Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра «Прикладная математика»

ОТЧЁТ ПО ЛАБОРАТОРНЫМ РАБОТАМ №1-4 ПО ДИСЦИПЛИНЕ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»

Выполнил
студент группы 3630102/70401

Кнодель Юлия Максимовна

Проверил
к. ф.-м. н., доцент

Баженов Александр Николаевич

Санкт-Петербург
2020

Содержание

1 Постановка задачи **4**
- 1.1 Задание 1 . . . 4
- 1.2 Задание 2 . . . 4
- 1.3 Задание 3 . . . 4
- 1.4 Задание 4 . . . 4

2 Теория **5**
- 2.1 Распределения . . . 5
- 2.2 Гистограмма . . . 5
  - 2.2.1 Определение . . . 5
  - 2.2.2 Графическое описание . . . 5
  - 2.2.3 Использование . . . 6
- 2.3 Вариационный ряд . . . 6
- 2.4 Выборочные числовые характеристики . . . 6
  - 2.4.1 Характеристики положения . . . 6
  - 2.4.2 Характеристики рассеяния . . . 7
- 2.5 Боксплот Тьюки . . . 7
  - 2.5.1 Определение . . . 7
  - 2.5.2 Описание . . . 7
  - 2.5.3 Построение . . . 7
- 2.6 Теоретическая вероятность выбросов . . . 7
- 2.7 Эмпирическая функция распределения . . . 8
  - 2.7.1 Статистический ряд . . . 8
  - 2.7.2 Определение . . . 8
  - 2.7.3 Описание . . . 8
- 2.8 Оценки плотности вероятности . . . 9
  - 2.8.1 Определение . . . 9
  - 2.8.2 Ядерные оценки . . . 9

3 Реализация **10**

4 Результаты **10**
- 4.1 Гистограммы и графики плотности распределения . . . 10
- 4.2 Характеристики положения и рассеяния . . . 12
- 4.3 Боксплот Тьюки . . . 13
- 4.4 Доля выбросов . . . 15
- 4.5 Теоретическая вероятность выбросов . . . 15
- 4.6 Эмпирическая функция распределения . . . 15
- 4.7 Ядерные оценки плотности распределения . . . 17

5 Обсуждение **22**
- 5.1 Гистограмма и график плотности распределения . . . 22
- 5.2 Характеристики положения и рассеяния . . . 22
- 5.3 Боксплот Тьюки и доля выбросов . . . 23

5.4 Ядерные оценки плотности распределения 23

6 Приложения **23**

Список иллюстраций

1 Нормальное распределение 10
2 Распределение Коши 10
3 Распределение Лапласа 11
4 Распределение Пуассона 11
5 Равномерное распределение 11
6 Нормальное распределение 13
7 Распределение Коши 13
8 Распределение Лапласа 14
9 Распределение Пуассона 14
10 Равномерное распределение 14
11 Нормальное распределение 15
12 Распределение Коши 16
13 Распределение Лапласа 16
14 Распределение Пуассона 16
15 Равномерное распределение 17
16 Нормальное распределение, $n = 20$ 17
17 Нормальное распределение, $n = 60$ 18
18 Нормальное распределение, $n = 100$ 18
19 Распределение Коши, $n = 20$ 18
20 Распределение Коши, $n = 60$ 19
21 Распределение Коши, $n = 100$ 19
22 Распределение Лапласа, $n = 20$ 19
23 Распределение Лапласа, $n = 60$ 20
24 Распределение Лапласа, $n = 100$ 20
25 Распределение Пуассона, $n = 20$ 20
26 Распределение Пуассона, $n = 60$ 21
27 Распределение Пуассона, $n = 100$ 21
28 Равномерное распределение, $n = 20$ 21
29 Равномерное распределение, $n = 60$ 22
30 Равномерное распределение, $n = 100$ 22

Список таблиц

1 Статистический ряд 8
2 Таблица распределения 9
3 Нормальное распределение 12
4 Распределение Коши 12
5 Распределение Лапласа 12

6 Распределение Пуассона 12
7 Равномерное распределение 13
8 Доля выбросов 15
9 Теоретическая вероятность выбросов 15

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

1.1 Задание 1

Сгенерировать выборки размером 10, 50 и 1000 элементов.
Построить на одном рисунке гистограмму и график плотности распределения.

1.2 Задание 2

Сгенерировать выборки размером 10, 100 и 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных: $\overline{x}, medx, z_R, z_Q, z_{tr}$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

$$E(z) = \overline{z} \tag{1}$$

Вычислить оценку дисперсии по формуле:

$$D(z) = \overline{z^2} - \overline{z}^2 \tag{2}$$

Представить полученные данные в виде таблиц.

1.3 Задание 3

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

1.4 Задание 4

Сгенерировать выборки размером 20, 60 и 100 элементов. 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.

2 Теория

2.1 Распределения

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Лапласа

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

- Распределение Пуассона

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

- Равномерное распределение

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

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 Вариационный ряд

Вариационным рядом называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются. Запись вариационного ряда: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Элементы вариационного ряда $x_{(i)} (i = 1, 2, \ldots, n)$ называются порядковыми статистиками.

2.4 Выборочные числовые характеристики

С помощью выборки образуются её числовые характеристики. Это числовые характеристики дискретной случайной величины X^*, принимающей выборочные значения $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

2.4.1 Характеристики положения

- Выборочное среднее
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{8}$$

- Выборочная медиана
$$medx = \begin{cases} x_{(l+1)} & n = 2l + 1 \\ \frac{x_{(l)} + x_{(l+1)}}{2} & n = 2l \end{cases} \tag{9}$$

- Полусумма экстремальных выборочных элементов
$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{10}$$

- Полусумма квартилей
Выборочная квартиль z_p порядка p определяется формулой
$$z_p = \begin{cases} x_{([np]+1)} & np\text{−дробное} \\ x_{(np)} & np\text{−целое} \end{cases} \tag{11}$$
Полусумма квартилей
$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{12}$$

- Усечённое среднее

$$z_{tr} = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}, r \approx \frac{n}{4} \tag{13}$$

2.4.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{14}$$

2.5 Боксплот Тьюки

2.5.1 Определение

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

2.5.2 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.

2.5.3 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

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

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.

2.6 Теоретическая вероятность выбросов

Встроенными средствами языка программирования R в среде разработки RStudio можно вычислить теоретические первый и третий квартили распределений (Q_1^T и Q_3^T соответственно). По формуле (15) можно вычислить теоретические нижнюю и

верхнюю границы уса (X_1^T и X_2^T соответственно). Выбросами считаются величины x, такие что:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix}$$

Теоретическая вероятность выбросов для непрерывных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T))$$

где $F(X) = P(x \leq X)$ - функция распределения. Теоретическая вероятность выбросов для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > x_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T))$$

где $F(X) = P(x \leq X)$ - функция распределения

2.7 Эмпирическая функция распределения

2.7.1 Статистический ряд

Статистическим рядом называется последовательность различных элементов выборки z1,z2, ...,zk, расположенных в возрастающем порядке с указанием частот n1,n2, ...,nk, с которыми эти элементы содержатся в выборке. Статистический ряд обычно записывается в виде таблицы

z	z_1	z_2	...	z_k
n	n_1	n_2	...	n_k

Таблица 1: Статистический ряд

2.7.2 Определение

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

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

2.7.3 Описание

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

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

$F^*(x)$ — функция распределения дискретной случайной величины X^*, заданной таблицей распределения

X^*	z_1	z_2	...	z_k
P	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$

Таблица 2: Таблица распределения

Эмпирическая функция распределения является оценкой, т. е. приближённым значением, генеральной функции распределения

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

2.8 Оценки плотности вероятности

2.8.1 Определение

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

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

2.8.2 Ядерные оценки

Представим 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 вероятности, $x_1, ..., x_n$ — элементы выборки, h_n — любая последовательность положительных чисел, обладающая свойствами

$$h_n \xrightarrow[n\to\infty]{} 0; \frac{h_n}{n^{-1}} \xrightarrow[n\to\infty]{} \infty$$

Такие оценки называются непрерывными ядерными.

