Санкт-Петербургский политехнический университет имени Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

Математическая статистика
Отчёт по лабораторным работам №1-4

Выполнил:

Студент: Габдушев Рушан
Группа: 5030102/90201

Принял:

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

2022 г.

Содержание

Список иллюстраций

Список таблиц

1. Постановка задачи

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

Выполнить следующие задачи:

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: $\overline{x}, medx, z_R, z_Q, z_{tr}$. 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.
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 на отрезке [-4; 4] для непрерывных распределений и на отрезке [6; 14] для распределения Пуассона.
5. Сгенерировать выборки размером 20, 60 и 100 элементов.
   Построить на них 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 распределений:

- Нормальное распределение

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

- Распределение Коши

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

- 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

$$U(x,-\sqrt{3},\sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}}, & x \in [-\sqrt{3},\sqrt{3}] \\ 0, & x \notin [-\sqrt{3},\sqrt{3}] \end{cases}. \tag{7}$$

2.2. Гистограммы

2.2.1. Определение и описание

Гистограмма - функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него. Используются гистограммы для визуализации данных на начальном этапе статистической обработки. Построение гистограмм используется для получения эмпирической оценки плотности распределения случайной величины.

2.2.2. Построение гистограммы

Гистограммы строятся следующим образом: все множество значений, которые могут принимать элементы выборки, разбивается на несколько интервалов. Чаще всего, эти интервалы делают одинакового размера, но это не обязательно (в данной лабораторной работе интервалы будут одинакового размера). Если интервалы одинакового размера, то высота каждого прямоугольника гистограммы будет прямо пропорционален числу элементов выборки, попавших в этот интервал. Если же интервалы разного размера, то высота прямоугольников выбирается так, чтоб их площадь была пропорциональна числу элементов выборки, попавших в этот интервал.

2.3. Вариационный ряд

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

2.4. Выборочные числовые характеристики

2.4.1. Характеристики положения

- Выборочное среднее

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

- 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

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

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

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. Построение

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

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

где X_1 — нижняя граница уса, X_2— верхняя граница уса, Q_1— первый квартиль, Q_2 — третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков [3].

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:

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

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

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

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:

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

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

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

$F^*(x)$- 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 оценки в виде суммы с числом слагаемых, равным объёму выборки:

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

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

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

Такие оценки называются непрерывными ядерными.

Гауссово (нормальное) 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. Реализация

Данная лабораторная работа была выполнена с использованием языка программирования Python 3.9 в среде разработки PyCharm 2021.3.2 с использованием следующих библиотек:

- scipy версии 1.8.0
- numpy версии 1.22.3
- matplotlib версии 3.5.1
- seaborn версии 0.11.2

4. Результаты

4.1. Гистограммы

- Нормальное распределение

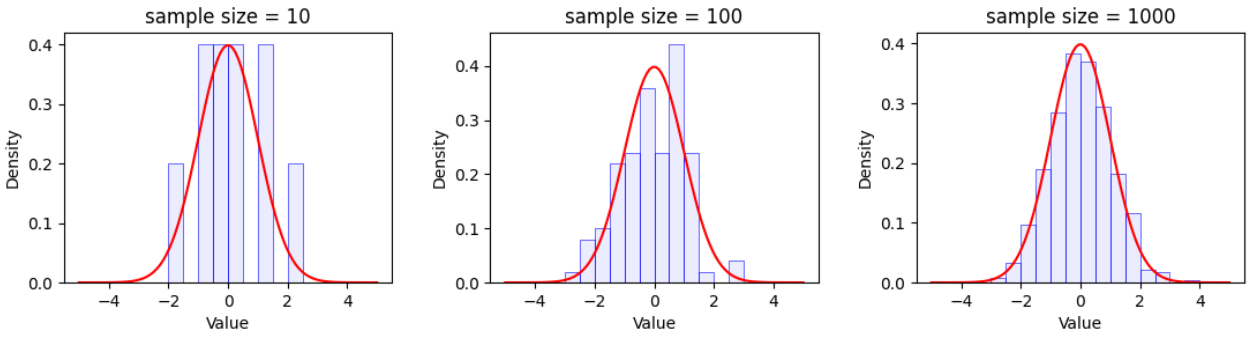

Рис. 1. Гистограмма и плотность вероятности для нормального распределения [N = 10, 100, 1000]

