Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и вычислительной физики, ФизМех

Направление подготовки
«01.03.02 Прикладная математика и информатика»

Отчет по лабораторной работе № 1
дисциплина "Математическая статистика"

Выполнил студент гр. 5030102/00101 Кочетков И.Д.
Преподаватель: Баженов А.Н.

Санкт-Петербург

2023

Содержание

1 Постановка задачи **5**

2 Теория **6**
2.1 Рассматриваемые распределения 6
2.2 Гистограмма 6
2.2.1 Построение гистограммы 6
2.3 Вариационный ряд 6
2.4 Выборочные числовые характеристики 7
2.4.1 Характеристики положения 7
2.4.2 Характеристики рассеяния 7
2.5 Боксплот Тьюки 7
2.5.1 Построение 7
2.6 Теоретическая вероятность выбросов 8
2.7 Эмпирическая функция распределения 8
2.7.1 Статический ряд 8
2.7.2 Эмпирическая функция распределения 8
2.7.3 Нахождение эмпирической функции распределения 9
2.8 Оценки плотности вероятности 9
2.8.1 Определение 9
2.8.2 Ядерные оценки 9

3 Реализация **11**

4 Результаты **12**
4.1 Гистограммы и графики плотности распределения 12
4.2 Характеристики положения и рассеяния 14
4.3 Боксплот Тьюки 17
4.4 Доля выбросов 19
4.5 Теоретическая вероятность выбросов 20
4.6 Эмпирическая функция распределения 20
4.7 Ядерные оценки плотности распределения 23

5 Обсуждение **31**
5.1 Гистограммы 31
5.2 Характеристики положения и рассеяния 31
5.3 Доля и теоретическая вероятность выбросов 31
5.4 Эмпирическая функция распределения. Ядерные оценки плотности 31

6 Литература **33**

Список иллюстраций

1 Нормальное распределение 12
2 Распределение Коши 12
3 Распределение Лапласа 13
4 Распределение Пуассона 13
5 Равномерное распределение 14
6 Нормальное распределение 17
7 Распределение Коши 17
8 Распределение Лапласа 18
9 Распределение Пуассона 18
10 Равномерное распределение 19
11 Нормальное распределение 20
12 Распределение Коши 21
13 Распределение Лапласа 21
14 Распределение Пуассона 22
15 Равномерное распределение 22
16 Нормальное распределение, $n = 10$ 23
17 Нормальное распределение, $n = 50$ 23
18 Нормальное распределение, $n = 100$ 24
19 Распределение Коши, $n = 10$ 24
20 Распределение Коши, $n = 50$ 25
21 Распределение Коши, $n = 100$ 25
22 Распределение Лапласа, $n = 10$ 26
23 Распределение Лапласа, $n = 50$ 26
24 Распределение Лапласа, $n = 100$ 27
25 Распределение Пуассона, $n = 10$ 27
26 Распределение Пуассона, $n = 50$ 28
27 Распределение Пуассона, $n = 100$ 28
28 Равномерное распределение, $n = 10$ 29
29 Равномерное распределение, $n = 50$ 29
30 Равномерное распределение, $n = 100$ 30

1 Постановка задачи

Для 4 распределений:

- Нормальное распределение $N(x, 0, 1)$
- Распределение Коши $C(x, 0, 1)$
- Распределение Пуассона $P(k, 10)$
- Равномерное распределение $U(x, -\sqrt{3}, \sqrt{3})$

1. Сгенерировать выборки размером 10, 50 и 100 элементов. Построить на одном рисунке гистограмму и график плотности распределения.

2. Сгенерировать выборки размером 10, 50 и 100 элементов. Для каждой выборки вычислить следующие характеристики положения данных: $\overline{x}$ (выборочное среднее) , $medx$ (выборочная медиана), z_R (полусумма экстремальных выборочных элементов), z_Q (полусумма квартелей), z_{tr} (усечённое среднее). Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

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

   Вычислить оценку дисперсий по формуле:

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

   Представить полученные данные в виде таблиц

3. Сгенерировать выборки размером 20 и 100 элементов.
   Построить для них боксплот Тьюки. 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 10, 50 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 Рассматриваемые распределения

- Нормальное распределение

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

- Распределение Коши

$$C(x,0,1) = \frac{1}{\pi} \frac{1}{x^2+1} \tag{4}$$

- Распределение Пуассона

$$P(k,10) = \frac{10^k}{k!} e^{-10} \tag{5}$$

- Равномерное распределение

$$U(x,-\sqrt{3},\sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & , |\mathrm{x}| \leq \sqrt{3} \\ 0 & , |\mathrm{x}| > \sqrt{3} \end{cases} \tag{6}$$

2.2 Гистограмма

2.2.1 Построение гистограммы

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.

2.3 Вариационный ряд

Вариационным ряд - последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются.

2.4 Выборочные числовые характеристики

2.4.1 Характеристики положения

- Выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

- Выборочная медиана

$$medx = \begin{cases} x_{(l+1)} & , n = 2l+1 \\ \frac{x_{(l)}+x_{(l+1)}}{2} & , n = 2l \end{cases} \tag{8}$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{9}$$

- Полусумма квартилей
  Выборочная квартиль z_p порядка p определяется формулой:

$$z_p = \begin{cases} x_{(|np|+1)} & , np \text{ дробное} \\ x_{(np)} & , np \text{ целое} \end{cases} \tag{10}$$

  Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{11}$$

$$z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_{(i)}, r \approx \frac{n}{4} \tag{12}$$

2.4.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{13}$$

2.5 Боксплот Тьюки

2.5.1 Построение

Границами ящика – первый и третий квартили, линия в середине ящика- медиана. Концы усов — края статистически значимой выборки (без выбросов). Длина

«усов»:

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1), \tag{14}$$

где X_1 — нижняя граница уса, $X2$— верхняя граница уса, Q_1— первый квартиль, Q_2 — третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков.

2.6 Теоретическая вероятность выбросов

Можно вычислить теоретические первый и третий квартили распределений $-Q_1^T$ и $-Q_3^T$. По формуле (14) – теоретические нижнюю и верхнюю границы уса $-X_1^T$ и $-X_2^T$. Выбросы- величины x:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix} \tag{15}$$

Теоретическая вероятность выбросов:

- для непрерывных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T)) \tag{16}$$

- для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > x_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T)) \tag{17}$$

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

2.7 Эмпирическая функция распределения

2.7.1 Статический ряд

Статистическим ряд- последовательность различных элементов выборки $z_1, z_2, ..., z_k$ положенных в возрастающем порядке с указанием частот $n_1, n_2, ..., n_k$, с которыми эти элементы содержатся в выборке. Обычно записывается в виде таблицы.

2.7.2 Эмпирическая функция распределения

Эмпирическая (выборочная) функция распределения (э.ф.р)- относительная частота события $X < x$, полученная по данной выборке:

$$F_n^* = P^*(X < x) \tag{18}$$

2.7.3 Нахождение эмпирической функции распределения

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

$$F^*(x) = \frac{1}{n}\sum_{z_i<x} n_i. \tag{19}$$

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

X^*	z_1	z_2	...	z_k
P	n_1/n	n_2/n	...	n_k/n

Таблица 1: Таблица распределения

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

$$F_n^*(x) \approx F_X(x). \tag{20}$$

2.8 Оценки плотности вероятности

2.8.1 Определение

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

$$\widehat{f}(x) \approx f(x). \tag{21}$$

2.8.2 Ядерные оценки

Представим оценки в виде суммы с числом слагаемых, равным объёму выборки:

$$\widehat{f_n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right). \tag{22}$$

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

$$h_n \xrightarrow[n\to\infty]{} 0; \quad \frac{h_n}{n^-1} \xrightarrow[n\to\infty]{} \infty. \tag{23}$$

Такие оценки называются непрерывными ядерными [2, с. 421-423].

