

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. ПЕТРА ВЕЛИКОГО

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ И МЕХАНИКИ

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ

ОТЧЁТ ПО ЛАБОРАТОРНЫМ РАБОТАМ
№1 – 4
ИССЛЕДОВАНИЕ РАСПРЕДЕЛЕНИЙ

Студент группы 3630102/70301

Камянский Д.В.

Преподаватель

Баженов А. Н.

САНКТ-ПЕТЕРБУРГ
2020 г.

Содержание

1. Список иллюстраций	3
2. Список таблиц	4
3. Постановка задачи	5
4. Теория	5
4.1. Плотности распределения вероятностей	5
4.2. Характеристики положения	6
4.3. Боксплот Тьюки	6
4.4. Эмпирические функции и ядерные оценки	6
5. Реализация	7
6. Результаты	8
6.1. Плотности распределения вероятностей	8
6.2. Характеристики положения	12
6.3. Боксплот Тьюки	14
6.4. Ядерные функции	20
6.5. Эмпирические функции распределения	28
7. Обсуждение	30
7.1. Характеристики положения	30
8. Выводы	30
8.1. Плотности распределения вероятностей	30
8.2. Характеристики положения	31
8.3. Боксплот Тьюки	31
8.4. Эмпирические функции и ядерные оценки	31
9. Список литературы	32

1 Список иллюстраций

1	Нормальное распределение (1)	8
2	Распределение Лапласа (3)	9
3	Распределение Коши (2)	10
4	Распределение Пуассона (4)	11
5	Равномерное распределение (5)	12
6	Boxplot распределений для размера выборки $N=20$	14
7	Boxplot стандартное нормальное распределение	15
8	Boxplot стандартное распределение Лапласа	16
9	Boxplot стандартное распределение Коши	17
10	Boxplot распределение Пуассона	18
11	Boxplot равномерное распределение	19
12	Ядерная функция плотности для нормального распределения $n = 20$...	20
13	Ядерная функция плотности для нормального распределения $n = 60$...	21
14	Ядерная функция плотности для нормального распределения $n = 100$..	21
15	Ядерная функция плотности для распределения Лапласа $n = 20$	22
16	Ядерная функция плотности для распределения Лапласа $n = 60$	22
17	Ядерная функция плотности для распределения Лапласа $n = 100$	23
18	Ядерная функция плотности для распределения Коши $n = 20$	23
19	Ядерная функция плотности для распределения Коши $n = 60$	24
20	Ядерная функция плотности для распределения Коши $n = 100$	24
21	Ядерная функция плотности для распределения Пуассона $n = 20$	25
22	Ядерная функция плотности для распределения Пуассона $n = 60$	25
23	Ядерная функция плотности для распределения Пуассона $n = 100$	26
24	Ядерная функция плотности для равномерного распределения $n = 20$...	26
25	Ядерная функция плотности для равномерного распределения $n = 60$...	27
26	Ядерная функция плотности для равномерного распределения $n = 100$.	27
27	Эмпирическая функция для стандартного нормального распределения ..	28
28	Эмпирическая функция для стандартного распределения Лапласа	28
29	Эмпирическая функция для стандартного распределения Коши	29

30	Эмпирическая функция для распределения Пуассона	29
31	Эмпирическая функция для равномерного распределения	30

2 Список таблиц

1	Стандартное нормальное распределение.	12
2	Стандартное распределение Лапласа.	13
3	Распределение Коши.	13
4	Равномерное распределение.	13
5	Распределение Пуассона.	14
6	Теоретические оценки доли выбросов	19
7	Экспериментальные оценки доли выбросов	20

3 Постановка задачи

Для 5 распределений:

1.
$$N(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (1)$$

2.
$$C(x, 0, 1) = \frac{1}{\pi(1+x^2)} \quad (2)$$

3.
$$L\left(x, 0, \frac{1}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \quad (3)$$

4.
$$P(5, k) = \frac{5^k}{k!} e^{-5} \quad (4)$$

5.
$$M(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & |x| \leq \sqrt{3} \\ 0 & |x| > \sqrt{3} \end{cases} \quad (5)$$

необходимо было выполнить:

1. Сгенерировать выборки размером 10, 50, 1000. Построить на одном рисунке гистограмму и график плотности распределения.
2. Сгенерировать выборки размером 10, 100, 1000. Для каждой выборки вычислить следующие статистические характеристики положения данных: \bar{x} , $medx$, Z_R , Z_Q , Z_{tr} . Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов.
3. Сгенерировать выборки размером 20 и 1000. Для каждого распределения определить долю выбросов экспериментально (сгенерировав выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.
4. Сгенерировать выборки размером 20, 60, 100. Построить на них эмпирические функции распределения и ядерные оценки плотности распределения на отрезке $[-4; 4]$ для непрерывных распределений и на отрезке $[6; 14]$ для распределения Пуассона.

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.

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.

1. Выборочное среднее :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (6)$$

2. Выборочная медиана :

$$\text{med } x = \begin{cases} x_{k+1}, & n = 2k + 1 \\ \frac{1}{2}(x_k + x_{k+1}), & n = 2k \end{cases} \quad (7)$$

3. Полусумма экстремальных значений :

$$Z_R = \frac{1}{2}(x_1 + x_n) \quad (8)$$

4. Полусумма квантилей :

$$Z_Q = \frac{1}{2}(Z_{\frac{1}{4}} + Z_{\frac{3}{4}}) \quad (9)$$

5. Усечённое среднее :

$$Z_{tr} = \frac{1}{n - 2r} \sum_{i=r+1}^{n-r} x_i \quad (10)$$

4.3 Боксплот Тьюки

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

Границами ящика служат первый и третий квантили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину «усов» определяют разность первого квантиля и полутора межквантильных расстояний и сумма третьего квантиля и полутора межквантильных расстояний.

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

Правая и левая границы: $X_1 = Q_1 - 1.5(Q_3 - Q_1)$, $X_2 = Q_3 + 1.5(Q_3 - Q_1)$

Теоретическая вероятность выбросов:

Для непрерывных распределений: $P_e^t = F(X_1^t) + (1 - F(X_2^t))$

Для дискретных распределений: $P_e^t = (F(X_1^t) - P(x = X_1^t)) + (1 - F(X_2^t))$

4.4 Эмпирические функции и ядерные оценки

Статистическим рядом называется последовательность различных элементов выборки z_1, z_2, \dots, z_k , расположенных в возрастающем порядке с указанием частот n_1, n_2, \dots, n_k , с которыми эти элементы содержатся в выборке.

Эмпирическая функция распределения:

$$F^*(x) = \frac{1}{n} \sum_{z_i \leq x} n_i \quad (11)$$

Ядерная оценка плотности:

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

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:

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

Параметр сглаживания будем выбирать по правилу Сильвермана

$$h_n = 1.06 * \sigma * n^{-1/5} \quad (14)$$

где σ - выборочное стандартное отклонение.

5 Реализация

Лабораторная работа была выполнена на языке *Python* 3.8.2. Для генерации выборки был использован модуль *stats* библиотеки *scipy* для генерации выборок различных распределений и библиотека *matplotlib* для построения графиков.

6 Результаты

6.1 Плотности распределения вероятностей

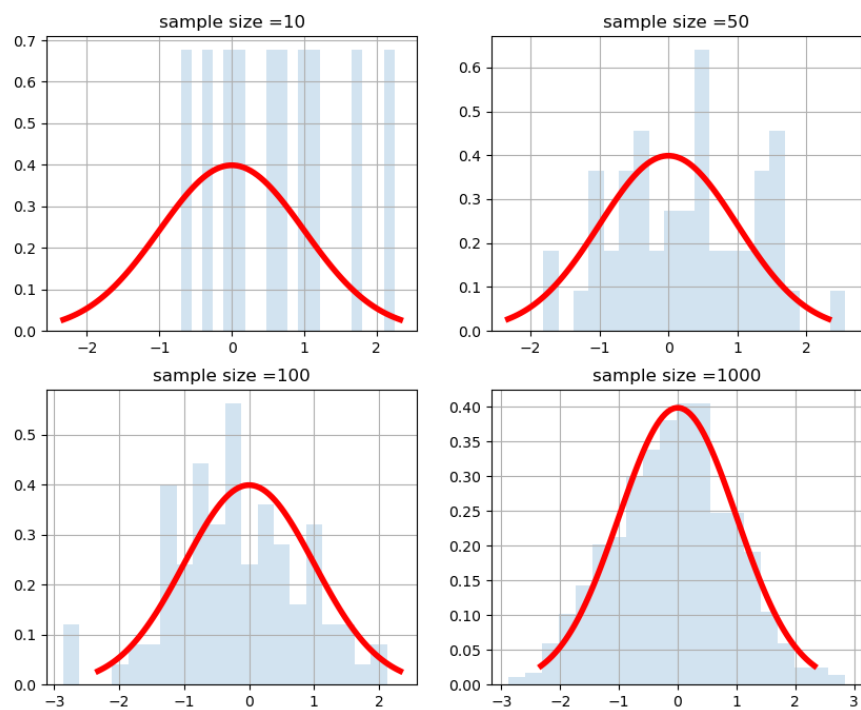


Рис. 1: Нормальное распределение (1)

