Санкт-Петербургский Политехнический
Университет
им. Петра Великого

Институт прикладной математики и механики

Кафедра "Прикладная математика"

Отчёт
Лабораторная работа №1 – 4
по дисциплине
"Математическая статистика"

Выполнил студент:
Мальцов Дмитрий Дмитриевич
группа: 3630102/70401

Проверил:
к.ф-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2020 год

Содержание

Стр.

Список иллюстраций

Список таблиц

1 Постановка задачи

Для 5-ти рапределений:
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 10, 50 и 1000 элементов. Построить на одном рисунке гистограмму и график плотности распределения.

2 Теория

2.1 Распределения

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

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

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

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \tag{4}$$

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

2.2 Гистограмма

2.2.1 Определение

Гистограмма в математической статистике — это функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него.

2.2.2 Графическое описание

Графически гистограмма строится следующим образом. Сначала множество значений, которое может принимать элемент выборки, разбивается на несколько интервалов. Чаще всего эти интервалы берут одинаковыми, но это не является строгим требованием. Эти интервалы откладываются на горизонтальной оси, затем над каждым рисуется прямоугольник. Если все интервалы были одинаковыми, то высота каждого прямоугольника пропорциональна числу элементов выборки, попадающих в соответствующий интервал. Если интервалы разные, то высота прямоугольника выбирается таким образом, чтобы его площадь была пропорциональна числу элементов выборки, которые попали в этот интервал.

2.2.3 Использование

Гистограммы применяются в основном для визуализации данных на начальном этапе статистической обработки. Построение гистограмм используется для получения эмпирической оценки плотности распределения случайной величины. Для построения гистограммы наблюдаемый диапазон изменения случайной величины разбивается на несколько интервалов и подсчитывается доля от всех измерений, попавшая в каждый из интервалов. Величина каждой доли, отнесенная к величине интервала, принимается в качестве оценки значения плотности распределения на соответствующем интервале.

2.3 Выборочные числовые характеристики

1. Выборочное среднее [?]:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

2. Выборочная медиана [?]:

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

3. Полусумма экстремальных значений [?]:

$$Z_R = \frac{1}{2}\left(x_1 + x_n\right) \tag{8}$$

4. Полусумма квартилей [?]:

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

5. Усечённое среднее [?]:

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

2.4 Боксплот Тьюки

2.4.1 Определение

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

2.4.2 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.

2.4.3 Использование

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:

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

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

2.5 Эмпирическая функция распределения

2.5.1 Статистический ряд

Статистическим рядом называется последовательность различных элементов выборки$z_1, ..., z_k$, расположенных в возрастающем порядке с указанием частот $n_1, ..., n_k$, с которыми эти элементы содержатся в выборке. Статистический ряд обычно записывается в виде таблицы.

2.5.2 Определение

Эмпирической (выборочной) функцией распределения (э. ф. р.) называется относительная частота события $X < x$, полученная по данной выборке:

$$F_n^*(x) = P^*(X < x)$$

.

2.5.3 Описание

Для получения относительной частоты $P^*(X < x)$ просуммируем в статистическом ряде, построенном по данной выборке, все частоты n_i , для которых элементы z_i статистического ряда меньше x. Тогда $P^*(X < x) = \frac{1}{n}\sum_{z_i<x} n_i$. Получаем

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

F^*(x) — функция распределения дискретной случайной величины X^* , заданной таблицей распределения. Эмпирическая функция распределения является оценкой, т. е. приближённым значением, генеральной функции распределения.

$$F_n^*(x) \approx F_X(x)$$

2.6 Оценки плотности вероятности

2.6.1 Определение

Оценкой плотности вероятности f(x) называется функция $\hat{f}$(x), построенная на основе выборки, приближенно равная f(x)

$$\hat{f}(x) \approx f(x)$$

2.6.2 Ядерные оценки

ёПредставим оценку в виде суммы с числом слагаемых, равным объёму выборки:

$$\hat{f_n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x - x_i}{h_n})$$

Здесь функция K(u), называемая ядерной (ядром), непрерывна и является плотностью вер-ти, $x_1, ..., x_n$ - элементы выборки, h_n - любая последовательноть положительных чисел, обладающая свойствами:
1)при n-> ∞ h_n->0
2)$\frac{h_n}{n^{-1}}$ -> ∞ ,когда n->∞.
Такие оценки называются непрерывными ядерными.

Замечание. Свойство, означающее сближение оценки с оцениваемой величиной при n -> ∞ в каком-либо смысле, называется состоятельностью оценки.
Если плотность f(x) кусочно-непрерывная, то ядерная оценка плотности является состоятельной при соблюдении условий, накладываемых на параметр сглаживания h_n, а также на ядро K(u).
Гауссово(нормальное) ядро

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

Правило Сильвермана

$$h_n = 1.06\hat{\sigma}n^{-0.2}$$

, где $\hat{\sigma}$ - выборочное среднее отклонение.

3 Реализация

Для генерации выборки был использован *Python* 3.7: модуль *random* библиотеки *numpy* для генерации случайных чисел с различными распределениями и библиотека *matplotlib* для построения графиков и гистограмм.

