Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт
Высшая школа прикладной математики и физики

Отчёт
по лабораторным работам №1-4
по дисциплине
«Математическая статистика»

Выполнила студентка:
Зинякова Екатерина
Группа: 5030102/00201
Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2023

1 Постановка задачи

Для 5 распределений:

- 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 $U(x, -\sqrt{3}, \sqrt{3})$

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

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

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

   Вычислить оценку дисперсии по формуле:

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

   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 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 Рассматриваемые распределения

Плотности:

- 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 Лапласа

$$L(x, 0, \frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \tag{5}$$

- Распределение Пуассона

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \tag{6}$$

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при } |x| \leq \sqrt{3} \\ 0 & \text{при } |x| > \sqrt{3} \end{cases} \tag{7}$$

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 [1].

2.3 Вариационный ряд

Вариационным ряд - последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются [2, с. 409].

2.4 Выборочные числовые характеристики

2.4.1 Характеристики положения

- Выборочное среднее

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

- Выборочная медиана

$$med\, x = \begin{cases} x_{(l+1)} & \text{при } n = 2l+1 \\ \frac{x_{(l)} + x_{(l+1)}}{2} & \text{при } n = 2l \end{cases} \tag{9}$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{10}$$

- Полусумма квартилей
  Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном,} \\ x_{(np)} & \text{при } np \text{ целом.} \end{cases} \tag{11}$$

  Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{12}$$

- Усечённое среднее

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

2.4.2 Характеристики рассеяния

Выборочная дисперсия

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

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»:

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), \quad X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1) \tag{15}$$

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 [3].

2.6 Теоретическая вероятность выбросов

Выбросы − величины x:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix} \tag{16}$$

Теоретическая вероятность выбросов:

- для непрерывных распределений

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

- для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = \left(F(X_1^T) - P(x = X_1^T)\right) + \left(1 - F(X_2^T)\right). \tag{18}$$

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

2.7 Эмпирическая функция распределения

2.7.1 Статистический ряд

Статистическим ряд − последовательность различных элементов выборки z_1, z_2, ... , z_k, расположенных 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:

$$F_n^*(x) = P^*(X < x). \tag{19}$$

2.7.3 Нахождение э. ф. р.

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 $P^*(X < x) = \frac{1}{n}\sum\limits_{z_i<x} n_i$. Получаем

$$F^*(x) = \frac{1}{n}\sum_{z_i<x} n_i. \tag{20}$$

$F^*(x)-$ 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 1: Таблица распределения

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

$$F_n^*(x) \approx F_X(x). \tag{21}$$

2.8 Оценки плотности вероятности

2.8.1 Определение

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

$$\hat{f}(x) \approx f(x). \tag{22}$$

2.8.2 Ядерные оценки

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

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right). \tag{23}$$

Здесь функция $K(u)$, называемая ядерной (ядром), непрерывна и является плотностью вероятности, x_1, x_2, ... , x_n – элементы выборки, h_n – любая последовательность положительных чисел, обладающая свойствами

$$h_n \xrightarrow[n\to\infty]{} 0; \quad \frac{h_n}{n^{-1}} \xrightarrow[n\to\infty]{} \infty. \tag{24}$$

Гауссово (нормальное) ядро [4, с.38]

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{25}$$

Правило Сильвермана [4, с.44]

$$h_n = 1.06\hat{\sigma}n^{-1/5}, \tag{26}$$

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

3 Реализация

Лабораторная работы выполнена на языке программирования Python в среде разработке PyCharm с использованием встроенных библиотек(numpy, statsmodels, matplotlib, scipy, seaborn)

4 Результаты

4.1 Гистограмма и график плотности распределения

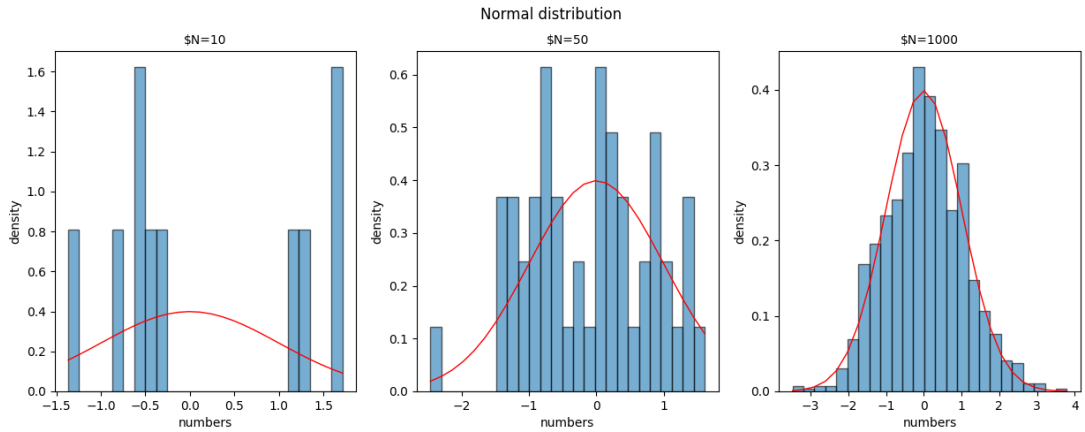

Рис. 1: Нормальное распределение.

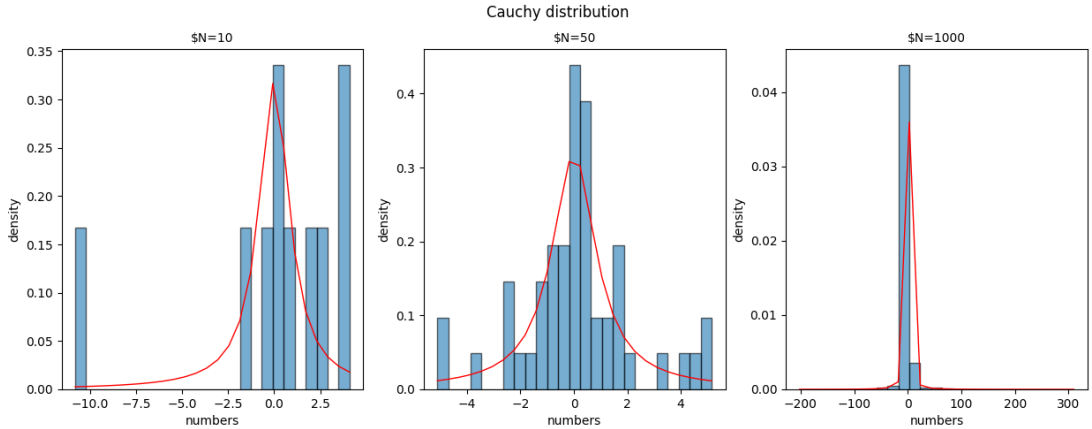

Рис. 2: Распределение Коши.

