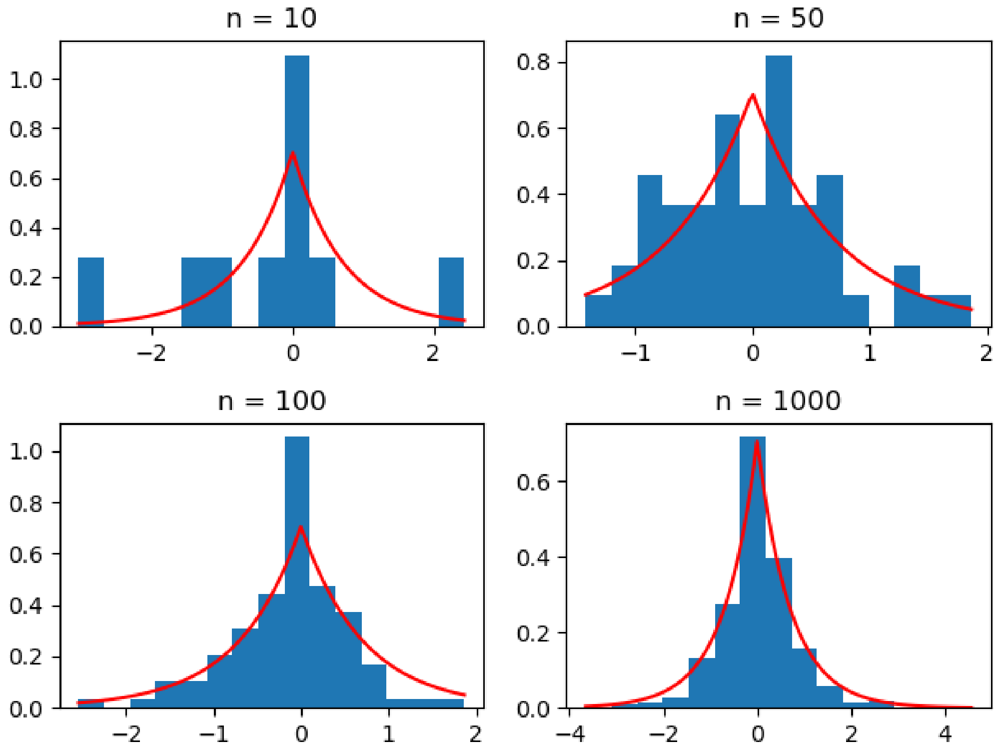


Рис. 2: Распределение Лапласа (3)

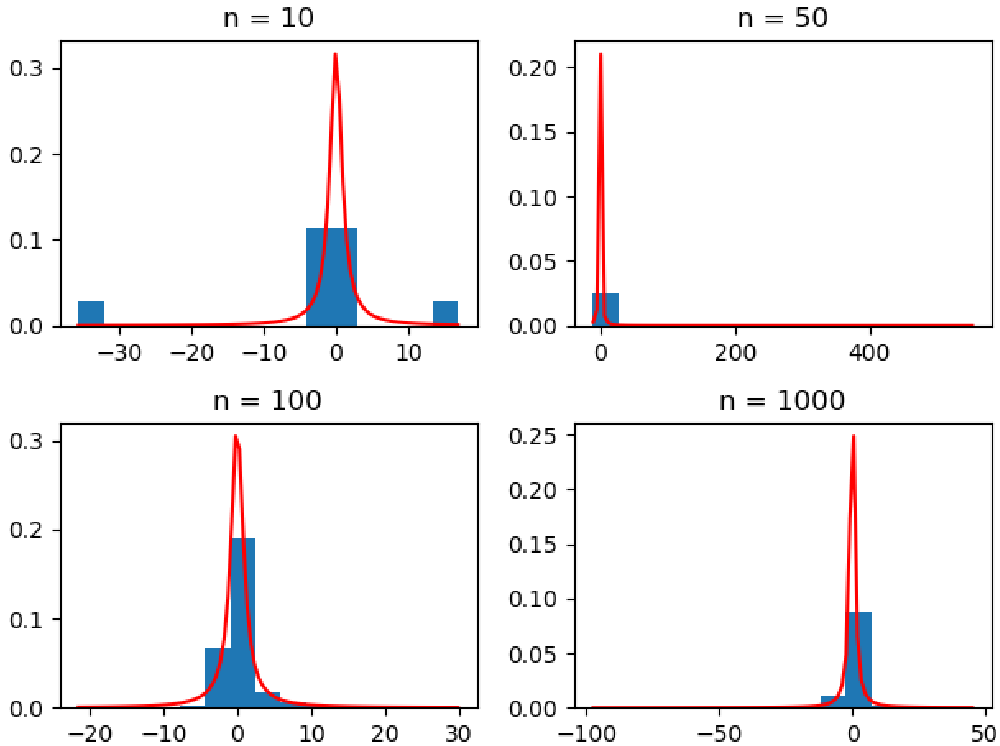


Рис. 3: Распределение Коши (2)

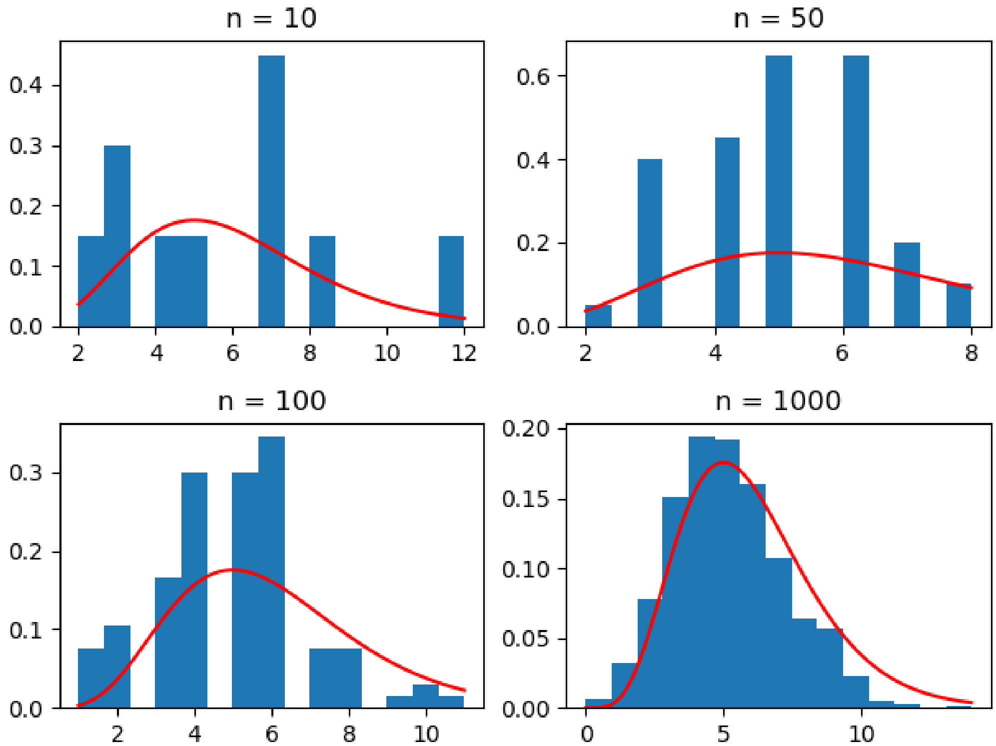


Рис. 4: Распределение Пуассона (4)