Замечание. Свойство, означающее сближение оценки с оцениваемой величиной при $n \to \infty$ 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

$$h_n = 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

где $\hat{\sigma}$ − выборочное стандартное отклонение.

3 Реализация

В приложении находится ссылка на репозиторий на GitHub, где находится исходный код лабораторной работы.

4 Результаты

4.1 Гистограммы и графики плотности распределения

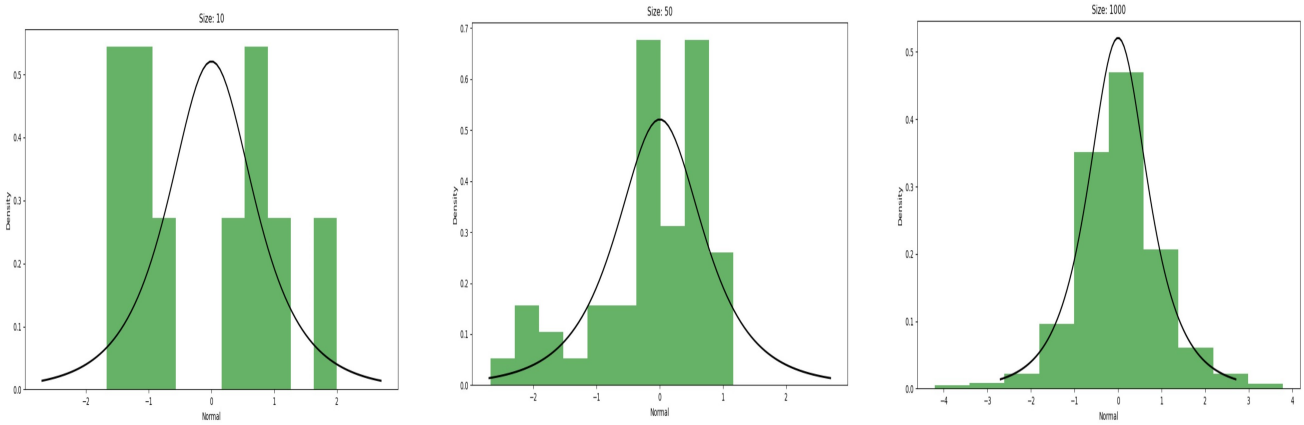

Рис. 1: Нормальное распределение

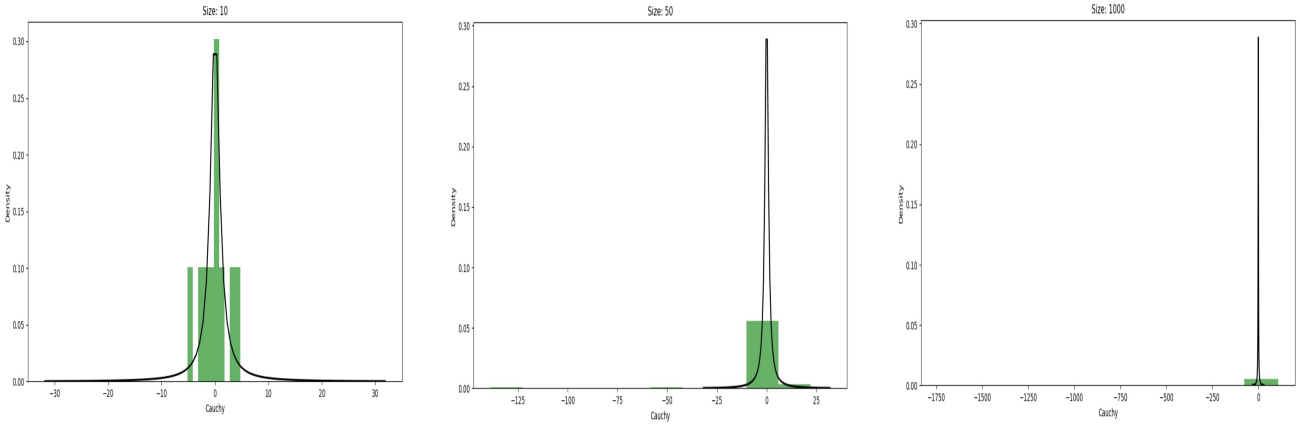

Рис. 2: Распределение Коши

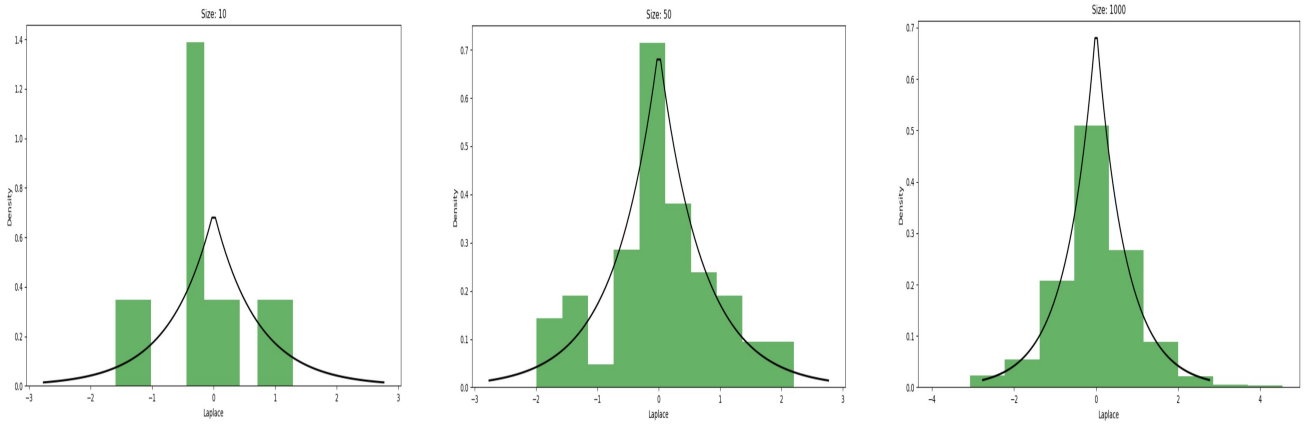

Рис. 3: Распределение Лапласа

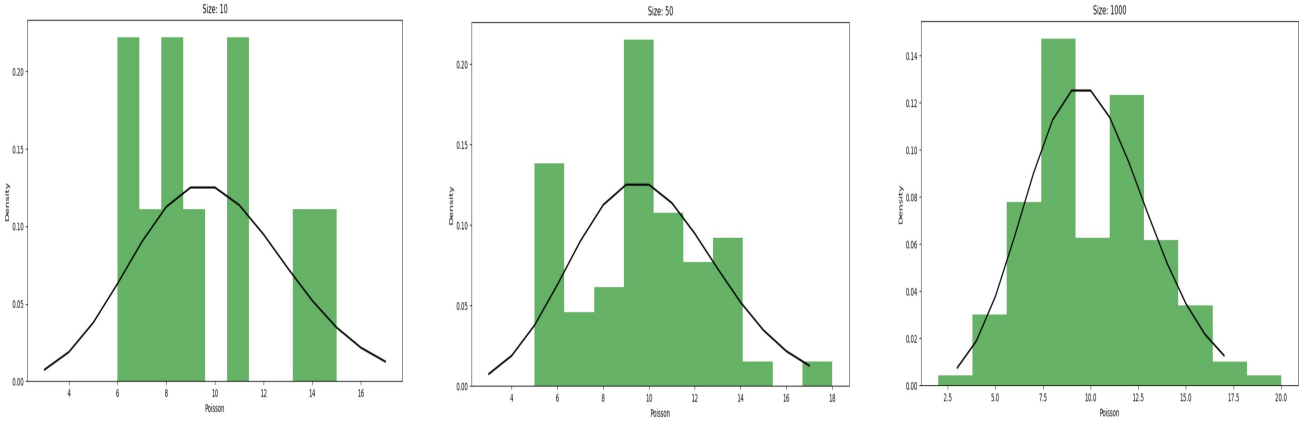

Рис. 4: Распределение Пуассона

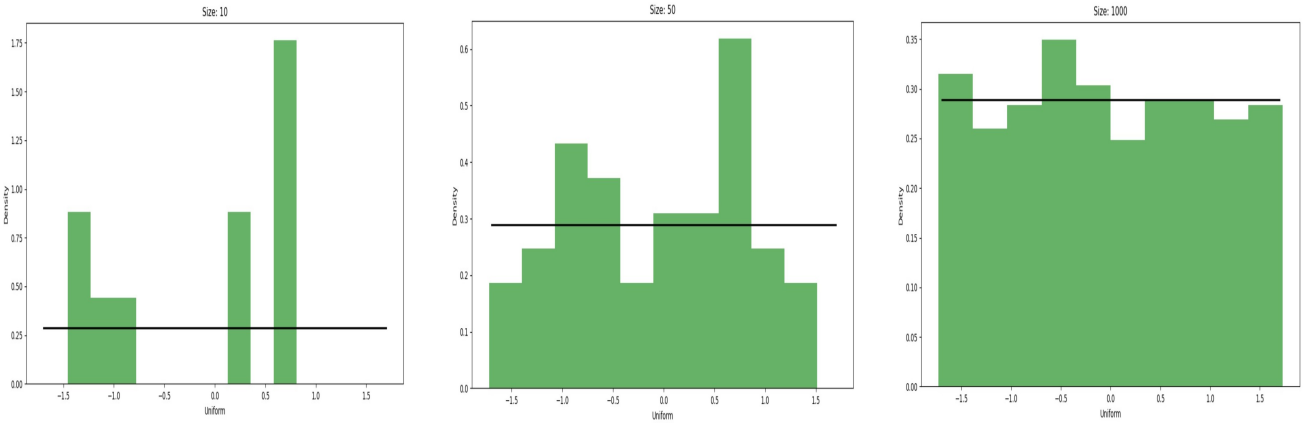

Рис. 5: Равномерное распределение

4.2 Характеристики положения и рассеяния

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Normal E(z) 10	0.025501	0.010119	0.030651	0.030570	0.025708
Normal D(z) 10	0.103765	0.092774	0.476065	0.505294	0.172139
Normal E(z) 100	0.001106	0.001414	-0.04678	0.008496	0.005806
Normal D(z) 100	0.010594	0.009886	0.461752	0.514966	0.018883
Normal E(z) 1000	-0.000306	0.00014	0.029287	0.017512	-0.001402
Normal D(z) 1000	0.0010219	0.00090	0.499531	0.498409	0.0020510

Таблица 3: Нормальное распределение

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Cauchy E(z) 10	-1.82930	-0.01278	-3.6080	-0.6565	-1.73308
Cauchy D(z) 10	987.1422	0.34462	8934.589	854.6997	1708.300
Cauchy E(z) 100	-8.04716	-0.002323	2.58638	3.76963	-17.4805