Гауссово (нормальное) ядро

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{24}$$

Правило Сильвермана

$$h_n = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{1/5} \approx 1.06 \hat{\sigma} n^{-1/5}, \tag{25}$$

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

3 Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python с использованием библиотек Numpy, Scipy, Seaborn,Statsmodels, Matplotlib. Исходный код лабораторнойработы приведён в приложении.

4 Результаты

4.1 Гистограммы и графики плотности распределения

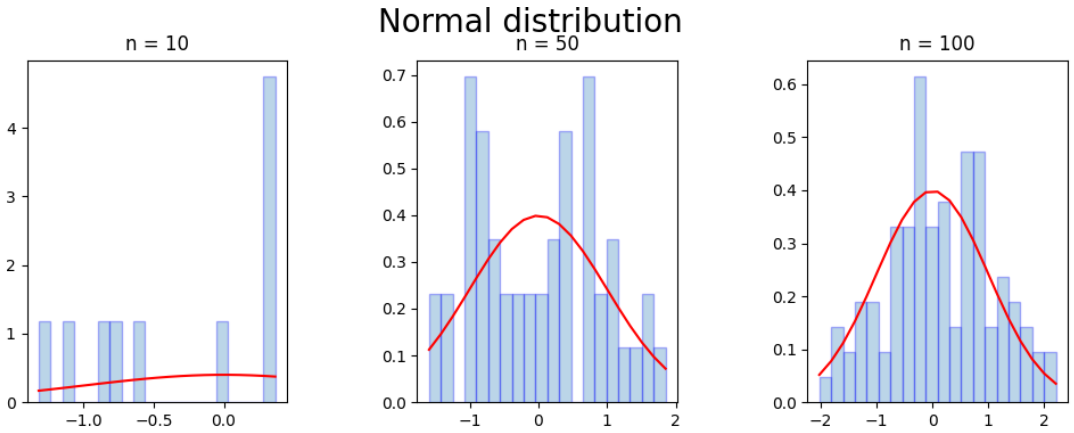

Рис. 1: Нормальное распределение

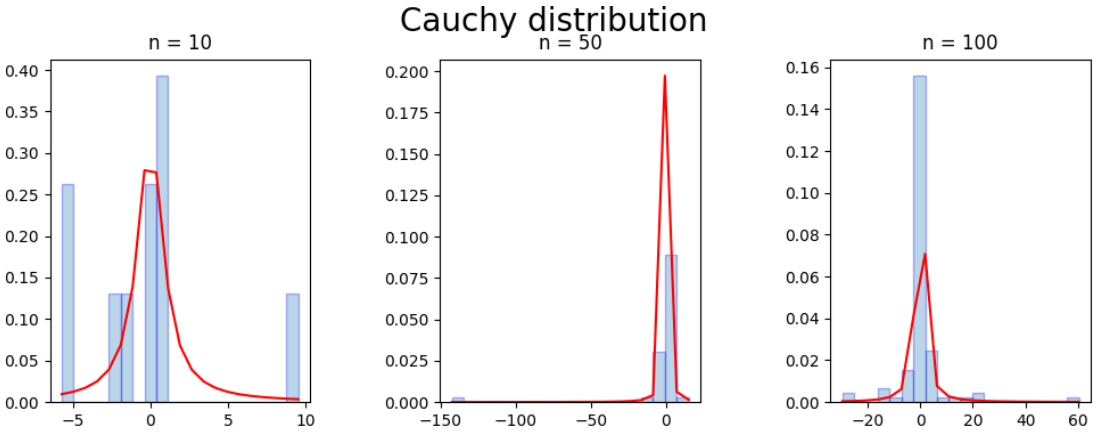

Рис. 2: Распределение Коши

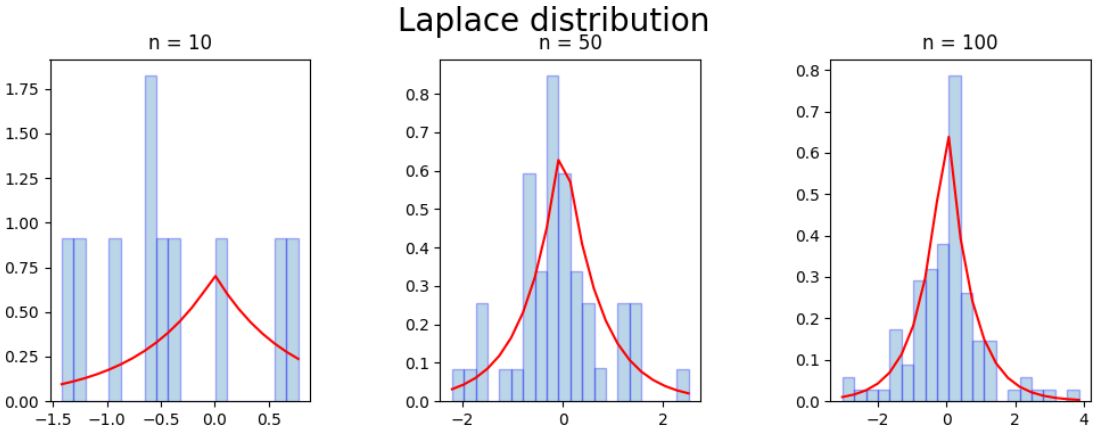

Рис. 3: Распределение Лапласа

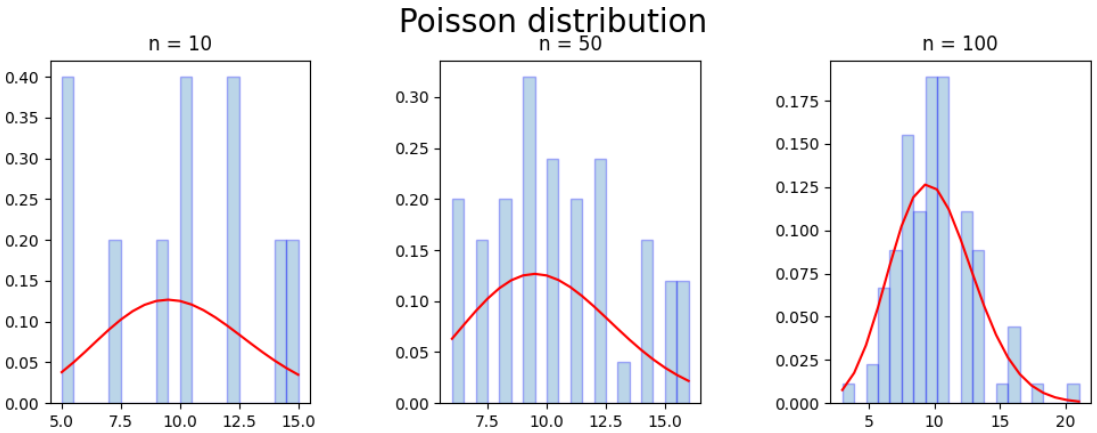

Рис. 4: Распределение Пуассона

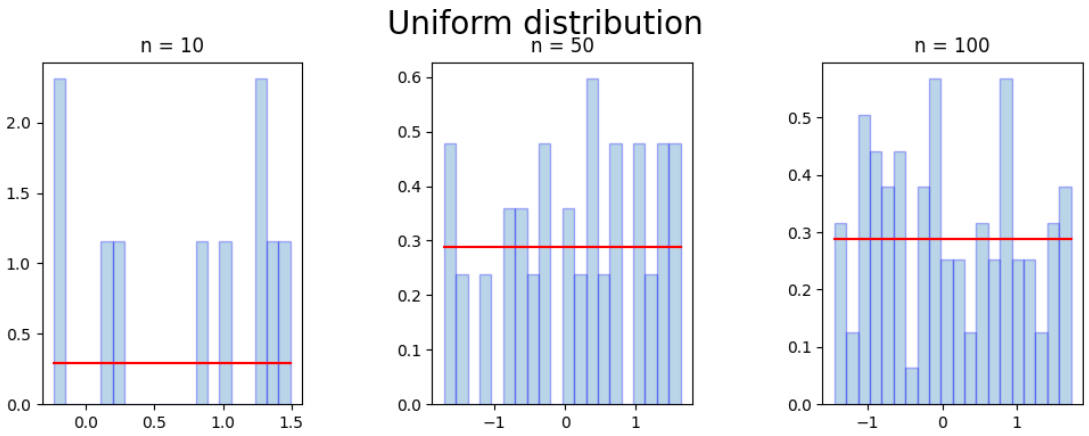

Рис. 5: Равномерное распределение

4.2 Характеристики положения и рассеяния

Normal n = 10					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.0025	-0.0028	-0.011	0.0029	0.11
D(z)	0.095	0.14	0.18	0.10	0.081
Normal n = 50					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.0029	0.0017	-0.00048	0.0029	0.028
D(z)	0.019	0.028	0.11	0.023	0.021
Normal n = 100					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.0032	0.0010	-0.015	-0.0017	0.013
D(z)	0.0097	0.015	0.090	0.012	0.011

Таблица 2: Нормальное распределение

Cauchy n = 10					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.83	0.020	3.88	0.018	0.23
D(z)	181.74	0.37	4246.80	1.28	0.38
Cauchy n = 50					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.088	0.0025	-2.27	0.0040	0.043
D(z)	375.13	0.052	228338.25	0.11	0.053
Cauchy n = 100					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.50	0.0085	-23.98	0.0094	0.029
D(z)	360.10	0.027	885426.18	0.053	0.027

Таблица 3: Распределение Коши

Laplace n = 10					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.016	0.0085	0.039	0.0097	0.098
D(z)	0.11	0.082	0.45	0.095	0.055
Laplace n = 50					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.0070	0.0019	0.043	0.0038	0.020
D(z)	0.019	0.012	0.41	0.019	0.011
Laplace n = 100					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.0019	0.0021	0.030	0.00032	0.010
D(z)	0.010	0.0060	0.45	0.0097	0.0061

Таблица 4: Распределение Лапласа

Poisson n = 10					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	10.037	9.89	10.30	9.96	8.62
D(z)	1.034	1.49	1.86	1.19	0.90
Poisson n = 50					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	9.99	9.82	10.70	9.92	9.56
D(z)	0.19	0.38	1.18	0.26	0.21
Poisson n = 100					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	9.99	9.84	10.91	9.90	9.69
D(z)	0.097	0.20	0.93	0.15	0.11

Таблица 5: Распределение Пуассона

Uniform n = 10					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	0.0097	0.016	0.0070	0.0087	0.14
D(z)	0.097	0.23	0.043	0.14	0.12
Uniform n = 50					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.0050	-0.0070	-0.0022	-0.0031	0.028
D(z)	0.019	0.056	0.0023	0.029	0.035
Uniform n = 100					
	Mean	Median	z_R	z_Q	z_{tr}
E(z)	-0.0045	-0.0064	0.0011	-0.0072	0.0098
D(z)	0.010	0.029	0.00064	0.016	0.019