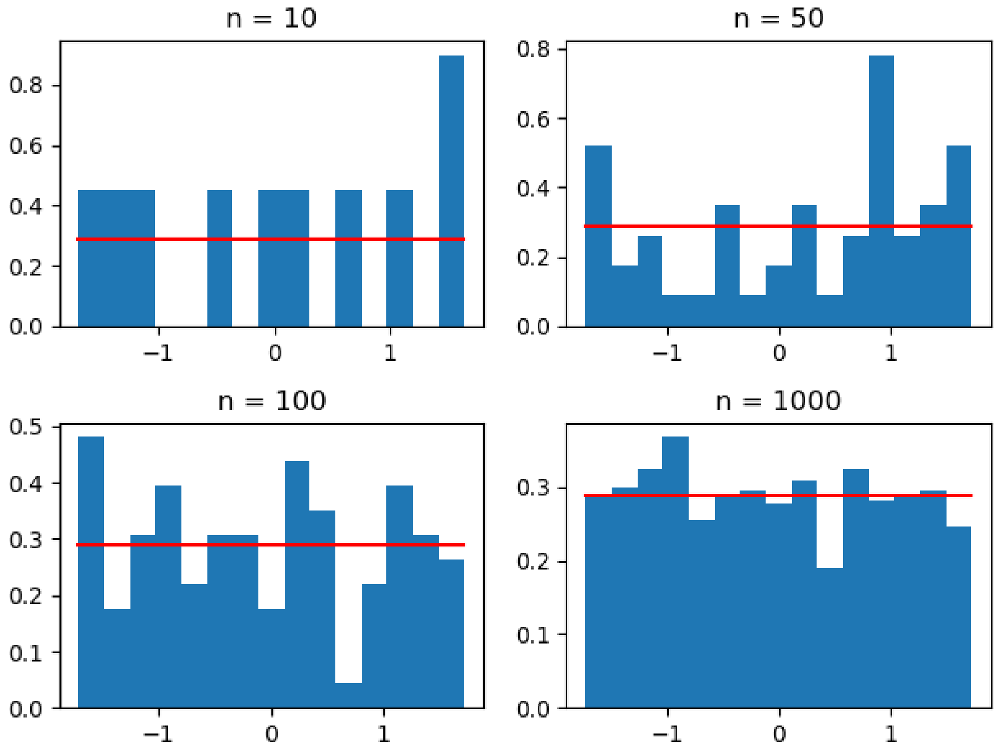


Рис. 5: Равномерное распределение (5)

6.2 Характеристики положения

Таблица 1: Стандартное нормальное распределение (1).

$n = 20$	average	med	Z_R	Z_Q	Z_{tr}, $r = \frac{n}{4}$
$E =$	−0.00	−0.00	0.0	0.00	−0.01
$D =$	0.051763	0.072023	0.149073	0.058764	0.059323
$n = 60$	average	med	Z_R	Z_Q	Z_{tr}, $r = \frac{n}{4}$
$E =$	0.00	0.00	−0.0	−0.00	−0.00
$D =$	0.016322	0.026125	0.110309	0.020489	0.020389
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}, $r = \frac{n}{4}$
$E =$	0.000	−0.00	−0.00	−0.00	−0.00
$D =$	0.009563	0.015479	0.096256	0.012569	0.012224

Таблица 2: Стандартное распределение Коши (2).

$n = 20$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	3.414266	−0.012522	−5.262327	−0.027313	0.020763
$D =$	12977.113689	0.130212	12626.531931	0.371684	0.150542
$n = 60$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	−1.180135	−0.008143	−69.304468	0.007723	−0.006201
$D =$	1424.446598	0.038886	6240734.028690	0.086309	0.042207
$n = 100$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.811064	−0.003271	30.434787	0.002690	−0.008544
$D =$	350.070803	0.023004	1465401.042218	0.054476	0.025649

Таблица 3: Распределение Лапласа (3).

$n = 20$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	−0.01	−0.01	−0.0	0.00	−0.00
$D =$	0.047943	0.031492	0.436478	0.047849	0.031600
$n = 60$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.00	−0.00	−0.0	−0.00	−0.00
$D =$	0.017707	0.010050	0.455793	0.014958	0.010114
$n = 100$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.005	0.004	0.0	0.00	−0.000
$D =$	0.009733	0.006307	0.409248	0.010233	0.006262

Таблица 4: Равномерное распределение (5).

$n = 20$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	−0.00	−0.0	0.00	−0.00	0.00
$D =$	0.046929	0.126544	0.013805	0.070389	0.098176
$n = 60$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.00	−0.00	−0.000	0.00	0.00
$D =$	0.018149	0.048531	0.001583	0.023893	0.032310
$n = 100$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.00	0.00	0.0010	0.00	0.00
$D =$	0.010555	0.029134	0.000562	0.014285	0.020388

Таблица 5: Распределение Пуассона (4).

$n = 20$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	1.99	1.8	2.5	1.9	1.8
$D =$	0.096696	0.201928	0.325744	0.126251	0.125845
$n = 60$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	1.99	1.92	2.9	1.94	1.84
$D =$	0.032178	0.063700	0.245694	0.033469	0.046742
$n = 100$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	1.99	1.96	3.1	1.96	1.83
$D =$	0.020435	0.030204	0.239758	0.018016	0.026728

6.3 Боксплот Тьюки

Рис. 6: Boxplot нормальное распределение (1)

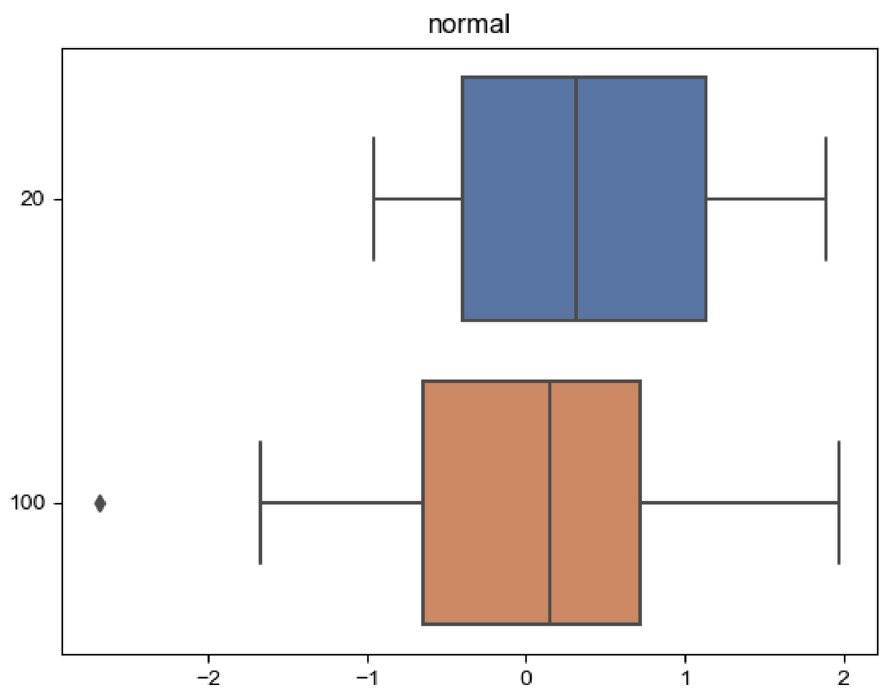


Рис. 7: Boxplot стандартное распределение Лапласа (3)