Cauchy D(z) 100	90333.65	0.024408	3996.35	5978.85	359703.9
Cauchy E(z) 1000	0.0993672	0.000122	-2.16597	-0.68649	-0.033700
Cauchy D(z) 1000	1228.695	0.0024604	3185.117	1023.5756	4370.5033

Таблица 4: Распределение Коши

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Laplace E(z) 10	-0.006466	0.000874	-0.017804	-0.029233	-0.010286
Laplace D(z) 10	0.0967693	0.068329	0.529119	0.4894004	0.159409
Laplace E(z) 100	-0.002621	-0.002186	0.0325949	0.000488	-0.007330
Laplace D(z) 100	0.010093	0.006105	0.5098014	0.528626	0.020044
Laplace E(z) 1000	-0.000461	4.5834-05	0.005827	0.008741	-0.0003342
Laplace D(z) 1000	0.000950	0.000526	0.503384	0.483016	0.001967

Таблица 5: Распределение Лапласа

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Poisson E(z) 10	10.0305	9.887	9.9375	10.1255	10.0543
Poisson D(z) 10	1.05715	1.45523	5.1613	5.0939	1.70460
Poisson E(z) 100	9.978119	9.8255	9.8905	9.777	9.98496
Poisson D(z) 100	0.097413	0.1972	5.11775	5.105771	0.198463
Poisson E(z) 1000	10.00179	9.997	9.905	9.988	10.0054
Poisson D(z) 1000	0.009803	0.001991	5.31547	5.15535	0.018978

Таблица 6: Распределение Пуассона

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Uniform E(z) 10	0.017459	0.016779	0.043340	-0.011148	0.001145
Uniform D(z) 10	0.098445	0.219330	0.487425	0.5378816	0.166308
Uniform E(z) 100	-0.002799	-0.00437	0.017764	0.001040	0.002232
Uniform D(z) 100	0.010377	0.030492	0.443175	0.514648	0.020929
Uniform E(z) 1000	0.001501	0.002397	-0.033368	0.026533	0.001575
Uniform D(z) 1000	0.001005	0.002988	0.5136426	0.495917	0.002077

