Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Кафедра "Прикладная математика"

Отчет по лабораторным работам №1-4 по дисциплине "Математическая статистика"

Выполнил студент:

Кротиков Сергей Ильич

группа: 5030102/90101

Проверил:

к.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург, 2022

СОДЕРЖАНИЕ

СПИСОК ИЛЛЮСТРАЦИЙ

СПИСОК ТАБЛИЦ

1 Постановка задачи

Для 5 распределений:

- $N(x, 0, 1)$ – нормальное распределение
- $C(x, 0, 1)$ – распределение Коши
- $L(x, 0, \frac{1}{\sqrt{2}})$ – распределение Лапласа
- $P(k, 10)$ – распределение Пуассона
- $U(x, -\sqrt{3}, \sqrt{3})$ – равномерное распределение

1. Сгенерировать выборки размером 10, 100 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 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

2.3.1 Определение

Вариационным рядом называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются. За-

пись 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

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

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.

2.6 Теоретическая вероятность выбросов

Встроенными средствами языка программирования Python в среде разработки PyCharm 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:

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

Теоретическая вероятность выбросов

- для непрерывных распределений

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

- для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > x_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T)) \tag{18}$$

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

2.7 Эмпирическая функция распределения

2.7.1 Статистический ряд

Статистическим рядом назовем совокупность, состоящую из последовательности $\{z_i\}_{i=1}^k$ попарно различных элементов выборки, расположенных по возрастанию, и последовательности $\{n_i\}_{i=1}^k$ частот, с которыми эти элементы содержатся в выборке.

2.7.2 Эмпирическая функция распределения

Эмпирическая 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 $P^*(X < x) = \frac{1}{n}\sum_{z_i<x} n_i$. 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	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

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

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 $f(x)$

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

2.8.2 Ядерные оценки

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

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

$K(u)$ - ядро, т. е. непрерывная функция, являющаяся плотностью вероятности, $x_1, ..., x_n$ – элементы выборки, а $\{h_n\}_{n\in\mathbb{N}}$ - последовательность элементов из $\mathbb{R}_+$

такая, что

$$h_n \xrightarrow[n\to\infty]{} 0; \quad nh_n \xrightarrow[n\to\infty]{} \infty. \tag{24}$$

Такие оценки называются непрерывными ядерными.

Гауссово ядро:

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

Правило Сильвермана:

$$h_n = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5}, \tag{26}$$

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

3 Программная реализация

Лабораторная работа выполнена на языке Python вресии 3.7 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. scipy
2. numpy
3. matplotlib
4. math

В приложении находится ссылка на GitHub репозиторий с исходныи кодом.

4 Результаты

4.1 Гистограммы и графики плотности распределения

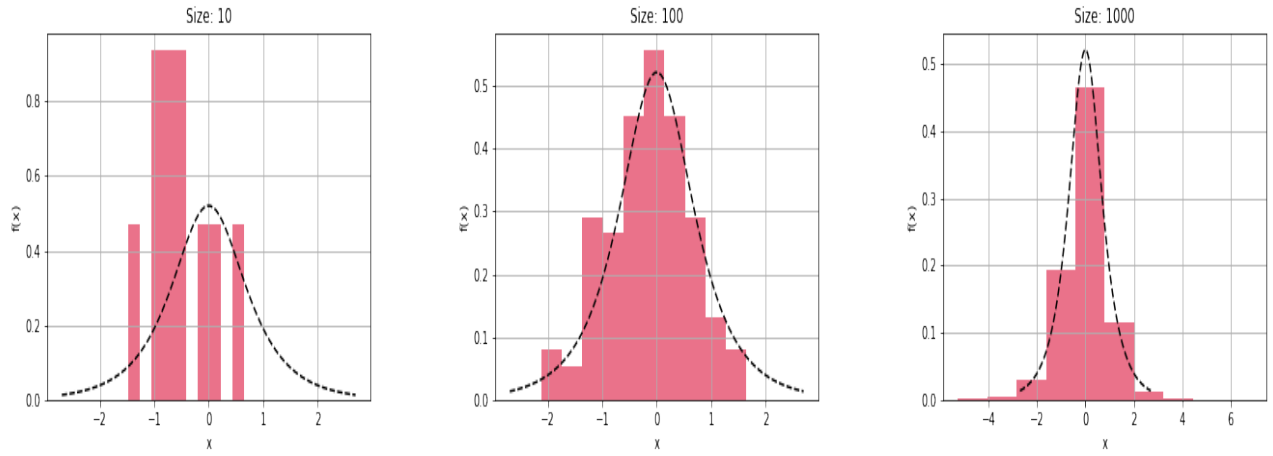

Рис. 1: Нормальное распределение (3)

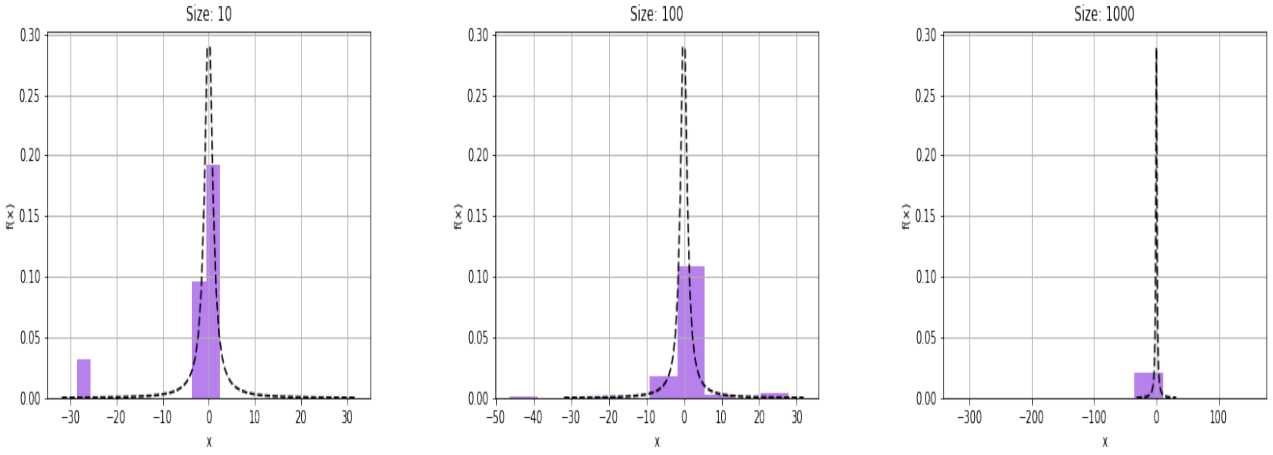

Рис. 2: Распределение Коши (4)