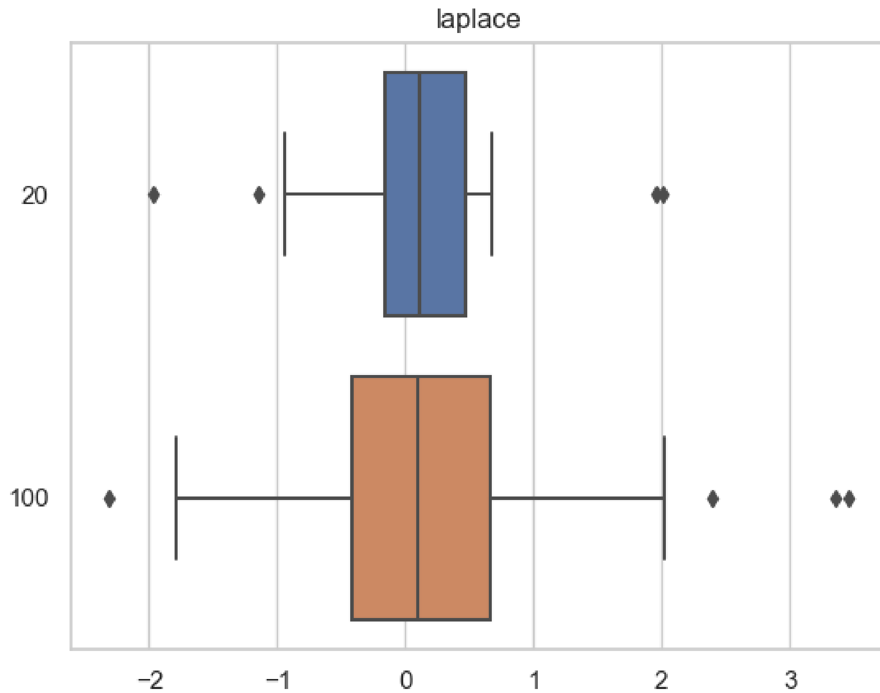


Рис. 8: Boxplot стандартное распределение Коши (2)

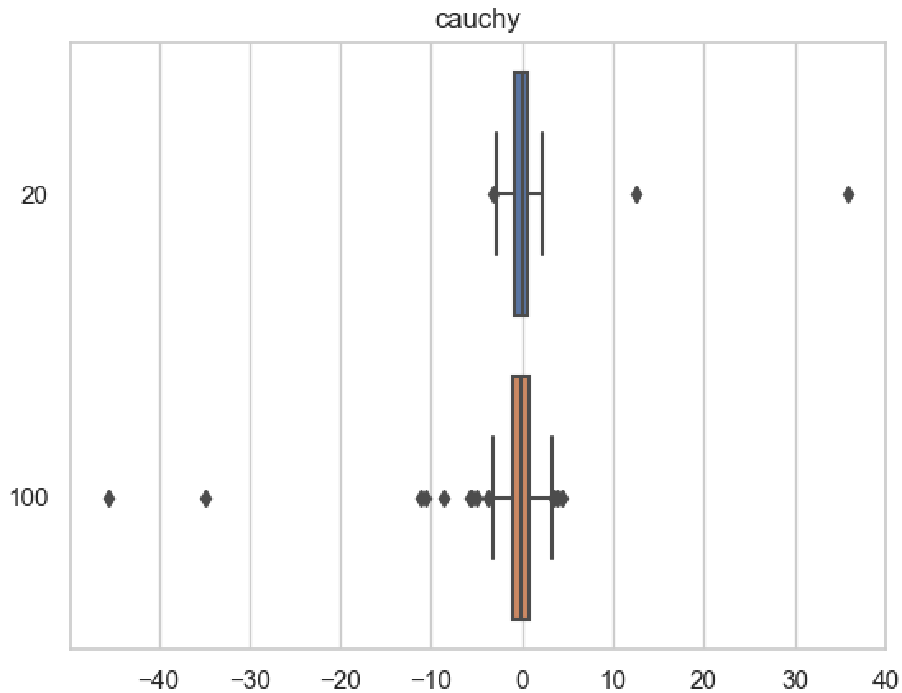


Рис. 9: Boxplot распределение Пуассона (4)

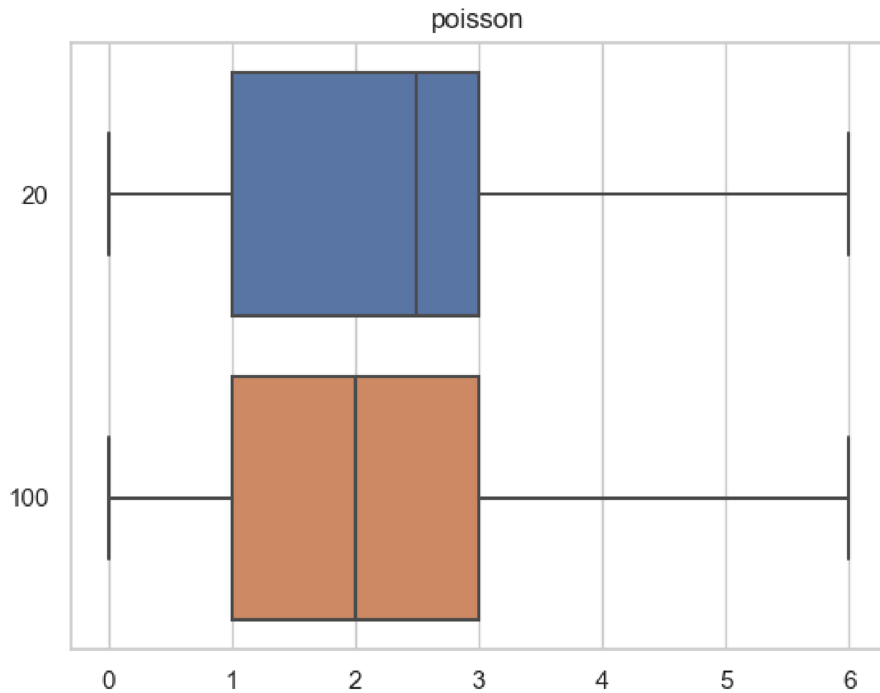


Рис. 10: Boxplot равномерное распределение (5)