Таблица 6: Равномерное распределение

4.3 Боксплот Тьюки

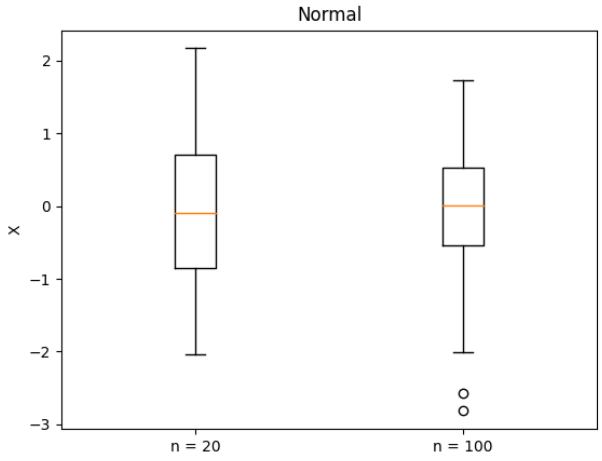

Рис. 6: Нормальное распределение

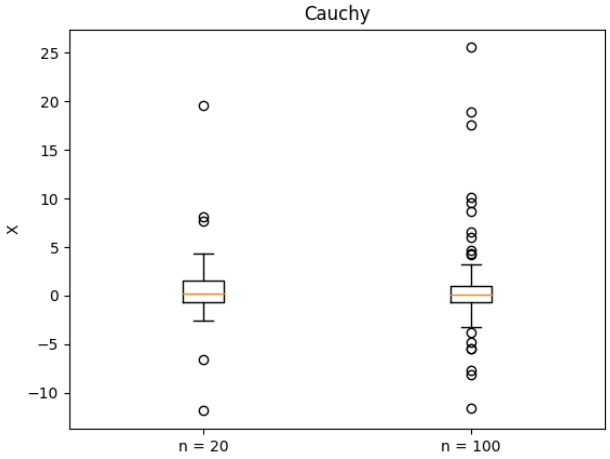

Рис. 7: Распределение Коши

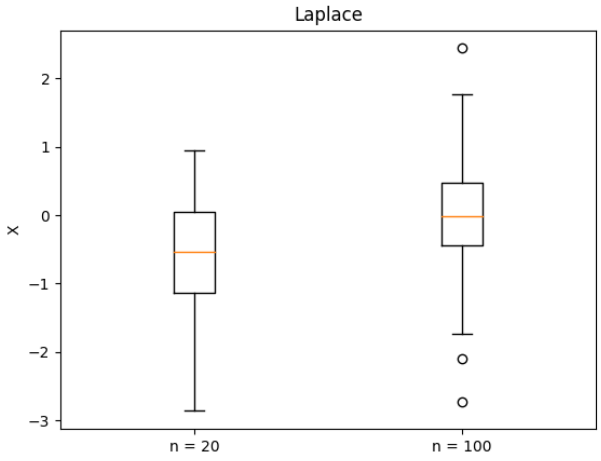

Рис. 8: Распределение Лапласа

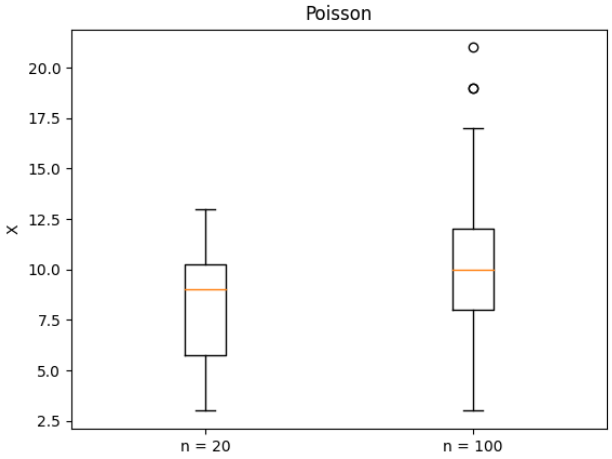

Рис. 9: Распределение Пуассона

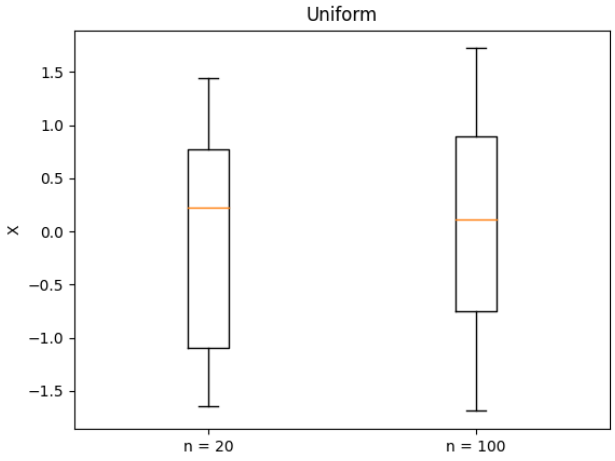

Рис. 10: Равномерное распределение

4.4 Доля выбросов

Округление доли выбросов:
Выборка случайна, поэтому в качестве оценки рассеяния можно взять дисперсию пуассоновского потока: $D_n \approx \sqrt{n}$
Доля $p_n = \frac{D_n}{n} = \frac{1}{\sqrt{n}}$
Доля $n = 20 : p_n = \frac{1}{\sqrt{20}}$ - примерно 0.2 или 20%
Для $n = 100 : p_n = \frac{1}{\sqrt{100}}$ - примерно 0.1 или 10%
Исходя из этого можно решить, сколько знаков оставлять в доле выброса.

Выборка	Доля выбросов
Normal n=20	0.022
Normal n=100	0.01
Cauchy n=20	0.151
Cauchy n=100	0.156
Laplace n=20	0.072
Laplace n=100	0.066
Poisson n=20	0.023
Poisson n=100	0.01
Uniform n=20	0.003
Uniform n=100	0.0

Таблица 7: 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 Пуассона	8	12	2	18	0.008
Равномерное распределение	-0.866	0.866	-3.464	3.464	0