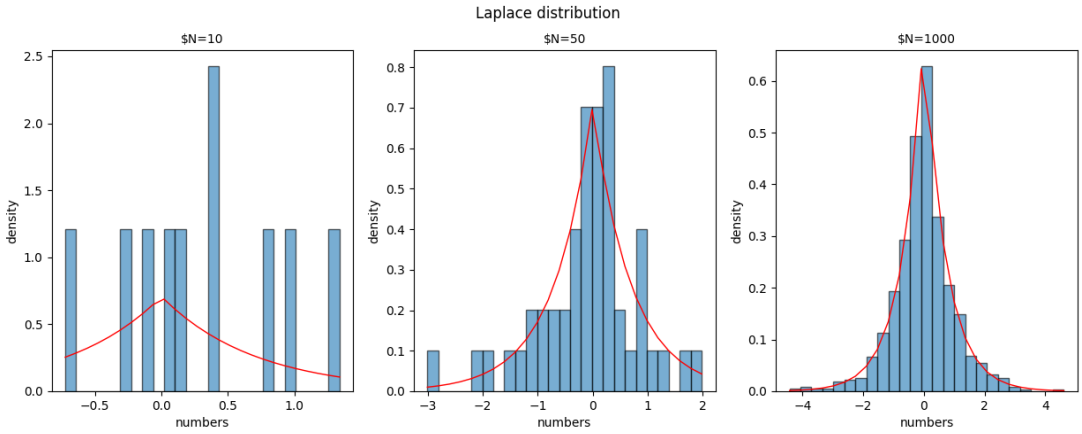

Рис. 3: Распределение Лапласс.

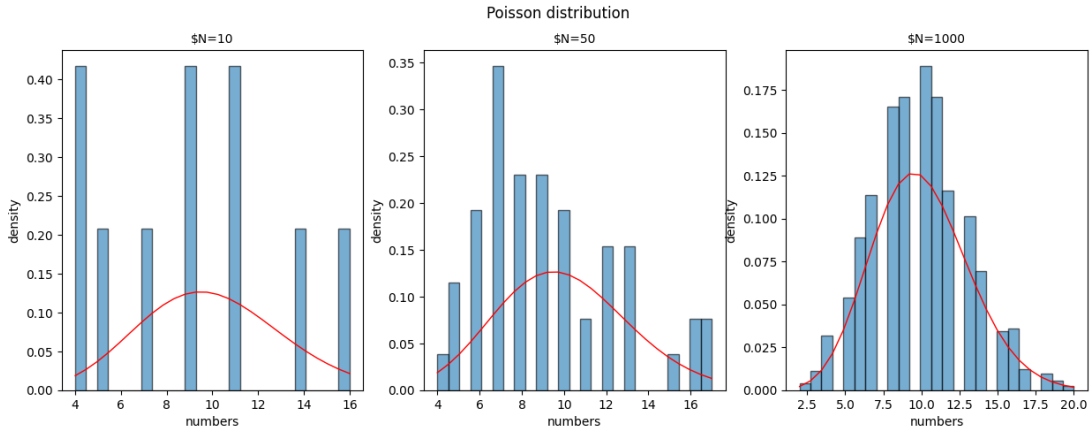

Рис. 4: Распределение Пуассона.

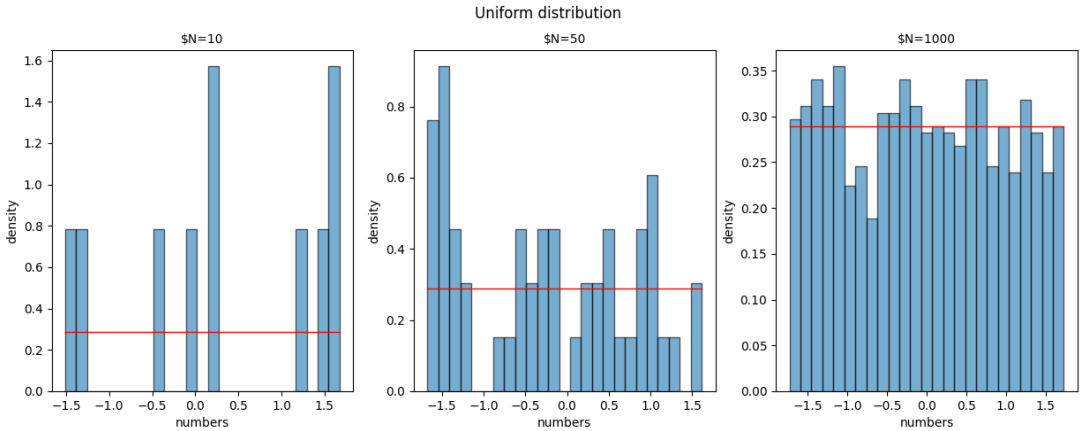

Рис. 5: Равномерное распределение.

4.2 Характеристики положения и рассеяния

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0012	0.0012	-0.0345	0.3086	0.0161
D(z)	0.0994	0.0994	0.5267	0.123	0.1708

Таблица 2: Нормальное распределение n = 10.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0006	0.0006	-0.027	0.1861	0.0035
D(z)	0.0595	0.0595	0.5036	0.0879	0.1042

Таблица 3: Нормальное распределение n = 50.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0	0.0	-0.0167	0.1243	0.0029
D(z)	0.04	0.04	0.4953	0.0666	0.0702

Таблица 4: Нормальное распределение n = 1000.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.2169	0.2169	0.0299	0.3718	0.1618
D(z)	144.3643	144.3643	141.3867	1.5929	336.6468

Таблица 5: Распределение Коши n = 10.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.1471	0.1471	-0.0774	0.3263	-0.2792
D(z)	290.2251	290.2251	150.8196	1.3051	623.1905

Таблица 6: Распределение Коши n = 50.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0636	0.0636	-0.4299	0.2724	-0.4007
D(z)	383.4691	383.4691	813.6458	1.1029	1029.9574

Таблица 7: Распределение Коши n = 1000.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0559	0.0559	-0.3656	0.2769	-0.3418
D(z)	328.7027	328.7027	697.5107	0.9625	882.8661

Таблица 8: Распределение Лапласа n = 10.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0487	0.0487	-0.3164	0.25	-0.2992
D(z)	287.6177	287.6177	610.4052	0.8499	772.5251

Таблица 9: Распределение Лапласа n = 50.

		mediana	z_r	z_Q	$z_t r$
E(z)	0.0434	0.0434	-0.2788	0.2226	-0.2659
D(z)	255.6605	255.6605	542.6472	0.7616	686.6981

Таблица 10: Распределение Лапласа n = 1000.

		mediana	z_r	z_Q	$z_t r$
E(z)	1.0363	1.0363	0.7448	1.2926	0.7553
D(z)	239.0636	239.0636	498.2958	11.1272	627.5751

Таблица 11: Распределение Пуассона n = 10.

		mediana	z_r	z_Q	$z_t r$
E(z)	1.8505	1.8505	1.5918	2.0931	1.5951
D(z)	223.9773	223.9773	460.6414	16.5504	577.6113

Таблица 12: Распределение Пуассона n = 50.

		mediana	z_r	z_Q	$z_t r$
E(z)	2.5294	2.5294	2.2849	2.7516	2.2953
D(z)	210.3835	210.3835	427.9188	19.9411	534.8715

Таблица 13: Распределение Пуассона n = 1000.

		mediana	z_r	z_Q	$z_t r$
E(z)	2.3336	2.3336	2.1087	2.5634	2.1171
D(z)	194.6675	194.6675	395.4103	18.8418	494.1211

Таблица 14: Равномерное распределение n = 10.

		mediana	z_r	z_Q	$z_t r$
E(z)	2.1663	2.1663	1.9578	2.3845	1.9653
D(z)	181.1281	181.1281	367.4996	17.9141	459.1292

Таблица 15: Равномерное распределение n = 50.

		mediana	z_r	z_Q	$z_t r$
E(z)	2.0219	2.0219	1.8266	2.2256	1.8344
D(z)	169.3449	169.3449	343.2738	17.0731	428.7605

Таблица 16: Равномерное распределение n = 1000.