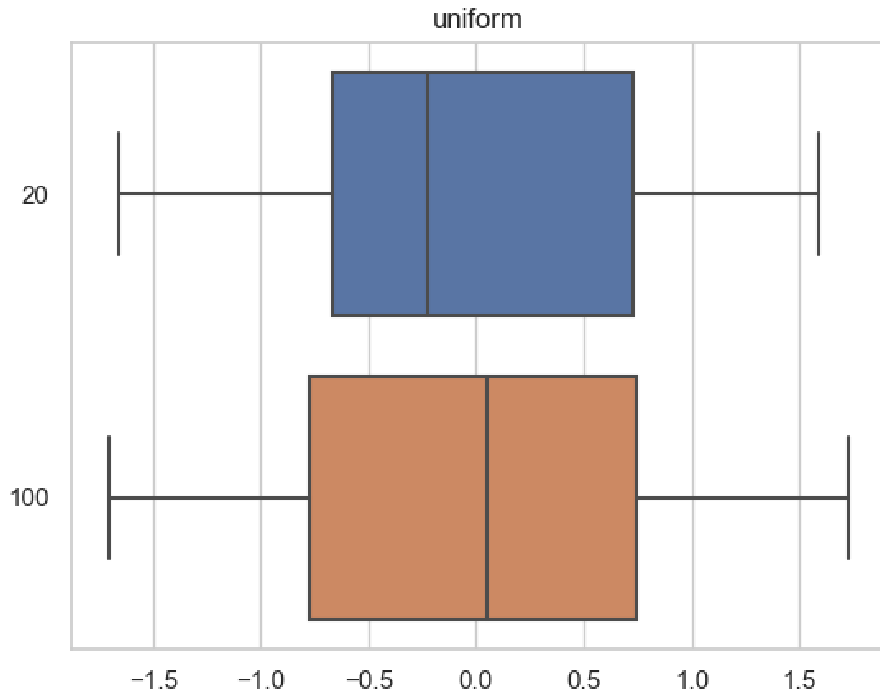


Таблица 6: Зависимость выбросов от размера выборки

Выборка	Процент выбросов
normal	
n = 20	2
n = 100	1
cauchy	
n = 20	15
n = 100	16
laplace	
n = 20	7
n = 100	6
uniform	
n = 20	0
n = 100	0
poisson	
n = 20	3
n = 100	0

6.4 Ядерная оценка плотности

Рис. 11: Эмпирическая функция для нормального стандартного распределения (1)

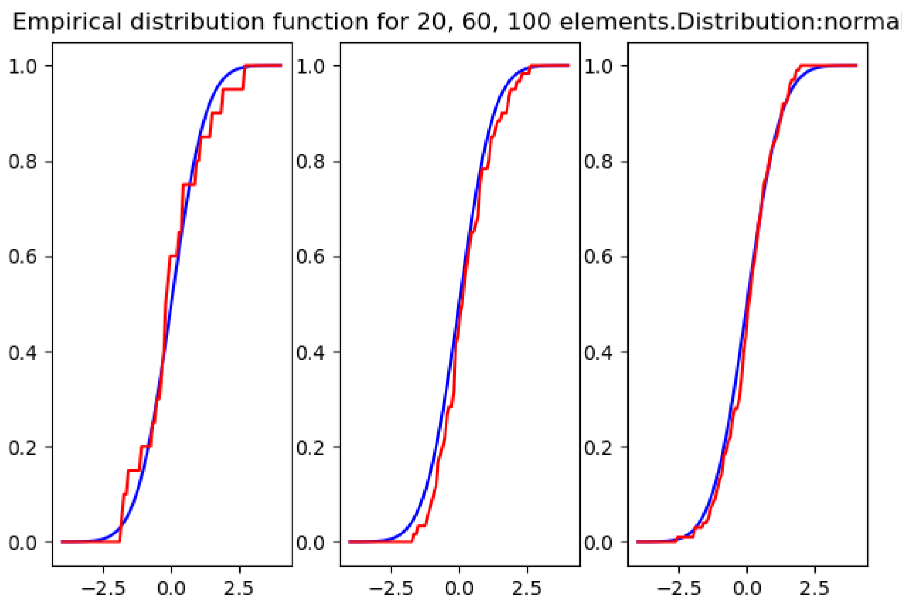


Рис. 12: Эмпирическая функция для стандартного распределения Лапласа (3)

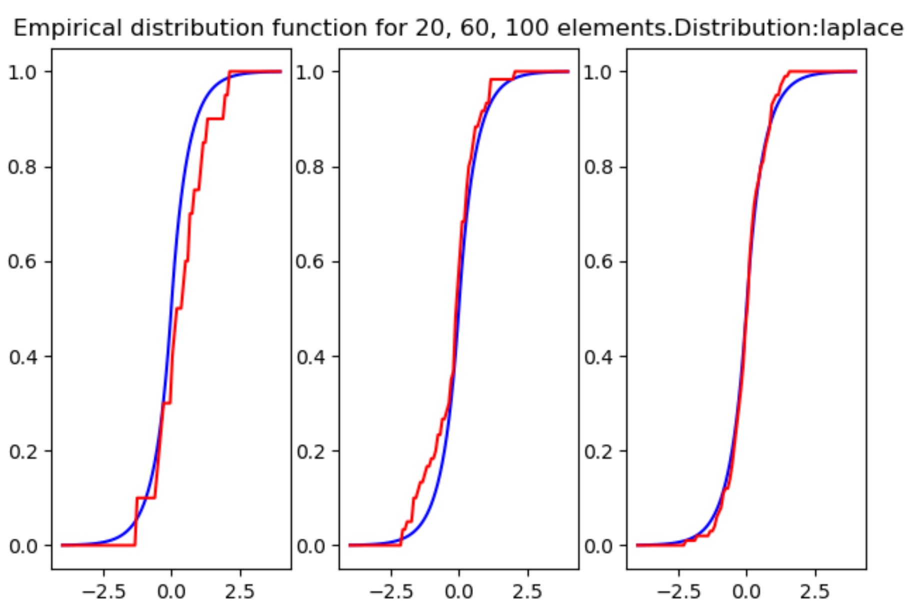


Рис. 13: Эмпирическая функция для стандартного распределения Коши (2)

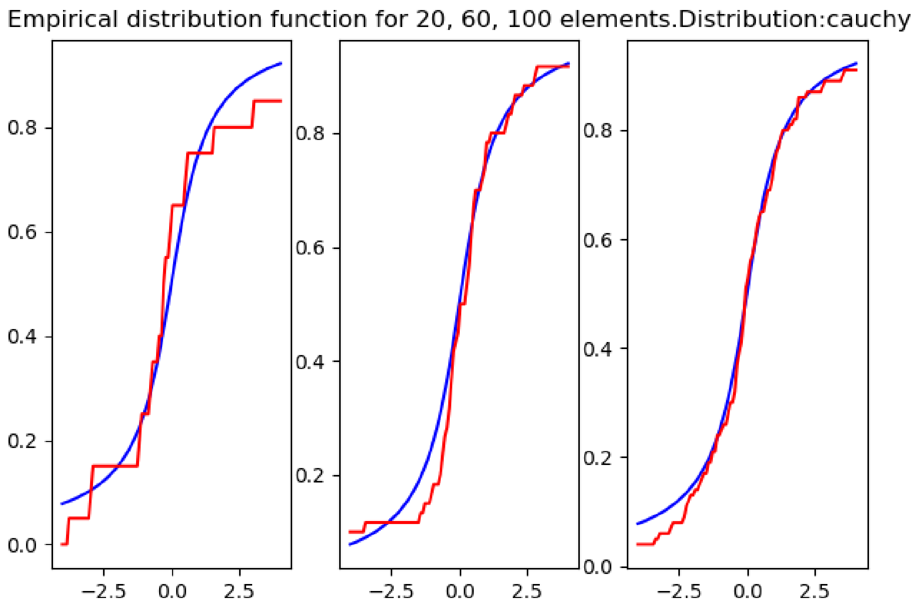


Рис. 14: Эмпирическая функция для распределения Пуассона (4)