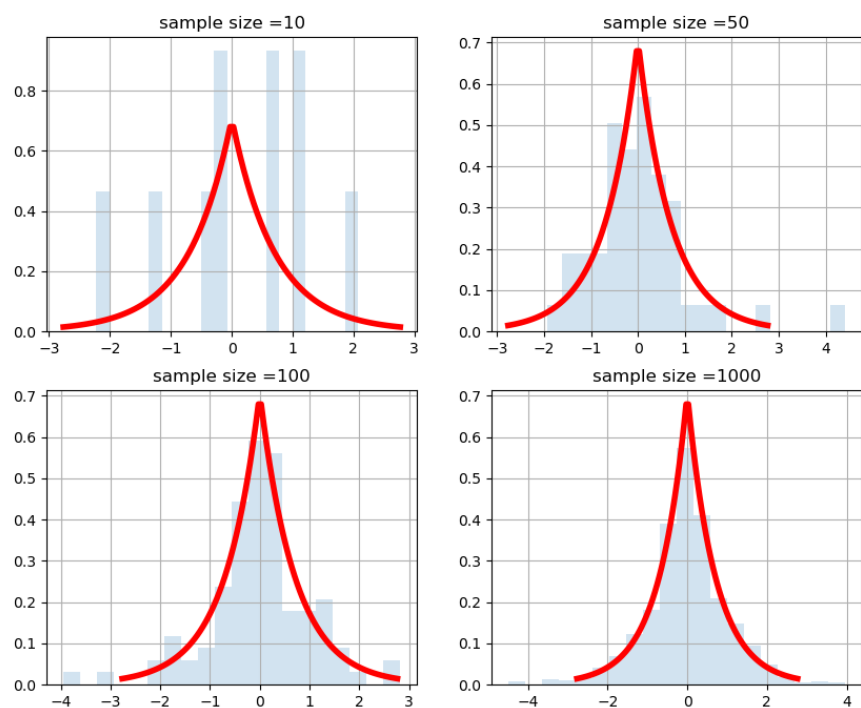


Рис. 2: Распределение Лапласа (3)

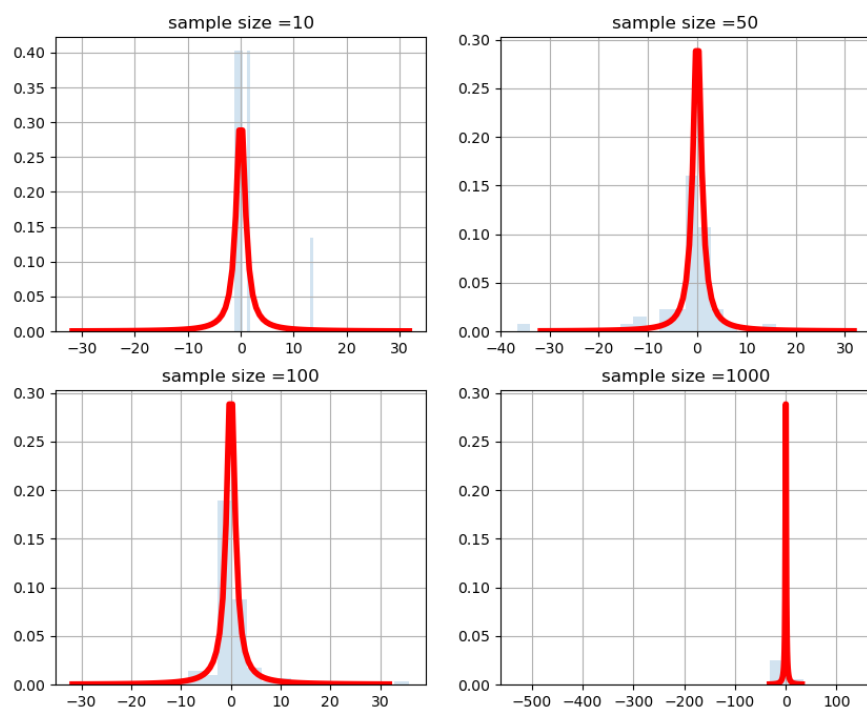


Рис. 3: Распределение Коши (2)

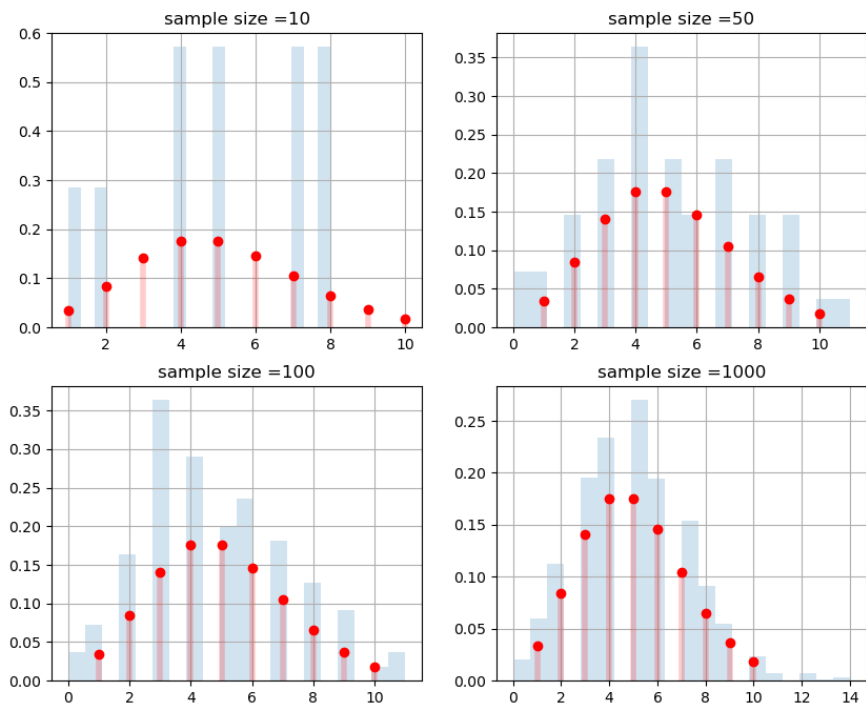


Рис. 4: Распределение Пуассона (4)

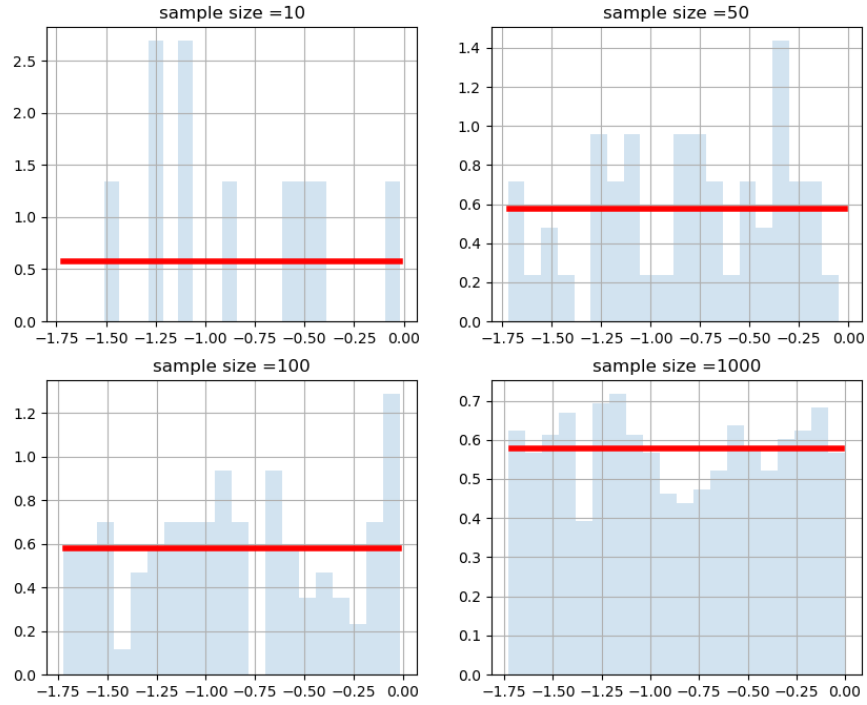


Рис. 5: Равномерное распределение (5)

6.2 Характеристики положения

Таблица 1: Стандартное нормальное распределение.

$n = 10$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.0	-0.02	0.0	-0.0	-0.0
$D =$	0.106931	0.145363	0.186191	0.111027	0.165807
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.0	0.01	0.02	0.0	0.0
$D =$	0.010365	0.015172	0.092881	0.011974	0.017426
$n = 1000$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	0.0	0.0	-0.0	0.0	0.0
$D =$	0.00096	0.001594	0.060543	0.001256	0.001888

Таблица 2: Стандартное распределение Лапласа.