4.3 Боксплот Тьюки

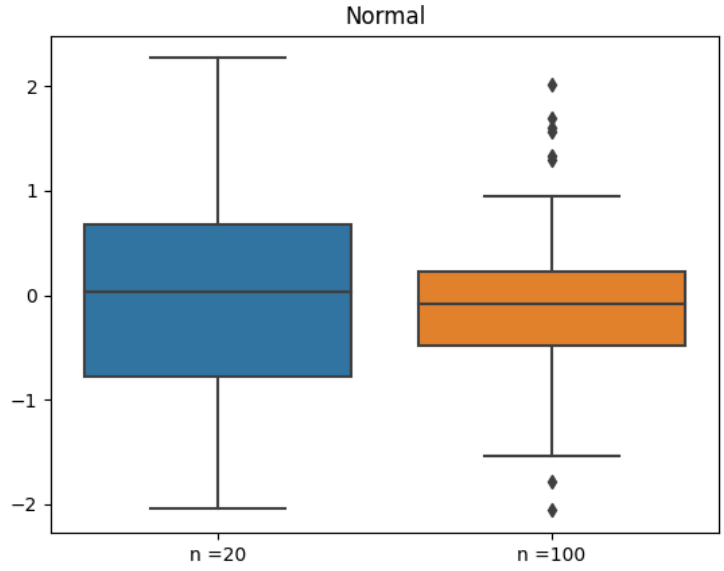

Рис. 6: Нормальное распределение.

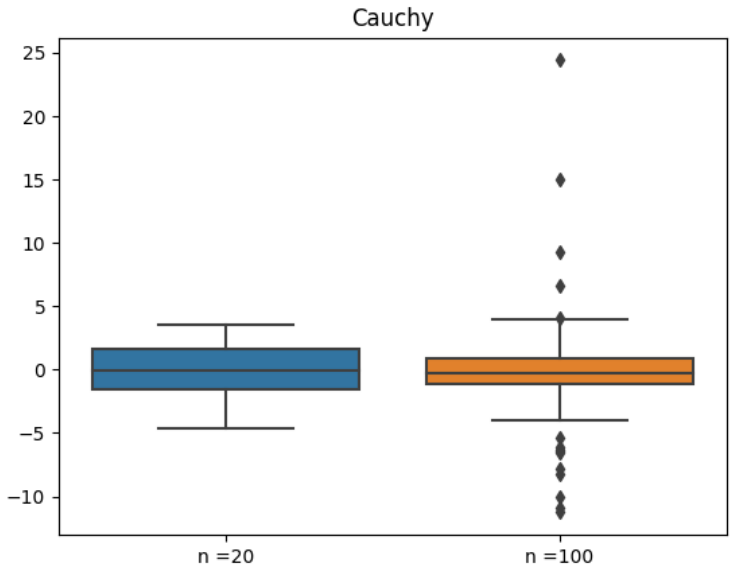

Рис. 7: Распределение Коши.

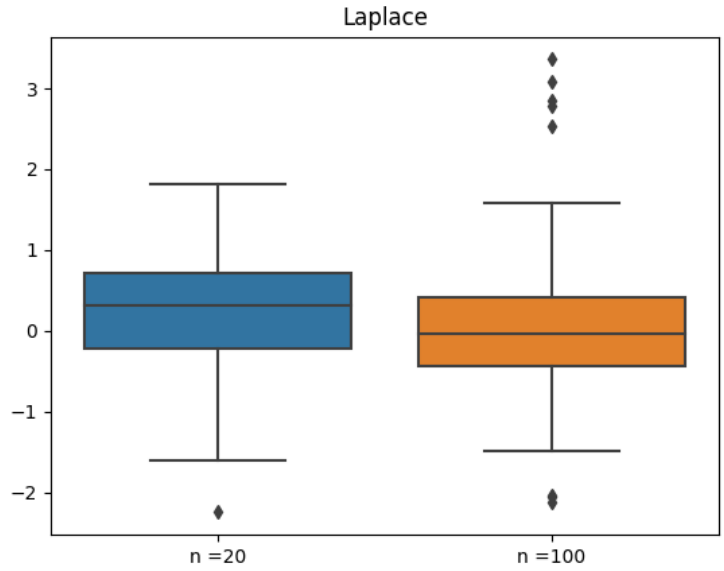

Рис. 8: Распределение Лапласа.

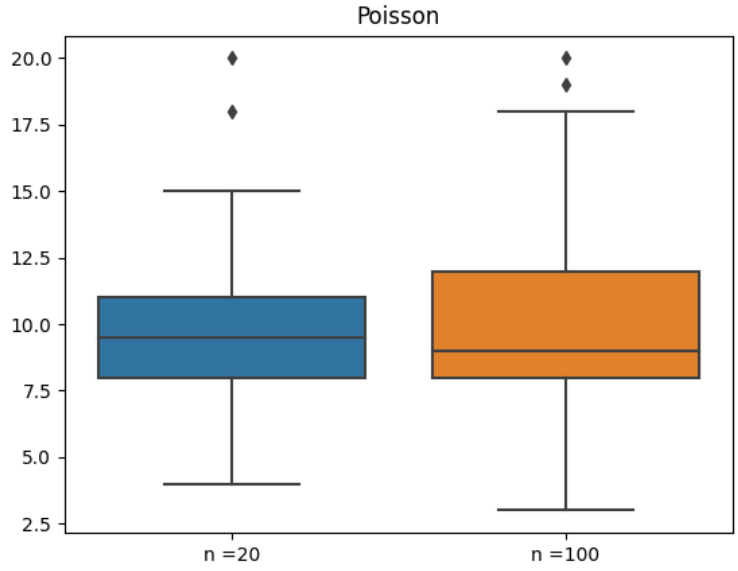

Рис. 9: Распределение Пуассона.