Таблица 7: Равномерное распределение

4.3 Боксплот Тьюки

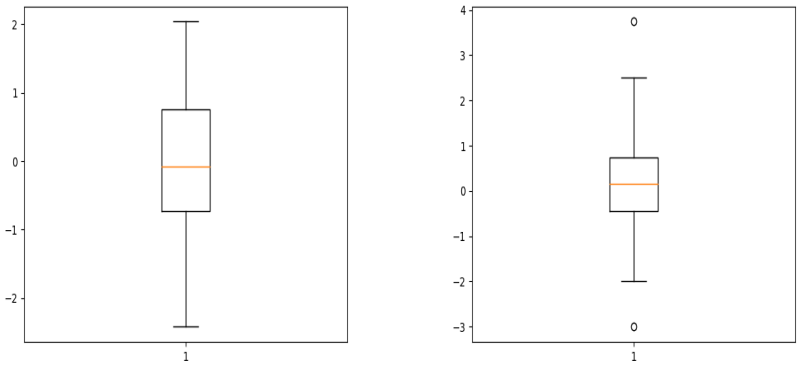

Рис. 6: Нормальное распределение

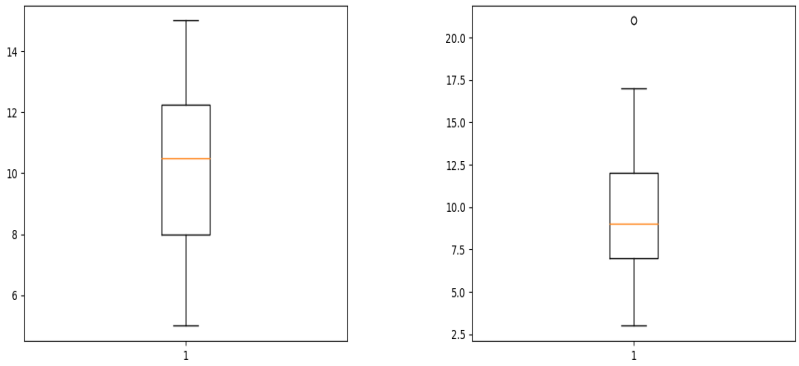

Рис. 7: Распределение Коши

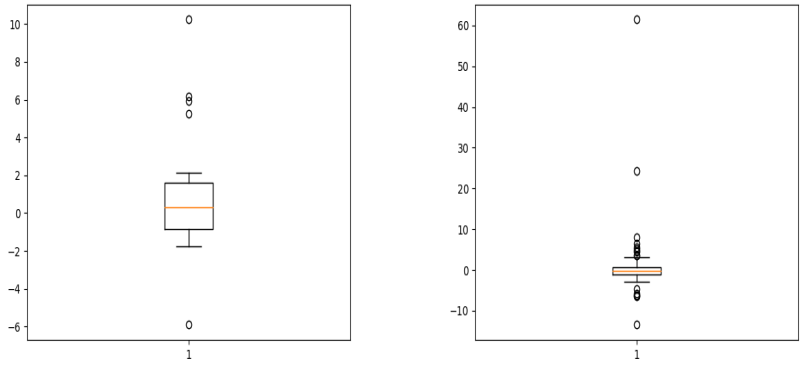

Рис. 8: Распределение Лапласа

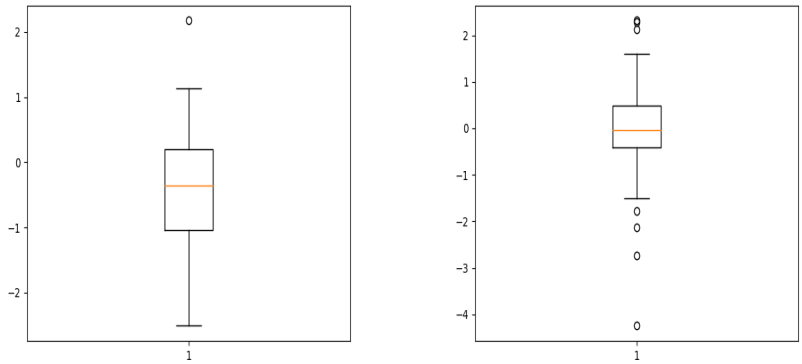

Рис. 9: Распределение Пуассона

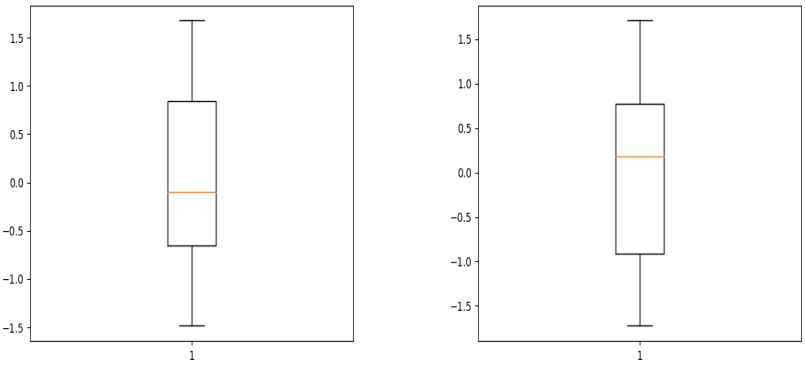

Рис. 10: Равномерное распределение

4.4 Доля выбросов

Выборка	Доля выбросов
Normal n=20	0.06
Normal n=100	0.05
Cauchy n=20	0.15
Cauchy n=100	0.16
Laplace n=20	0.08
Laplace n=100	0.07
Poisson n=20	0.02
Poisson n=100	0.01
Uniform n=20	0
Uniform n=100	0

Таблица 8: Доля выбросов

4.5 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 9: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