4 Результаты

4.1 Гистограмма и график плотности распределения

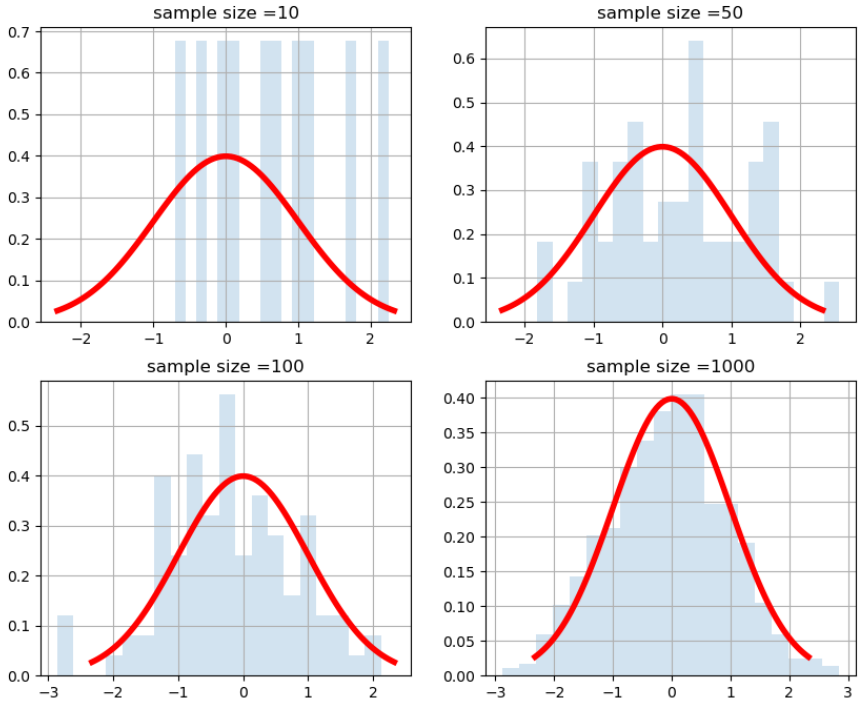

Рис. 1: Нормальное распределение

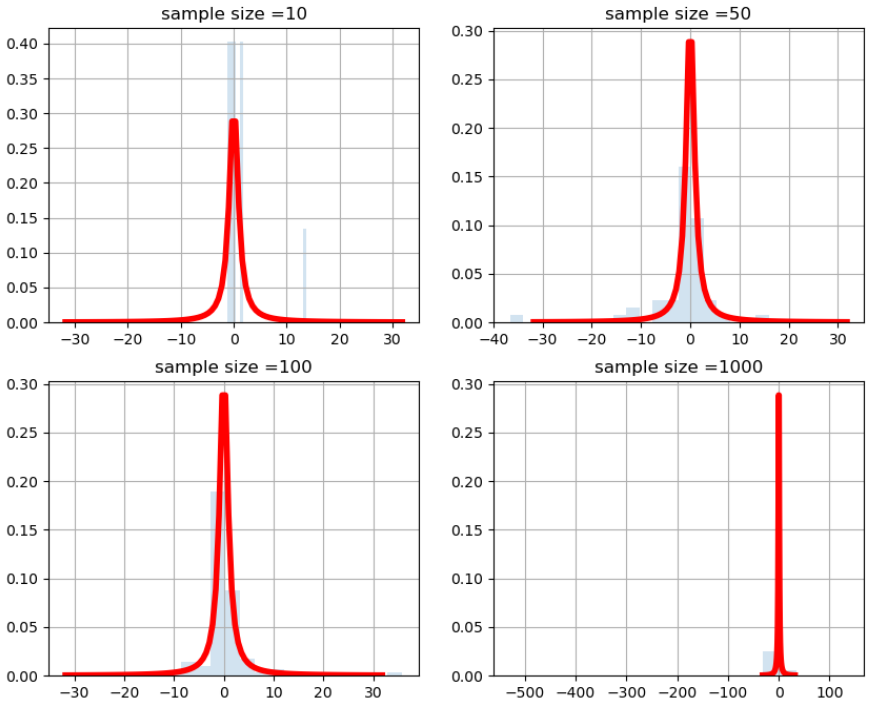

Рис. 2: Распределение Коши

4.2 Характеристики полложения и рассеяния

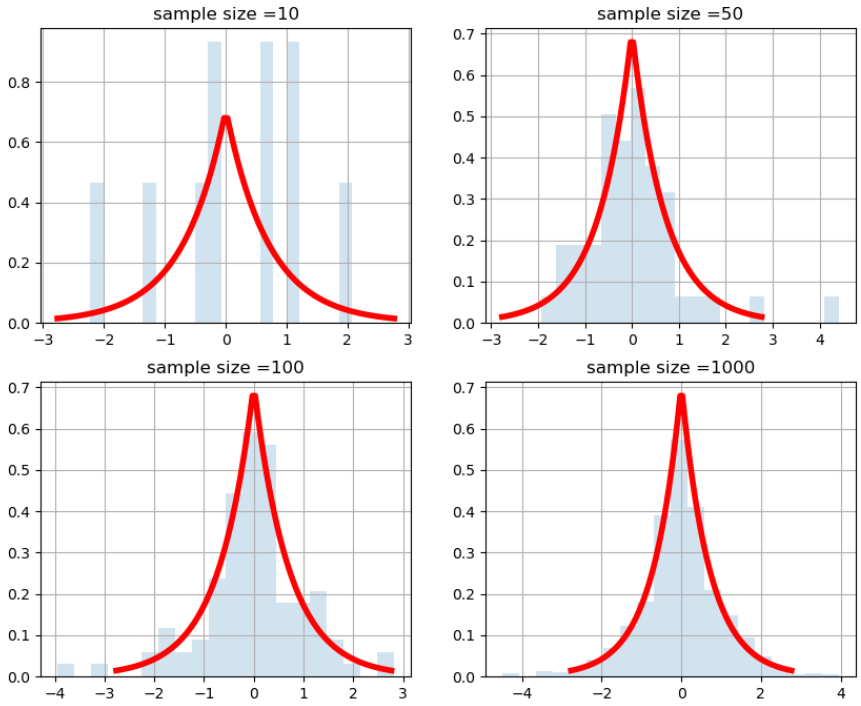

Рис. 3: Распределение Лапласа

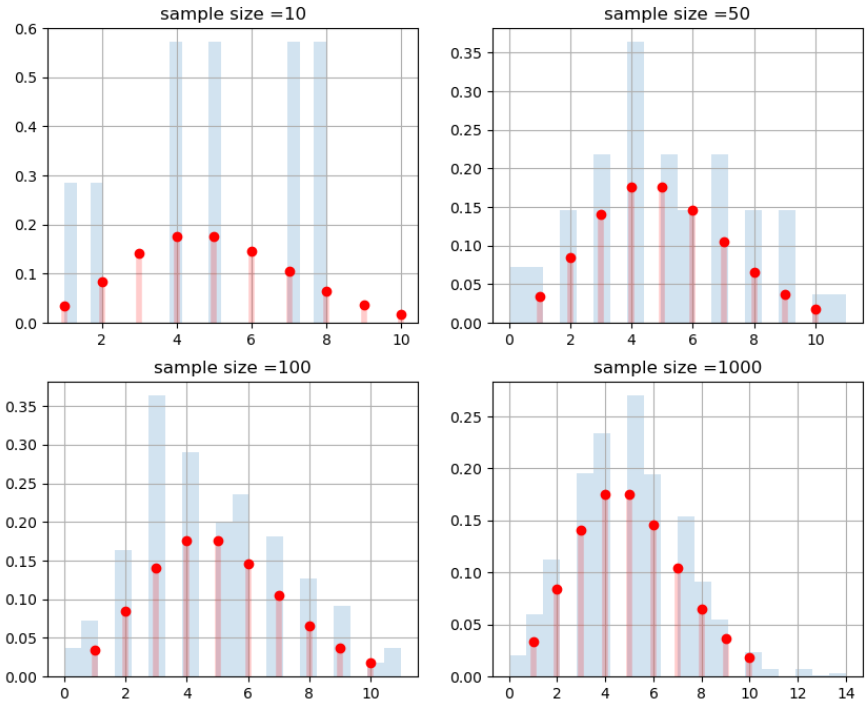

Рис. 4: Распределение Пуассона

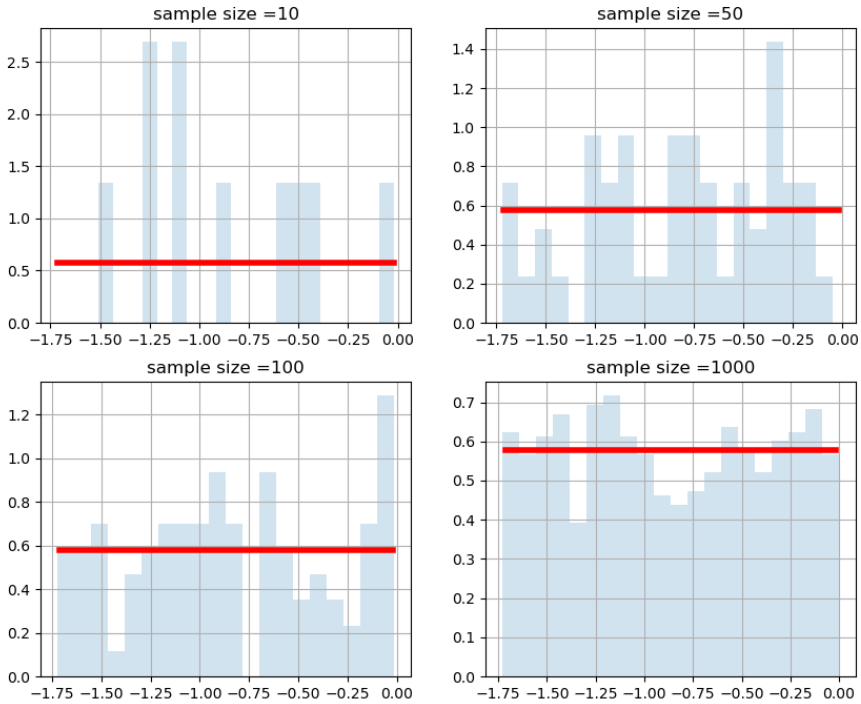

Рис. 5: Равномерное распределение

Таблица 1: Стандартное нормальное распределение.

$n = 10$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.01	-0.01	-0.01	0.0	-0.01
$D =$	0.099	0.136	0.180	0.117	0.113
$n = 50$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.00	-0.01	$0,00$	-0.01	0.01
$D =$	0.0197	0.03147	0.1196	0.0243	0.0230
$n = 1000$	Z_R	$Z_{tr},\ r = \frac{n}{4}$	Z_Q	average	med
$E =$	0.007	0.000	-0.001	-0.002	-0.002
$D =$	0.06869	0.00115	0.00131	0.00100	0.00154

Таблица 2: Стандартное распределение Коши.

$n = 10$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	-0.68	0.02	-2.11	-0.00	-0.02
$D =$	412.916	0.341	8385.756	0.803	0.477
$n = 50$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	-22.042	-0.005	8.862	-0.010	0.003
$D =$	474719.9636	0.0522	1746206.4756	0.1028	0.0612
$n = 1000$	Z_R	average	med	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	1150.2818	0.2832	0.0003	-0.0015	-0.0029
$D =$	2196736948.38102	272.36030	0.00257	0.00495	0.00257

Таблица 3: Распределение Лапласа.

$n = 10$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	-0.00	-0.01	0.03	-0.02	-0.00
$D =$	0.100	0.069	0.425	0.094	0.071
$n = 50$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	-0.004	-0.000	0.022	0.005	0.004
$D =$	0.0190	0.0135	0.4011	0.0201	0.0125
$n = 1000$	Z_Q	average	med	$Z_{tr},\ r = \frac{n}{4}$	Z_R
$E =$	0.0012	0.0000	0.0000	0.0000	-0.0361
$D =$	0.00095	0.00102	0.00052	0.00060	0.40623

Таблица 4: Равномерное распределение.

$n = 10$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.01	-0.00	-0.00	-0.02	0.00
$D =$	0.098	0.220	0.041	0.139	0.158
$n = 50$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	0.001	0.005	0.000	-0.004	-0.009
$D =$	0.0211	0.0560	0.0023	0.0299	0.0342
$n = 1000$	Z_Q	average	Z_R	$Z_{tr},\ r = \frac{n}{4}$	med
$E =$	0.0021	0.0001	0.0000	0.0000	-0.0012
$D =$	0.00000	0.00103	0.00143	0.00202	0.00294

Таблица 5: Распределение Пуассона.