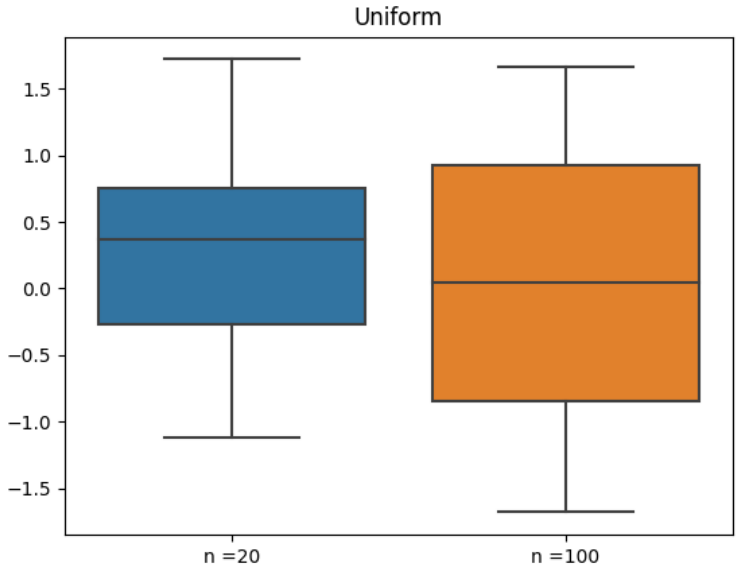

Рис. 10: Равномерное распределение.

4.4 Доля выбросов

Sample	Share of emissions
Normal n = 20	0.109
Normal n = 100	0.096
Cauchy n = 20	0.215
Cauchy n = 100	0.224
Laplace n = 20	0.157
Laplace n = 100	0.158
Poisson n = 20	0.111
Poisson n = 100	0.092
Uniform n = 20	0.052
Uniform n = 100	0.023

Таблица 17: Экспериментальная доля выбросов.

4.5 Теоретическая вероятность выбросов

Распределение	$P_B^T(17),(18)$
Нормальное распределение	0.007
Распределение Коши	0.156
Распределение Лапласа	0.063
Распределение Пуассона	0.008
Равномерное распределение	0

Таблица 18: Теоретическая вероятность выбросов.

4.6 Эмпирическая функция распределения

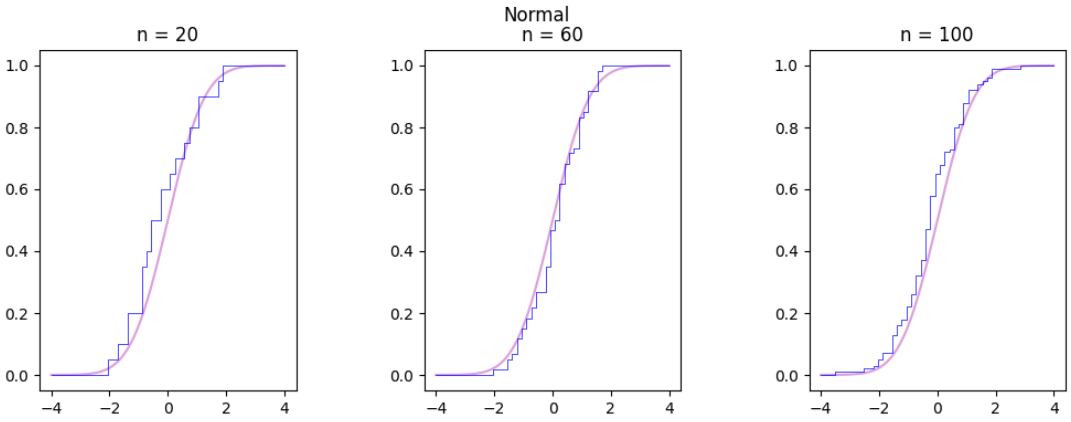

Рис. 11: Нормальное распределение.

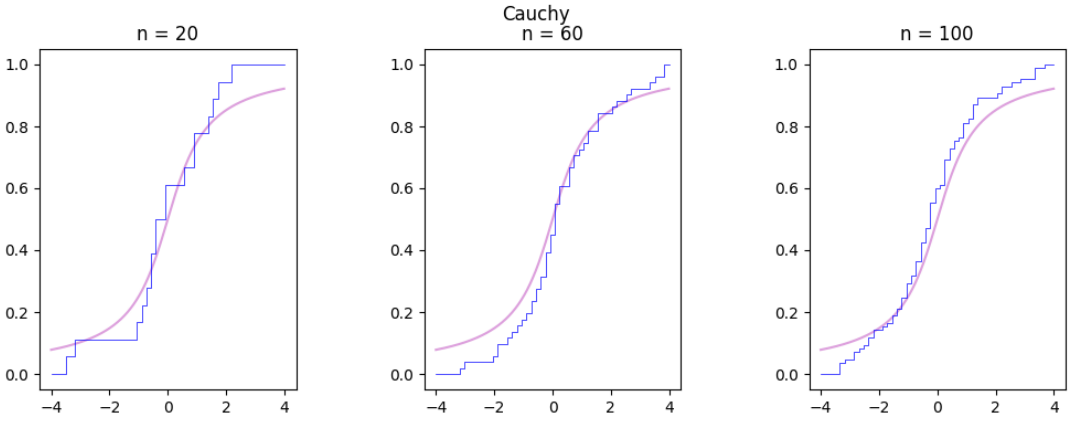

Рис. 12: Распределение Коши.

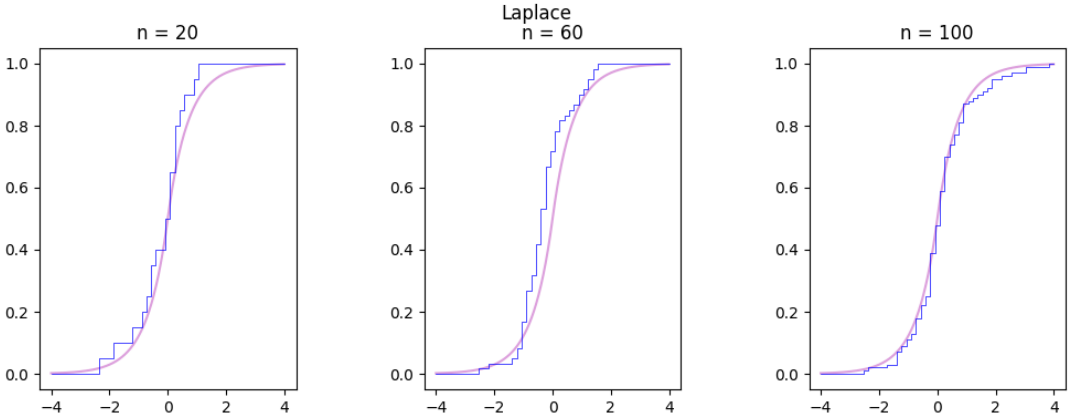

Рис. 13: Распределение Лапласа.

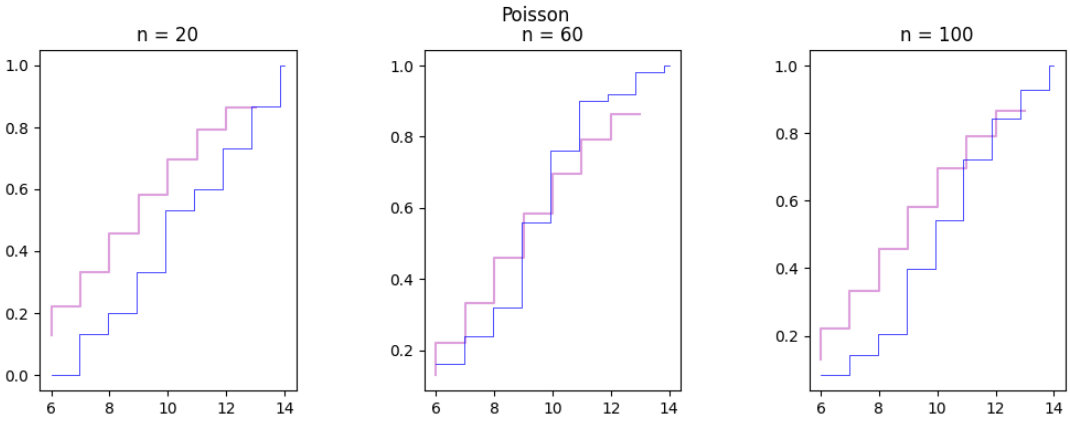

Рис. 14: Распределение Пуассона.