$n = 10$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.01	0.01	-0.0	-0.0	-0.01
$D =$	0.097868	0.069859	0.390201	0.093469	0.167235
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.0	0.01	0.01	0.0	-0.0
$D =$	0.009778	0.005878	0.429639	0.009072	0.020247
$n = 1000$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	0.0	0.0	-0.02	-0.0	0.0
$D =$	0.001009	0.000531	0.428975	0.000997	0.001992

Таблица 3: Распределение Коши.

$n = 10$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	0.31	-0.01	-5.87	0.03	3.5
$D =$	260.978036	0.347366	26462.674037	0.774488	6529.2454
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	2.26	0.01	22.79	-0.0	1.71
$D =$	2122.106445	0.023531	1741103.211025	0.05289	1778.363557
$n = 1000$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	1.14	0.0	-1845.52	-0.0	-1.64
$D =$	706.431432	0.002497	21117320297.542297	0.004864	1675.403679

Таблица 4: Равномерное распределение.

$n = 10$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.02	0.0	0.0	-0.0	-0.02
$D =$	0.095612	0.227685	0.043735	0.135551	0.171826
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	0.0	-0.0	-0.0	0.0	0.01
$D =$	0.009971	0.029194	0.00058	0.014725	0.019723
$n = 1000$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	-0.0	0.0	-0.0	0.0	-0.0
$D =$	0.000959	0.002864	$6e - 06$	0.001502	0.00205

Таблица 5: Распределение Пуассона.

$n = 10$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	5.01	4.84	5.3	4.88	5.02
$D =$	0.515302	0.739579	0.942244	0.586539	0.832244
$n = 100$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	5.0	4.9	5.97	4.89	4.99
$D =$	0.049874	0.102596	0.499128	0.143341	0.101492
$n = 1000$	average	med	Z_R	Z_Q	Z_{tr}
$E =$	5.0	5.0	6.83	4.67	5.0
$D =$	0.005329	0.0	0.341759	0.076798	0.010249