$n = 10$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	10.04	9.89	10.31	9.91	9.89
$D =$	1.005	1.493	1.793	1.111	1.144
$n = 50$	average	med	Z_R	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	9.969	9.835	10.748	9.928	9.874
$D =$	0.1963	0.3566	1.1819	0.2816	0.2671
$n = 1000$	Z_R	average	med	Z_Q	$Z_{tr},\ r = \frac{n}{4}$
$E =$	11.6572	10.0000	9.9971	9.9951	9.8530
$D =$	0.64385	0.00891	0.00299	0.00252	0.01182

4.3 Боксплот Тьюки

Ввелём на оси y следующие обозначения:
1 соответствует выборке из 20-ти элементов
2 - соответствует выборке из 100 элементов

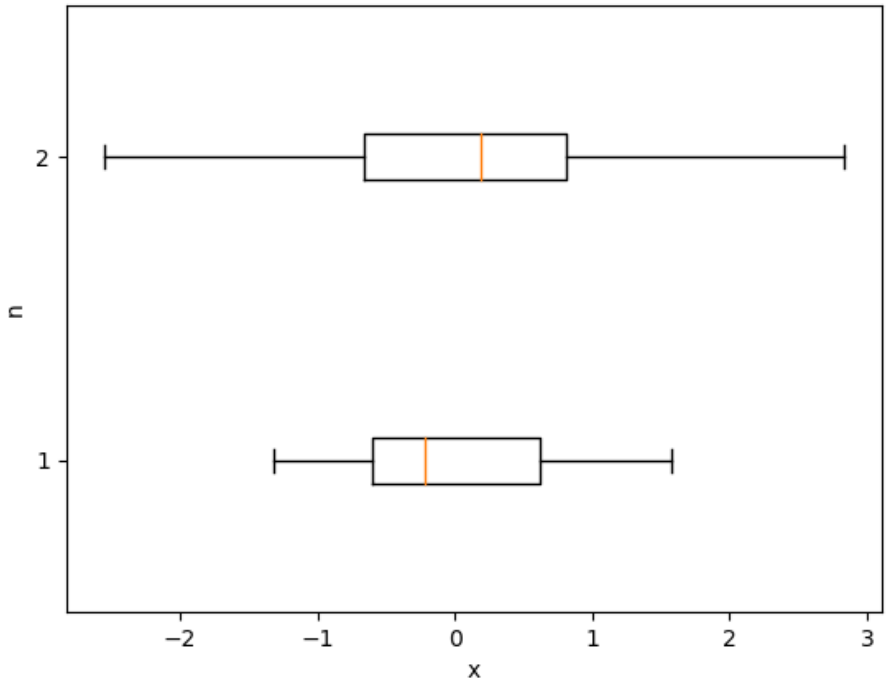

Рис. 6: Нормальное распределение

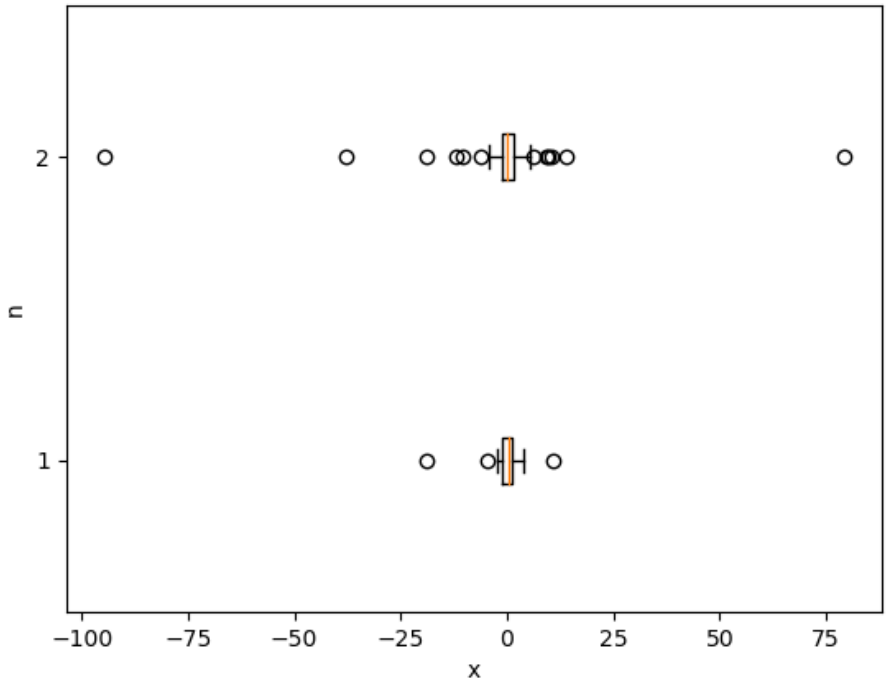

Рис. 7: Распределение Коши

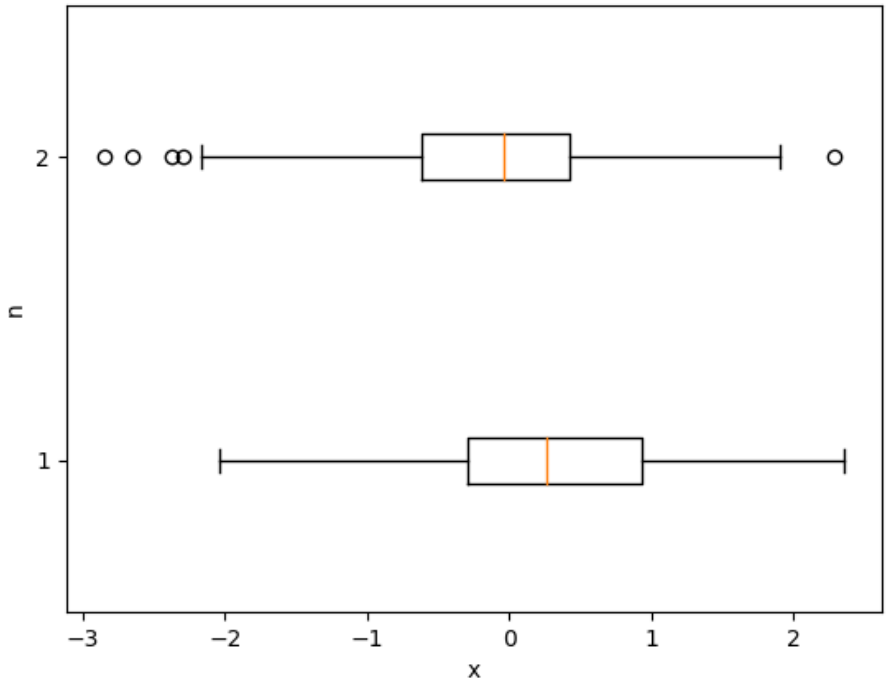

Рис. 8: Распределение Лапласа

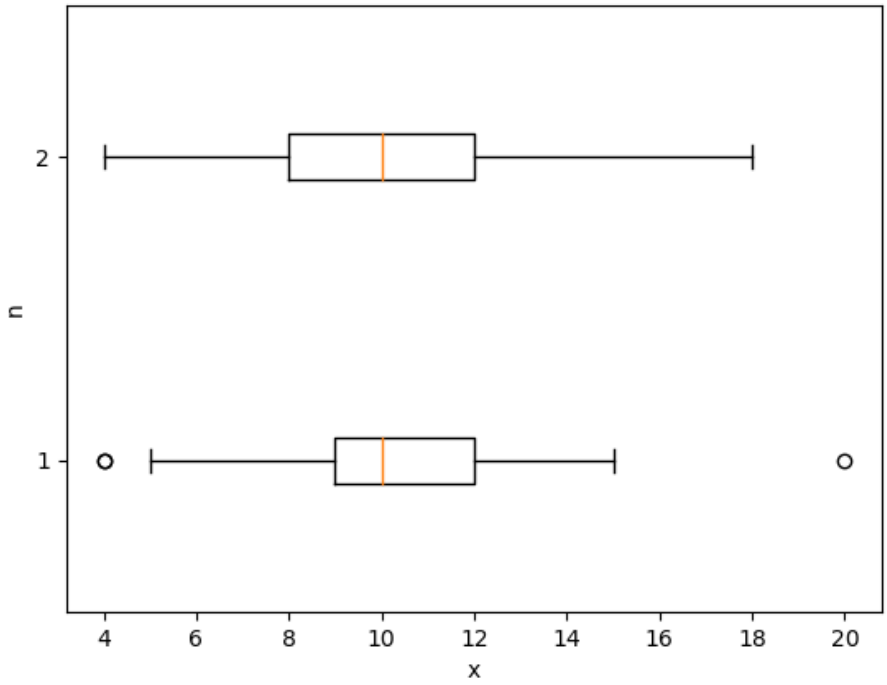

Рис. 9: Распределение Пуассона

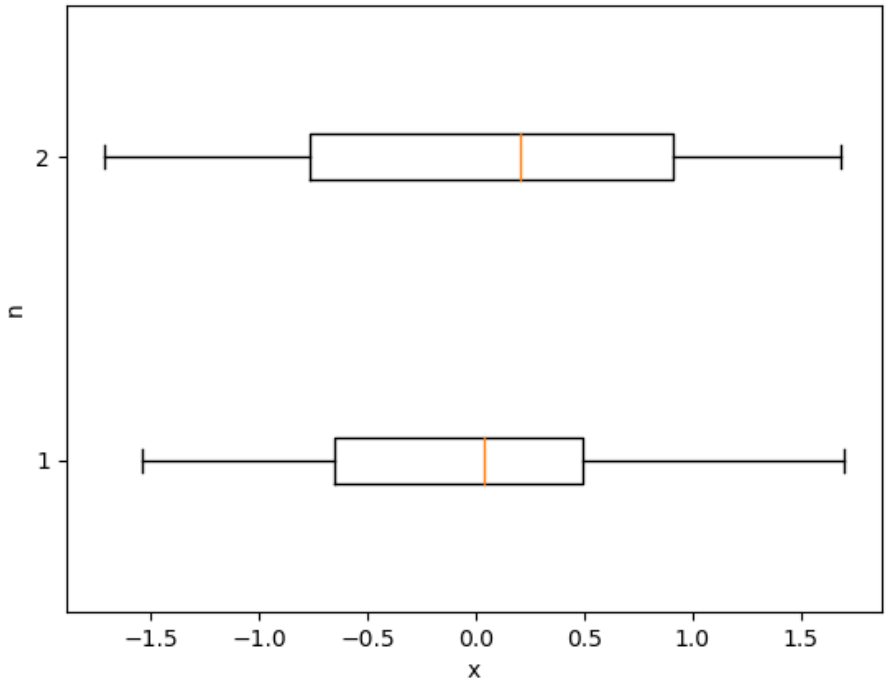

Рис. 10: Равномерное распределение

Таблица 6: Выбросы различных распределений в зависимости от выборки

Выборка	Процент выбросов	Дисперсии
Нормальное		
n = 20	2	0.0021
n = 100	1	0.0001
Коши		
n = 20	15	0.0053
n = 100	15	0.0011
Лапласа		
n = 20	7	0.0042
n = 100	6	0.0093
Пуассона		
n = 20	3	0.0019
n = 100	1	0.0002
Равномерное		
n = 20	0	0.0001
n = 100	0	0.0000

4.4 Эмпирическая функция распределений

Красный график - эталонная функция. Синий график - эмпирическая соответственно.