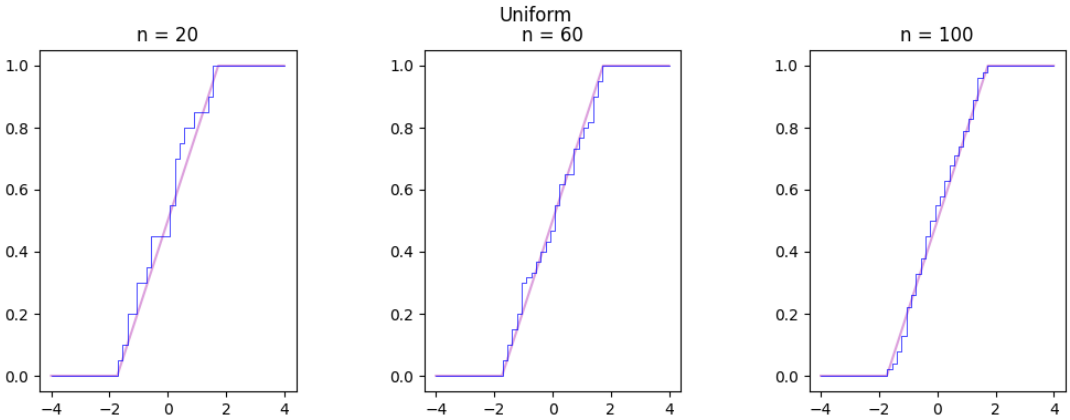

Рис. 15: Равномерное распределение.

4.7 Ядерные оценки плотности распределения

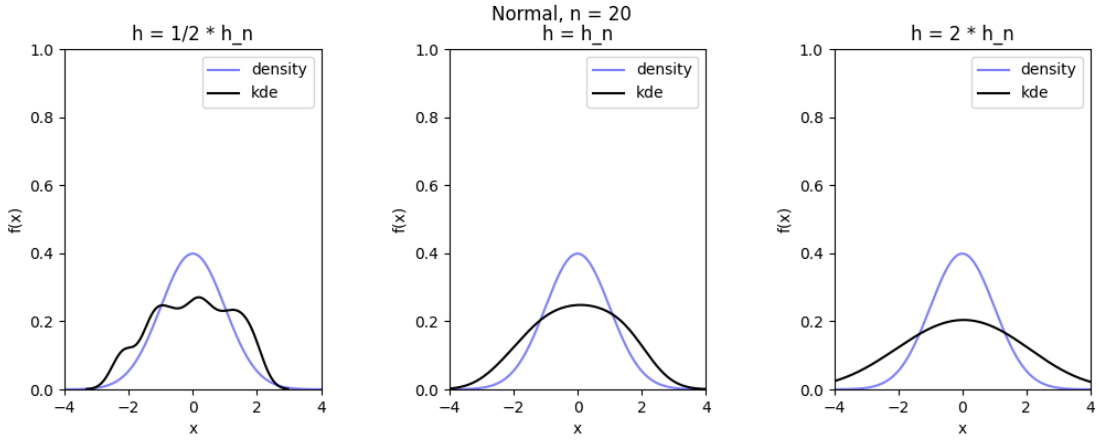

Рис. 16: Нормальное распределение n = 20.

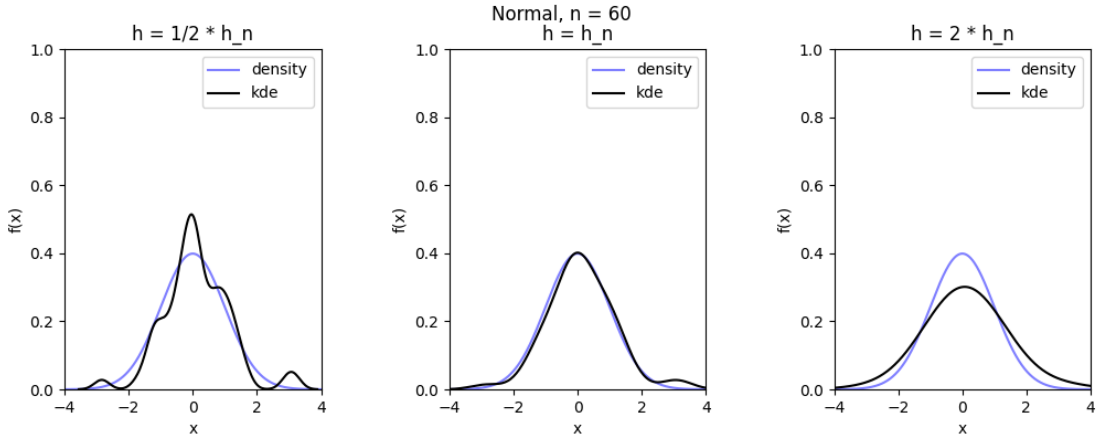

Рис. 17: Нормальное распределение n = 60.

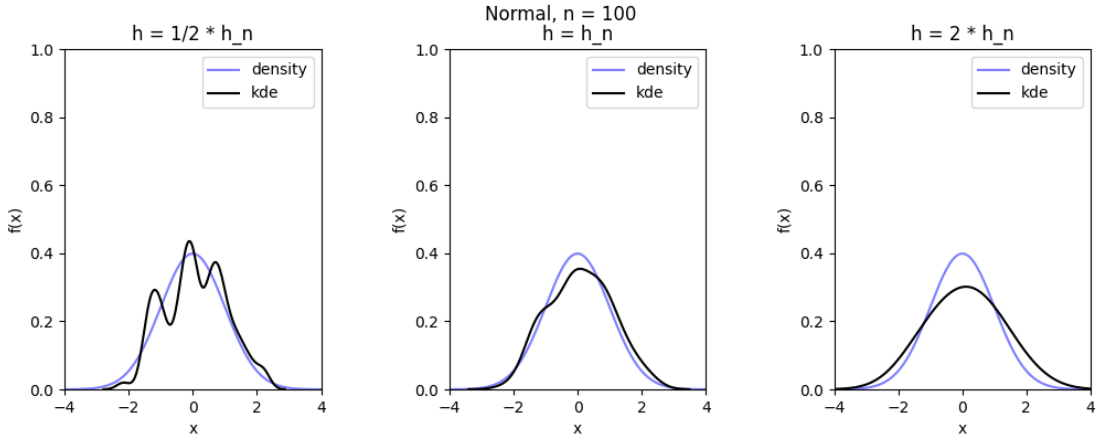

Рис. 18: Нормальное распределение n = 100.