6.3 Боксплот Тьюки

Рис. 6: Boxplot распределений для размера выборки $N=20$

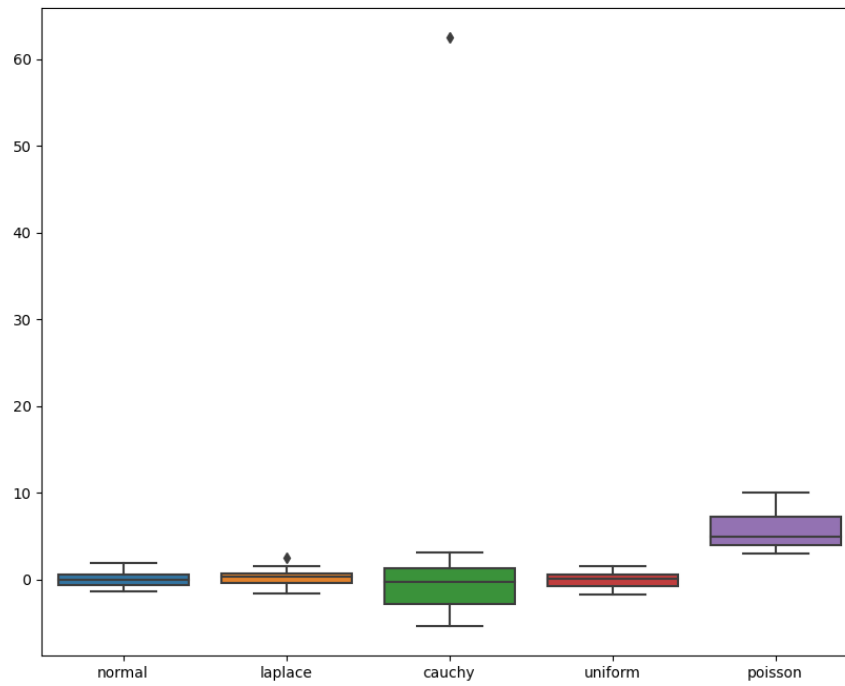


Рис. 7: Boxplot стандартное нормальное распределение

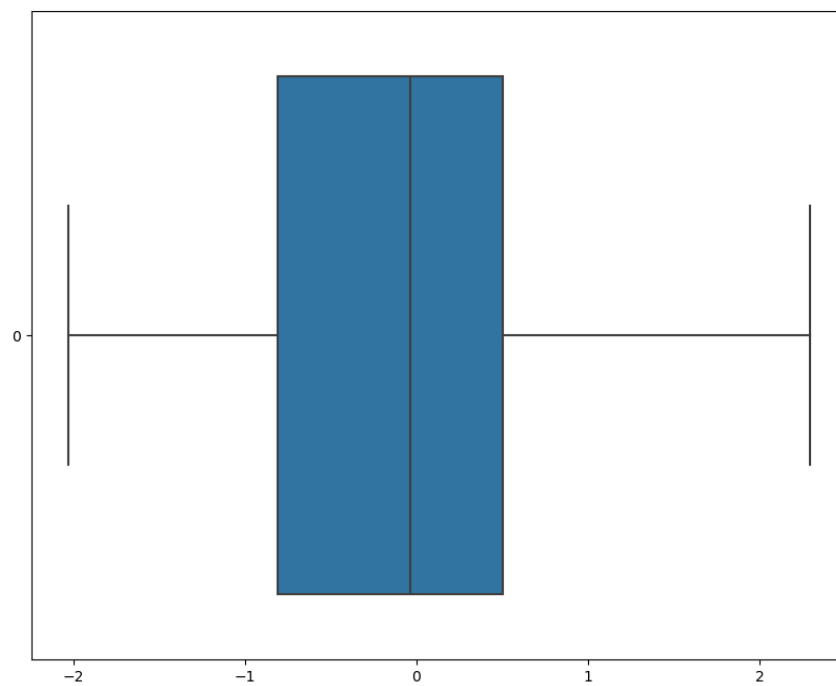


Рис. 8: Boxplot стандартное распределение Лапласа

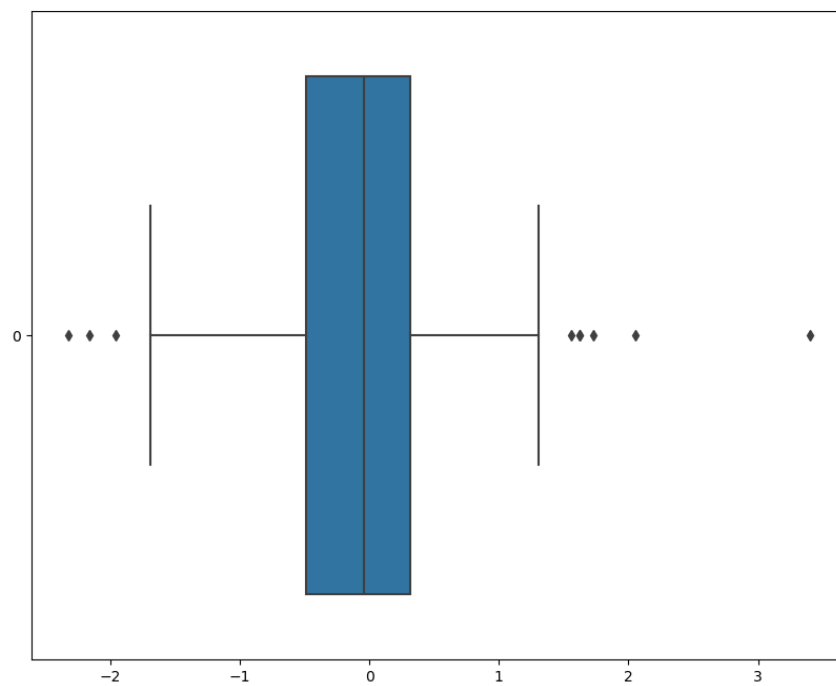


Рис. 9: Voxplot стандартное распределение Коши

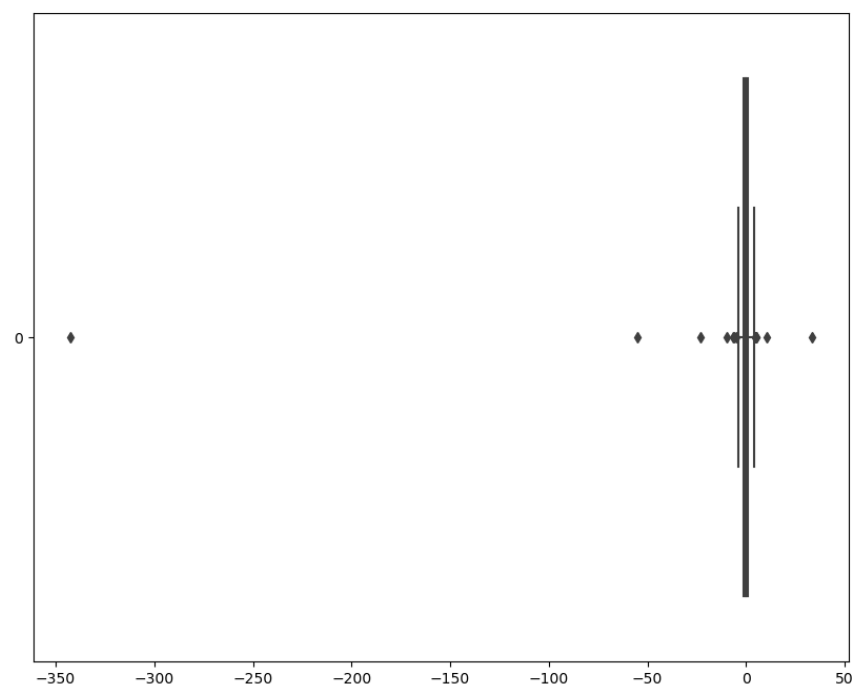


Рис. 10: Voxplot распределение Пуассона

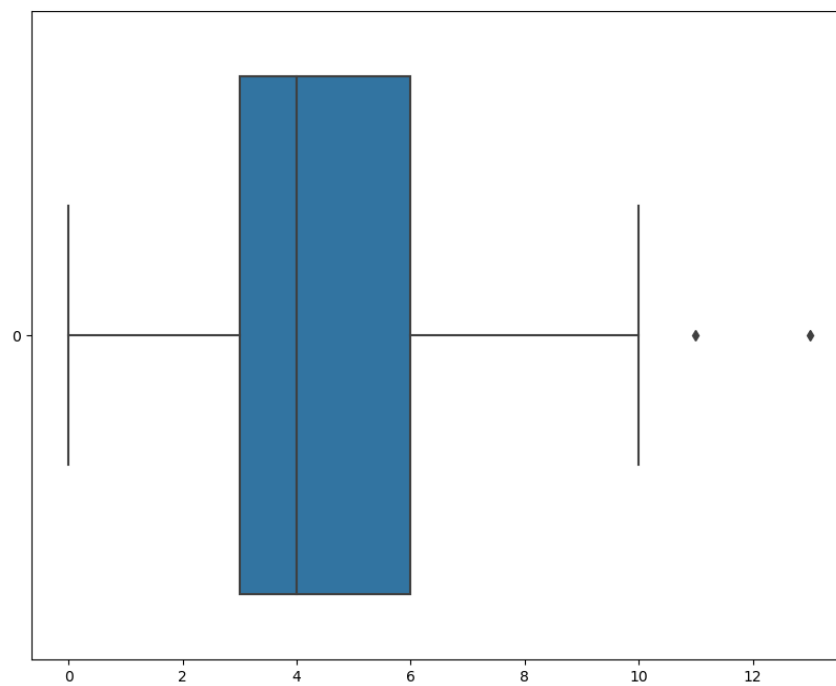


Рис. 11: Voxplot равномерное распределение

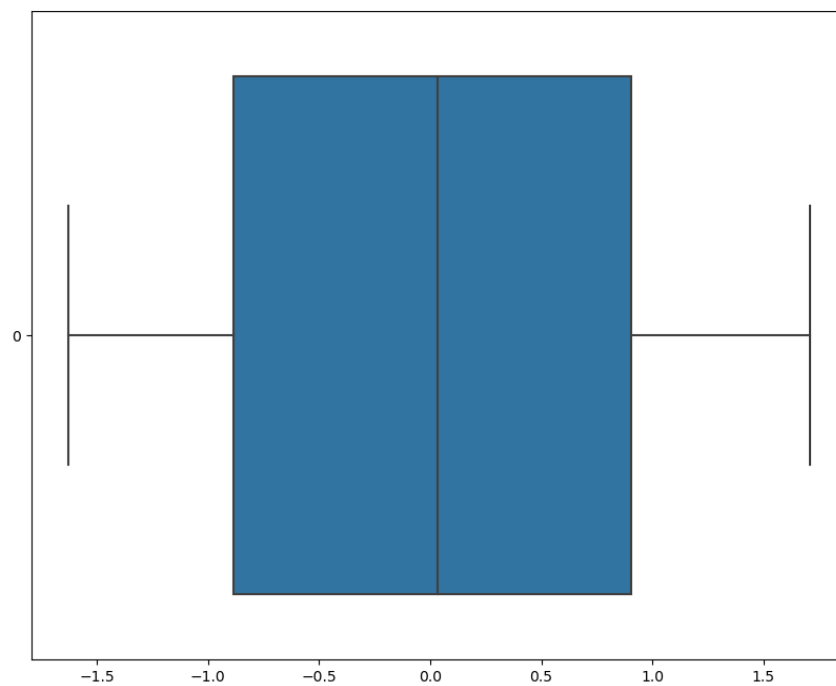


Таблица 6: Теоретические оценки доли выбросов

Распределение	Доля выбросов
normal	0.006977
cauchy	0.155958
laplace	0.0625
uniform	0.0
poisson	0.013695

Таблица 7: Экспериментальные оценки доли выбросов