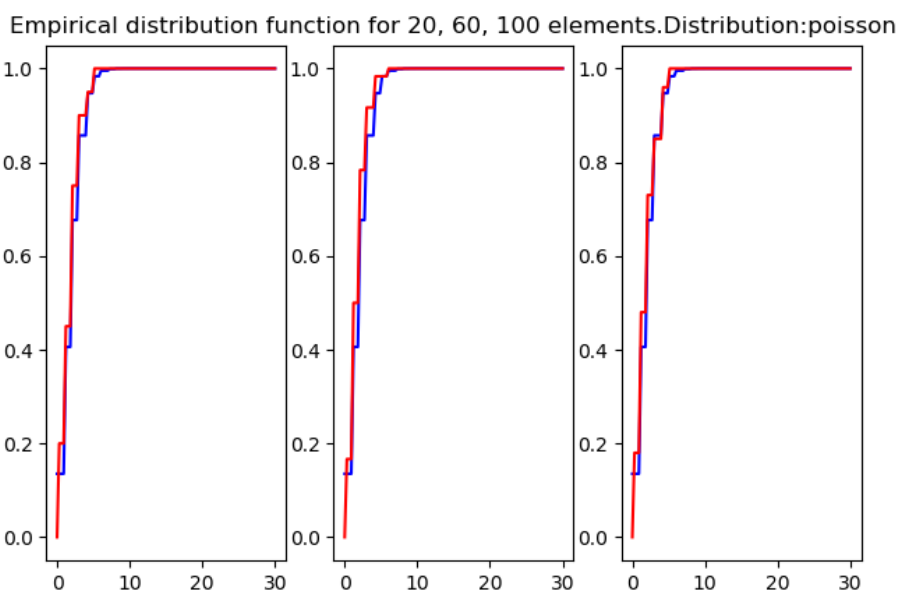
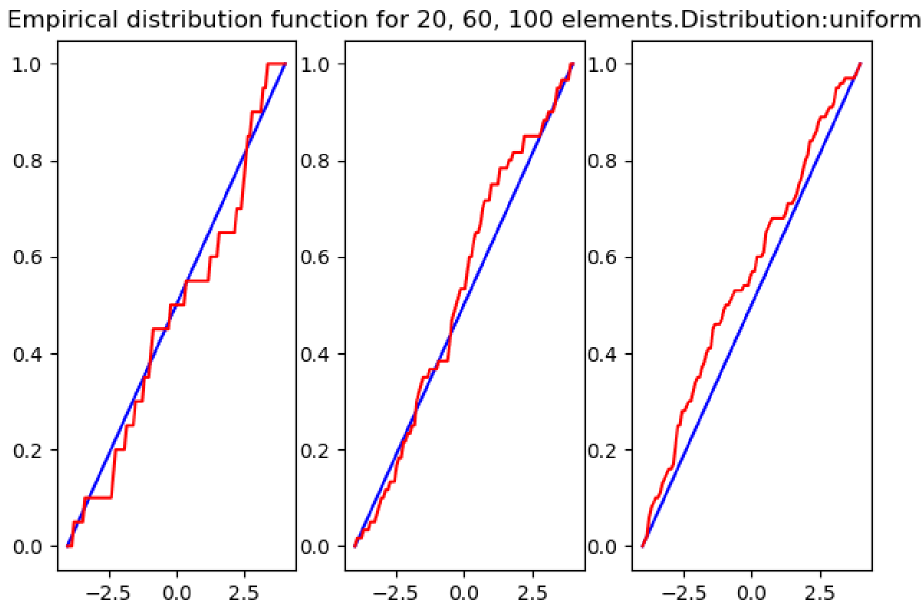


Рис. 15: Эмпирическая функция для равномерного распределения (5)