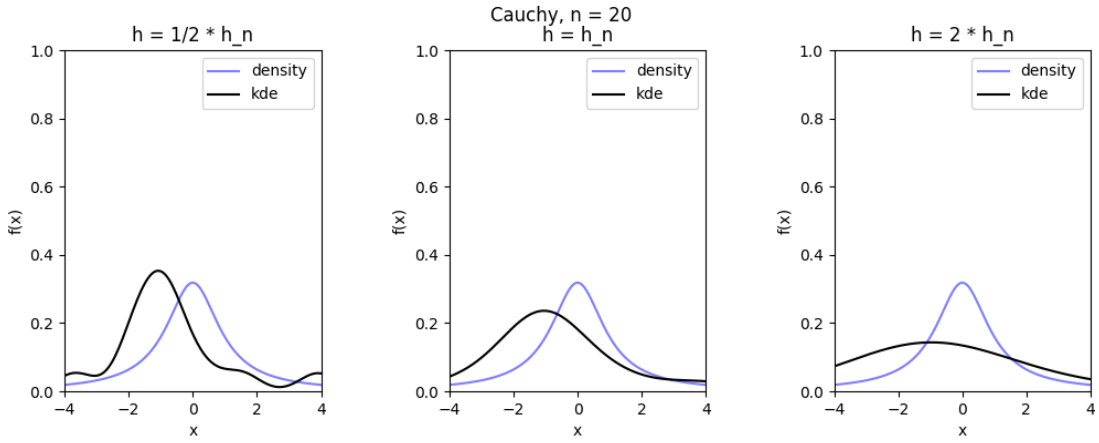

Рис. 19: Распределение Коши n = 20.

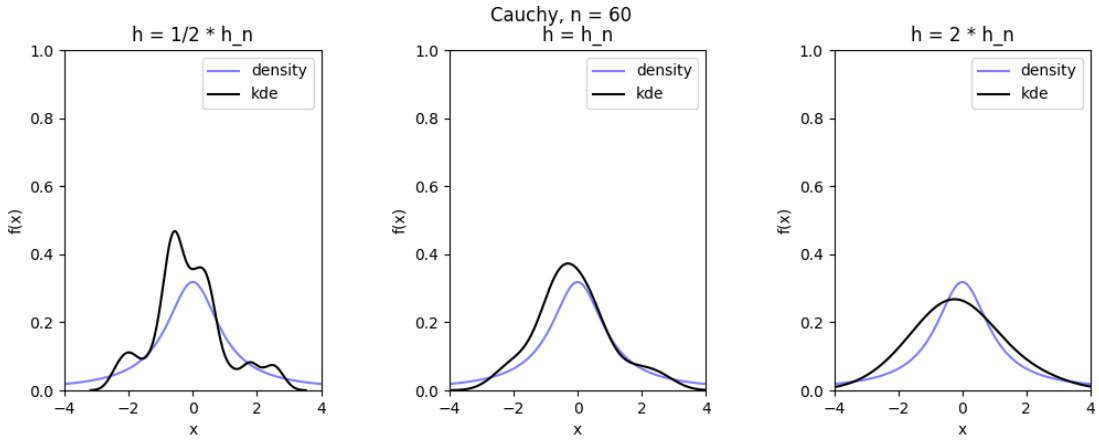

Рис. 20: Распределение Коши n = 60.

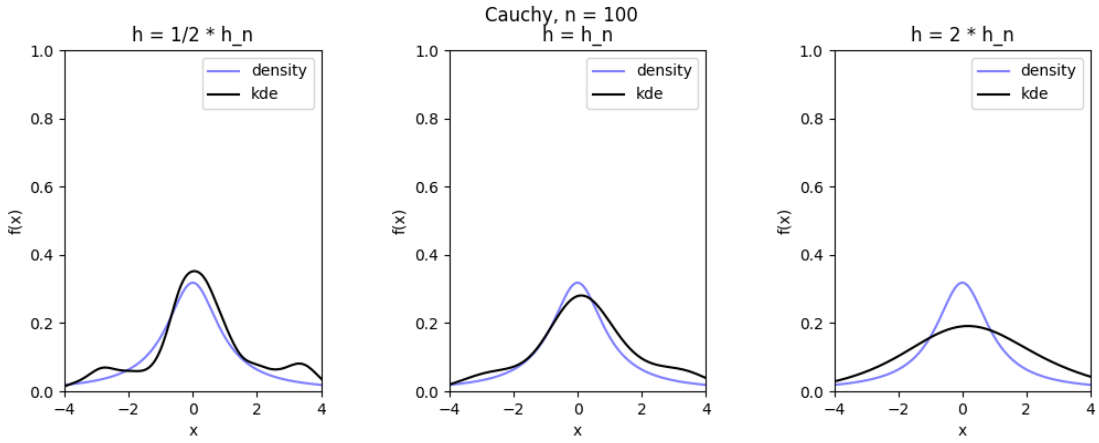

Рис. 21: Распределение Коши n = 100.

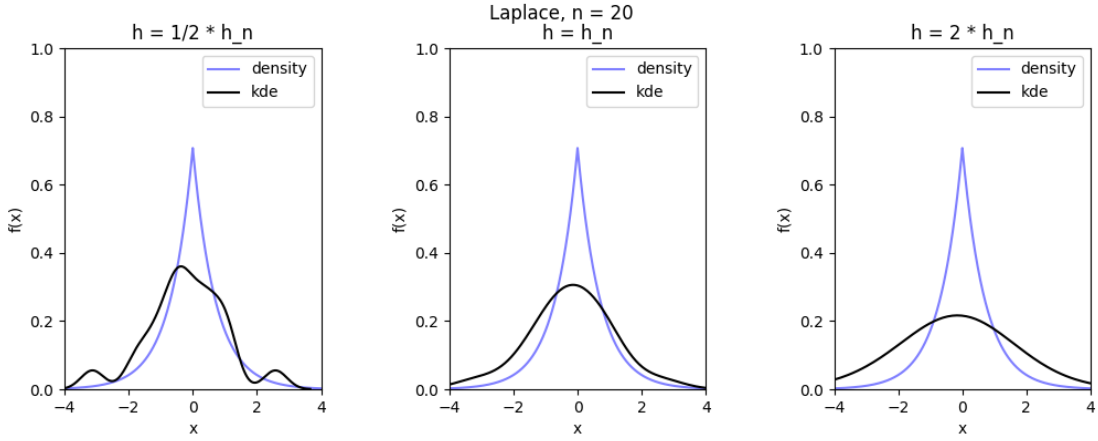

Рис. 22: Распределение Лапласа n = 20.