- Распределение Коши

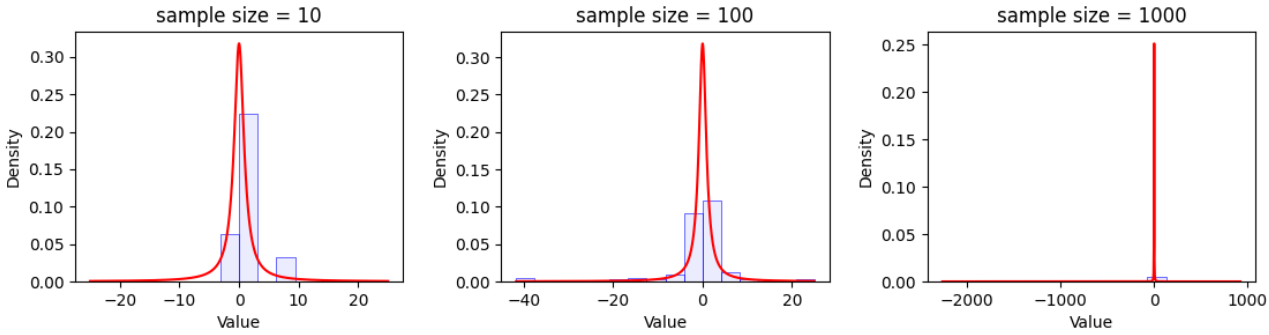

Рис. 2. Гистограмма и плотность вероятности для распределения Коши [N = 10, 100, 1000]

- Распределение Лапласа

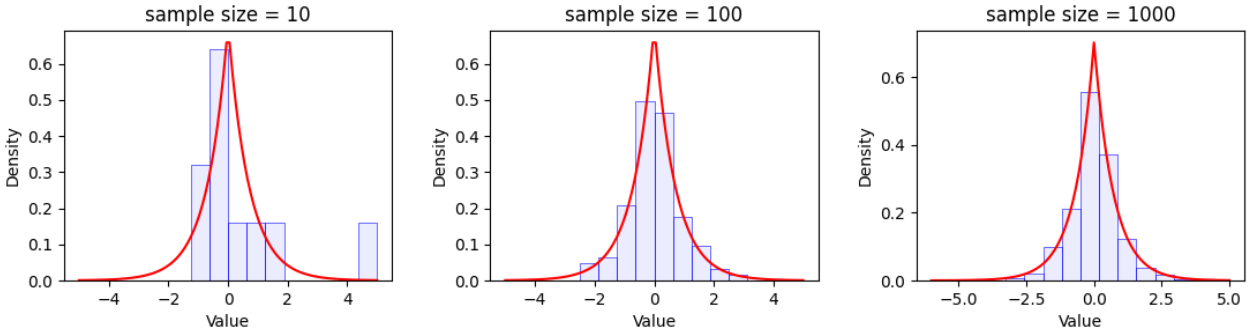

Рис. 3. Гистограмма и плотность вероятности для распределения Лапласа [N = 10, 100, 1000]