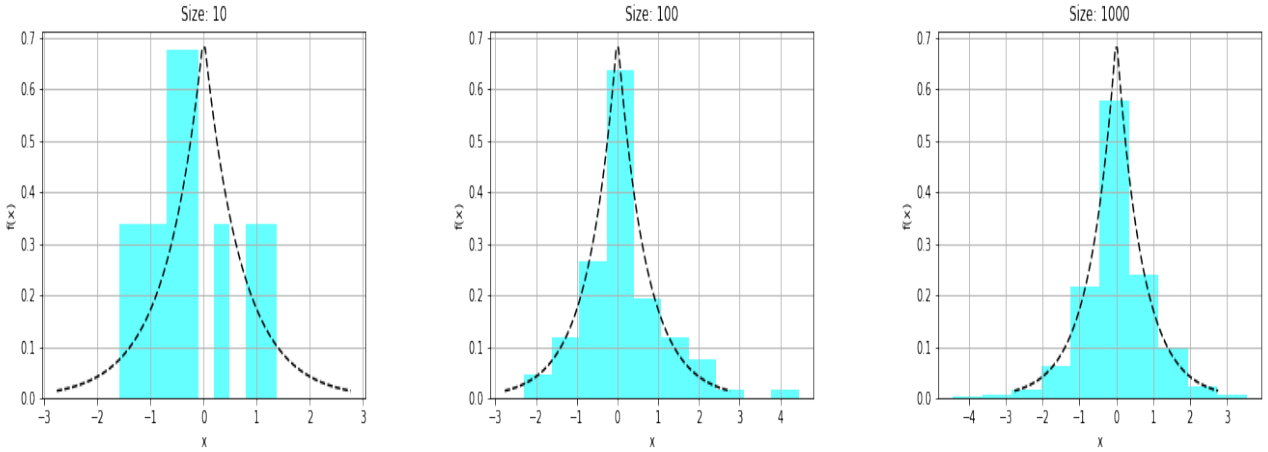

Рис. 3: Распределение Лапласа (5)

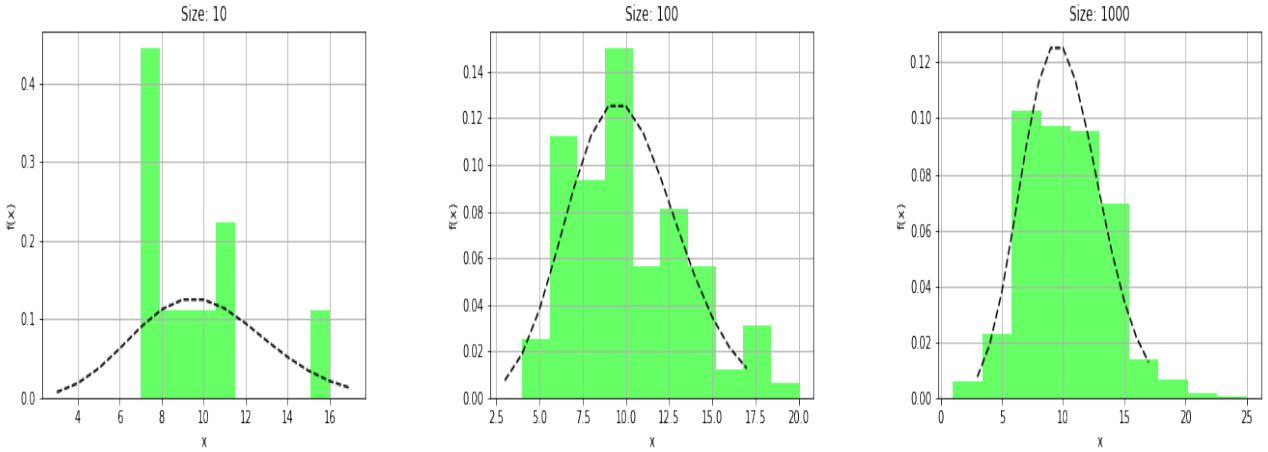

Рис. 4: Распределение Пуассона (6)

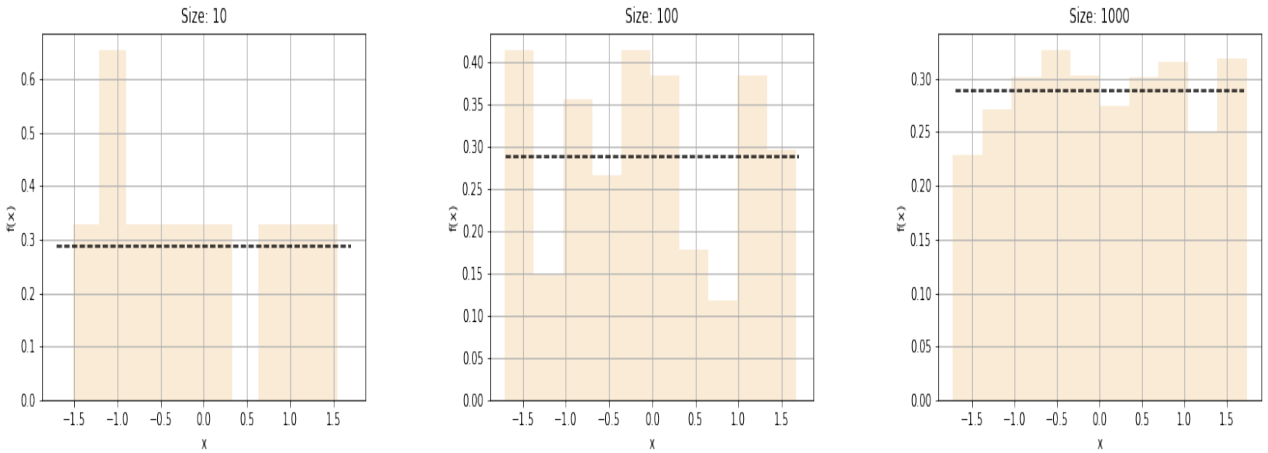

Рис. 5: Равномерное распределение (7)

4.2 Характеристики положения и рассеяния

Как было проведено округление:
В оценке $x = E \pm D$ вариации подлежит первая цифра после точки. В данном случае $x = 0.0 \pm 0.1k$, k - зависит от доверительной вероятности и вида распределения (рассматривается в дальнейшем цикле лабораторных работ) Округление сделано для $k = 1$

normal $n = 10$					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.025	0.030	0.015	0.337	0.306
$D(z)$	0.095	0.131	0.194	0.118	0.107
$E(z) \pm \sqrt{D(z)}$	[-0.284, 0.334]	[-0.331, 0.392]	[-0.425, 0.455]	[-0.006, 0.681]	[-0.020, 0.634]
normal $n = 100$					
$E(z)$	-0.002	-0.005	0.000	0.012	0.024
$D(z)$	0.010	0.016	0.096	0.012	0.012
$E(z) \pm \sqrt{D(z)}$	[-0.103, 0.098]	[-0.130, 0.119]	[-0.309, 0.310]	[-0.098, 0.123]	[-0.085, 0.134]
normal $n = 1000$					
$E(z)$	0.000888	-0.000149	-0.003741	0.002063	0.002709
$D(z)$	0.00102	0.00168	0.06555	0.00126	0.00126
$E(z) \pm \sqrt{D(z)}$	[-0.031081, 0.032857]	[-0.0411, 0.040802]	[-0.259772, 0.25229]	[-0.033448, 0.037574]	[-0.032886 0.038304]

Таблица 2: Нормальное распределение (3)