6.5 Эмпирические функции и ядерные оценки

Рис. 16: Ядерная функция плотности для нормального распределения (1), n = 20

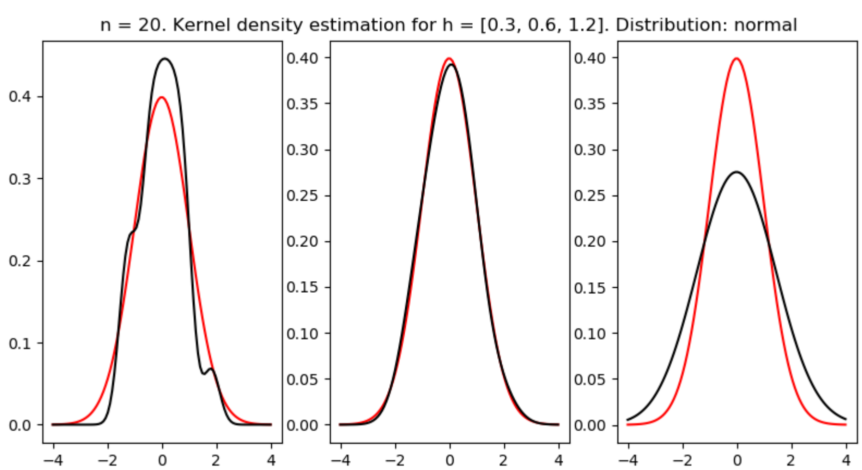


Рис. 17: Ядерная функция плотности для нормального распределения (1), n = 60

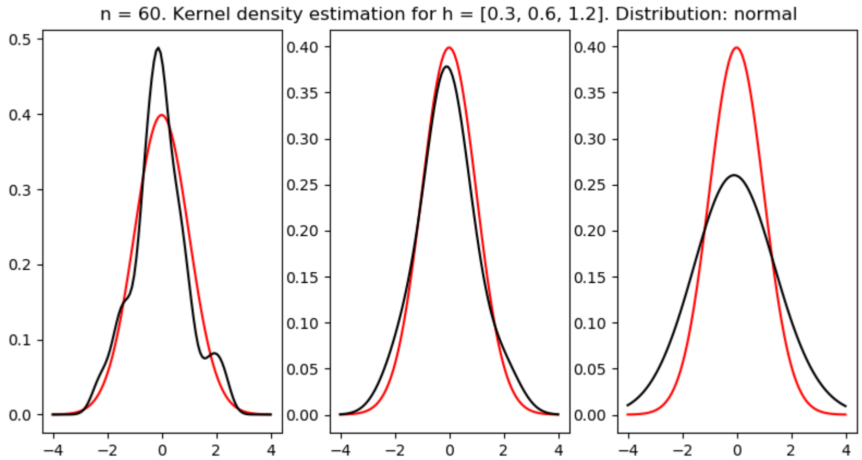


Рис. 18: Ядерная функция плотности для нормального распределения (1), n = 100

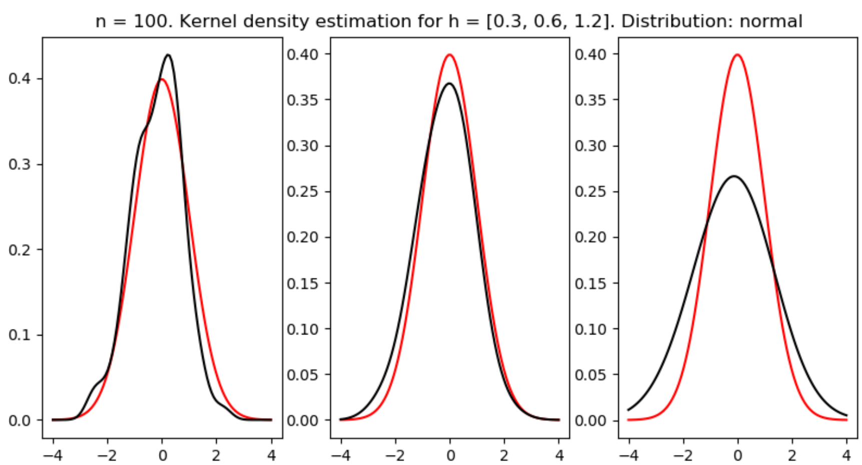


Рис. 19: Ядерная функция плотности для распределения Лапласа (3), n = 20

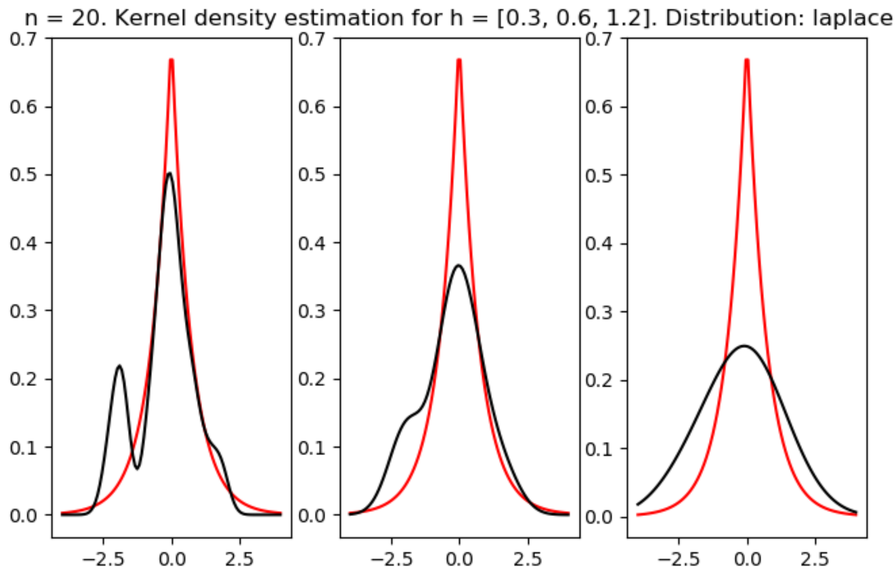


Рис. 20: Ядерная функция плотности для распределения Лапласа (3), n = 60

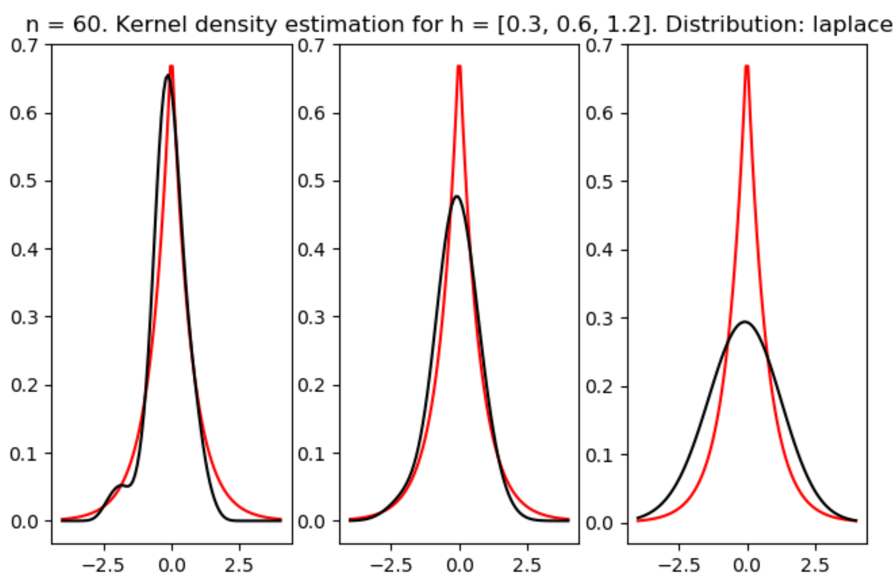


Рис. 21: Ядерная функция плотности для распределения Лапласа (3), n = 100

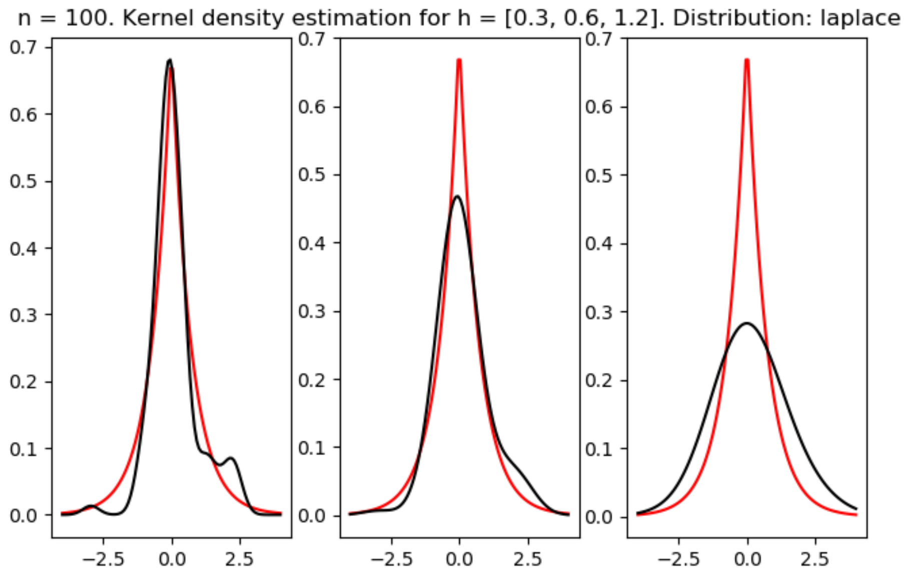


Рис. 22: Ядерная функция плотности для распределения Коши (2), n = 20

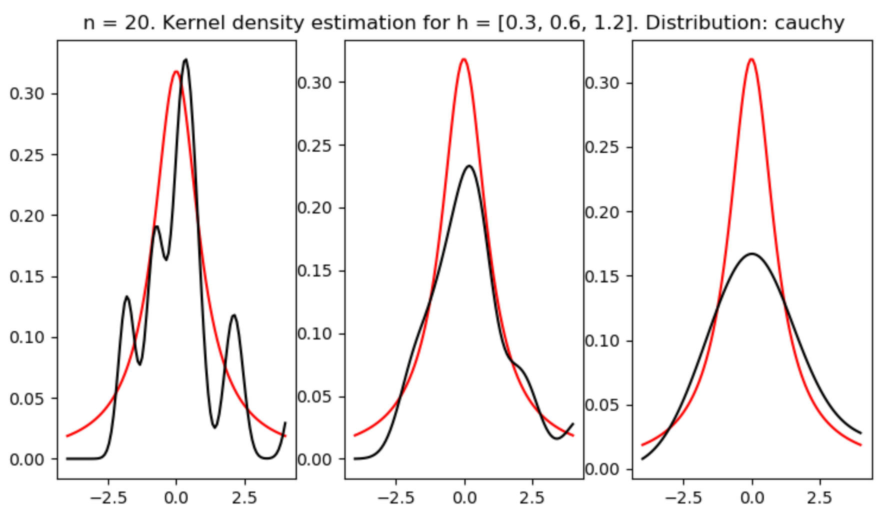


Рис. 23: Ядерная функция плотности для распределения Коши (2) 60

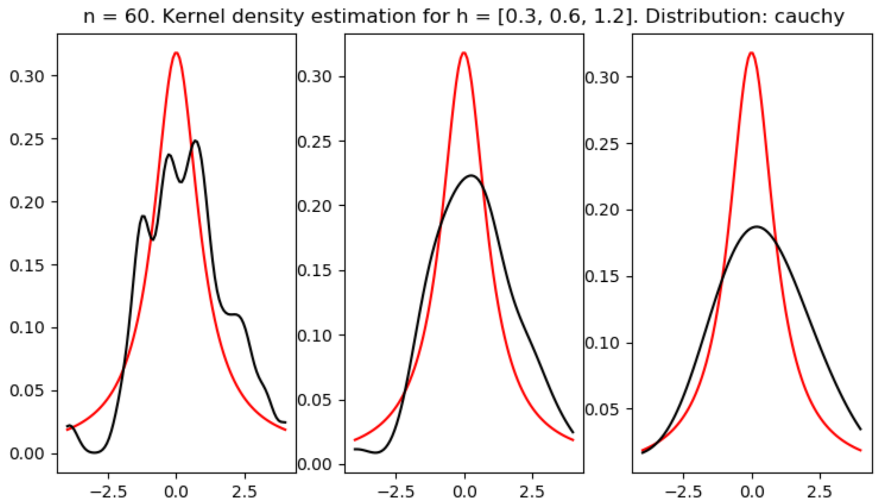


Рис. 24: Ядерная функция плотности для распределения Коши (2), n = 100

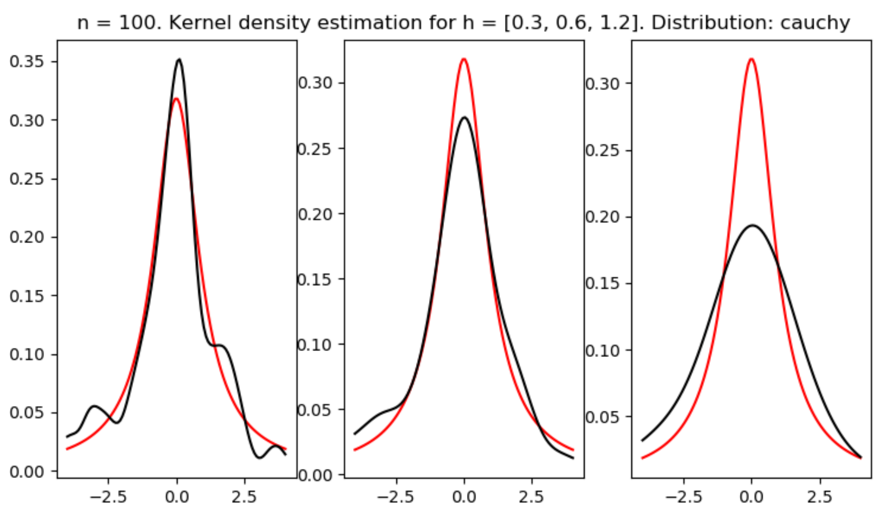


Рис. 25: Ядерная функция плотности для распределения Пуассона (4), n = 20

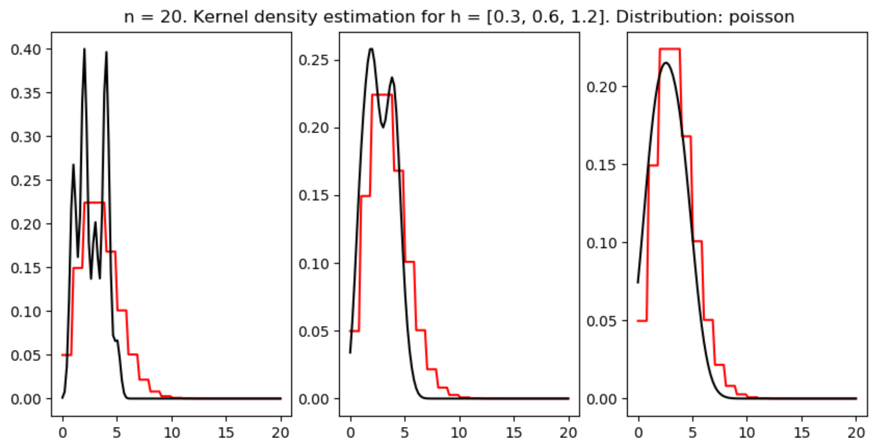


Рис. 26: Ядерная функция плотности для распределения Пуассона (4), n = 60

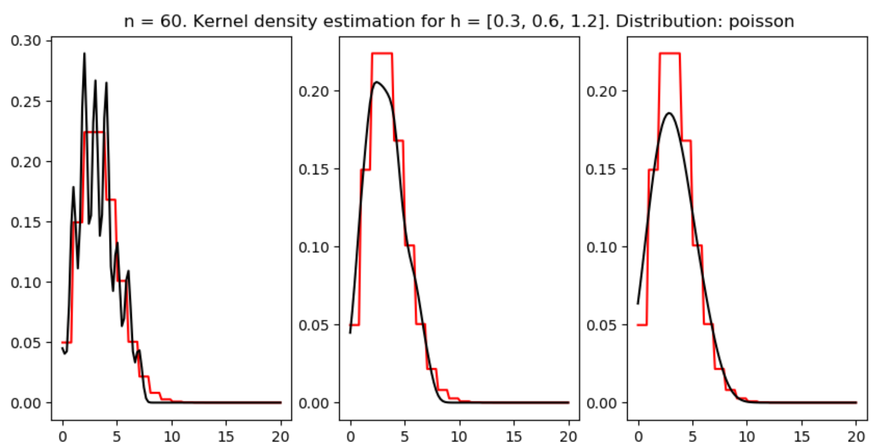


Рис. 27: Ядерная функция плотности для распределения Пуассона (4), n = 100

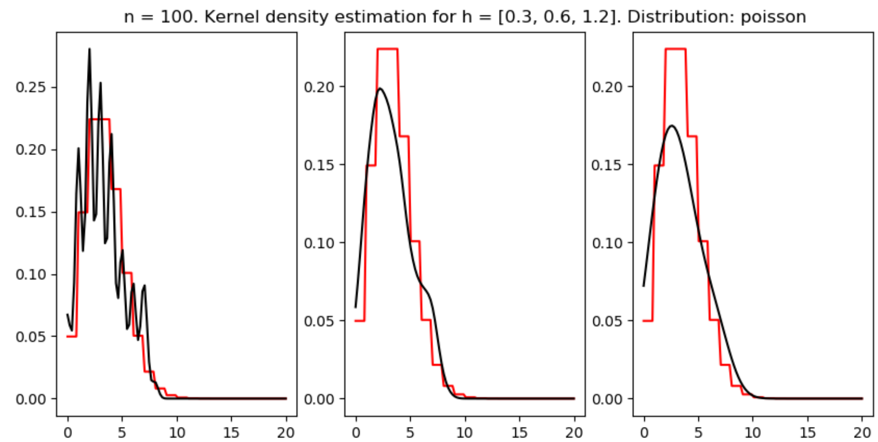


Рис. 28: Ядерная функция плотности для равномерного распределения (5), n = 20

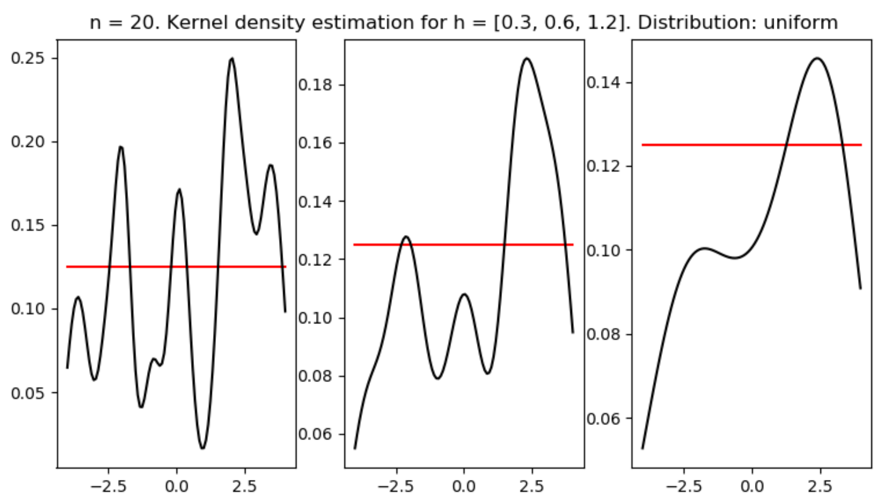


Рис. 29: Ядерная функция плотности для равномерного распределения (5), n = 60

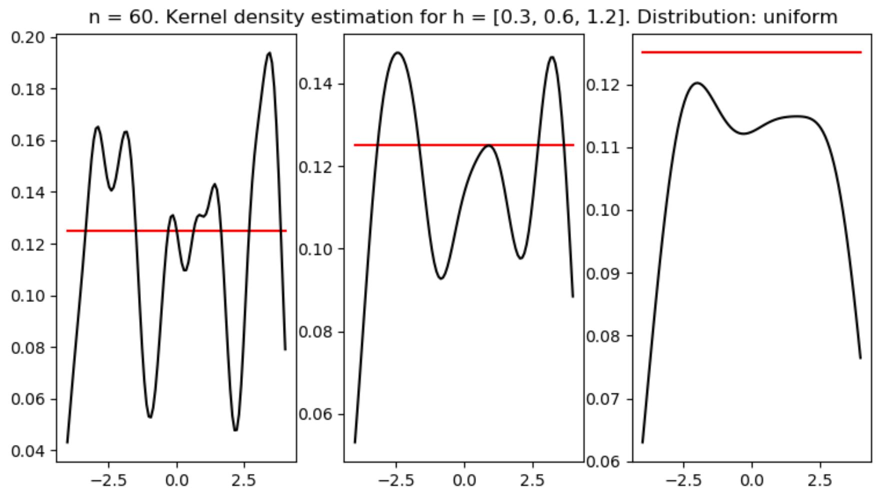