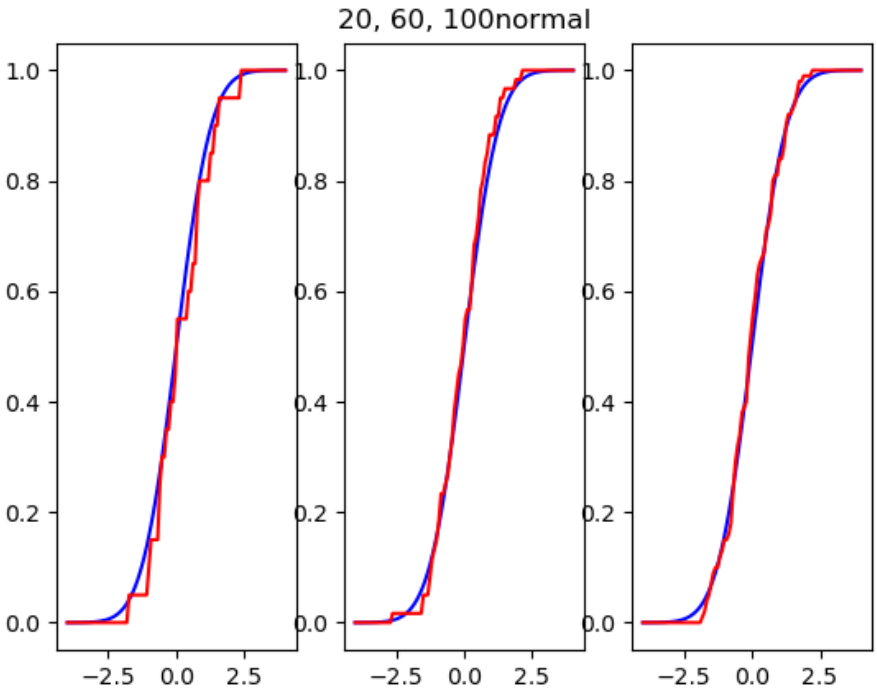

Рис. 11: Нормальное распределение

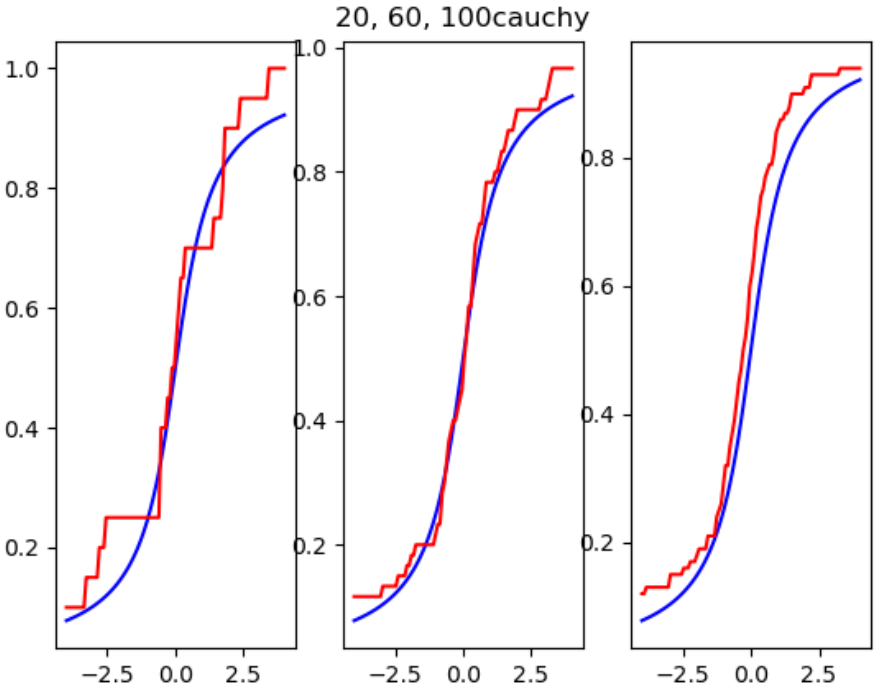

Рис. 12: Распределение Коши

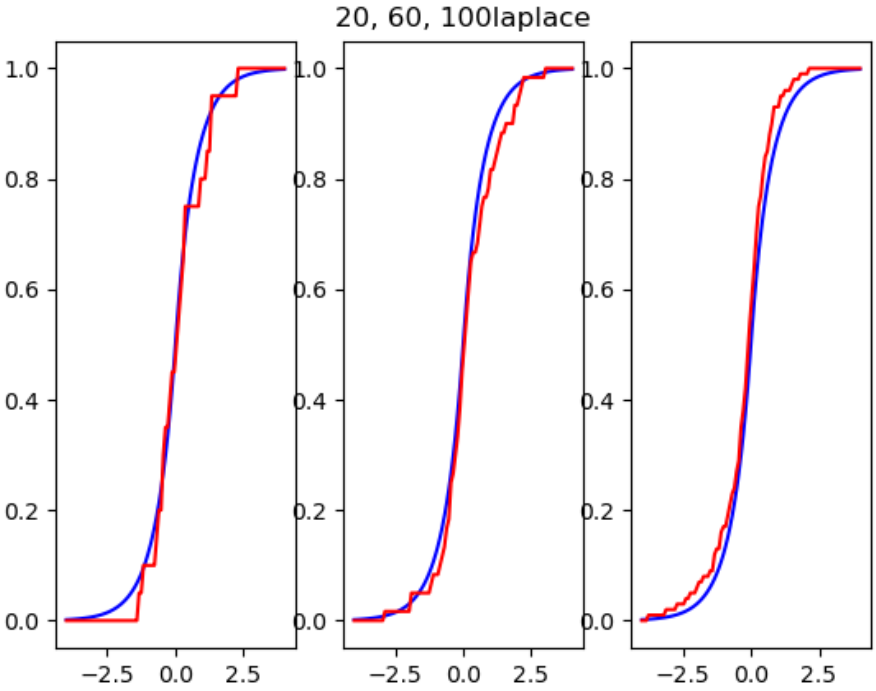

Рис. 13: Распределение Лапласа

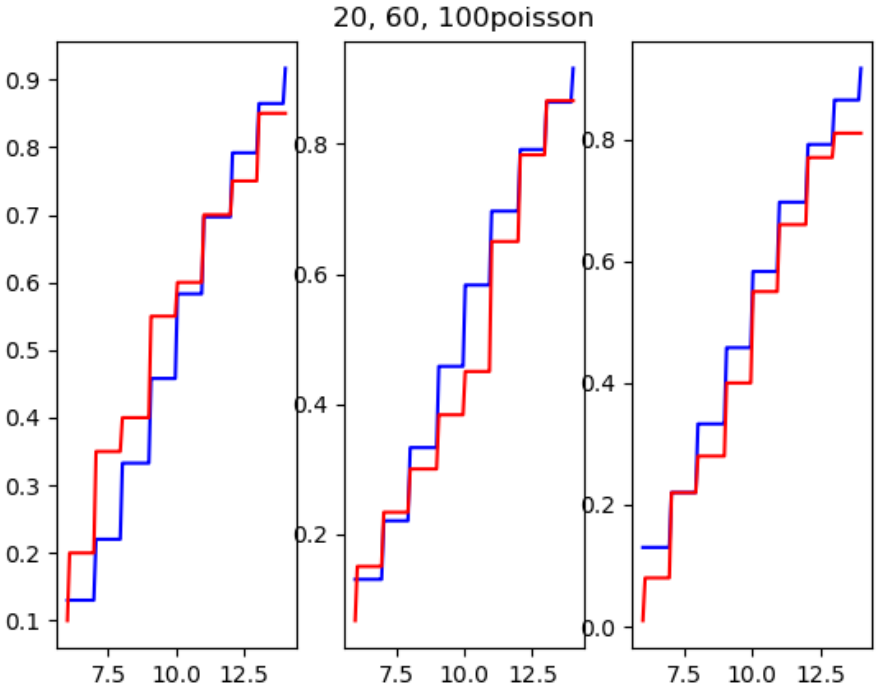

Рис. 14: Распределение Пуассона

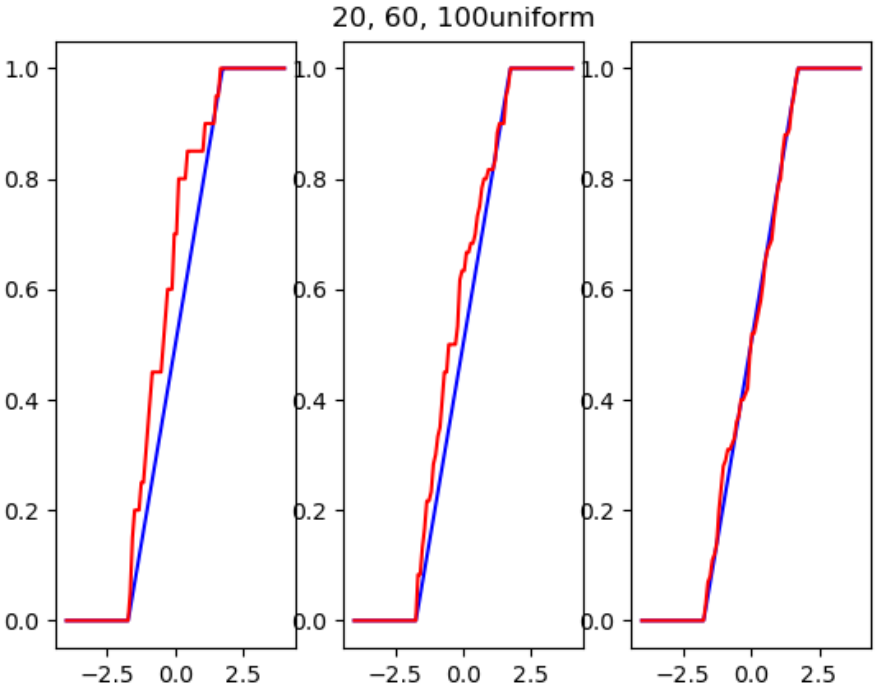

Рис. 15: Равномерное распределение

4.5 Ядерные оценки плотности распределений

Черный цвет - эталонная функция. Красный цвет - эмпирическая.