Таблица 8: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

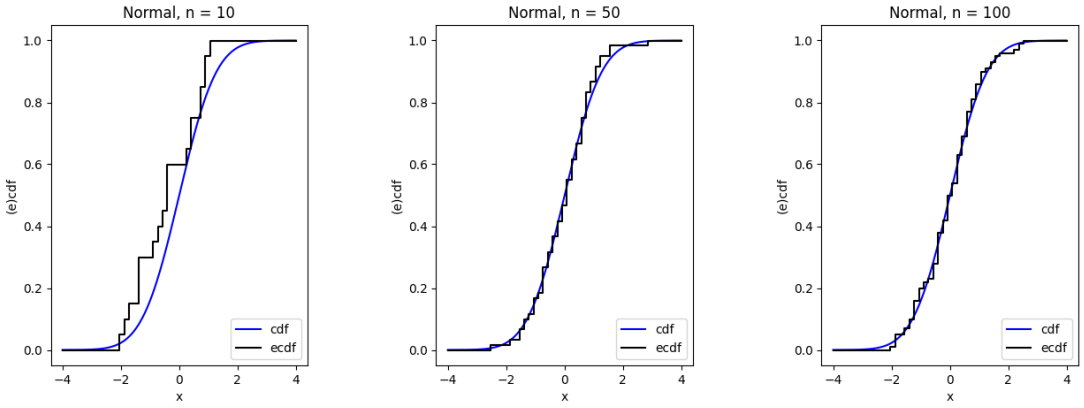

Рис. 11: Нормальное распределение

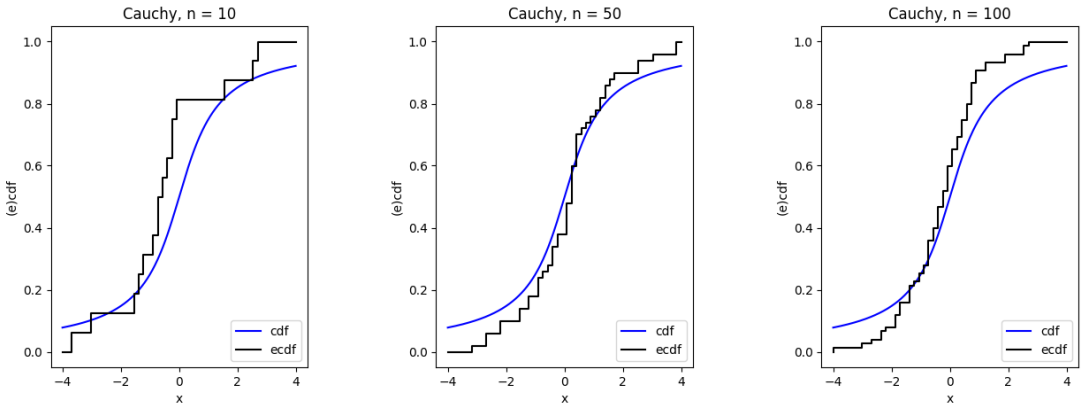

Рис. 12: Распределение Коши

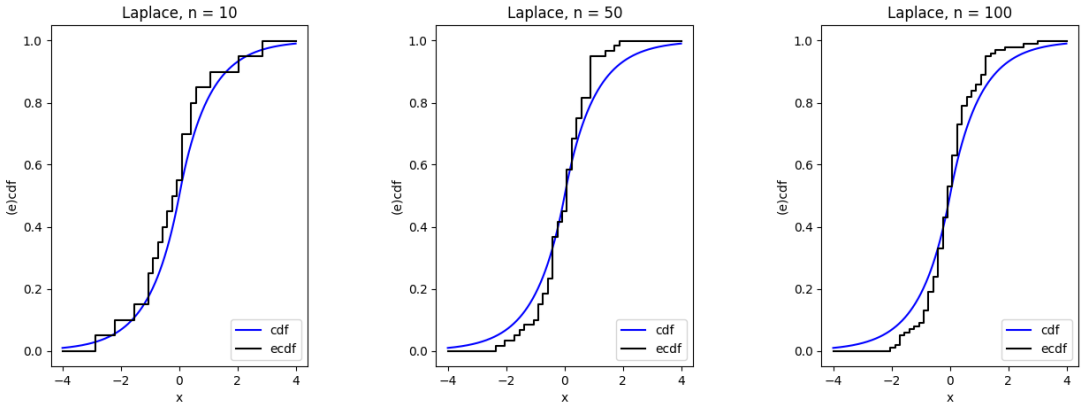

Рис. 13: Распределение Лапласа

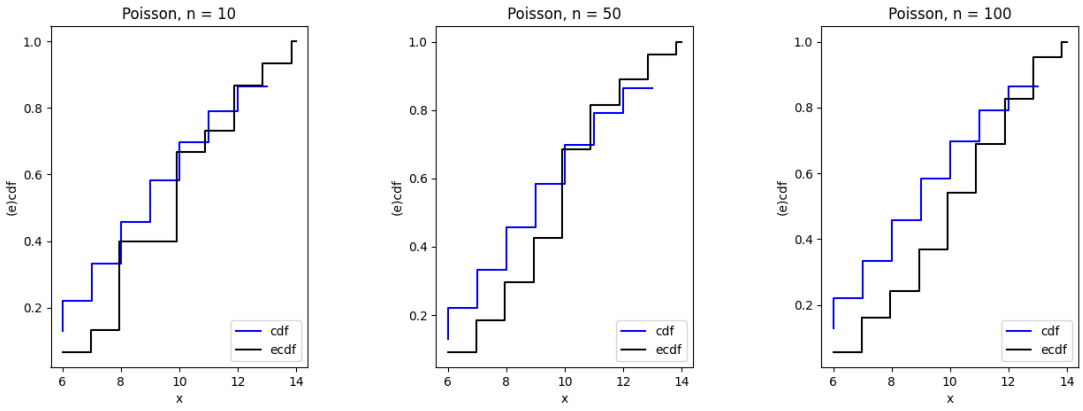

Рис. 14: Распределение Пуассона

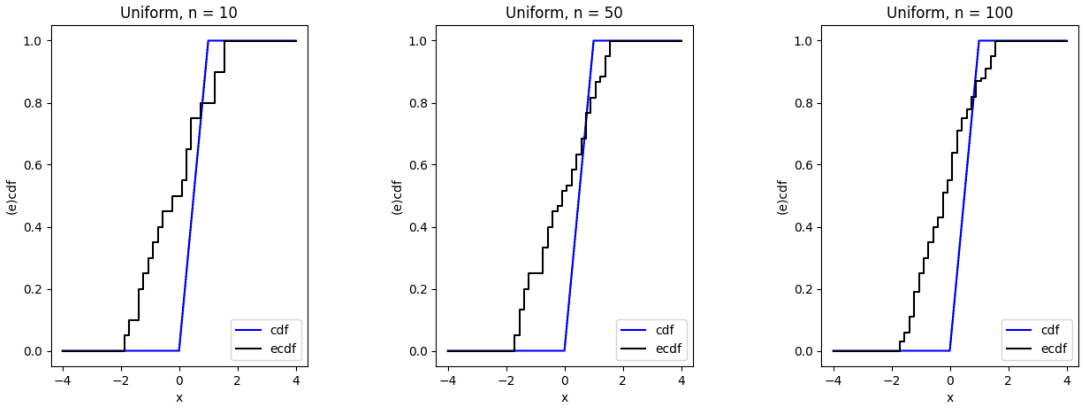

Рис. 15: Равномерное распределение

4.7 Ядерные оценки плотности распределения

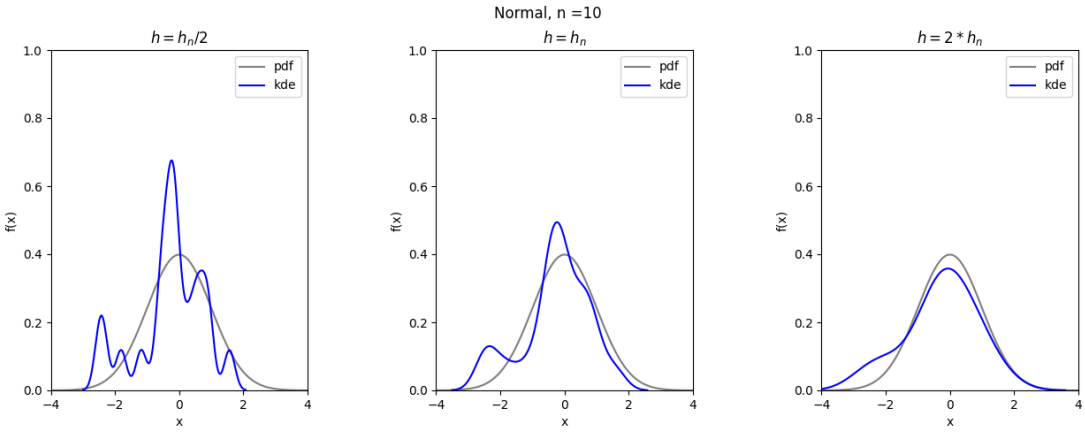

Рис. 16: Нормальное распределение, $n = 10$

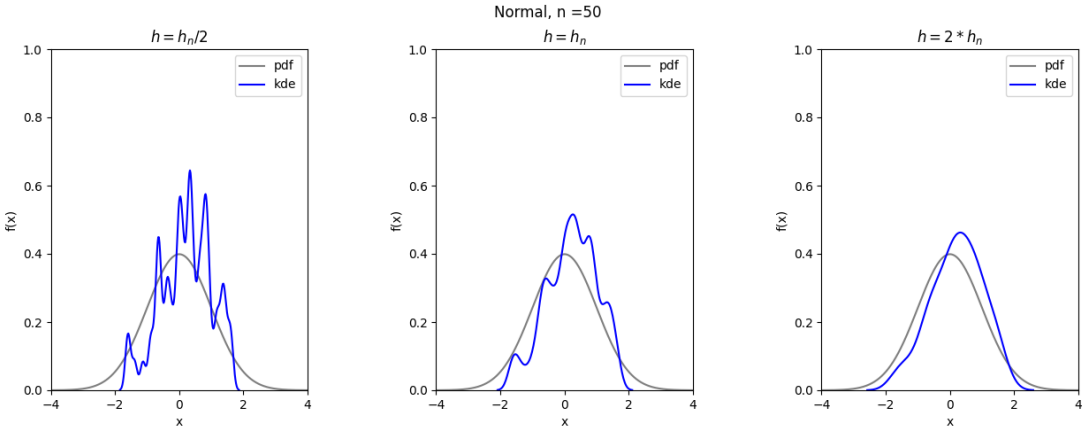

Рис. 17: Нормальное распределение, $n = 50$

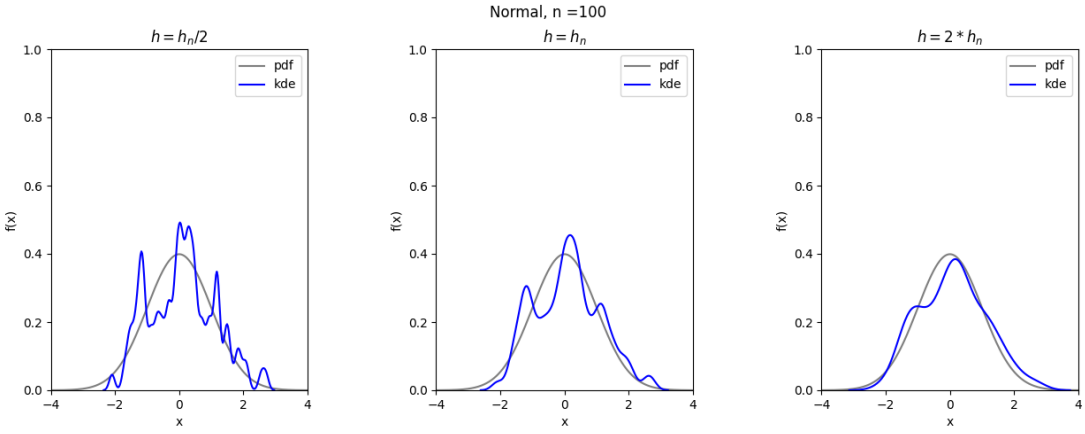

Рис. 18: Нормальное распределение, $n = 100$

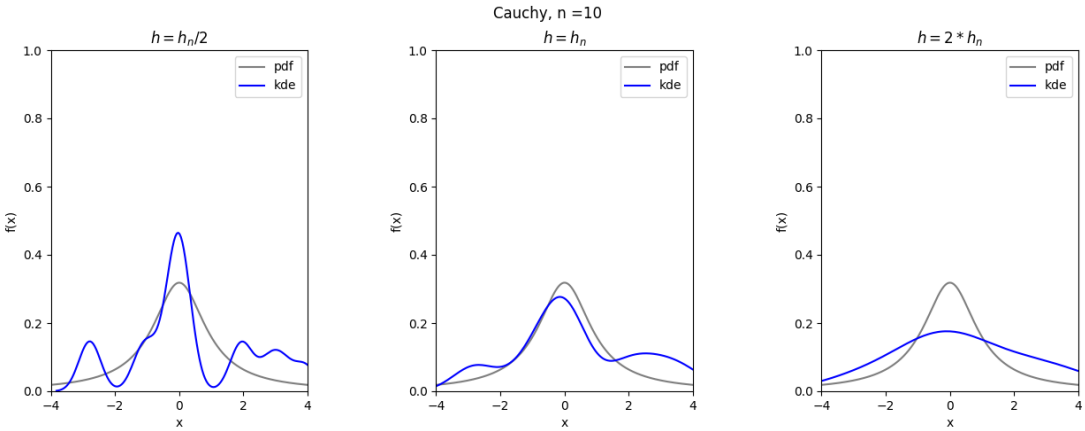

Рис. 19: Распределение Коши, $n = 10$