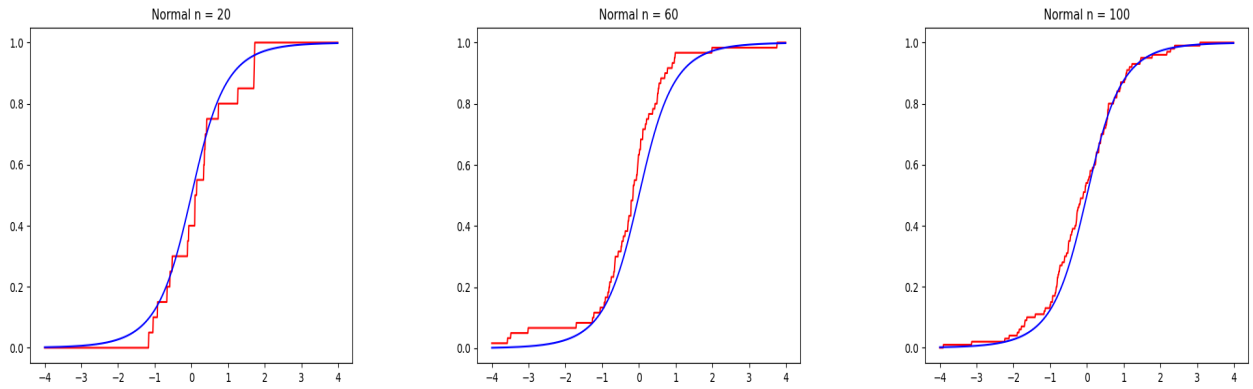

Рис. 11: Нормальное распределение

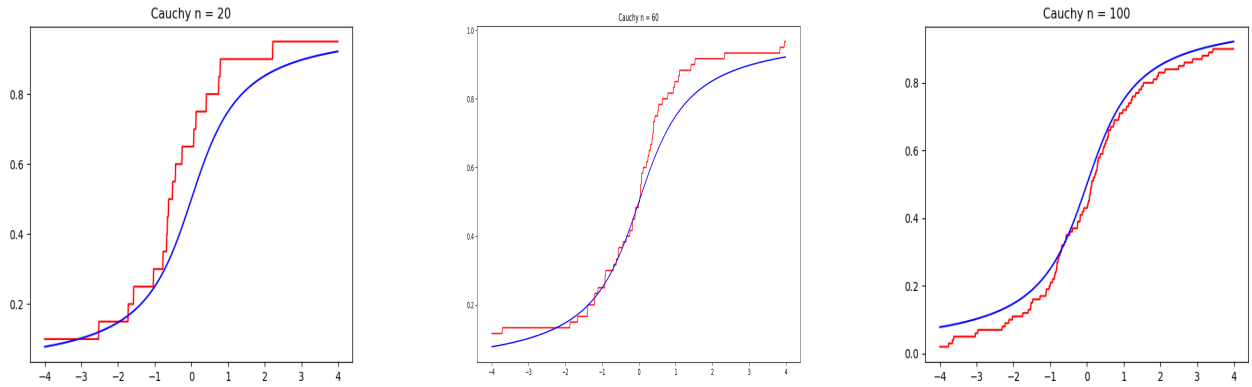

Рис. 12: Распределение Коши

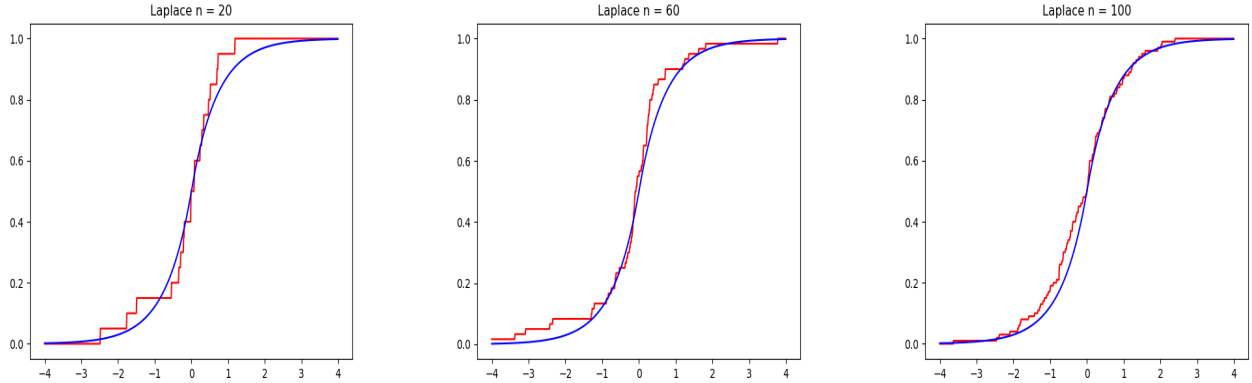

Рис. 13: Распределение Лапласа

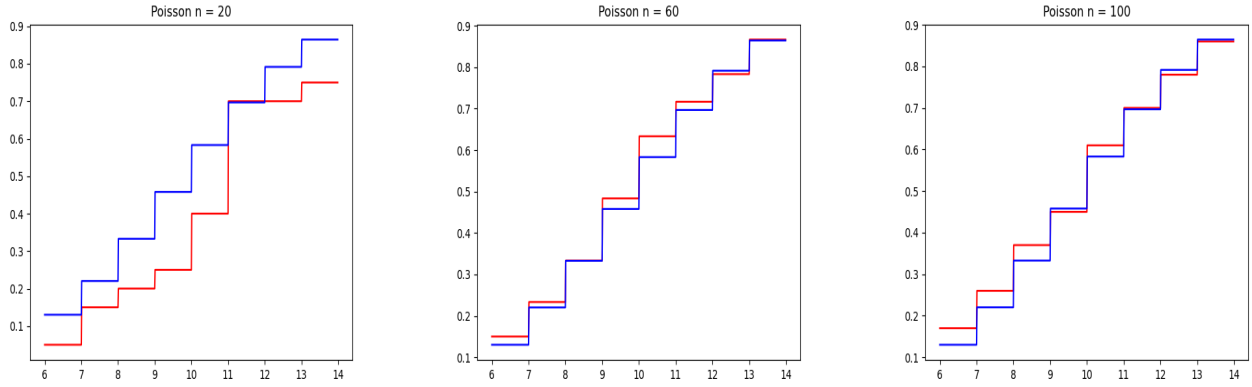

Рис. 14: Распределение Пуассона

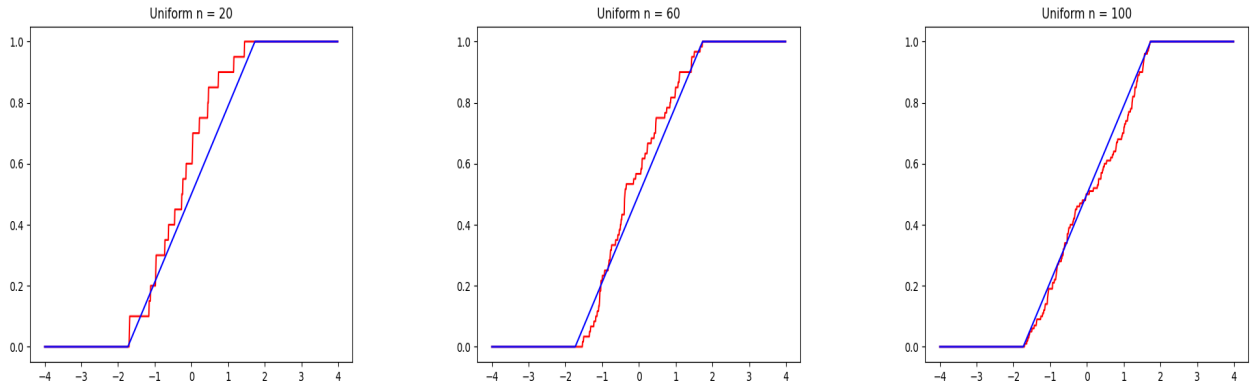

Рис. 15: Равномерное распределение

4.7 Ядерные оценки плотности распределения

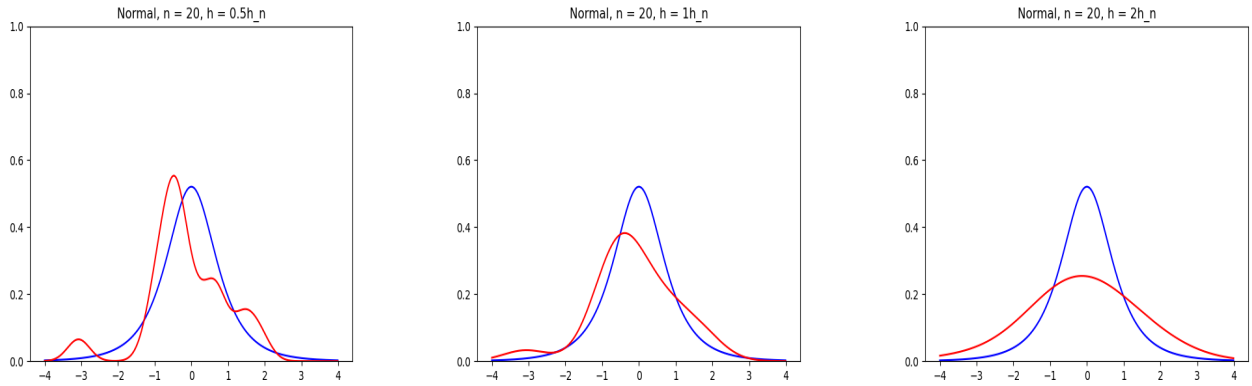

Рис. 16: Нормальное распределение, $n = 20$

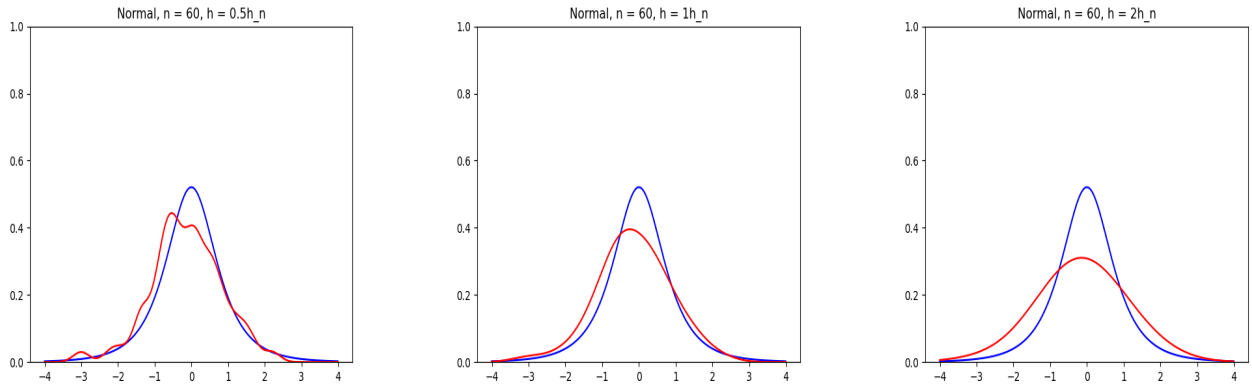

Рис. 17: Нормальное распределение, $n = 60$

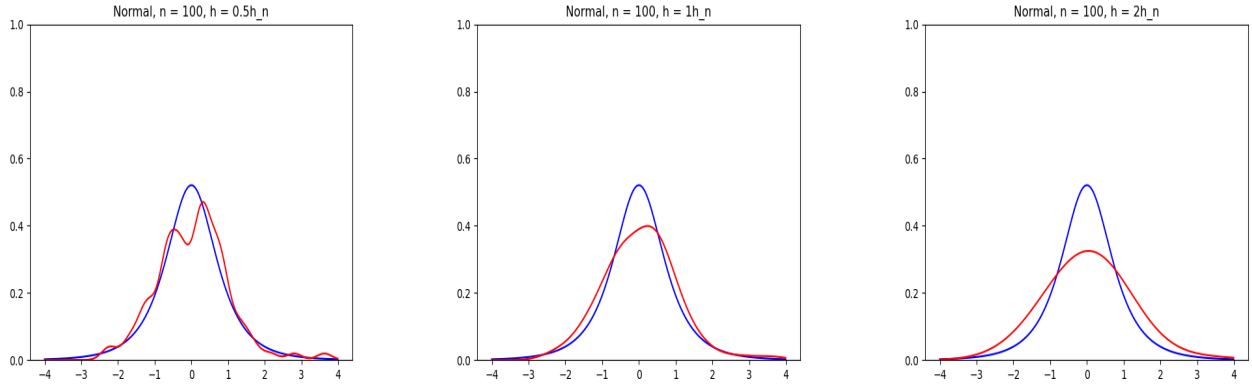

Рис. 18: Нормальное распределение, $n = 100$

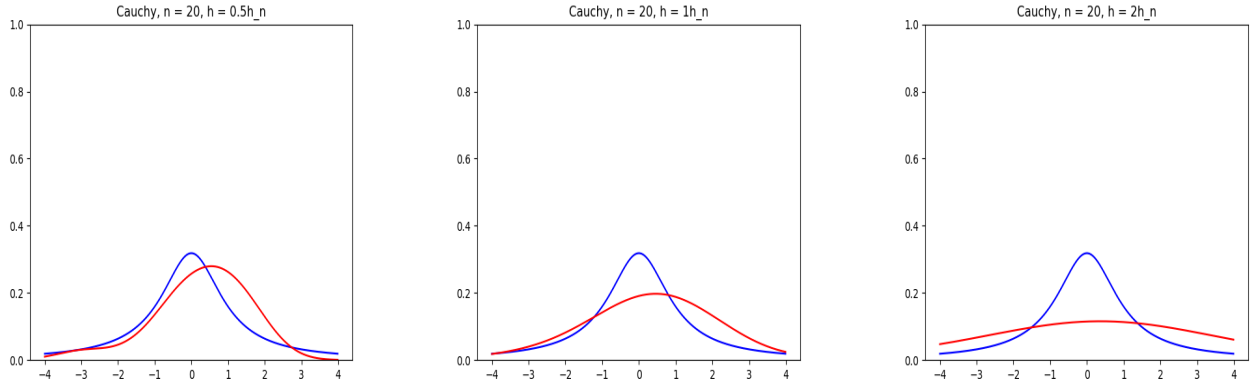

Рис. 19: Распределение Коши, $n = 20$

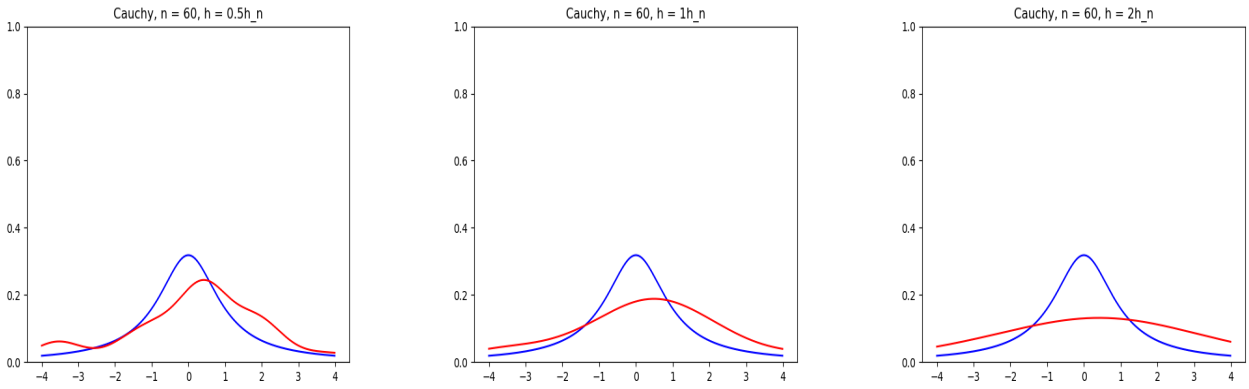

Рис. 20: Распределение Коши, $n = 60$

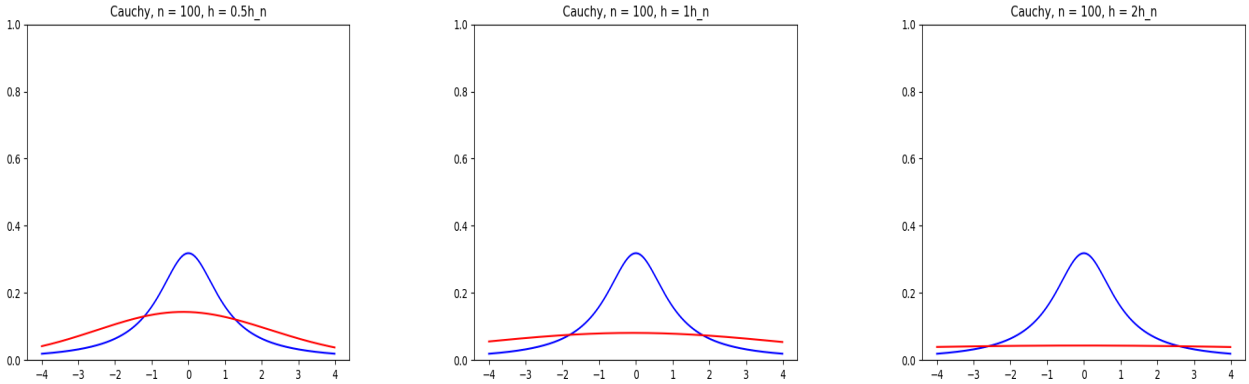

Рис. 21: Распределение Коши, $n = 100$

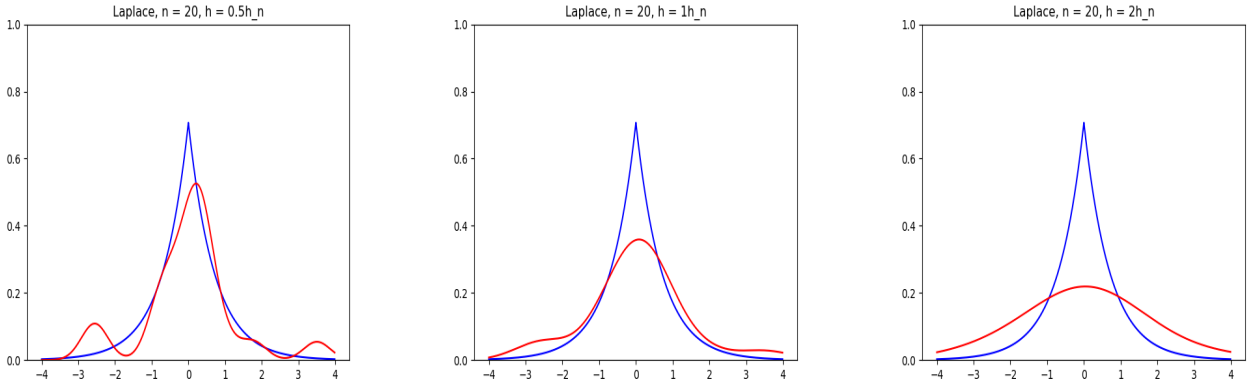

Рис. 22: Распределение Лапласа, $n = 20$

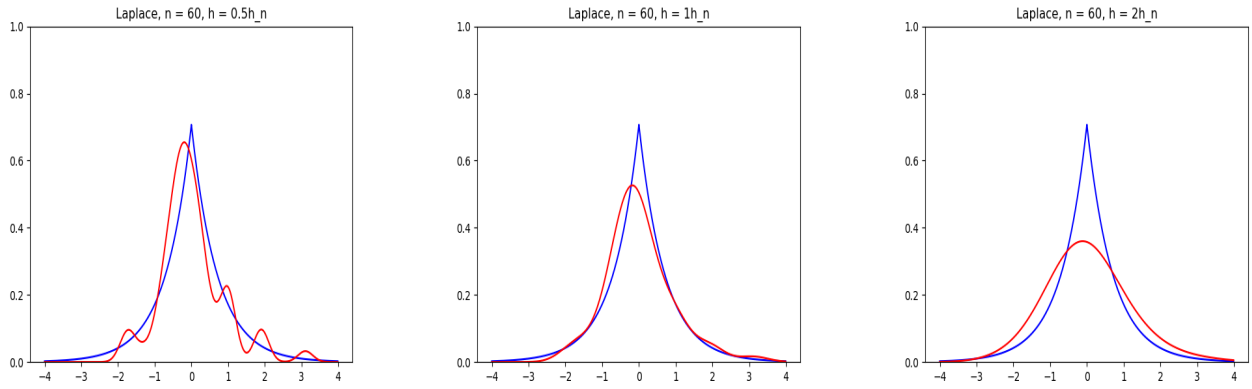

Рис. 23: Распределение Лапласа, $n = 60$