Cauchy $n = 10$					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	-7.345	0.023	-36.745	1.106	0.695
$D(z)$	47446.036	0.322	1186442.299	5.203	1.289
$E(z) \pm \sqrt{D(z)}$	[-225.166, 210.476]	[-0.543, 0.591]	[-1125.985, 1052.493]	[-1.174, 3.387]	[-0.439 1.831]
Cauchy $n = 100$					
$E(z)$	-1.3446	-0.007154	-65.8962	0.0223	0.0336
$D(z)$	846.790	0.022	2091172.999	0.050	0.024
$E(z) \pm \sqrt{D(z)}$	[-30.4443, 27.7549]	[-0.1584, 0.1441]	[-1511.9851, 1380.1926]	[-0.2019, 0.2466]	[-0.1213, 0.1887]
Cauchy $n = 1000$					
$E(z)$	-1.493	0.000	-737.525	0.004	0.004
$D(z)$	1692.9096	0.0023	422149639	0.0051	0.0024
$E(z) \pm \sqrt{D(z)}$	[-42.638, 39.651]	[-0.047, 0.048]	[-21283, 19808]	[-0.067, 0.076]	[-0.044 0.054]

Таблица 3: Распределение Коши (4)

Laplace $n = 10$					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.001	-0.001	0.017	0.300	0.231
$D(z)$	0.103	0.078	0.429	0.119	0.084
$E(z) \pm \sqrt{D(z)}$	[-0.318, 0.322]	[-0.281, 0.277]	[-0.637, 0.673]	[-0.045, 0.646]	[-0.059 0.520]
Laplace $n = 100$					
$E(z)$	0.003	0.000	0.027	0.016	0.021
$D(z)$	0.010	0.006	0.404	0.010	0.006
$E(z) \pm \sqrt{D(z)}$	[-0.095, 0.102]	[-0.076, 0.078]	[-0.608, 0.663]	[-0.083, 0.115]	[-0.057 0.100]
Laplace $n = 1000$					
$E(z)$	0.0003	-0.0008	-0.0084	0.0025	0.0018
$D(z)$	0.00105	0.00052	0.38054	0.00104	0.00062
$E(z) \pm \sqrt{D(z)}$	[-0.03219, 0.03280]	[-0.02377, 0.02214]	[-0.62537, 0.60839]	[-0.02976, 0.03473]	[-0.02302 0.02661]

Таблица 4: Распределение Лапласа (5)

Poisson $n = 10$					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	9.942	9.789	10.214	10.926	10.715
$D(z)$	1.008	1.387	1.971	1.407	1.209
$E(z) \pm \sqrt{D(z)}$	[8.938, 10.946]	[8.611, 10.967]	[8.810, 11.618]	[9.740, 12.113]	[9.616 11.815]
Poisson $n = 100$					
$E(z)$	10.000	9.841	10.953	9.955	9.939
$D(z)$	0.105	0.219	1.043	0.161	0.132
$E(z) \pm \sqrt{D(z)}$	[9.676, 10.323]	[9.373, 10.31]	[9.931, 11.974]	[9.553, 10.357]	[9.576 10.303]
Poisson $n = 1000$					
$E(z)$	10.001	9.997	11.651	9.997	9.869
$D(z)$	0.009	0.003	0.697	0.002	0.010
$E(z) \pm \sqrt{D(z)}$	[9.904, 10.099]	[9.942, 10.052]	[10.816, 12.486]	[9.952, 10.042]	[9.767 9.971]

Таблица 5: Распределение Пуассона (6)

Uniform $n = 10$					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.0215	0.0346	0.0098	0.3363	0.3429
$D(z)$	0.097	0.221	0.047	0.123	0.148
$E(z) \pm \sqrt{D(z)}$	[-0.289, 0.333]	[-0.435, 0.505]	[-0.208, 0.228]	[-0.014, 0.688]	[-0.042 0.728]
Uniform $n = 100$					
$E(z)$	0.008	0.012	0.002	0.023	0.045
$D(z)$	0.009	0.027	0.001	0.014	0.018
$E(z) \pm \sqrt{D(z)}$	[-0.089, 0.105]	[-0.151, 0.176]	[-0.021, 0.026]	[-0.097, 0.143]	[-0.091 0.181]
Uniform $n = 1000$					
$E(z)$	-0.0013	-0.0023	-0.0001	0.0002	0.0016
$D(z)$	0.0009	0.0028	0.0000	0.0014	0.0018
$E(z) \pm \sqrt{D(z)}$	[-0.0319, 0.0292]	[-0.0556, 0.0511]	[-0.0025, 0.0024]	[-0.0373, 0.0377]	[-0.0414 0.0446]

Таблица 6: Нормальное распределение (7)