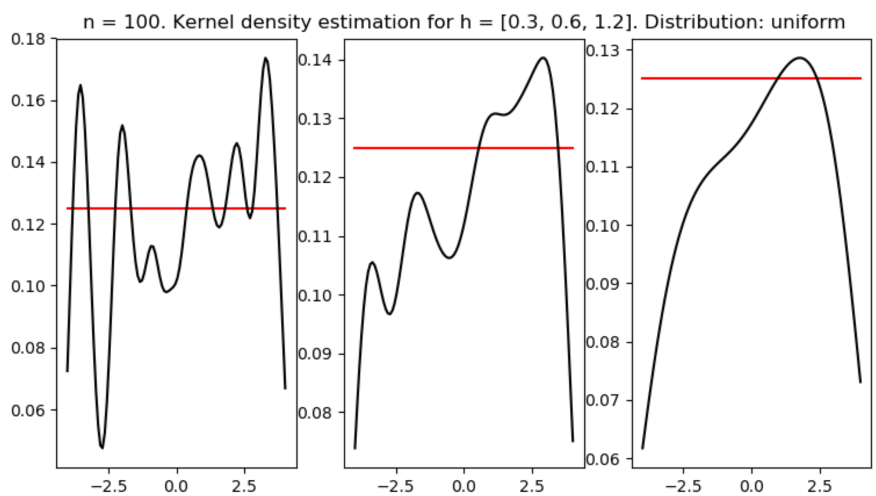


Рис. 30: Ядерная функция плотности для равномерного распределения (5), n = 100



7 Обсуждение

7.1 Характеристики положения

При вычислении средних значений пришлось отбрасывать некоторое число знаков после запятой, так как получаемая дисперсия не могла гарантировать получаемое точное значение.

Иными словами дисперсия может гарантировать порядок точности среднего значения только до первого значащего знака после запятой в дисперсии включительно.

Единственным исключением [в отбрасывании знаков после запятой] стало стандартное распределение Коши, так как оно имеет бесконечную дисперсию, а значит не может гарантировать никакой точности.

8 Выводы

8.1 Плотности распределения вероятностей

Как видно из графиков (1), (2), (3), (4), (5), при увеличении размера выборки построенная гистограмма точнее приближает график соответствующего распределения.

8.2 Характеристики положения

В процессе работы вычислены значения характеристик положения для определённых распределений на выборках фиксированной мощности и получено следующее ранжирование характеристик положения:

1. Стандартное нормальное распределение
$$\overline{x} < Z_{tr} < Z_Q < med\ x < Z_R$$
2. Стандартное распределение Коши
$$med\ x < Z_Q < Z_{tr} < \overline{x} < Z_R$$
3. Распределение Лапласа (коэффициент масштаба $\sqrt{2}$ коэффициент сдвига равен нулю)
$$med\ x < Z_{tr} < \overline{x} < Z_Q < Z_R$$
4. Равномерное распределение на отрезке $\left[-\sqrt{3}, \sqrt{3}\right]$
$$Z_R < \overline{x} < Z_{tr} < Z_Q < med\ x$$
5. Распределение Пуассона (значение мат ожидания равно 3)
$$\overline{x} < Z_{tr} < Z_Q < med\ x < Z_R$$