- Распределение Пуассона

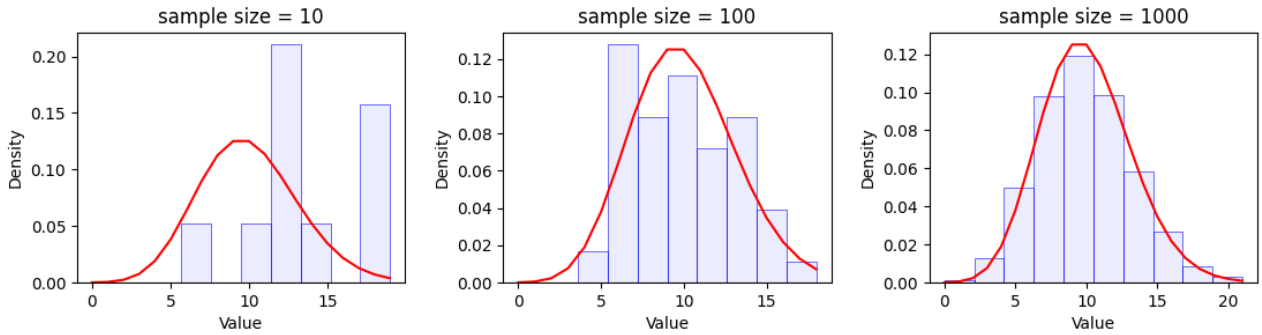

Рис. 4. Гистограмма и плотность вероятности для распределения Пуассона [N = 10, 100, 1000]

- Равномерное распределение

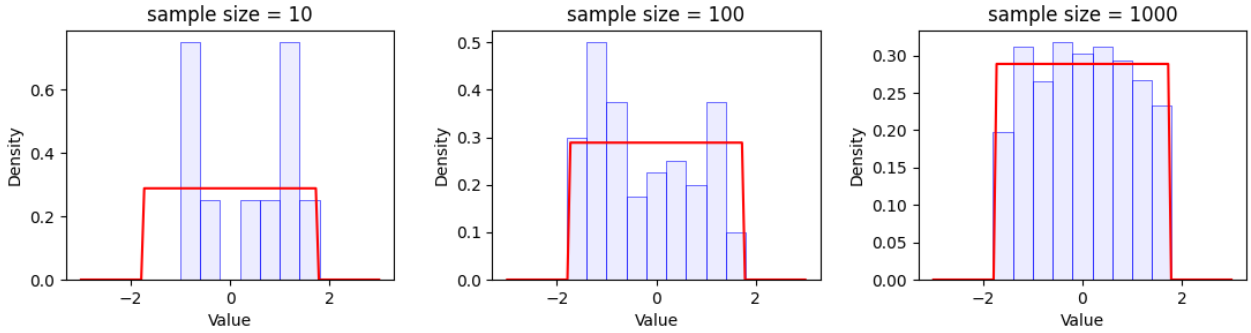

Рис. 5. Гистограмма и плотность вероятности для равномерного распределения [N = 10, 100, 1000]

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					
$size = 10$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.011120	-0.014057	-0.010067	-0.009290	-0.011004
$D(z)$	0.103748	0.140171	0.176979	0.125273	0.119265
$E(z) \pm \sqrt{D(z)}$	[-0.333219; 0.310979]	[-0.388451; 0.360337]	[-0.430756; 0.410622]	[-0.363229; 0.344649]	[-0.356352; 0.334344]
$\widehat{E}(z)$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$
$size = 100$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.000371	0.002745	0.003805	0.014744	0.000684
$D(z)$	0.010318	0.015320	0.091150	0.012351	0.011826
$E(z) \pm \sqrt{D(z)}$	[-0.101949; 0.101207]	[-0.121029; 0.126519]	[-0.298106; 0.305716]	[-0.096391; 0.125879]	[-0.108063; 0.109431]
$\widehat{E}(z)$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$
$size = 1000$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.000319	-0.000242	0.003349	0.002351	0.000426
$D(z)$	0.001040	0.001520	0.056056	0.001231	0.001176
$E(z) \pm \sqrt{D(z)}$	[-0.031930; 0.032568]	[-0.039229; 0.038745]	[-0.233412; 0.240110]	[-0.032735; 0.037437]	[-0.033867; 0.034719]
$\widehat{E}(z)$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$