Распределение	Средняя доля выбросов	Дисперсия доли выбросов
normal		
n = 20	0.02	0.00152
n = 100	0.01	0.000161
cauchy		
n = 20	0.15	0.004827
n = 100	0.155	0.001067
laplace		
n = 20	0.07	0.004469
n = 100	0.065	0.000995
uniform		
n = 20	0.002	0.000212
n = 100	0.0	0.0
poisson		
n = 20	0.02	0.002284
n = 100	0.014	0.000285

6.4 Ядерные функции

Рис. 12: Ядерная функция плотности для нормального распределения $n = 20$

$n=20$ with window width $h=[0.5, 1, 2]$

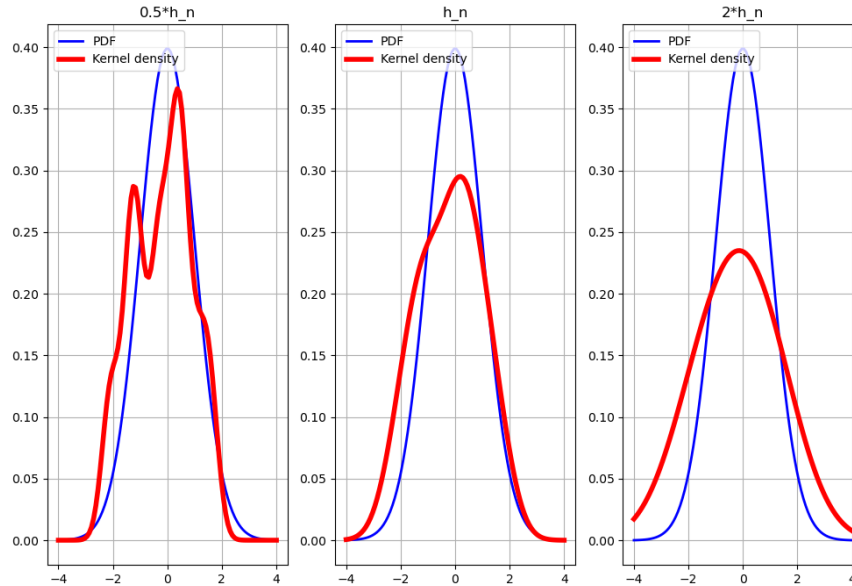


Рис. 13: Ядерная функция плотности для нормального распределения $n = 60$
 $n=60$ with window width $h=[0.5, 1, 2]$

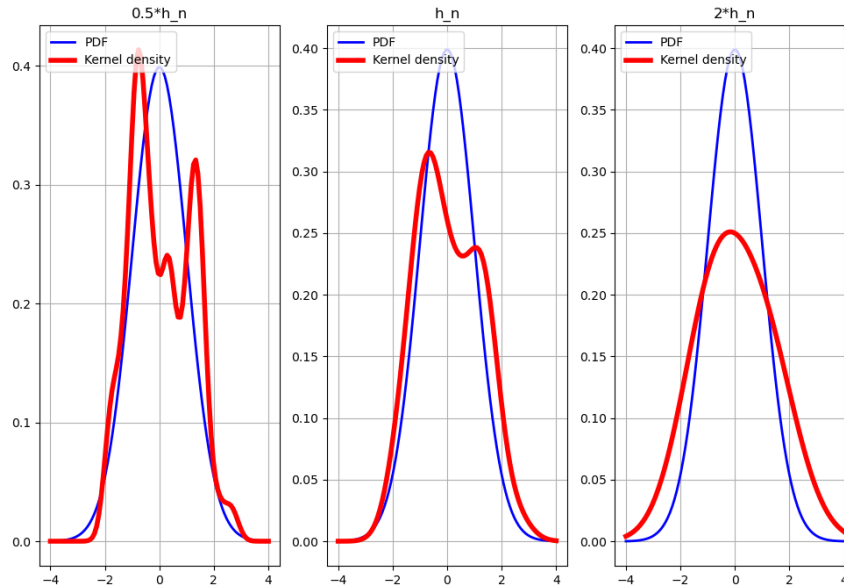


Рис. 14: Ядерная функция плотности для нормального распределения $n = 100$
 $n=100$ with window width $h=[0.5, 1, 2]$