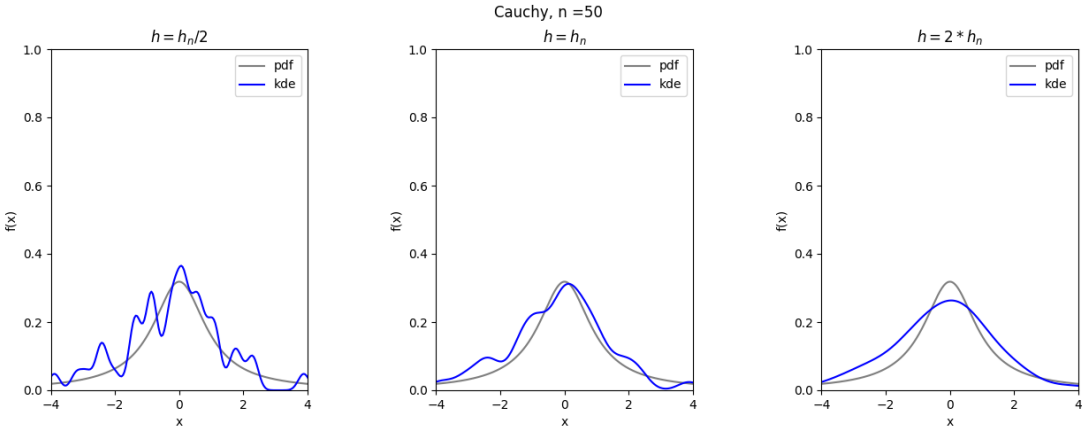

Рис. 20: Распределение Коши, $n = 50$

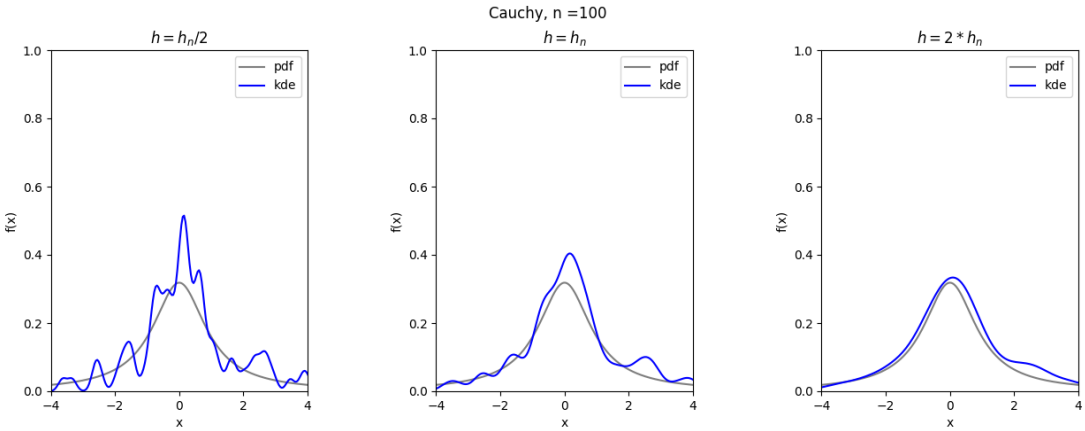

Рис. 21: Распределение Коши, $n = 100$

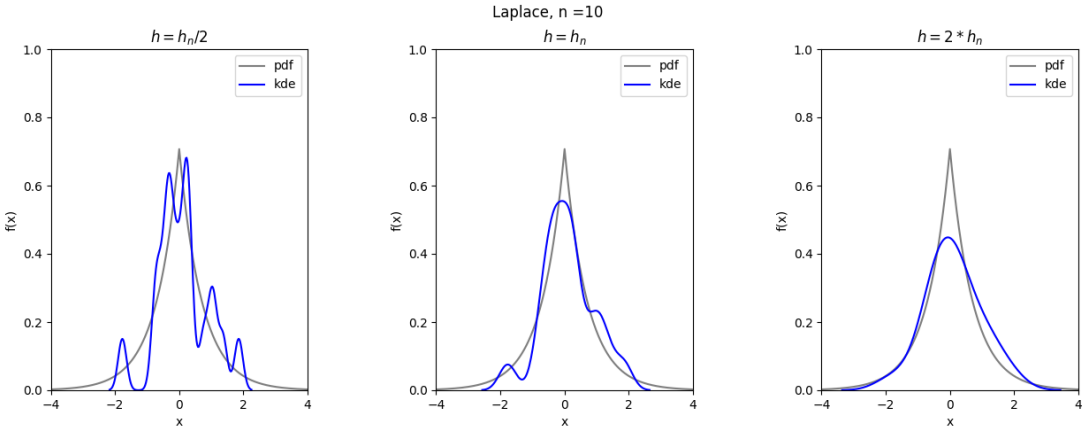

Рис. 22: Распределение Лапласа, $n = 10$

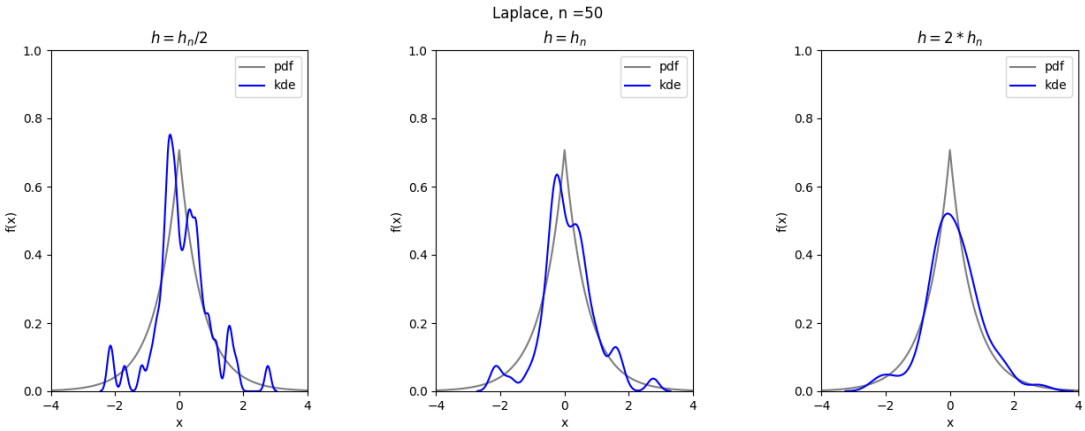

Рис. 23: Распределение Лапласа, $n = 50$

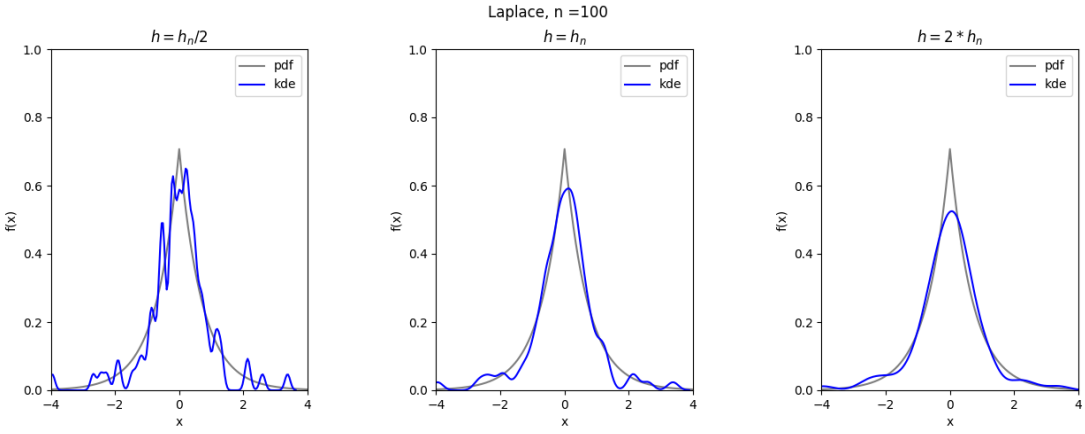

Рис. 24: Распределение Лапласа, $n = 100$

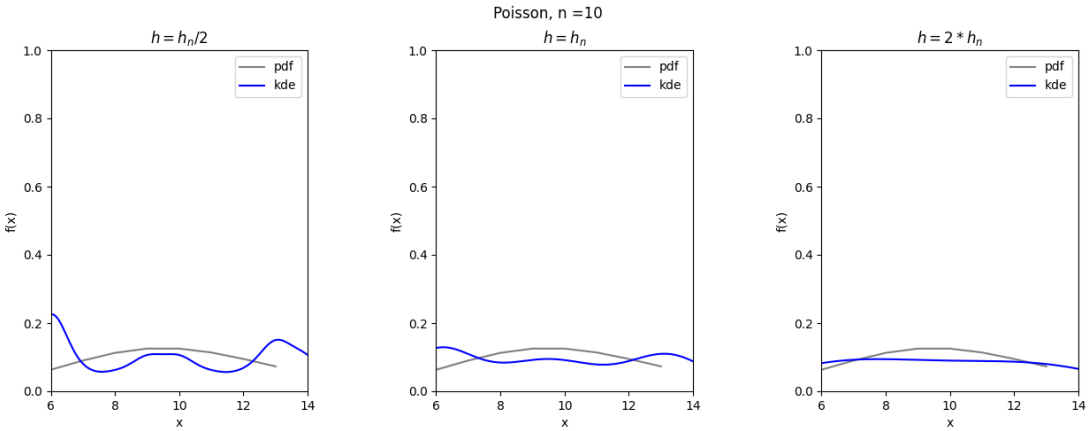

Рис. 25: Распределение Пуассона, $n = 10$

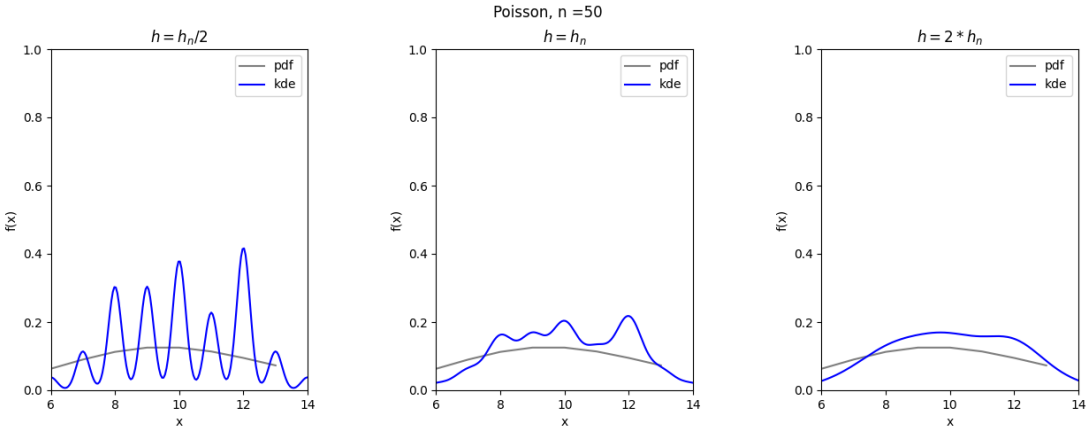

Рис. 26: Распределение Пуассона, $n = 50$

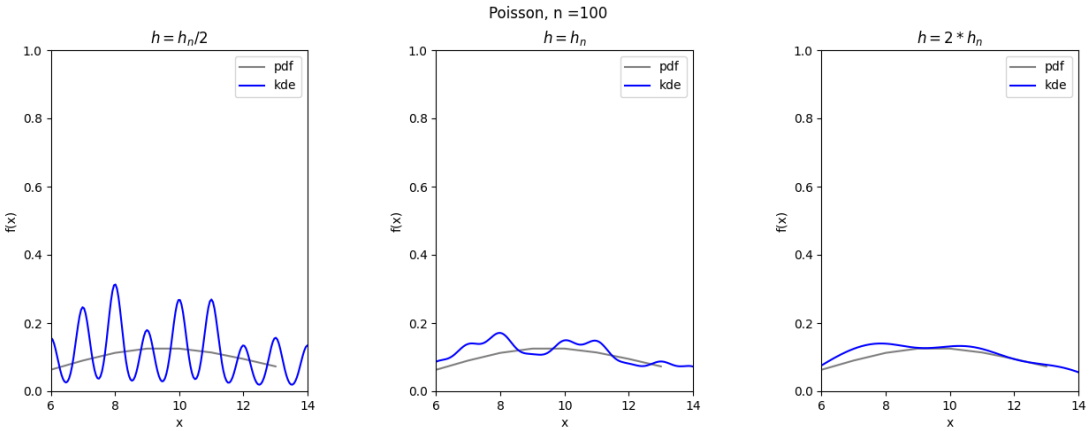

Рис. 27: Распределение Пуассона, $n = 100$

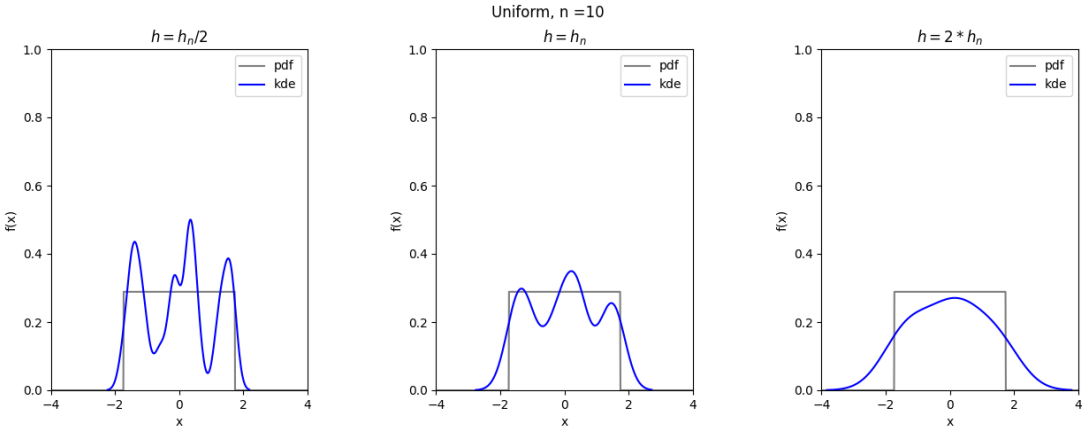

Рис. 28: Равномерное распределение, $n = 10$

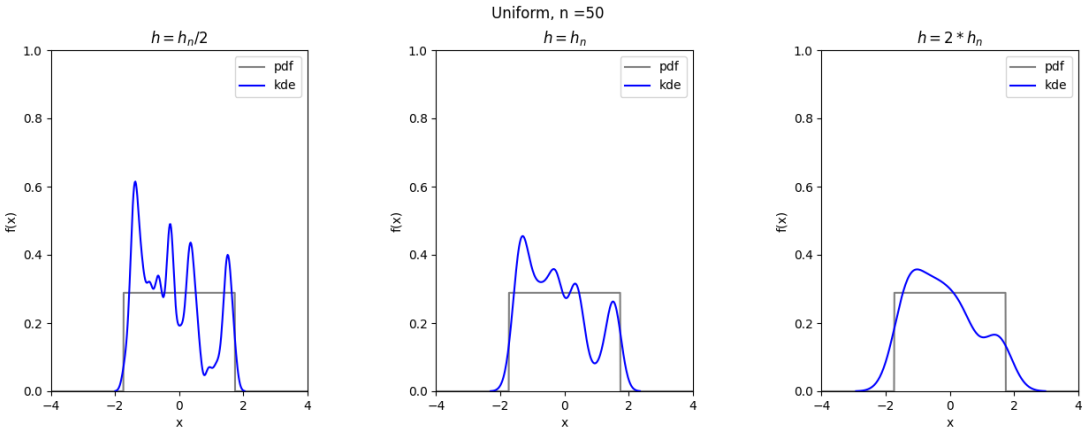

Рис. 29: Равномерное распределение, $n = 50$