4.3 Боксплот Тьюки

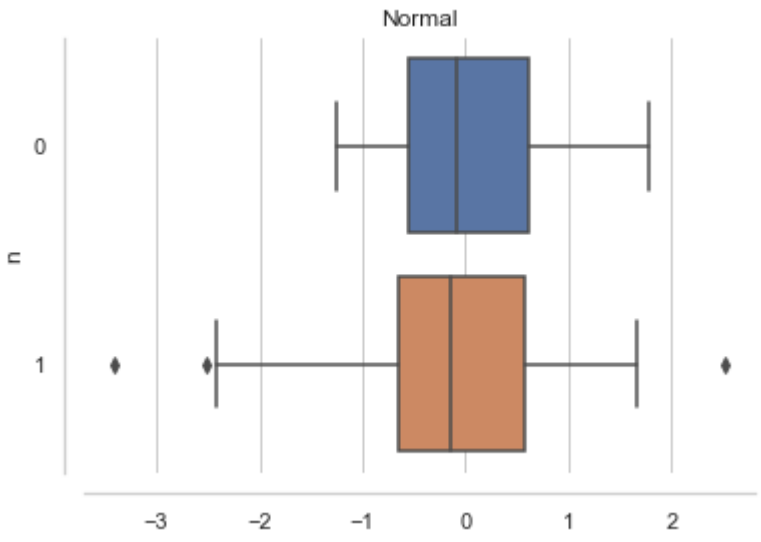

Рис. 6: Нормальное распределение

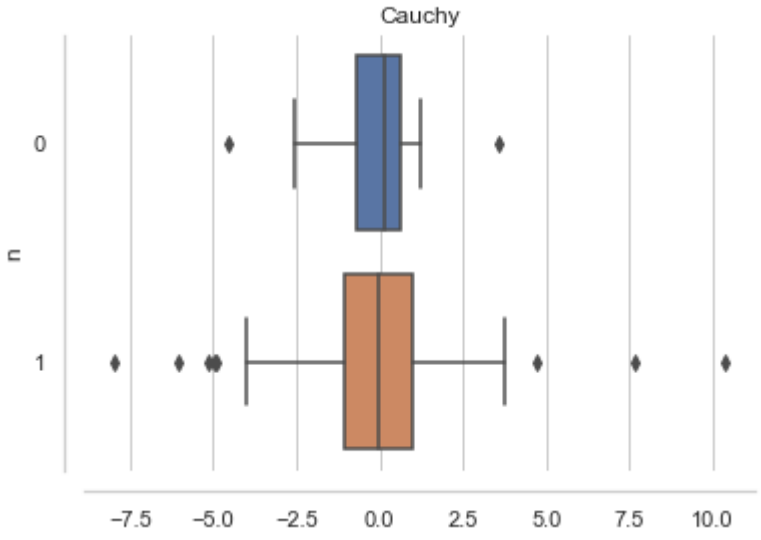

Рис. 7: Распределение Коши

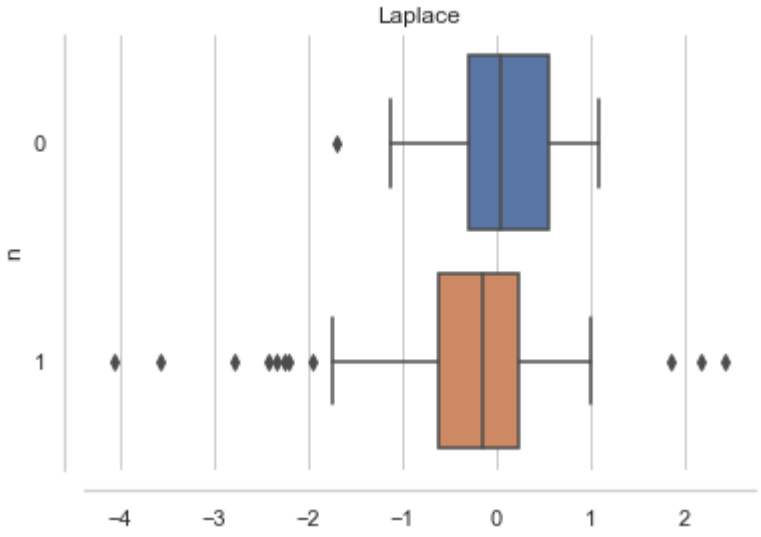

Рис. 8: Распределение Лапласа

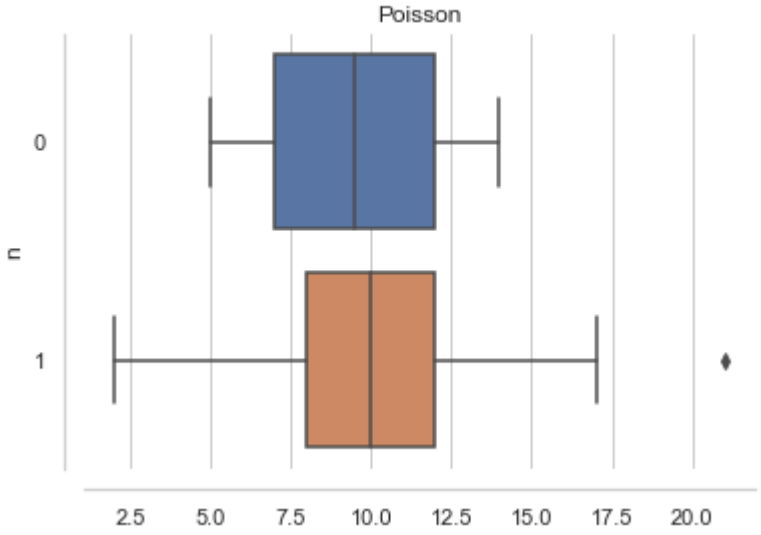

Рис. 9: Распределение Пуассона

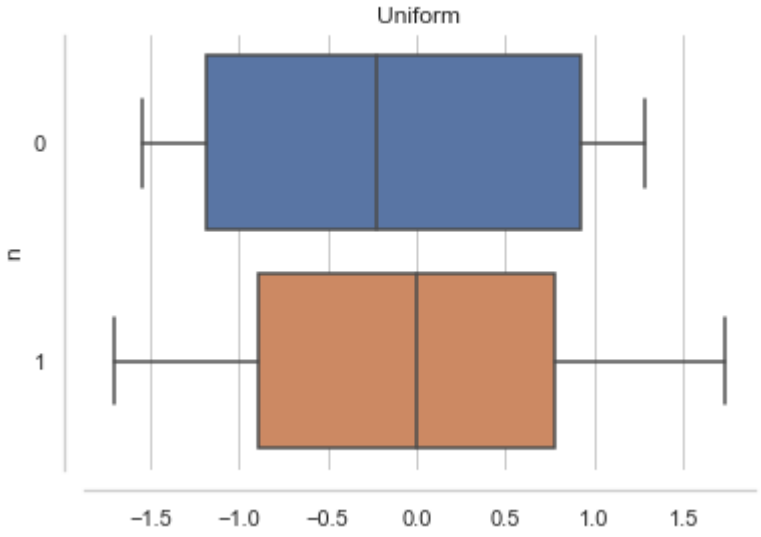

Рис. 10: Равномерное распределение

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	P_B^T
Normal n=20	0.023	0.007
Normal n=100	0.014	0.007
Cauchy n=20	0.152	0.156
Cauchy n=100	0.185	0.156
Laplace n=20	0.080	0.063
Laplace n=100	0.073	0.063
Poisson n=20	0.022	0.008
Poisson n=100	0.015	0.008
Uniform n=20	0.003	0
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 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 8: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

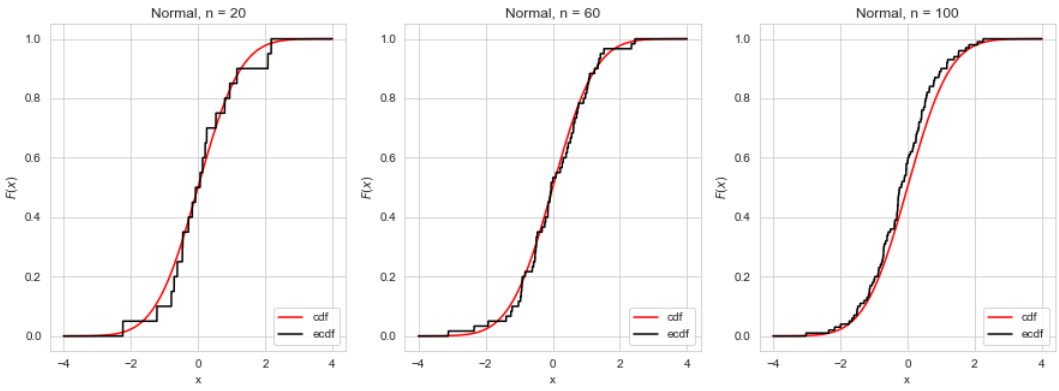

Рис. 11: Нормальное распределение (эмпирич. функция)

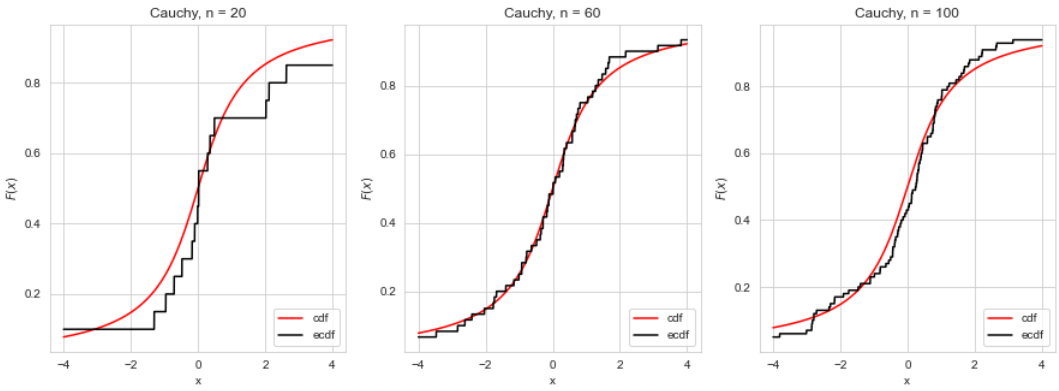

Рис. 12: Распределение Коши (эмпирич. функция)