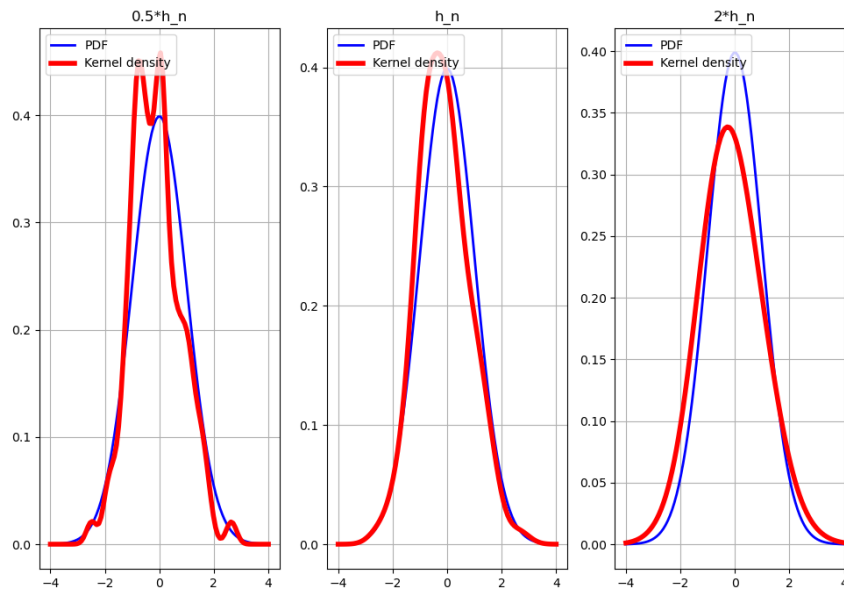


Рис. 15: Ядерная функция плотности для распределения Лапласа $n = 20$

$n=20$ with window width $h=[0.5, 1, 2]$

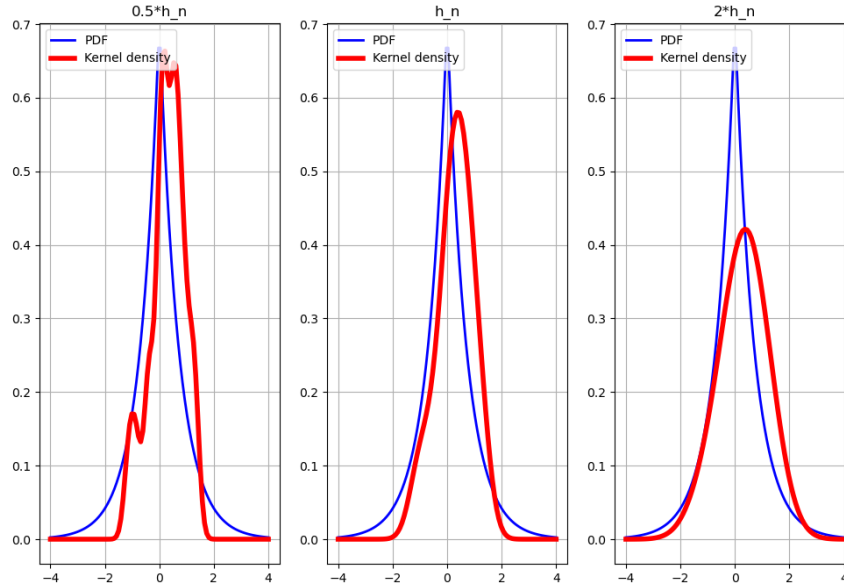


Рис. 16: Ядерная функция плотности для распределения Лапласа $n = 60$

$n=60$ with window width $h=[0.5, 1, 2]$

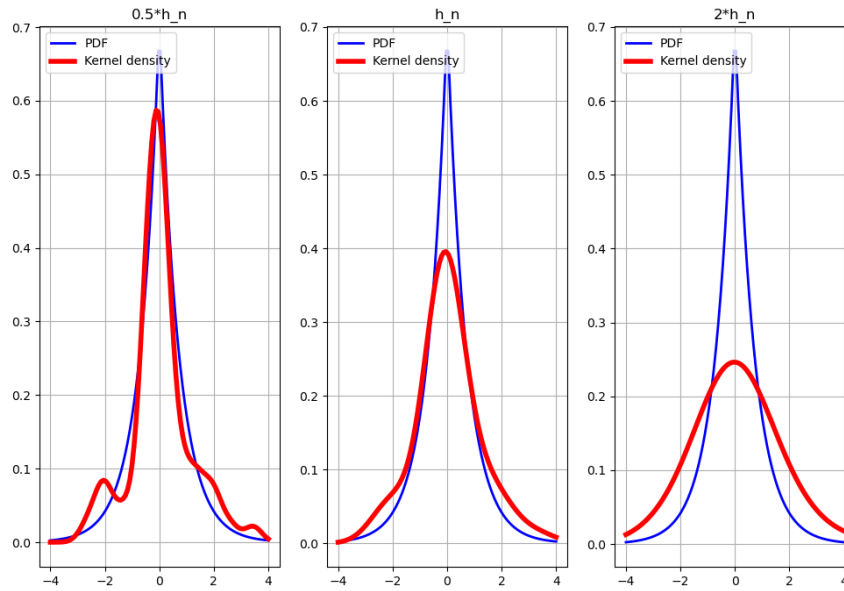


Рис. 17: Ядерная функция плотности для распределения Лапласа $n = 100$

$n=100$ with window width $h = [0.5, 1, 2]$

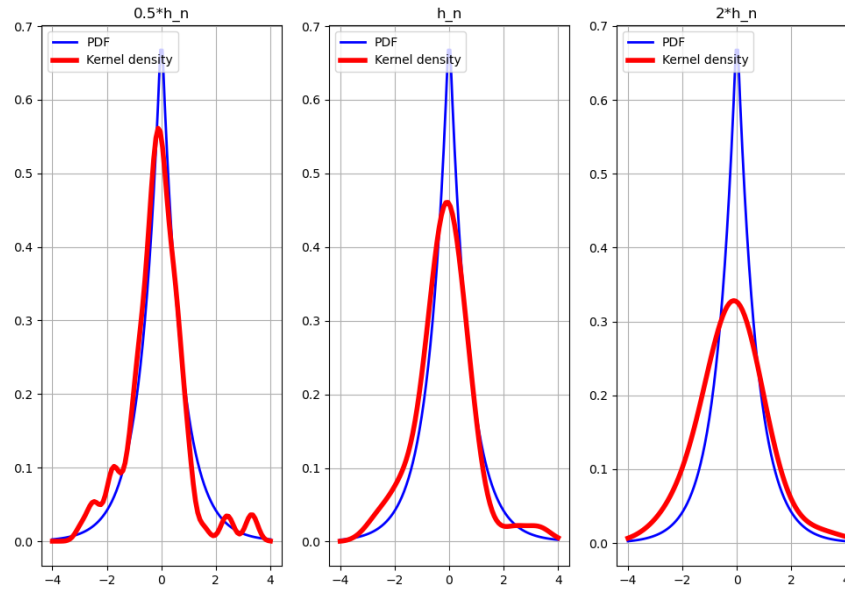


Рис. 18: Ядерная функция плотности для распределения Коши $n = 20$

$n=20$ with window width $h = [0.5, 1, 2]$

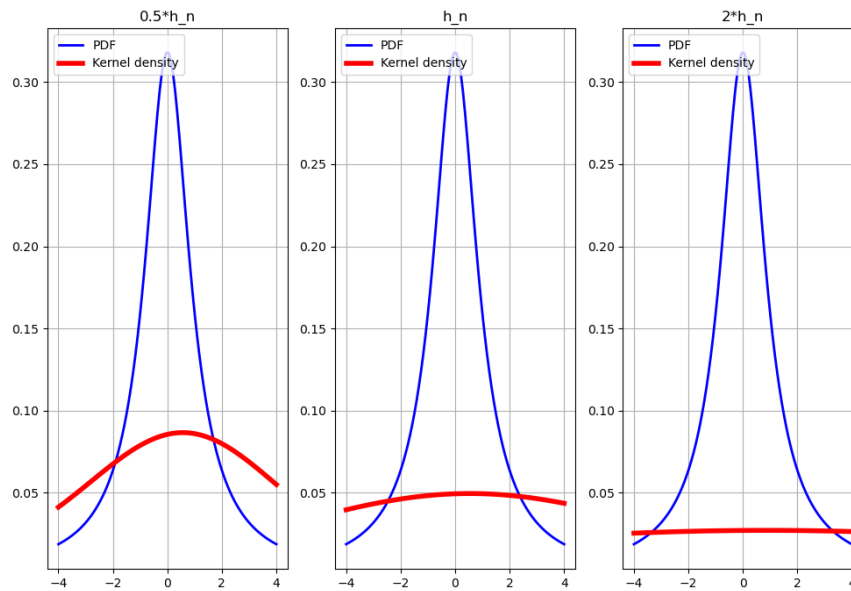


Рис. 19: Ядерная функция плотности для распределения Коши $n = 60$

$n=60$ with window width $h=[0.5, 1, 2]$

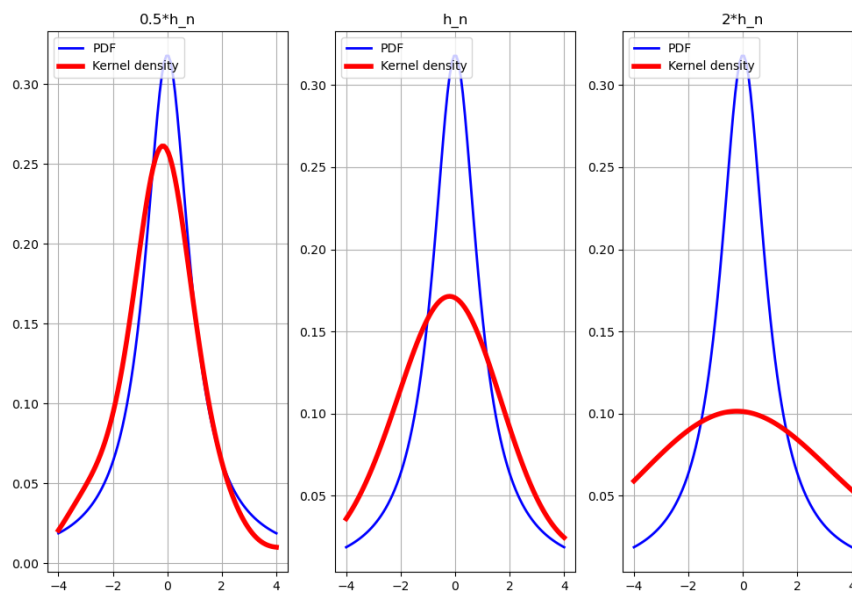


Рис. 20: Ядерная функция плотности для распределения Коши $n = 100$

$n=100$ with window width $h=[0.5, 1, 2]$

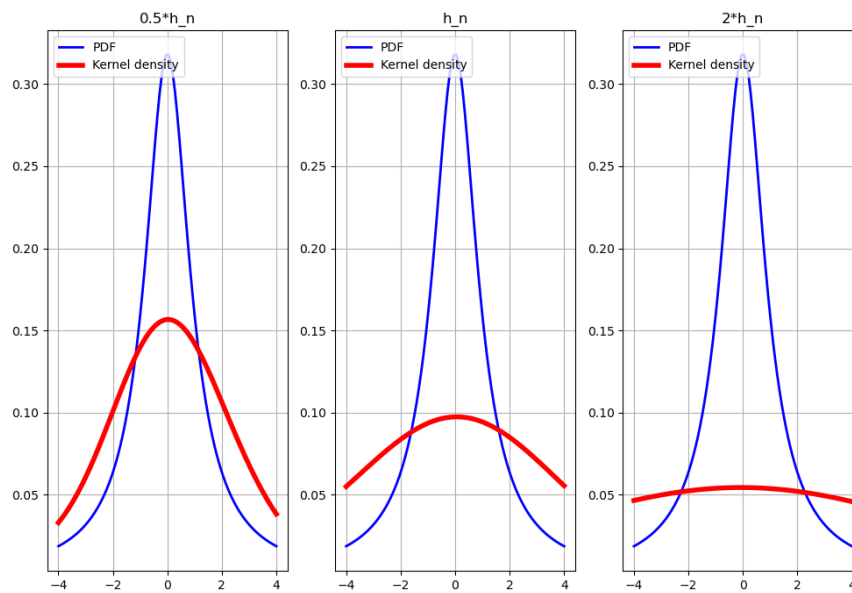