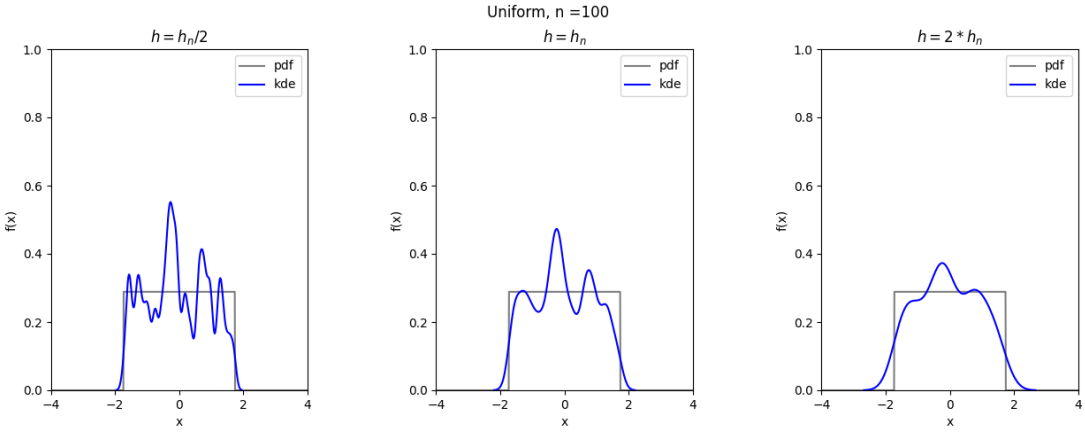

Рис. 30: Равномерное распределение, $n = 100$

5 Обсуждение

5.1 Гистограммы

Полученные результаты говорят нам о том, что с увеличением выборки для каждого из данных распределений гистограмма точнее описывается графиком плотности вероятности закона, по которому распределены величины полученной выборки. Чем меньше выборка, тем меньше выводов можно из нее сделать, т.е. по ней плохо определяется характер распределения величины.
Почти ни на одном из графиков максимумы гистограмм и плотностей распределения. Также наблюдаются всплески, которые лучше всего видны на графике распределении Коши.

5.2 Характеристики положения и рассеяния

Из приведенных таблиц можно выделить, что у распредления Коши значения дисперсии характеристик рассеяния довольно велики даже при увеличении размера выборки, что является следствием выбоpосов, наблюдаемых на гистограммах распределения.