8.3 Боксплот Тьюки

Экспериментально полученные проценты выбросок, близки к теоретическим Можно вывести соотношение между процентами выбросов:

$$uniform < normal < poisson < laplace < cauchy \tag{15}$$

8.4 Эмпирические функции и ядерные оценки

Эмпирическая функция лучше приближает эталонную функцию на больших выборках.

По визуальной оценке, наилучшее приближение функции распределения ядерной функции получено при увеличении ширины окна. При фиксированной ширине окна точнее приблизить функцию распределения позволяет увеличение выборки.

9 Список литературы

[1] Модуль numpy - https://physics.susu.ru/vorontsov/language/numpy.html

[2] Формулы распределений - https://vk.com/doc184549949_491827451

[3] Выборочное среднее - https://en.wikipedia.org/wiki/Sample_mean_and_covariance

[4] Выборочная медиана - http://femto.com.ua/articles/part_1/2194.html

[5] Полусумма экстремальных значений - https://studopedia.info/8-56888.html

[6] Квартили - https://studfiles.net/preview/2438125/page:13/

[7] Усечённое среднее - https://ole-olesko.livejournal.com/15773.html

[8] Выборочная медиана - http://femto.com.ua/articles/part_1/2194.html

[9] Квартили - https://studfiles.net/preview/2438125/page:13/

[10] Боксплот - https://en.wikipedia.org/wiki/Box_plot

[11] Модуль matplotlib - https://matplotlib.org/users/index.html

[12] Модуль scipy - https://docs.scipy.org/doc/scipy/reference/

[13] Формулы распределений - https://vk.com/doc184549949_491827451

[14] Н. И. Чернова, https://nsu.ru/mmf/tvims/chernova/ms/lec/node4.html, 2002

[15] Victor, https://www.mql5.com/ru/articles/396, 2012

[16] Nathaniel E. Helwig, http://users.stat.umn.edu/h̃elwig/notes/den-Notes.pdf, 2017