Рис. 21: Ядерная функция плотности для распределения Пуассона $n = 20$

$n=20$ with window width $h=[0.5, 1, 2]$

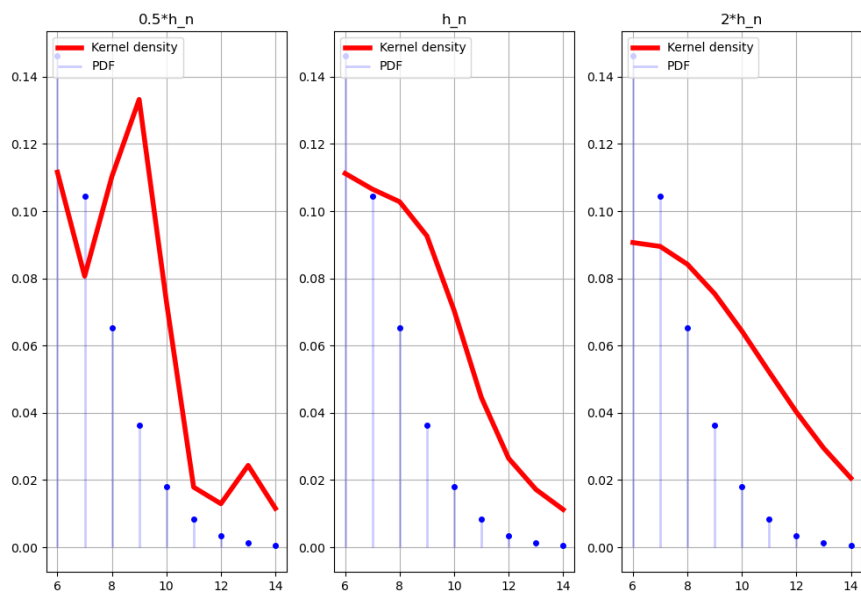


Рис. 22: Ядерная функция плотности для распределения Пуассона $n = 60$

$n=60$ with window width $h=[0.5, 1, 2]$

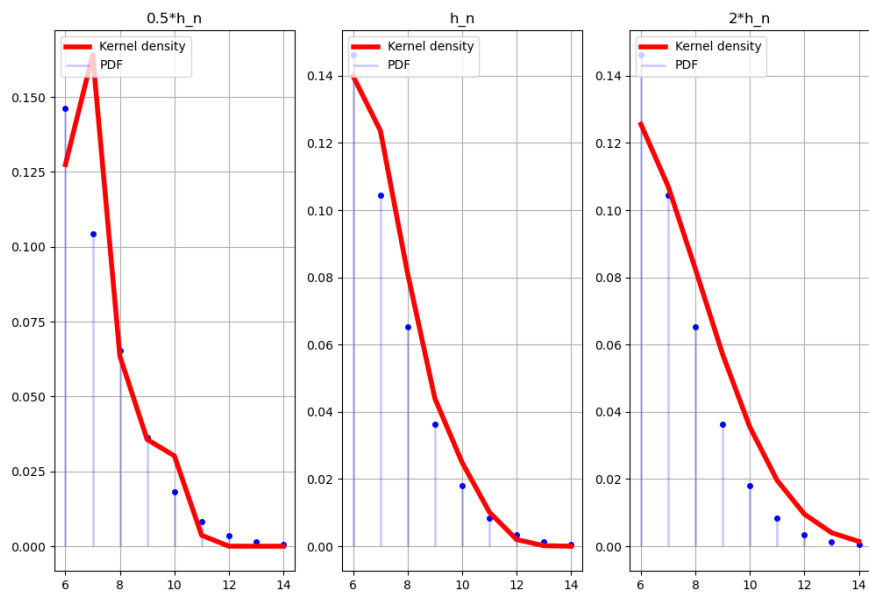


Рис. 23: Ядерная функция плотности для распределения Пуассона $n = 100$

$n=100$ with window width $h=[0.5, 1, 2]$

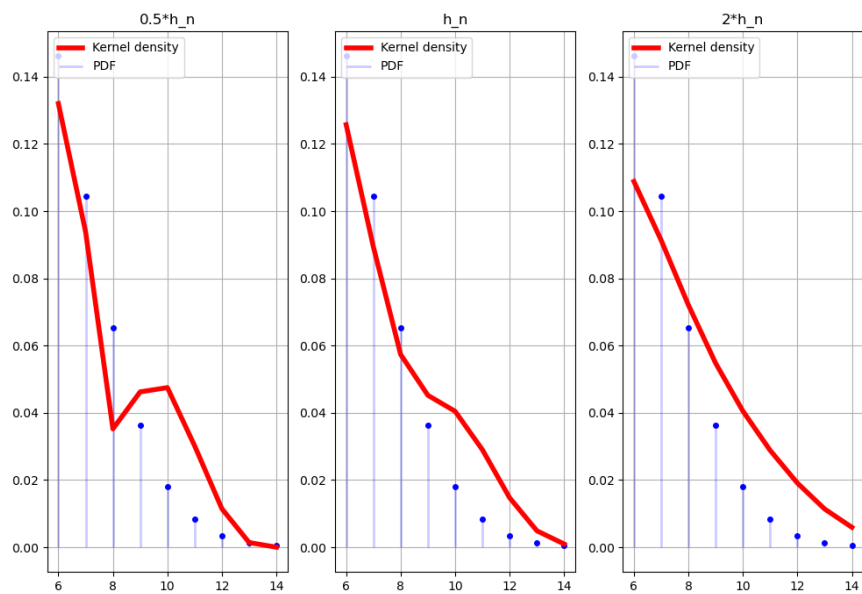


Рис. 24: Ядерная функция плотности для равномерного распределения $n = 20$

$n=20$ with window width $h=[0.5, 1, 2]$

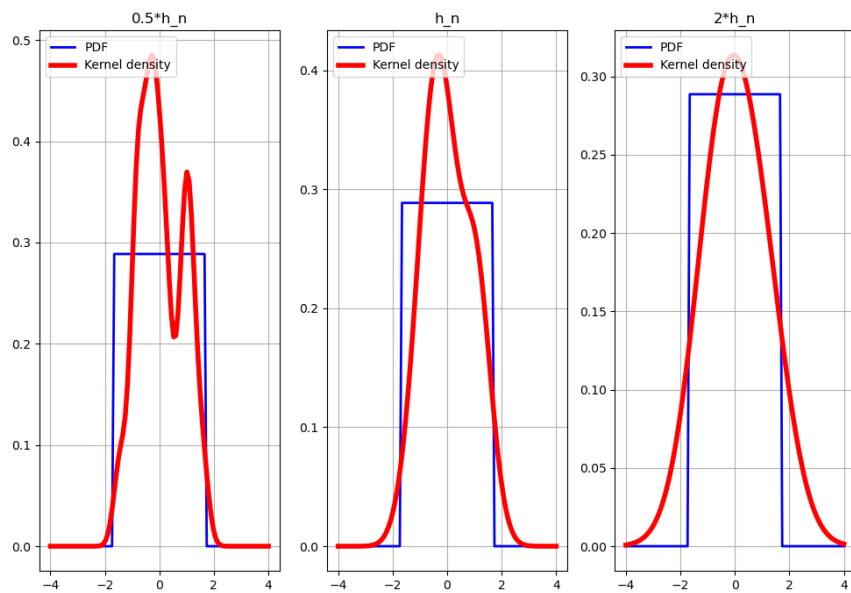


Рис. 25: Ядерная функция плотности для равномерного распределения $n = 60$
 $n=60$ with window width $h=[0.5, 1, 2]$

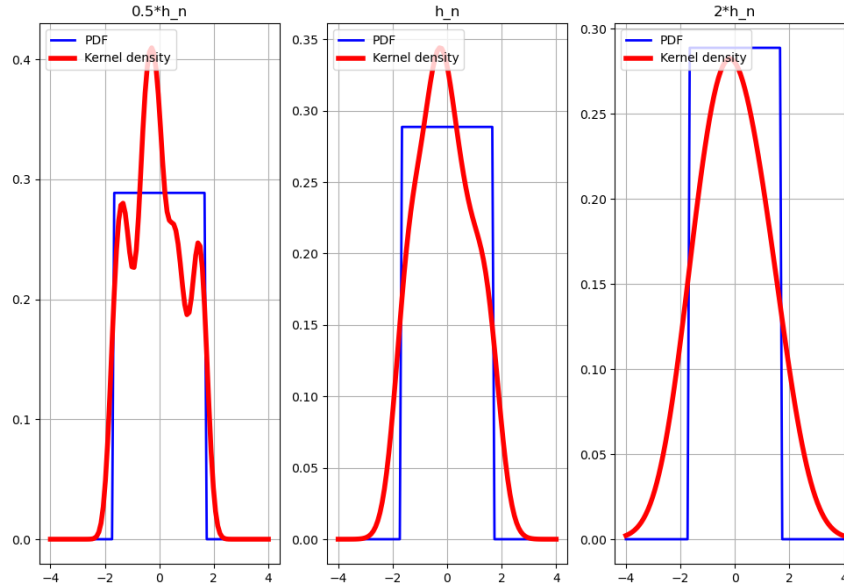
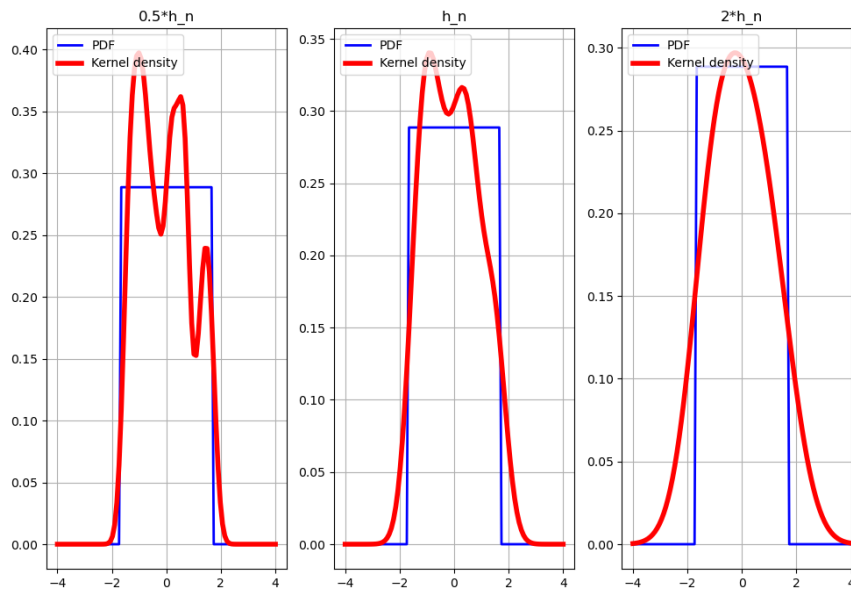


Рис. 26: Ядерная функция плотности для равномерного распределения $n = 100$
 $n=100$ with window width $h=[0.5, 1, 2]$