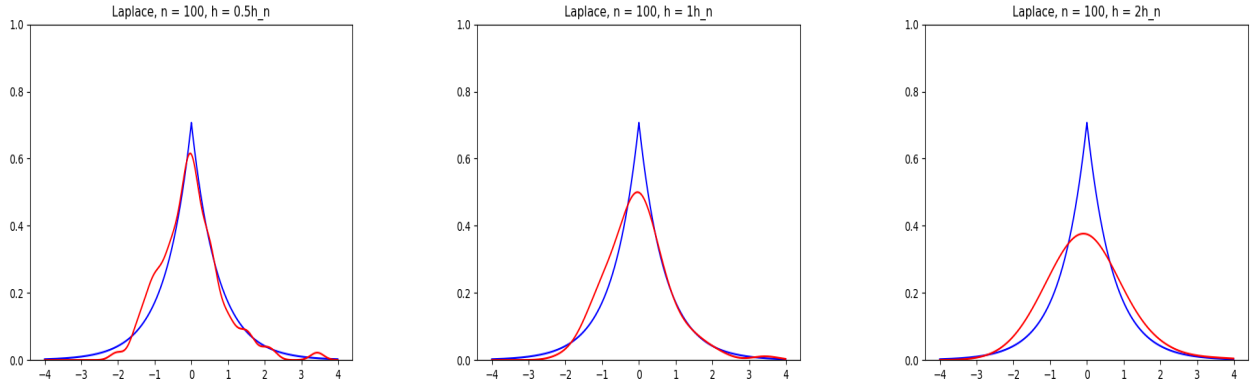

Рис. 24: Распределение Лапласа, $n = 100$

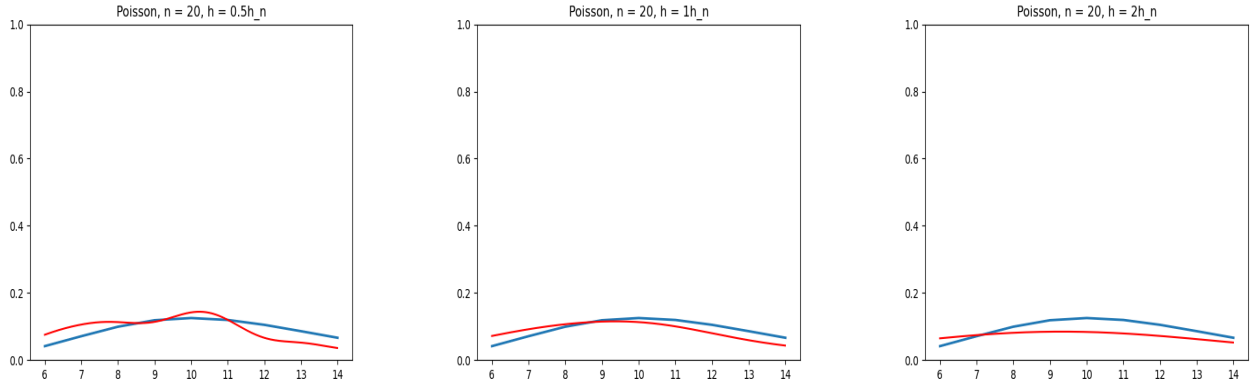

Рис. 25: Распределение Пуассона, $n = 20$

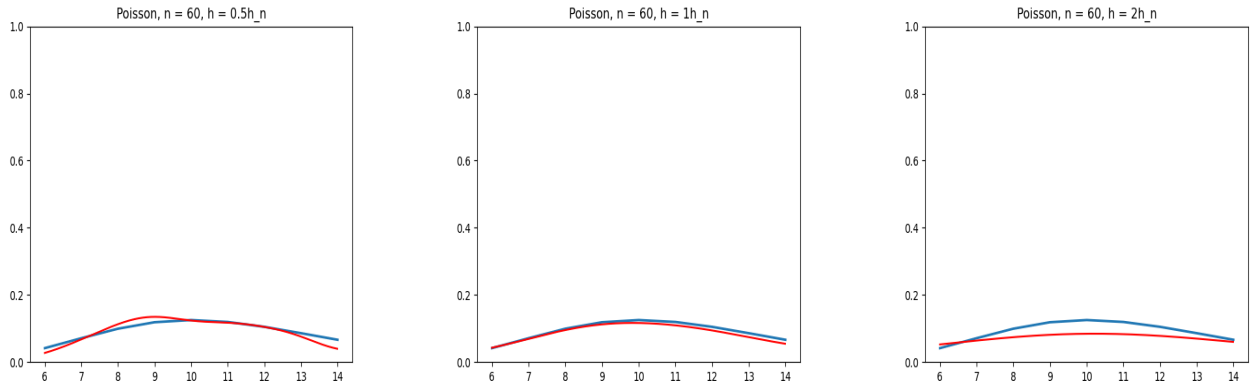

Рис. 26: Распределение Пуассона, $n = 60$

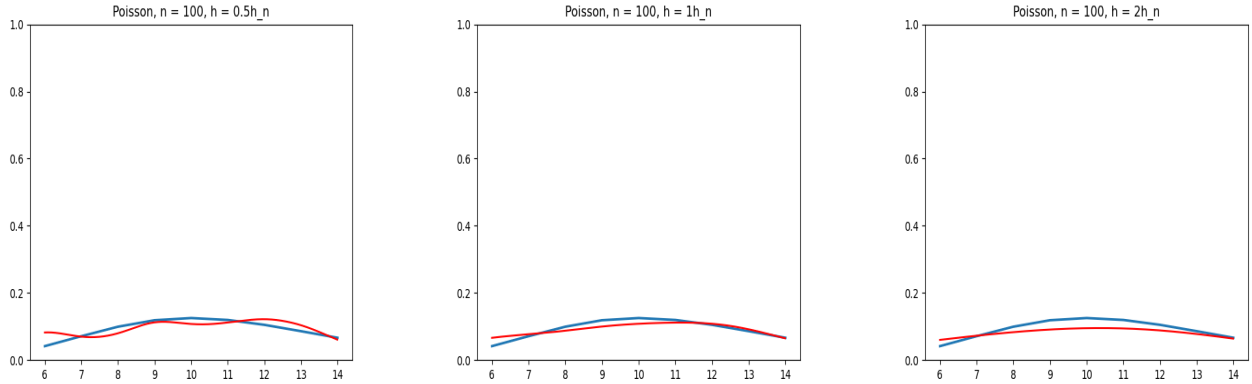

Рис. 27: Распределение Пуассона, $n = 100$

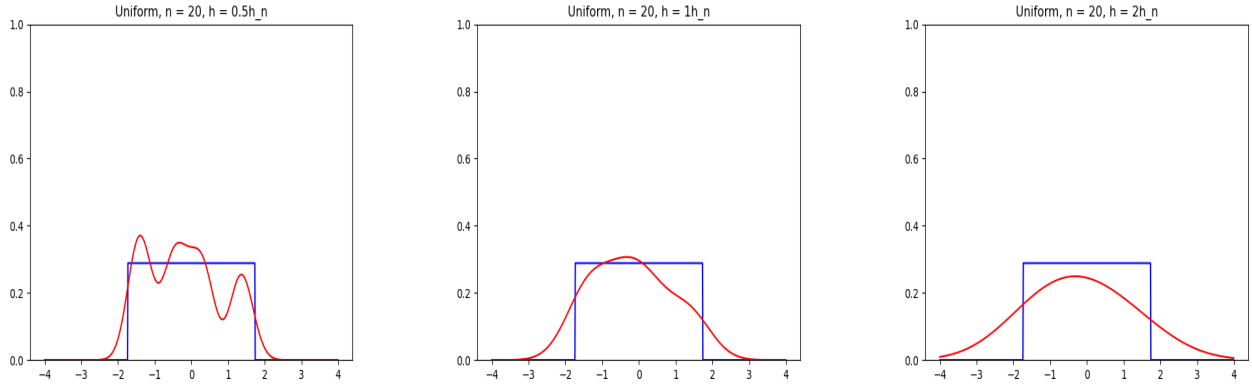

Рис. 28: Равномерное распределение, $n = 20$

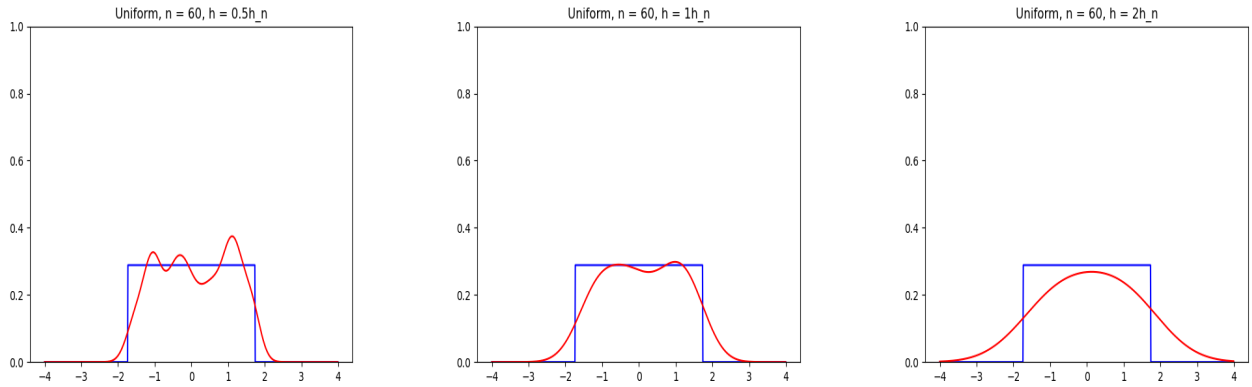

Рис. 29: Равномерное распределение, $n = 60$

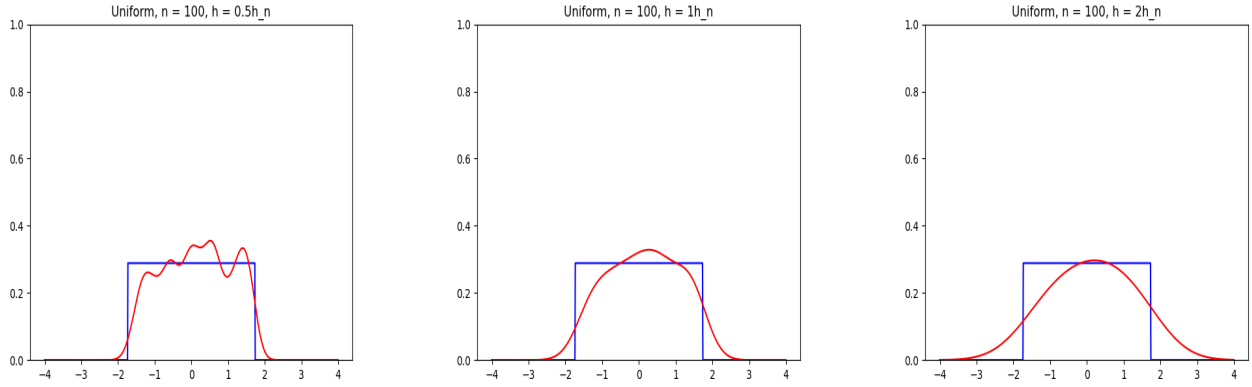

Рис. 30: Равномерное распределение, $n = 100$

5 Обсуждение

5.1 Гистограмма и график плотности распределения

По полученным графикам можно сделать вывод о том, что чем больше размер выборки, тем точнее гистограмма будет повторять график плотности вероятности того закона распределения, для которого была сгенерирована выборка. На маленьких выборках определить закон распределения почти невозможно.

5.2 Характеристики положения и рассеяния