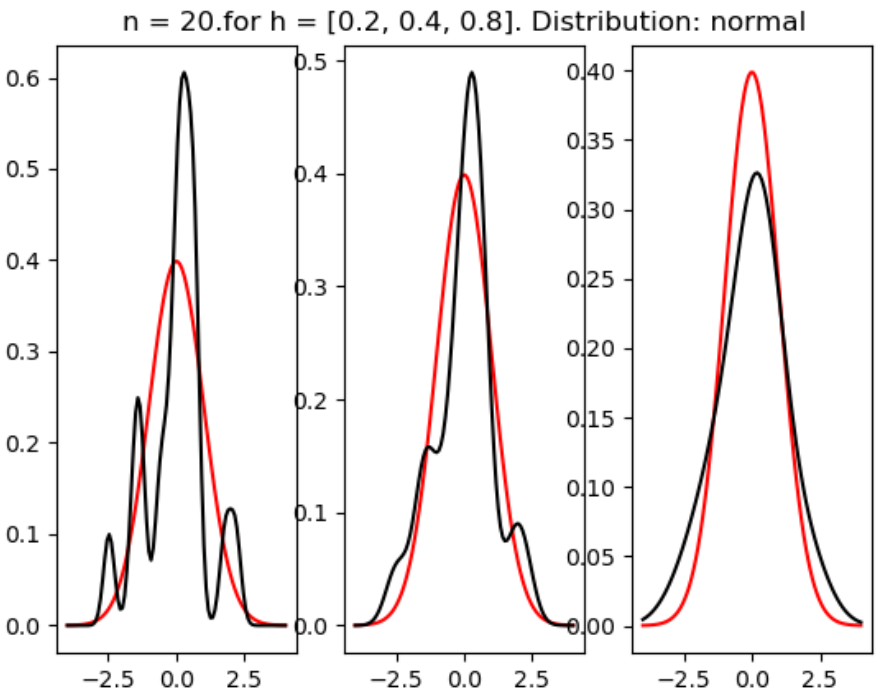

Рис. 16: Нормальное распределение n = 20

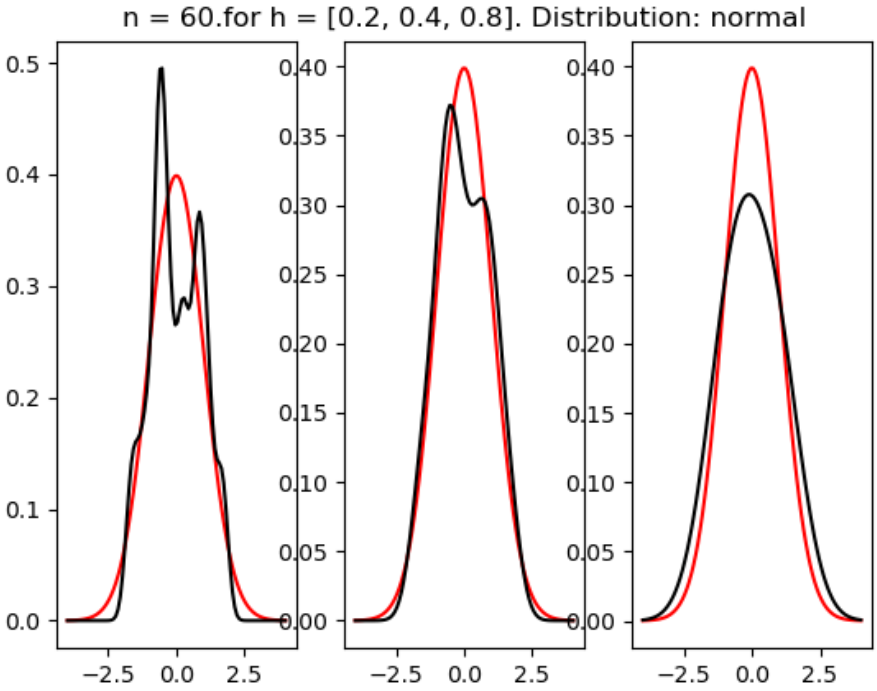

Рис. 17: Нормальное распределение n = 60

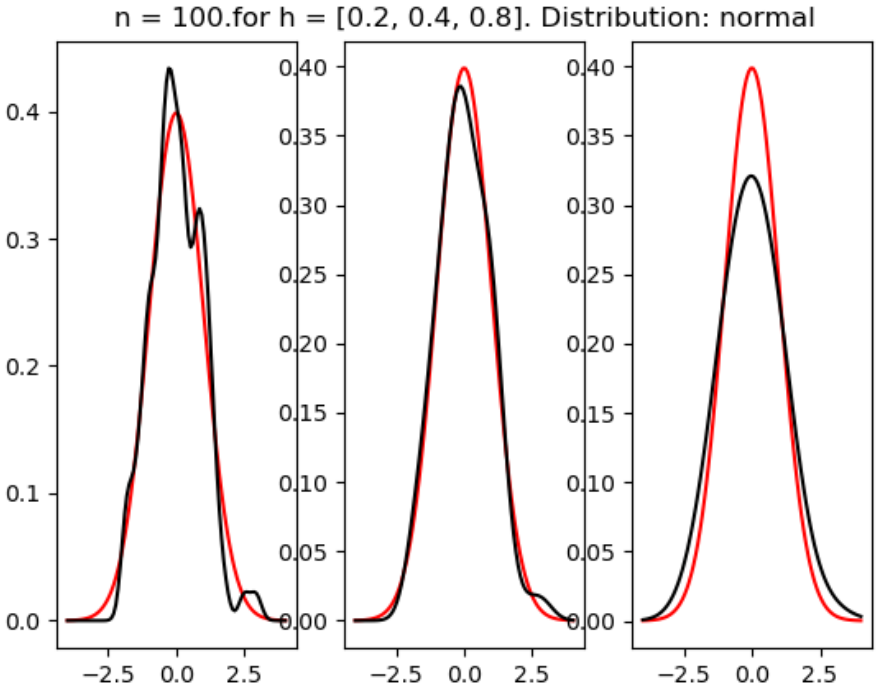

Рис. 18: Нормальное распределение n = 100

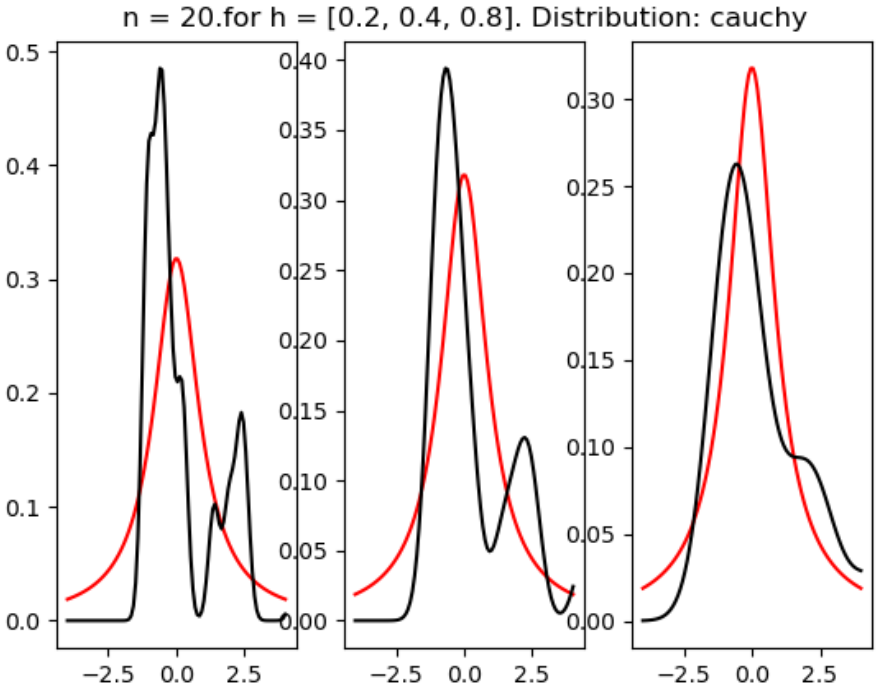

Рис. 19: Распределение Коши n = 20

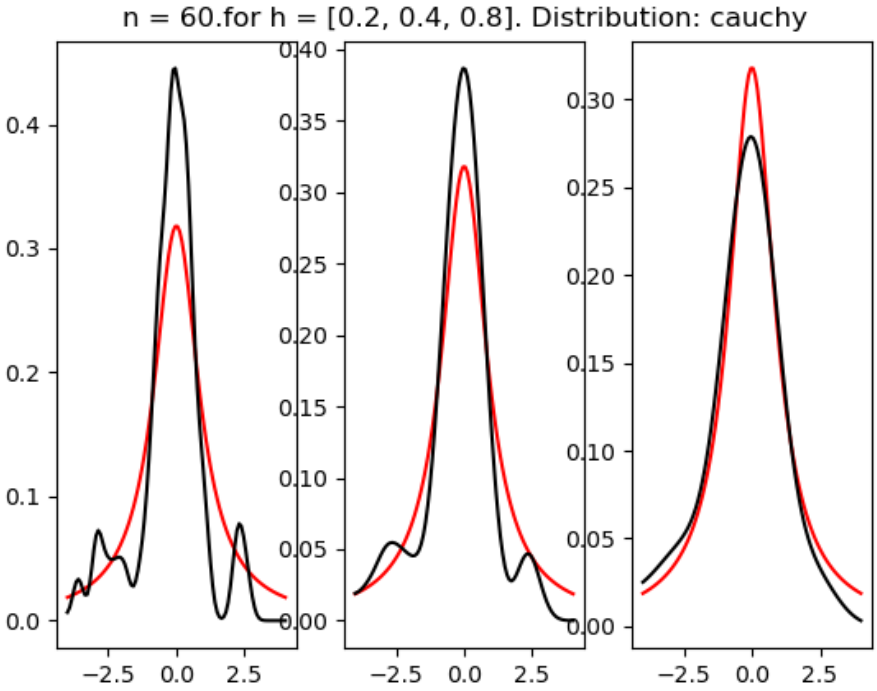

Рис. 20: Распределение Кошин = 60

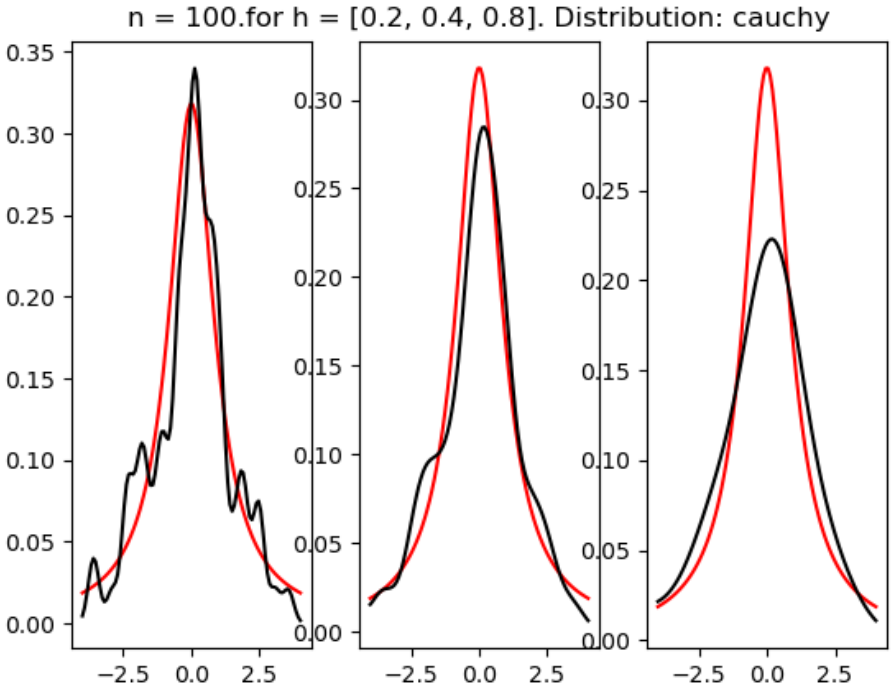

Рис. 21: Распределение Коши n = 100

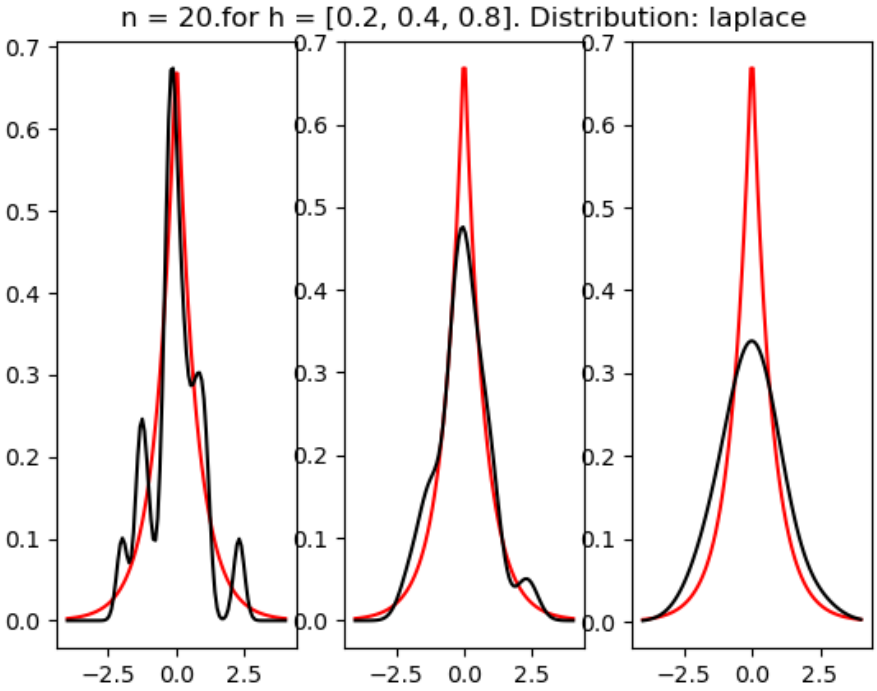

Рис. 22: Распределение Лапласа n = 20

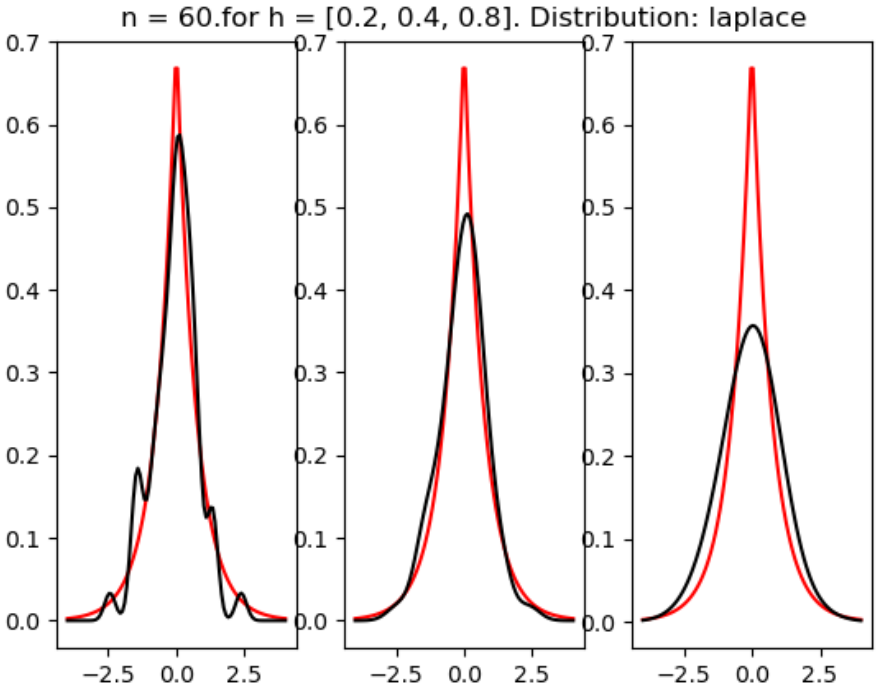

Рис. 23: Распределение Лапласа n = 60

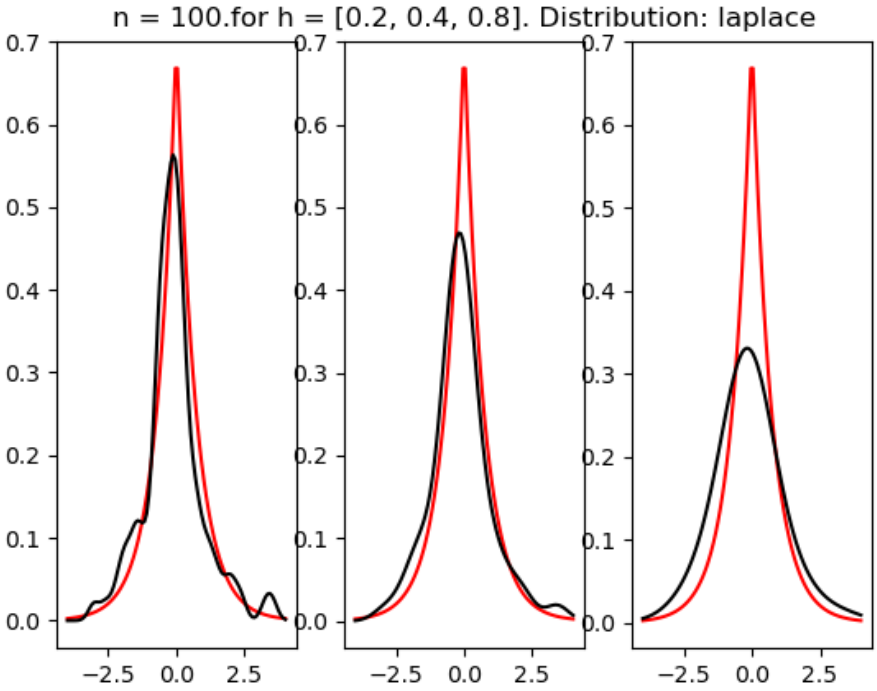