6.5 Эмпирические функции распределения

Рис. 27: Эмпирическая функция для стандартного нормального распределения

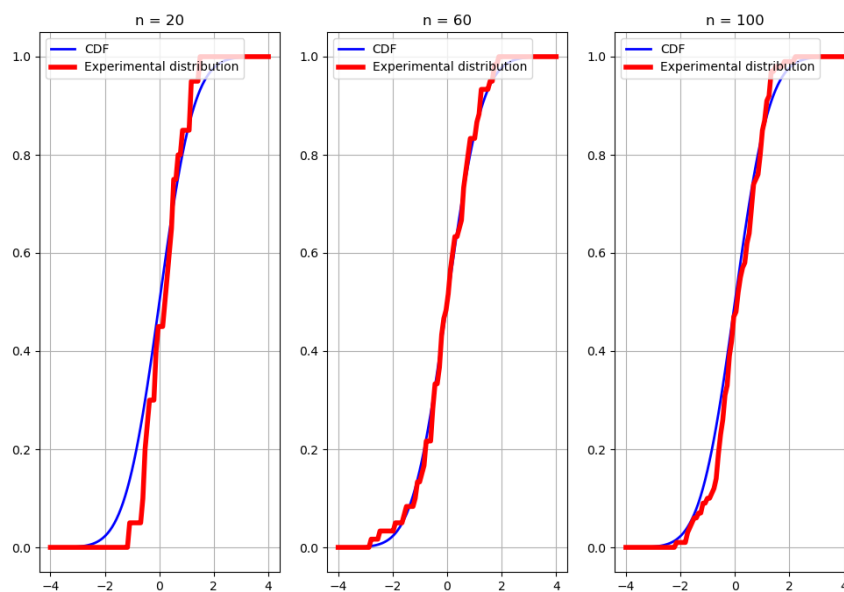


Рис. 28: Эмпирическая функция для стандартного распределения Лапласа

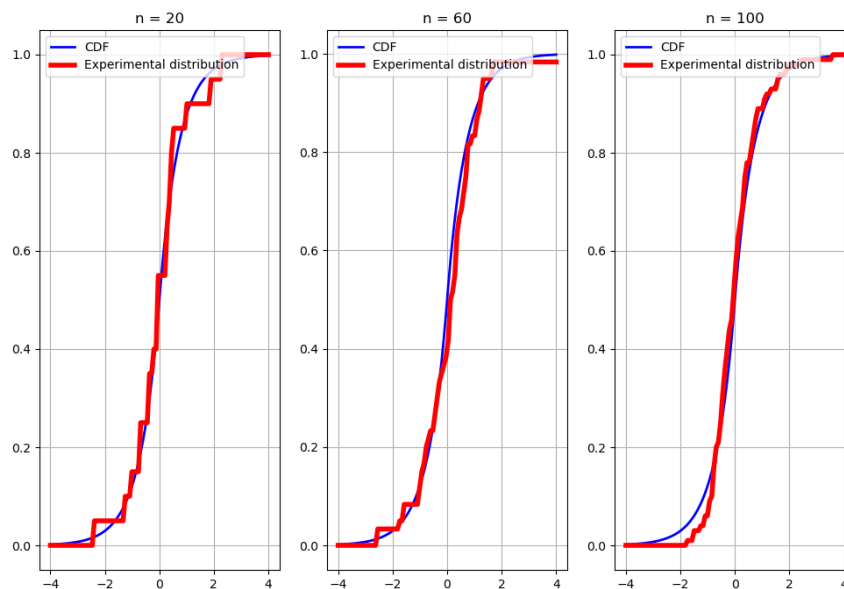


Рис. 29: Эмпирическая функция для стандартного распределения Коши

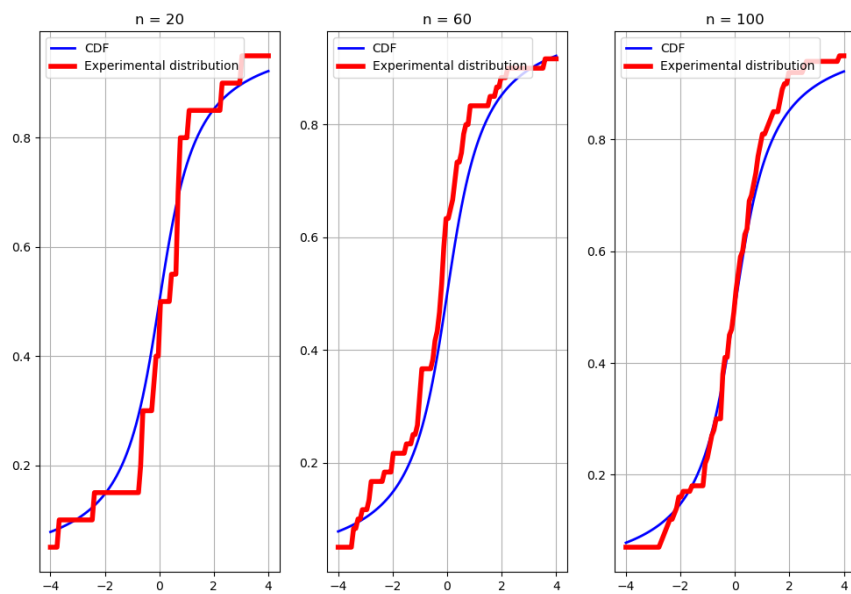


Рис. 30: Эмпирическая функция для распределения Пуассона

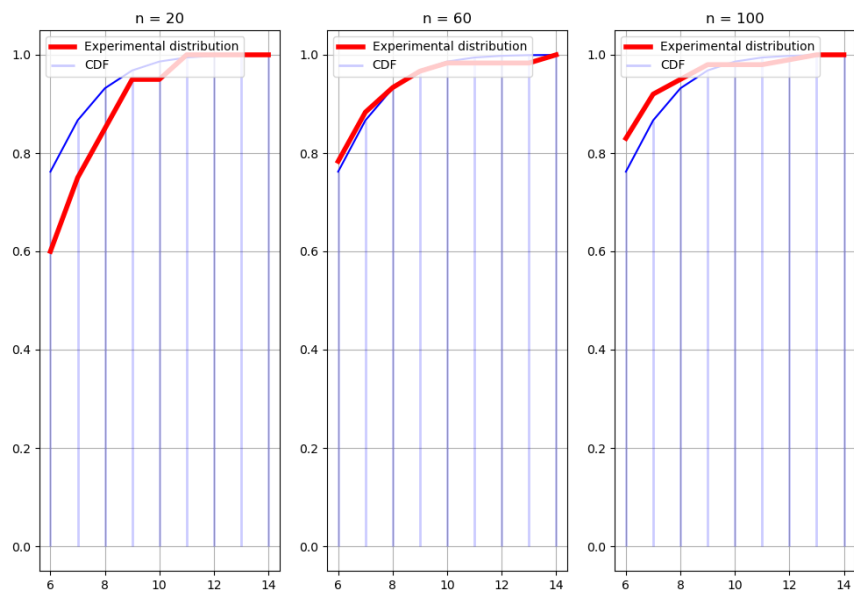
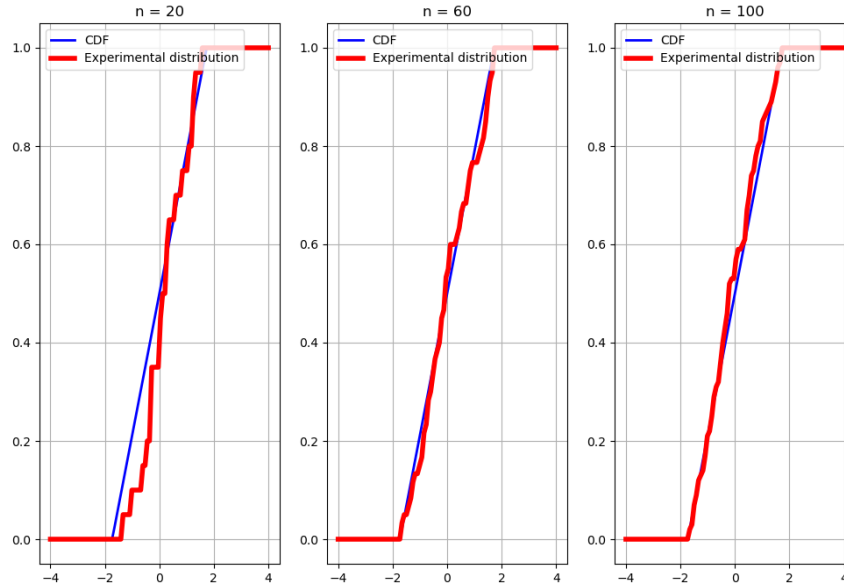


Рис. 31: Эмпирическая функция для равномерного распределения



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

Из графиков наглядно видно выполнение свойства гистограммы – площадь столбца гистограммы, построенного над произвольным интервалом группировки, с ростом объёма выборки сближается с площадью области под графиком плотности над этим же интервалом.

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

$$\bar{x} < Z_{tr} < Z_Q < med\ x < Z_R$$

2. Распределение Лапласа (коэффициент масштаба $\sqrt{2}$ коэффициент сдвига равен нулю)

$$Z_Q < Z_{tr} < \bar{x} < Z_R < med\ x$$

3. Стандартное распределение Коши

$$med\ x < Z_Q < \bar{x} < Z_{tr} < Z_R$$

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

$$Z_R < med\ x < Z_Q < Z_{tr} < \bar{x}$$

5. Распределение Пуассона (значение мат ожидания равно 5)

$$Z_Q < med\ x < \bar{x} < Z_{tr} < Z_R$$

8.3 Боксплот Тьюки

Экспериментально полученные оценки доли выбросов стремятся к теоретическим с ростом размера выборки, Значение для средней доли выбросов было ограничено первым значащим разрядом в значении дисперсии. Можно вывести соотношение между процентами выбросов:

$$uniform < normal < poisson < laplace < cauchy \quad (15)$$

По полученным данным видно, что наименьший процент выбросов у равномерного распределения, а наибольший процент выбросов у распределения Коши, при чем значения этих выбросов могут отклоняться от выборочного среднего на порядки.

8.4 Эмпирические функции и ядерные оценки

Эмпирическая функция лучше приближает эталонную функцию с ростом объёма выборки.

Ядерная оценка функции плотности вероятности с выбранным нормальным ядром лучше всего приближает распределения, близкие к нормальному, с ростом размера выборки качество оценки растёт. Исключением является распределение Коши, так как на относительно больших выборках крайне велики выбросы, которые сильно ухудшают приближение по вариационному ряду. Для улучшения качества приближения можно уменьшить ширину полосы.

Также существует проблема точного приближения функции плотности Лапласа с помощью нормального ядра, так как, несмотря на схожесть структуры этих двух функций (стандартного нормального и Лапласа к коэф. $1/\sqrt{2}$ распределений), разность скорости сходимости к точке мат. ожидания (излом Лапласа в этой точке) неизбежно будет давать погрешность.

С тем же связана проблема приближения и равномерного распределения, площадь графика которого равна прямоугольнику. Чем уже равномерное распределение, тем лучше ядерная оценка с увеличением размера выборки. Уменьшение ширины полосы скорее приводит к ухудшению результата, так как только увеличивает влияние неравномерности случайной выборки.

9 Список литературы

- [1] Модуль numpy - <https://physics.susu.ru/vorontsov/language/numpy.html>
- [2] Выборочное среднее - https://en.wikipedia.org/wiki/Sample_mean_and_covariance
- [3] Выборочная медиана - http://femto.com.ua/articles/part_1/2194.html
- [4] Полусумма экстремальных значений - <https://studopedia.info/8-56888.html>
- [5] Квартили - <https://studfiles.net/preview/2438125/page:13/>
- [6] Усечённое среднее - <https://ole-olesko.livejournal.com/15773.html>
- [7] Выборочная медиана - http://femto.com.ua/articles/part_1/2194.html
- [8] Квартили - <https://studfiles.net/preview/2438125/page:13/>
- [9] Боксплот - https://en.wikipedia.org/wiki/Box_plot
- [10] Модуль matplotlib - <https://matplotlib.org/users/index.html>
- [11] Модуль scipy - <https://docs.scipy.org/doc/scipy/reference/>
- [12] Н. И. Чернова, <https://nsu.ru/mmф/tvims/chernova/ms/lec/node4.html>, 2002
- [13] Victor, <https://www.mql5.com/ru/articles/396>, 2012
- [14] Nathaniel E. Helwig, <http://users.stat.umn.edu/~helwig/notes/den-Notes.pdf>, 2017