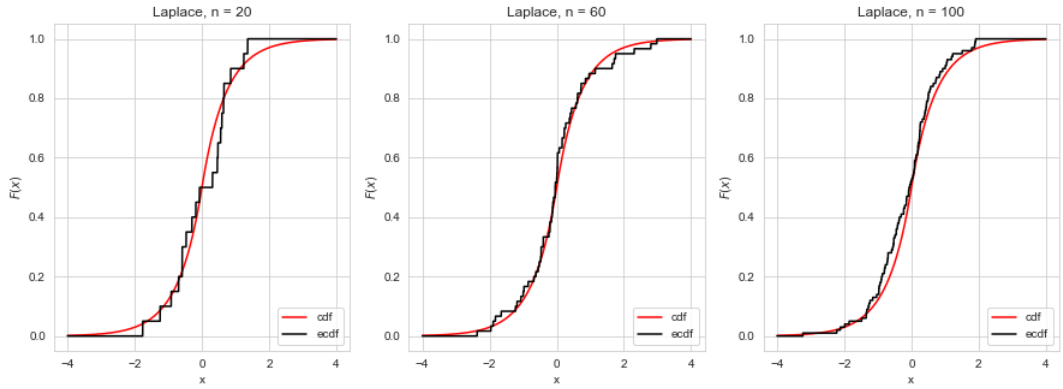

Рис. 13: Распределение Лапласа (эмпирич. функция)

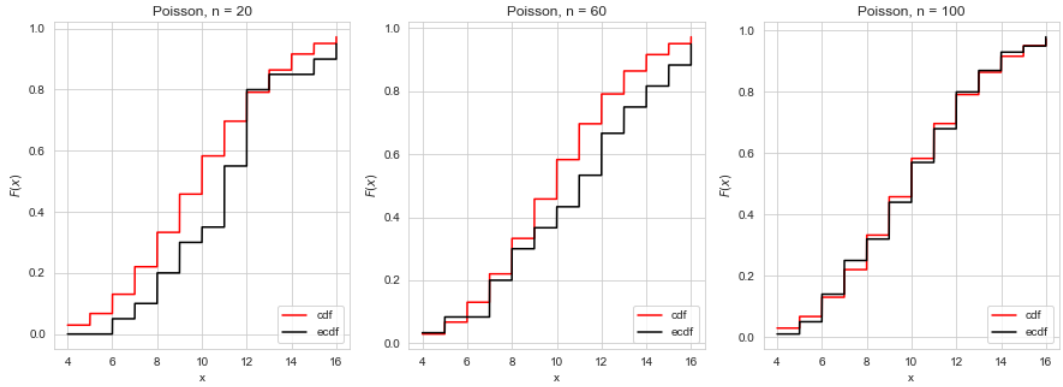

Рис. 14: Распределение Пуассона (эмпирич. функция)

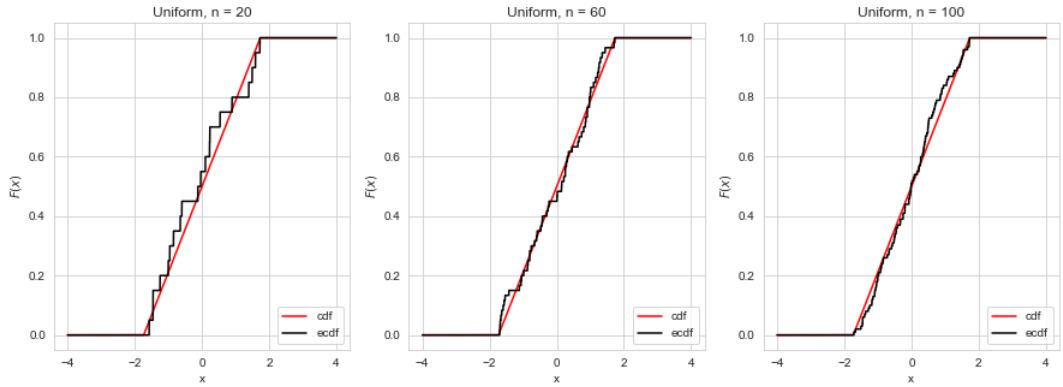

Рис. 15: Равномерное распределение (эмпирич. функция)

4.7 Ядерные оценки плотности распределения

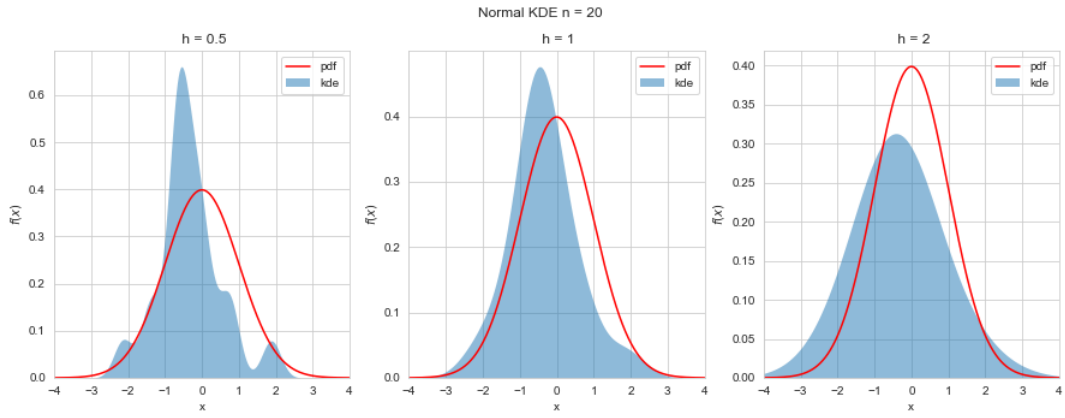

Рис. 16: нормальное распределение (ядерная оценка $n = 20$)

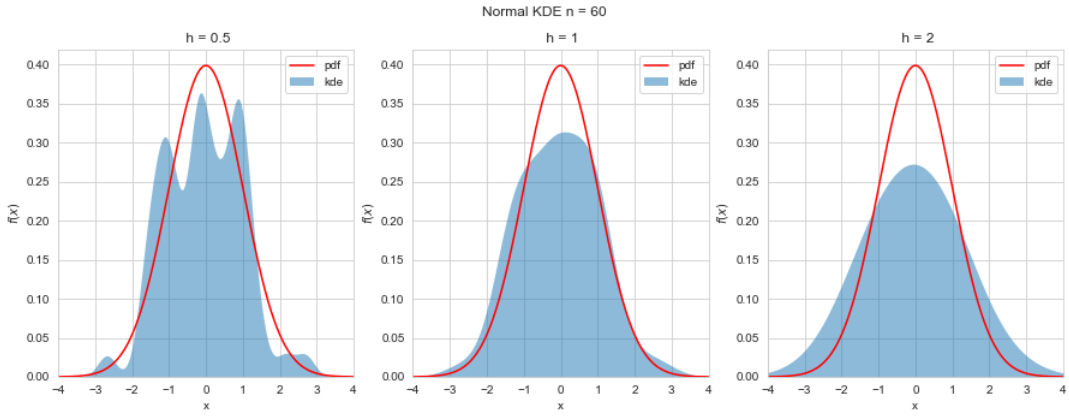

Рис. 17: нормальное распределение (ядерная оценка $n = 60$)