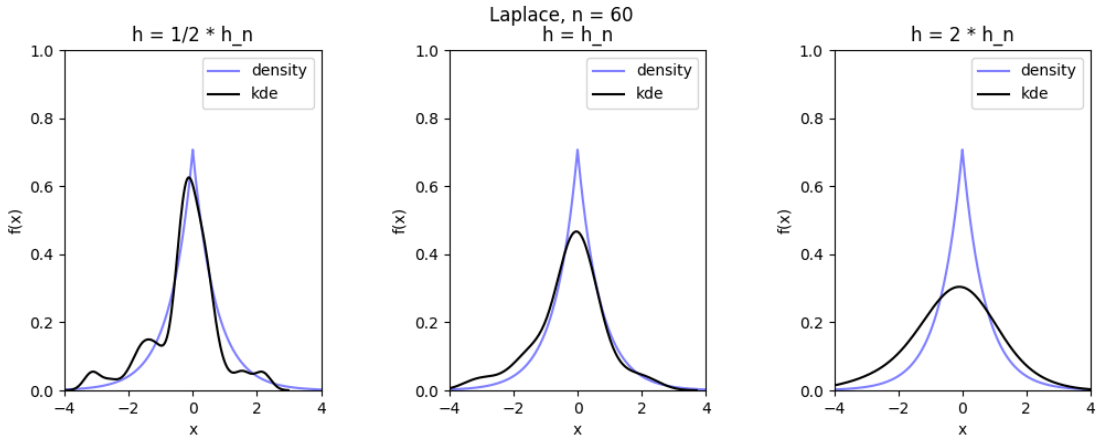

Рис. 23: Распределение Лапласа n = 60.

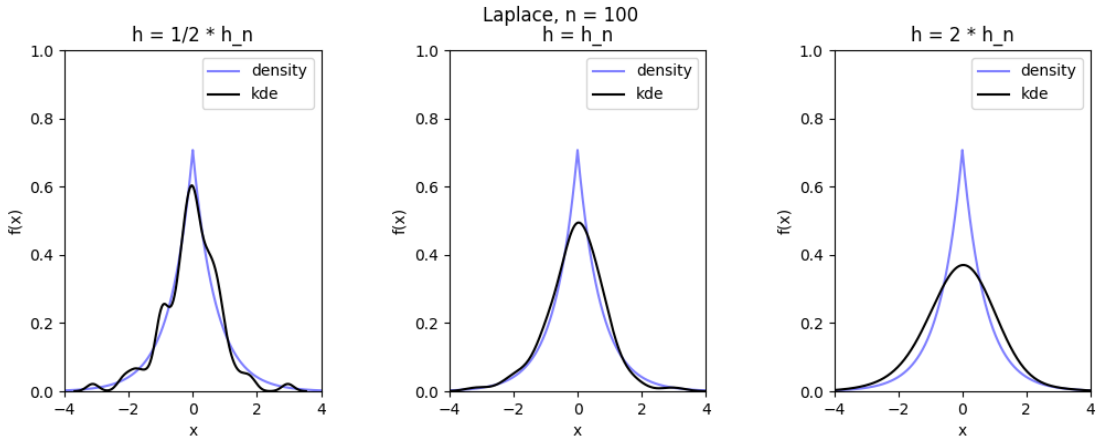

Рис. 24: Распределение Лапласа n = 100.

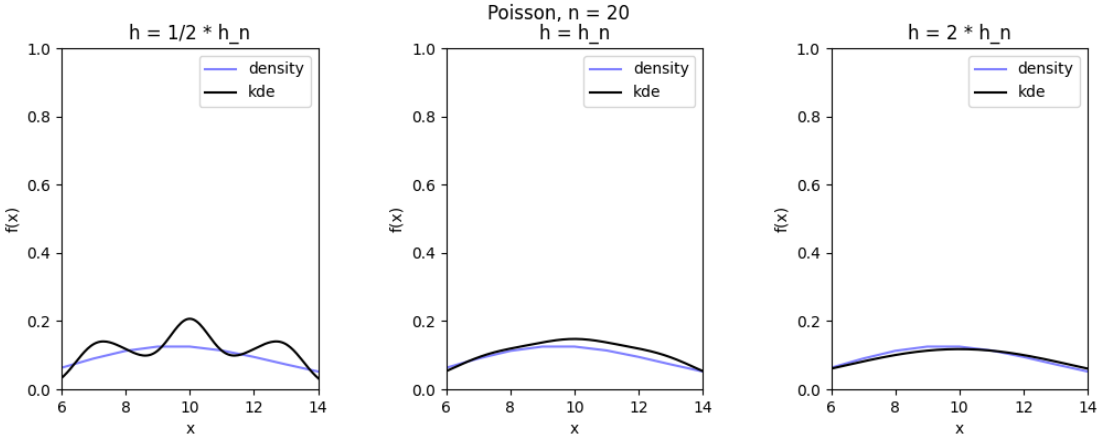

Рис. 25: Распределение Пуассона n = 20.

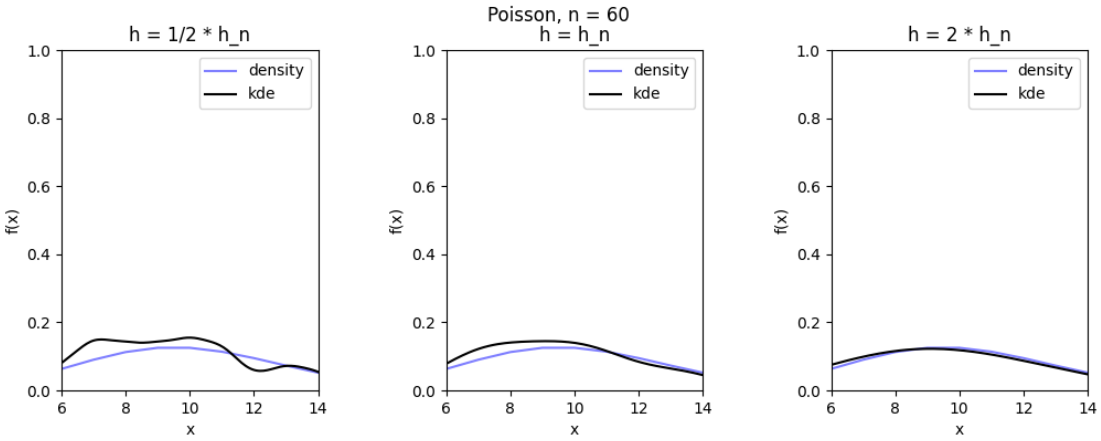

Рис. 26: Распределение Пуассона n = 60.

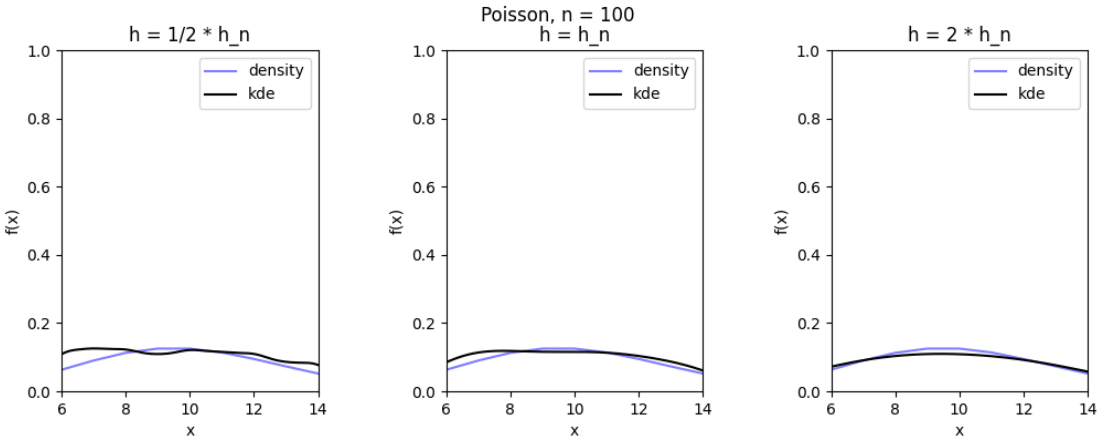

Рис. 27: Распределение Пуассона n = 100.