Рис. 24: Распределение Лапласа n = 100

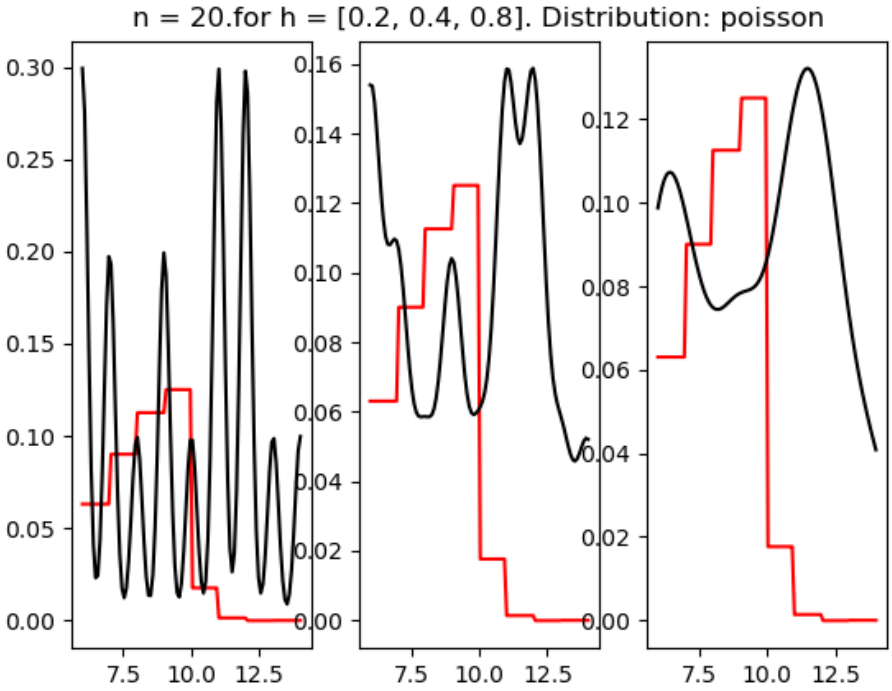

Рис. 25: Распределение Пуассона n = 20

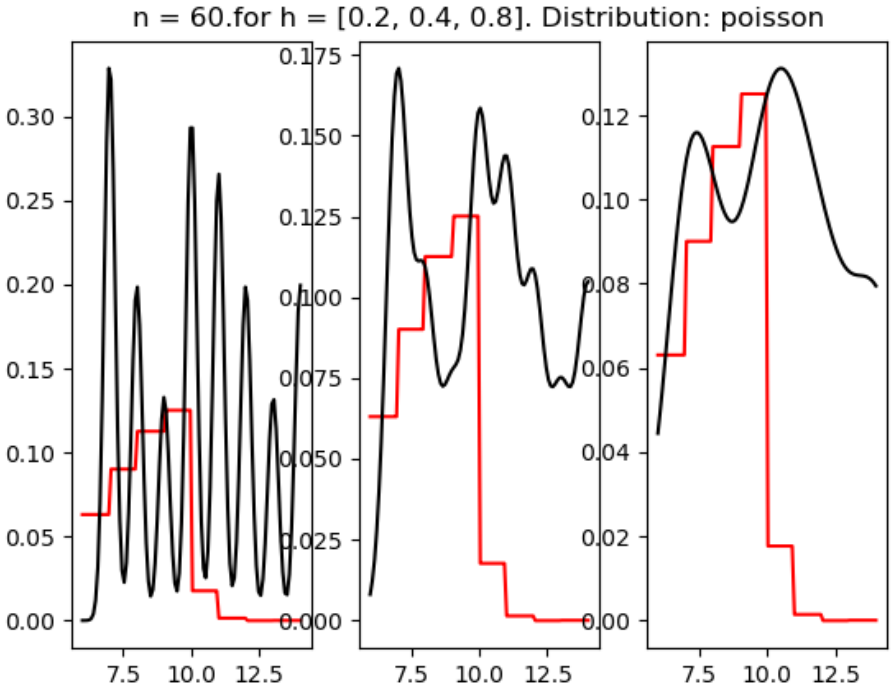

Рис. 26: Распределение Пуассона n = 60

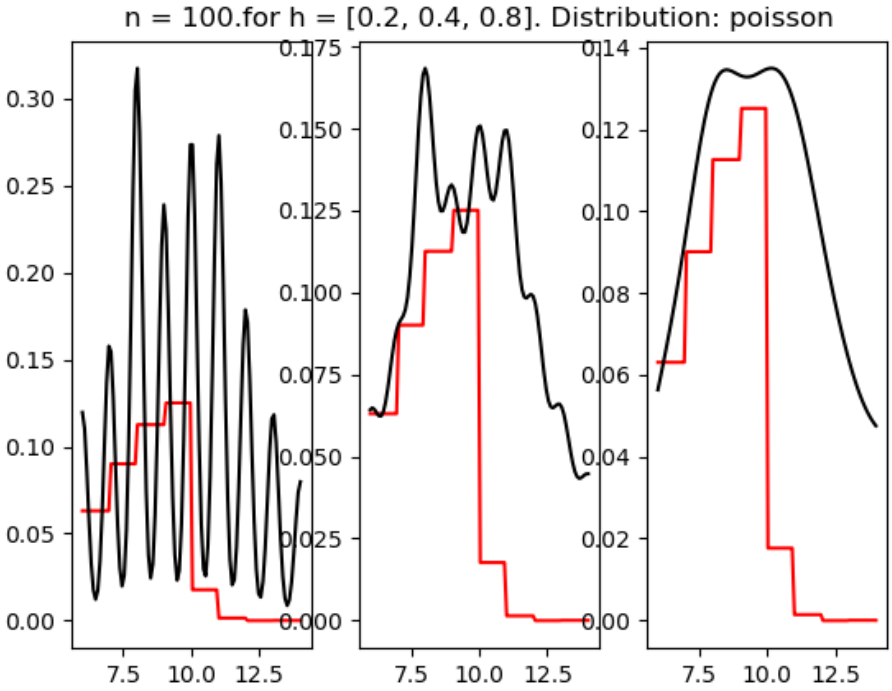

Рис. 27: Распределение Пуассона n = 100

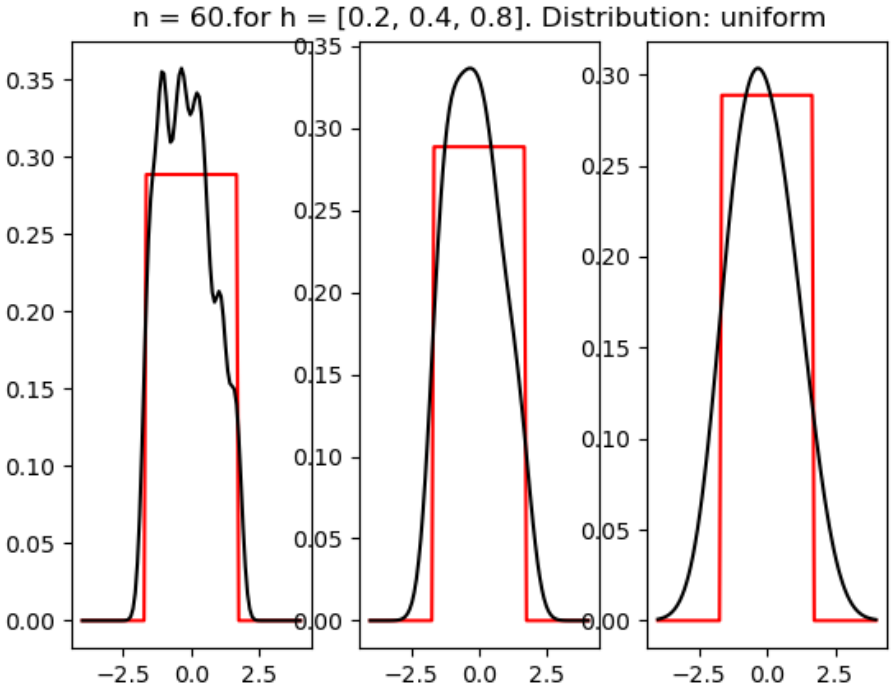

Рис. 28: Равномерное распределение n = 60

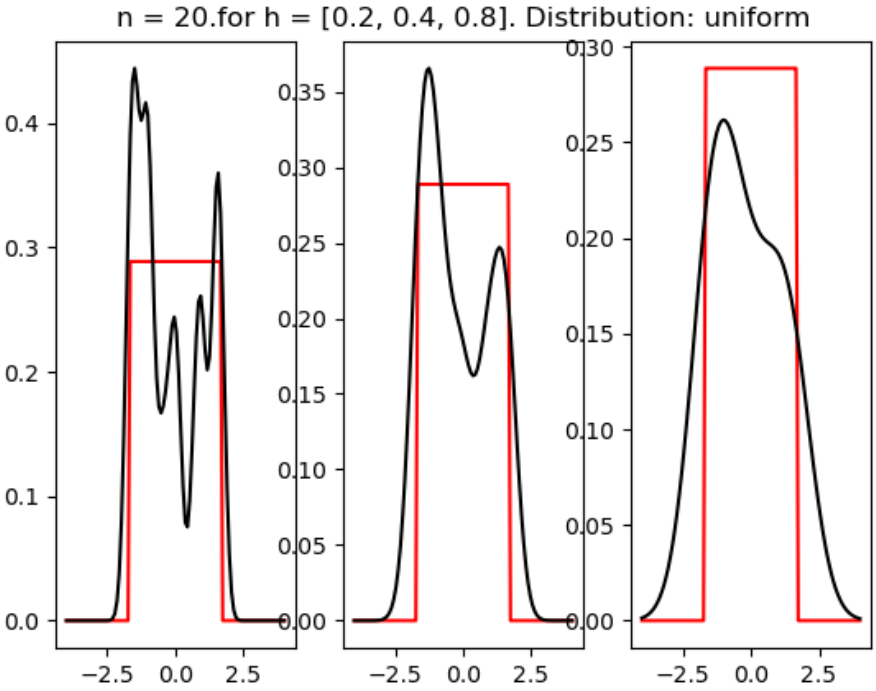

Рис. 29: Равномерное распределение n = 20

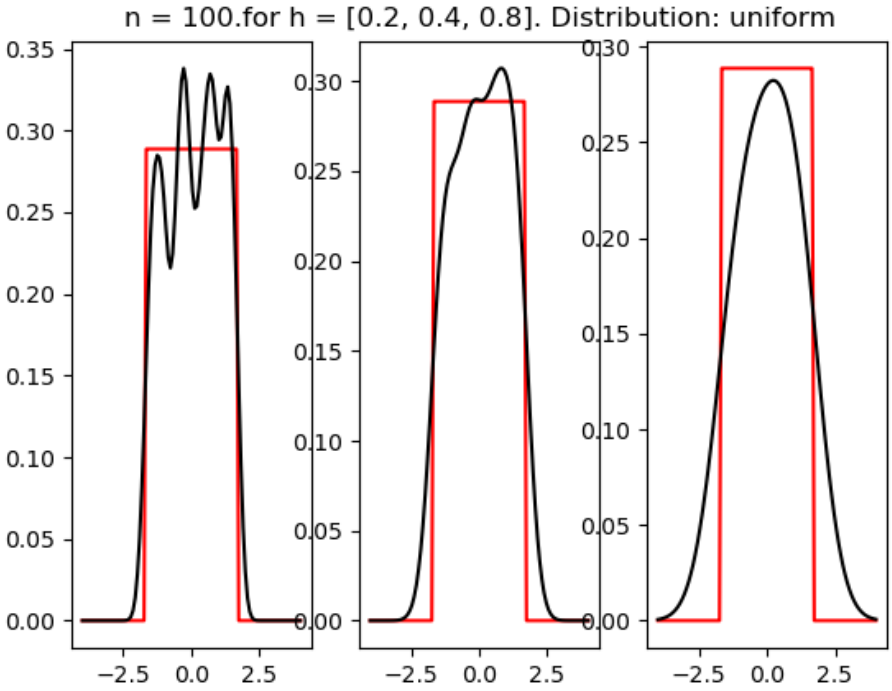

Рис. 30: Равномерное распределение n = 100

5 Обсуждение

5.1 Гистограмма и график распределения