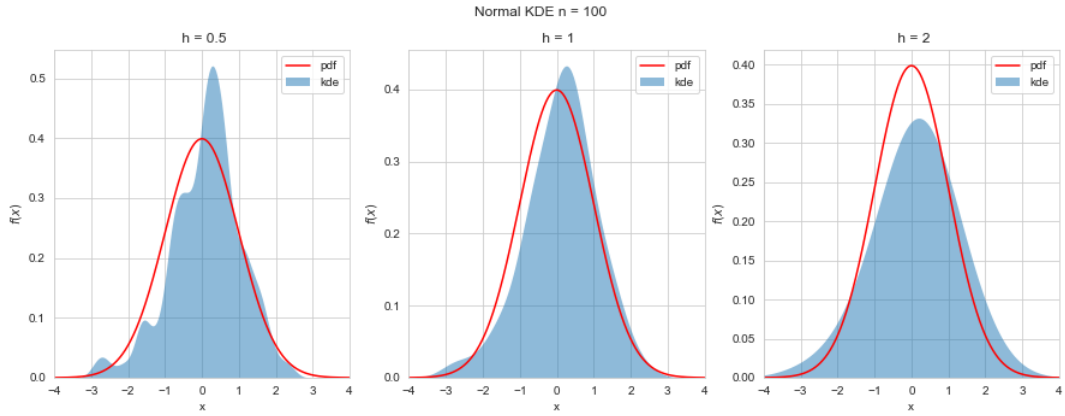

Рис. 18: нормальное распределение (ядерная оценка $n = 100$)

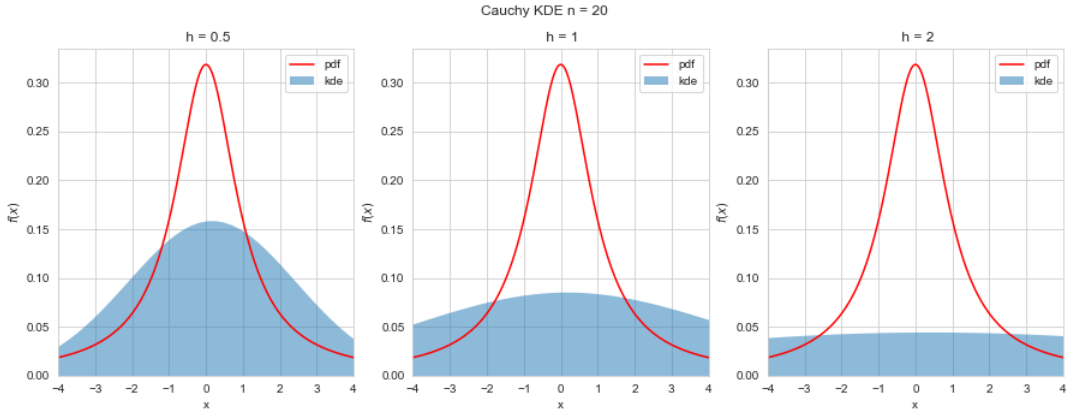

Рис. 19: Распределение Коши (ядерная оценка $n = 20$)

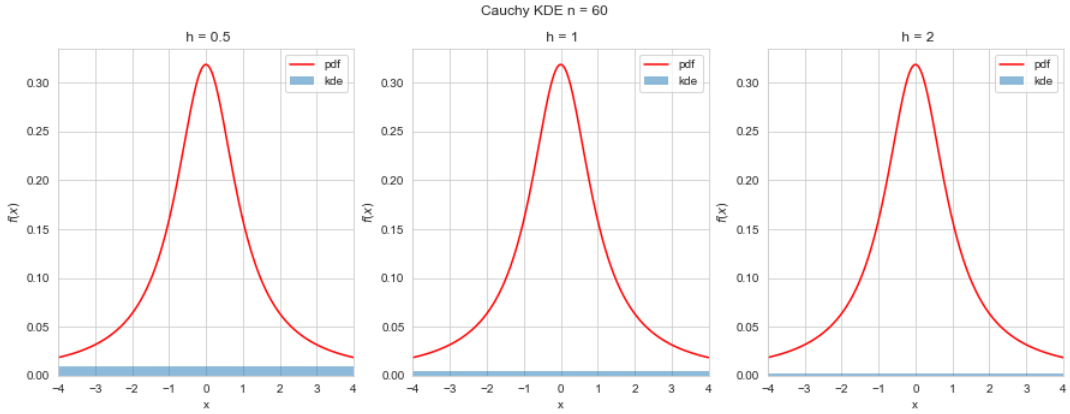

Рис. 20: Распределение Коши (ядерная оценка $n = 60$)

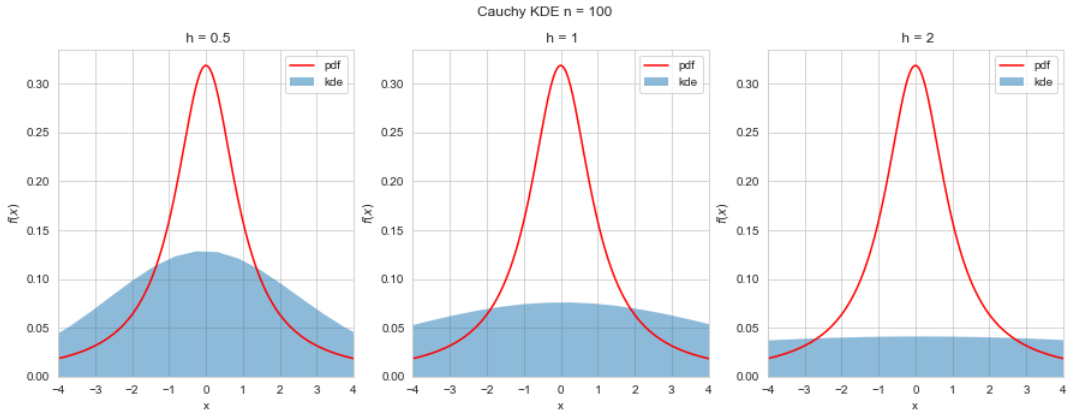

Рис. 21: Распределение Коши (ядерная оценка $n = 100$)

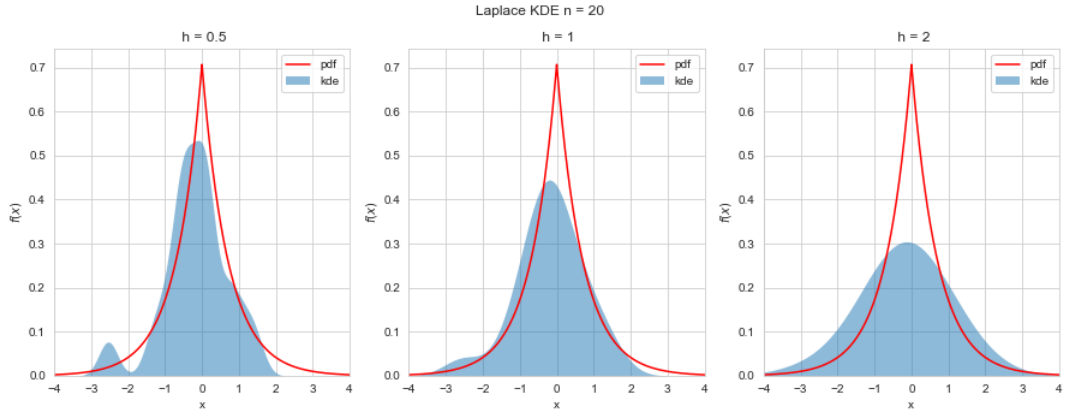

Рис. 22: Распределение Лапласа (ядерная оценка $n = 20$)