5.3 Доля и теоретическая вероятность выбросов

По полученным таблицам можно сделать вывод о том, что чем больше выборка, тем ближе доля выбросов будет к теоретической оценке. Для распределения Коши доля выбросов значительно выше, чем для остальных распределений.

5.4 Эмпирическая функция распределения. Ядерные оценки плотности

Можем наблюдать на иллюстрациях с эмпирической функцией распредления, что ступенчатая эмпирическая функция распределения тем лучше приближает функцию распределения реальной выборки, чем мощнее эта выборка. Заметим так же, что для распределения Пуассона и равномерного распределения отклонение функций друг от друга наибольшее.
Рисунки, посвященные ядерным оценкам, иллюстрируют сближение ядерной оценки и функции плотности вероятности для всех h с ростом размера выборки. Для распределения Пуассона наиболее ярко видно, как сглаживает отклонения увеличение параметра сглаживания h.
В зависимости от особенностей распределений для их описания лучше подходят разные параметры h в ядерной оценке: для равномерного распределения и распределения Пуассона лучше подойдет параметр $h = 2h_n$, для нормального и

Коши $h = h_n$. Такие значения дают вид ядерной оценки наиболее близкий к плотности, характерной данным распределениям.
Также можно увидеть, что чем больше коэффициент при параметре сглаживания , 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 $h = h_n/2$ по полученным приближениям становится сложно сказать плотность вероятности какого распределения они должны повторять, так как они очень похожи между собой.

6 Литература

Ссылка на Github: `https://github.com/IgorKochetkov-alg/Math_Stat_Labs`