Благодаря полученным графикам можно увидеть, что: чем больше выборка, тем график плотности более близок к гистограмме. При малой выборке, наблюдается скачок значений в гистограмме.

5.2 Характеристики положения и рассеяния

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.

5.3 Доля и теоретическая вероятность выбросов

5.3.1 Анализ данных

Из экспериментально полученных данных можно вывести соотношение между процентами выбросов:
равномерное распределение $<$ нормальное распределение $<$ распределение Пуассона $<$ распределение Лапласа $<$ распределение Коши

5.3.2 Сравнение с теоретическими значениями

Полученные экспериментально данные близки к теоретическим и видно, что наименьший процент выбросов у равномерного распределения ,а наибольший у распределения Коши

5.4 Эмпирическая функция и ядерные оценки плотности распределения

Видно, что эмпирическая функция лучше приближает эталонную функцию на больших выборках.

При фиксированной ширине окна точнее приблизить функцию распределения позволяет увеличение выборки.

Найлучшее приближение функции распределения ядерной функции для распределения Лапсласа, нормального распределения и распределения Коши достигается при $h = h_n$, в свою очередь для равномерного распределения наилучшее приближение достигатеся при $h = \frac{h_n}{2}$ и h_n. Для распределения Пуассона при $h = 2h_n$.

6 Литература

Модуль numpy
matplotlib boxplot
Боксплот Тьюки

7 Приложения