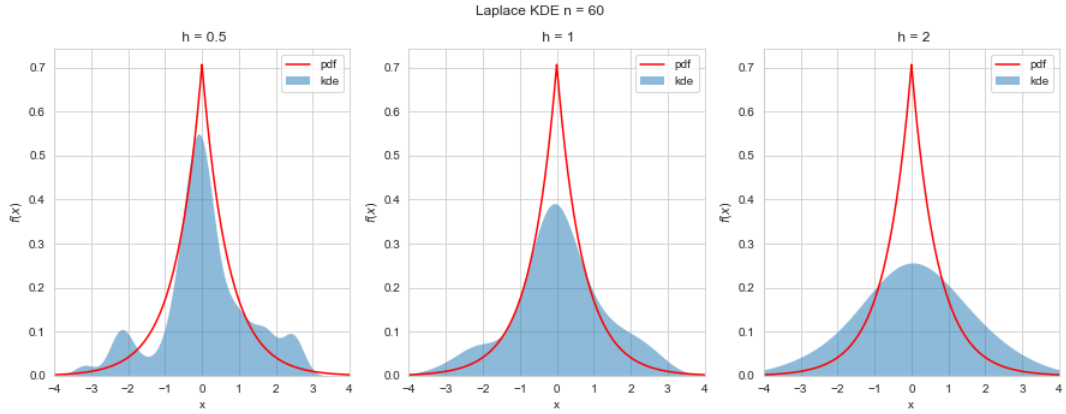

Рис. 23: Распределение Лапласа (ядерная оценка $n = 60$)

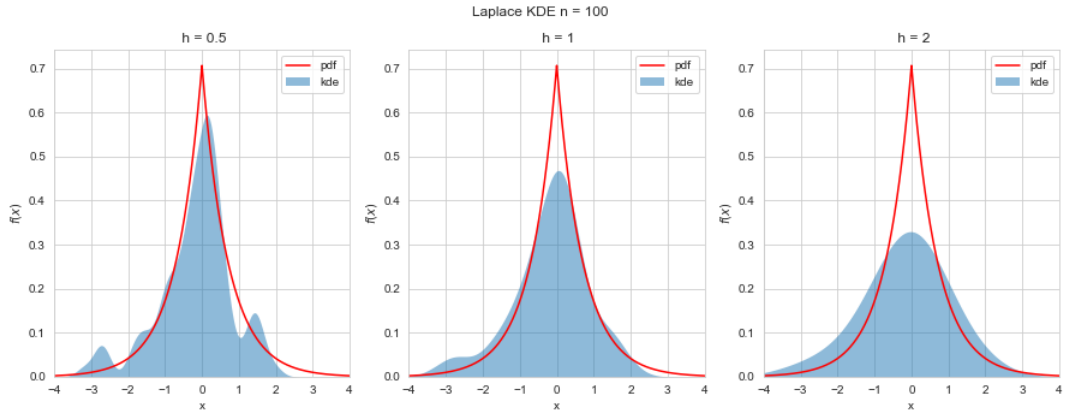

Рис. 24: Распределение Лапласа (ядерная оценка $n = 100$)

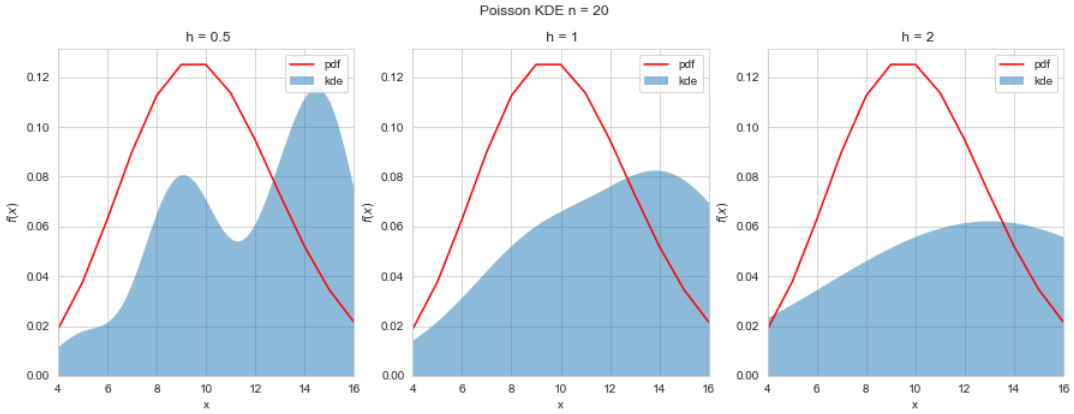

Рис. 25: Распределение Пуассона (ядерная оценка $n = 20$)

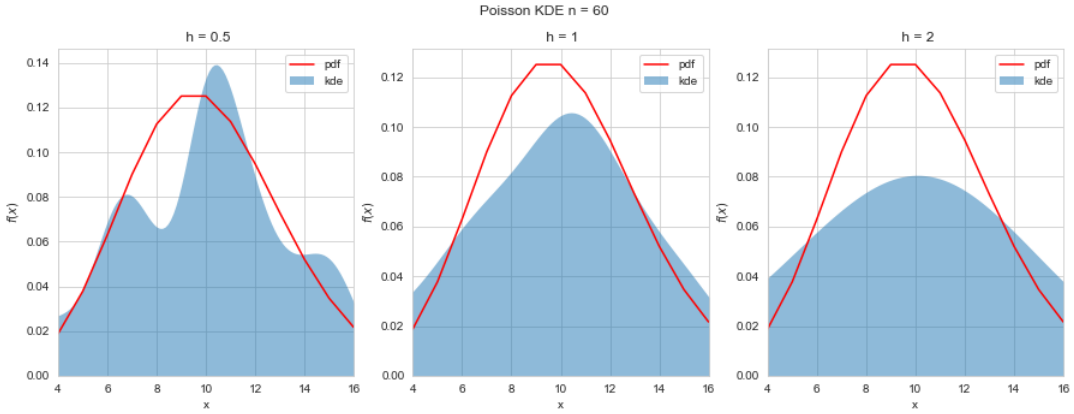

Рис. 26: Распределение Пуассона (ядерная оценка $n = 60$)