Исходя из данных, приведенных в таблицах, можно судить о том, что дисперсия характеристик рассеяния для распределения Коши является некой аномалией: значения слишком большие даже при увеличении размера выборки - понятно, что это результат выбросов, которые мы могли наблюдать в результатах предыдущего задания.

5.3 Боксплот Тьюки и доля выбросов

По данным, приведенным в таблице, можно сказать, что чем больше выборка, тем ближе доля выбросов будет к теоретической оценке. Снова доля выбросов для распределения Коши значительно выше, чем для остальных распределений. Равномерное распределение же в точности повторяет теоретическую оценку - выбросов мы не получали.

Боксплот Тьюки позволяет наглядно оценить наиболее важные характеристики распределений.

5.4 Ядерные оценки плотности распределения

По полученным графикам, можно сказать, что чем больше размер выборки - тем точнее полученные по ней оценки будут аппроксимировать реальные генеральную функцию распределения и плотность вероятности. Хотя даже на выборке $n = 20$ результат получается очень близким к действительности и можно понять, что это за распределение.

Также можно увидеть, что чем больше коэффициент при параметре сглаживания $\hat{h_n}$, тем меньше изменений знака производной у аппроксимирующей функции, вплоть до того, что при $h = 2h_n$ функция становится унимодальной на рассматриваемом промежутке. Также видно, что при $h = 2h_n$ по полученным приближениям становится сложно сказать плотность вероятности какого распределения они должны повторять, так как они очень похожи между собой.

Заметим, что для распределения Лапласа лучше использовать бóльшие значения h, так как исчезают скачки, которых нет у функции плотности данного распределения и, наоборот, для равномерного лучше брать мéньшие значения h, так как при бóльших получается, что плотность на концах промежутка меньше, чем в центре, что неверно. Именно при таком выборе значения h ядерная оценка будет лучше воспроизводить особенности данных распределений.

6 Приложения

Код программы - GitHub URL:
Задания 1-2: https://github.com/juliaknodel/MS-1-2
Задание 3: https://github.com/juliaknodel/MS-3
Задание 4: https://github.com/juliaknodel/MS-4

Отчет: https://github.com/juliaknodel/MS-4/tree/master/report1-4