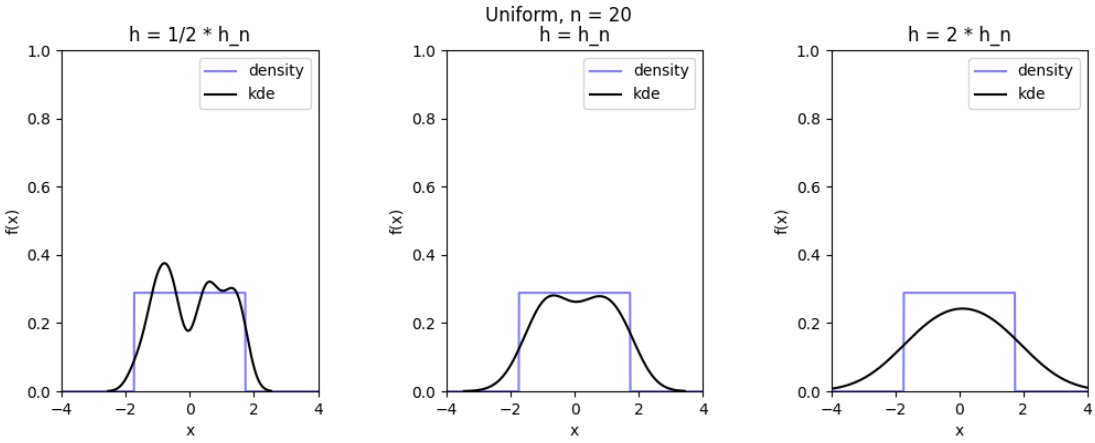

Рис. 28: Равномерное распределение n = 20.

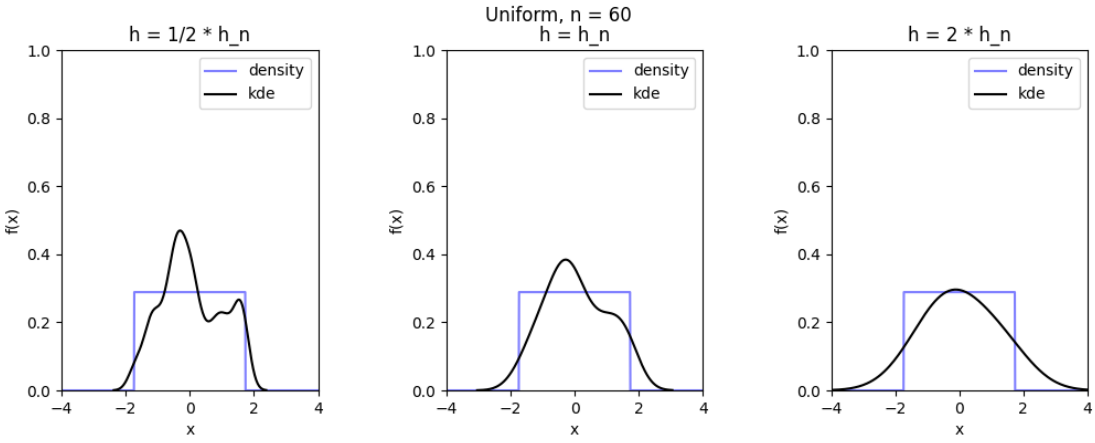

Рис. 29: Равномерное распределение n = 60.

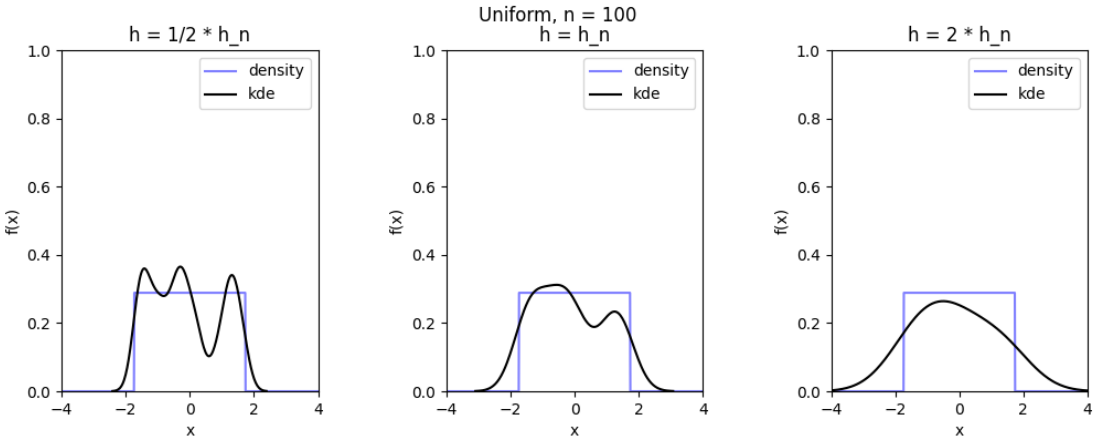

Рис. 30: Равномерное распределение n = 100.

5 Обсуждение

5.1 Гистограмма и график плотности распределения

Проделав лабораторную работу и проанализировов результаты,можем сделать вывод о том, что чем больше выборка для каждого из распределений, тем ближе ее гистограмма к графику плотности вероятности того закона, по которому распределены величины сгенерированной выборки. Чем меньше выборка, тем менее она показательна — тем хуже по ней определяется характер распределения величины.
Также можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Также наблюдаются всплески гистограмм, что наиболее хорошо прослеживается на распределении Коши.

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 Доля и теоретическая вероятность 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.
Боксплоты Тьюки действительно позволяют более наглядно и с меньшими усилиями оценивать важные характеристики распределений. Так, исходя из полученных рисунков, наглядно видно то, что мы довольно трудоёмко анализировали в предыдущих частях.

5.4 Эмпирическая функция и ядерные оценки плотности распределения

Можем наблюдать на иллюстрациях с э. ф. р., что ступенчатая эмпирическая функция распределения тем лучше приближает функцию распределения реальной выборки, чем мощнее эта выборка. Заметим так же, что для распределения Пуассона и распределения Коши 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 распределения Лапласа — $h = \frac{h_n}{2}$, а 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 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.

6 Ссылка на репозиторий

https://github.com/Katalien/Math$_S$*tatistics*

Список литературы

[1] Histogram. URL: https://en.wikipedia.org/wiki/Histogram.

[2] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. – Спб.: «Иван Федоров», 2001. – 592 с., илл.

[3] Box plot. URL: https://en.wikipedia.org/wiki/Box_plot.

[4] Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37-52.