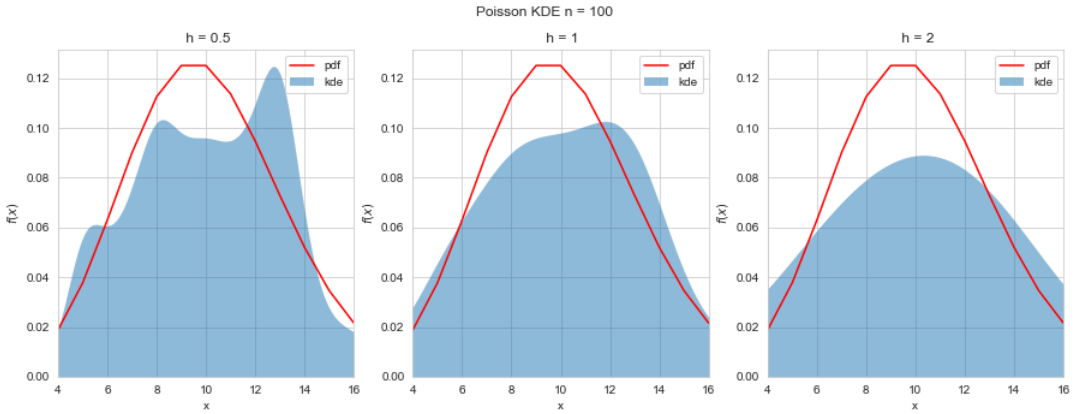

Рис. 27: Распределение Пуассона (ядерная оценка $n = 100$)

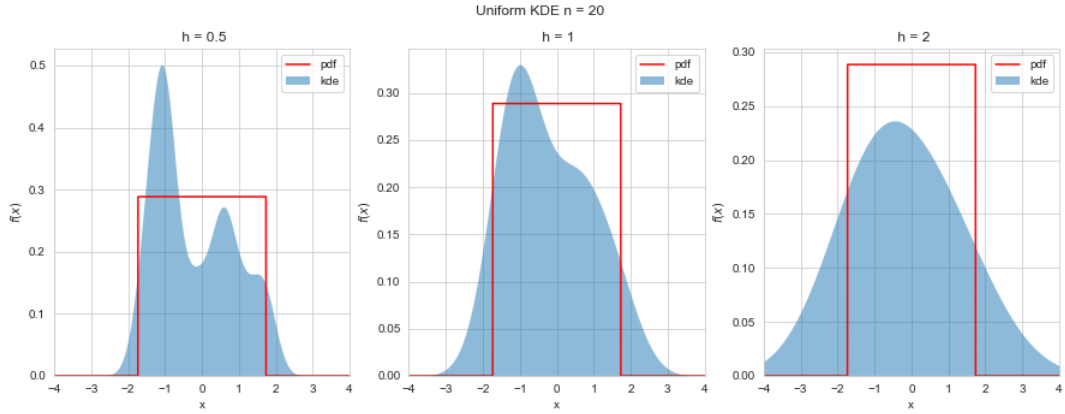

Рис. 28: Равномерное распределение (ядерная оценка $n = 20$)

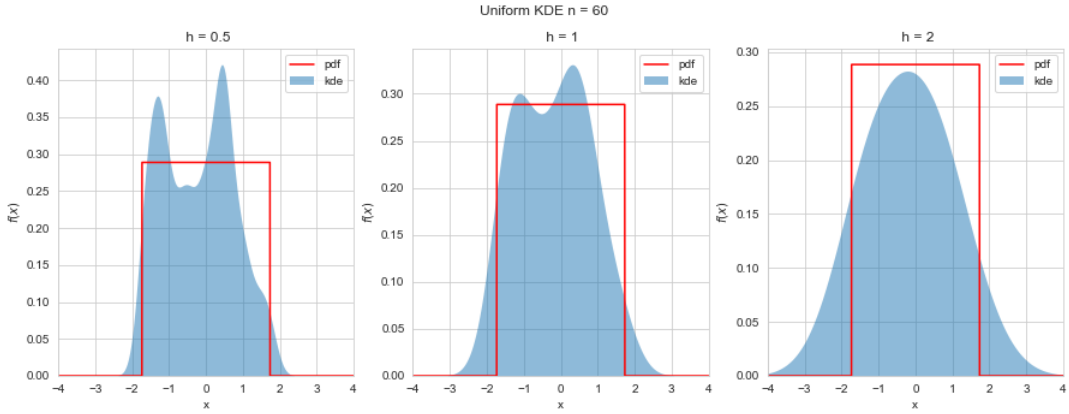

Рис. 29: Равномерное распределение (ядерная оценка $n = 60$)

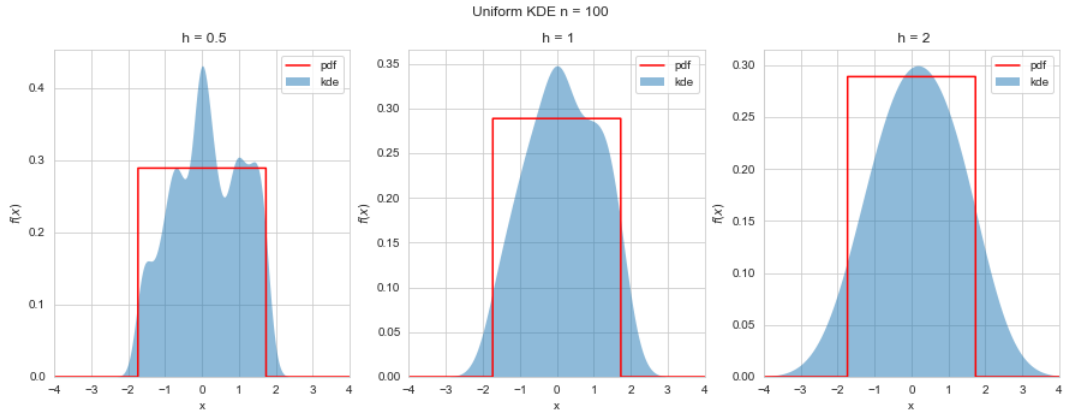

Рис. 30: Равномерное распределение (ядерная оценка $n = 100$)

5 Обсуждение

5.1 Гистограмма и график плотности распределения

По результатам проделанной работы 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.

Визуально очень трудно отличить гистограммы друг от друга, тем более при маленьких выборках. При выборке из 10 элементов вид гистограммы сильно отличается от плотности распределения. Чем больше выборка, тем точнее становится гистограмма. На выборке из 1000 элементов можем отличить и распознать с большей вероятностью равномерное распределение (все прямоугольники примерно на одном уровне), а также распределение Пуассона (оно визуально шире чем распределение Лапласа и нормальное). Однако, отличить между собой распределение Лапласа и нормальное тяжело.

Также можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Из полученных графиков можно увидеть, что только при распределении Пуассона на выборке из 1000 элементов, максимум графика плотности вероятности совпал с максимумом гистограммы. 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 (в нашем случае для 100 элементов), тем ближе доля выбросов будет к теоретической оценке. Снова доля выбросов для распределения Коши значительно выше, чем для остальных распределений. Для нормального, Лапласа и Пуассона погрешность при большой выборке составила не более 2 процентов. При увеличении выборки равномерное распределение показывает стремительный рост к теоретической оценке - выбросы практически не наблюдаются.

Боксплот Тьюки в удобной форме показывает многие важные характеристики выборки, такие как медиана, первый и третий квартили и другие. Исходя из которых можно делать выводы касательно природы входных данных.

5.4 Эмпирическая функция и ядерные оценки плотности распределения

Можем наблюдать на иллюстрациях с эмпирическими функциями, 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 и пуассоновского распределений оптимальным значением параметра является $h = 2h_n$, для распределений Лапласа - $h = h_n/2$, 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 $\hat{h_n}$, 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.

6 Приложение

Код программы GitHub URL:

https://github.com/Krotikov/matStat/tree/master/code