Таблица 2. Нормальное распределение

cauchy					
$size = 10$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.007256	0.027509	-0.233503	0.046041	0.030371
$D(z)$	188.63	0.338166	4572	1.369770	0.528894
$E(z) \pm \sqrt{D(z)}$	[-13.7413; 13.7268]	[-0.554011; 0.609029]	[-67.8511; 67.3841]	[-1.12433; 1.21641]	[-0.696880; 0.757622]
$\widehat{E}(z)$	–	$0.0^{+0.6}_{-0.6}$	–	–	$0.0^{+0.8}_{-0.8}$
$size = 100$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.586391	-0.007441	28.903241	0.034702	-0.002166
$D(z)$	765.999281	0.024822	1884899	0.051798	0.026260
$E(z) \pm \sqrt{D(z)}$	[-27.0903; 28.2631]	[-0.164991; 0.150109]	[-1344; 1401]	[-0.192890; 0.262294]	[-0.164215; 0.159883]
$\widehat{E}(z)$	–	$0.0^{+0.2}_{-0.2}$	–	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.0}_{-0.0}$
$size = 1000$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.229276	0.002852	-87.8948	0.003081	0.001475
$D(z)$	767.698	0.002567	188091672	0.005097	0.002651
$E(z) \pm \sqrt{D(z)}$	[-27.9366; 27.4780]	[-0.047814; 0.053518]	[-13802; 13626]	[-0.068312; 0.074474]	[-0.050013; 0.052963]
$\widehat{E}(z)$	–	$0.0^{+0.1}_{-0.1}$	–	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$

Таблица 3. Распределение Коши

laplace					
$size = 10$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.004731	-0.003830	0.015058	0.003482	-0.000079
$D(z)$	0.094743	0.071067	0.389967	0.097006	0.071709
$E(z) \pm \sqrt{D(z)}$	[-0.303073; 0.312535]	[-0.270414; 0.262754]	[-0.609415; 0.639531]	[-0.307976; 0.314940]	[-0.267864; 0.267706]
$\widehat{E}(z)$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.6}_{-0.6}$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.3}_{-0.3}$
$size = 100$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.009117	-0.008533	0.006783	0.005079	-0.010324
$D(z)$	0.010272	0.005881	0.380249	0.010553	0.006311
$E(z) \pm \sqrt{D(z)}$	[-0.110468; 0.092234]	[-0.085221; 0.068155]	[-0.609860; 0.623426]	[-0.097649; 0.107807]	[-0.089766; 0.069118]
$\widehat{E}(z)$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.6}_{-0.6}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$
$size = 1000$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.000130	0.000873	0.016769	0.001779	0.000541
$D(z)$	0.001020	0.000508	0.424275	0.001019	0.000600
$E(z) \pm \sqrt{D(z)}$	[-0.031807; 0.032067]	[-0.021666; 0.023412]	[-0.634595; 0.668133]	[-0.030143; 0.033701]	[-0.023954; 0.025036]
$\widehat{E}(z)$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.7}_{-0.7}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$

Таблица 4. Распределение Лапласа

poisson					
$size = 10$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	9.986800	9.835500	10.279000	9.917000	9.869833
$D(z)$	0.954046	1.440190	1.843159	1.174111	1.123307
$E(z) \pm \sqrt{D(z)}$	[9.010047; 10.963553]	[8.635421; 11.035579]	[8.921370; 11.636630]	[8.833436; 11.000564]	[8.809971; 10.929695]
$\widehat{E}(z)$	10^{+1}_{-1}	10^{+2}_{-2}	10^{+2}_{-2}	10^{+1}_{-1}	10^{+1}_{-1}
$size = 100$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	9.972320	9.831500	10.912500	9.937500	9.832440
$D(z)$	0.109885	0.212358	0.995094	0.165844	0.129318
$E(z) \pm \sqrt{D(z)}$	[9.640831; 10.303809]	[9.370677; 10.292323]	[9.914956; 11.910044]	[9.530261; 10.344739]	[9.472832; 10.192048]
$\widehat{E}(z)$	10^{+1}_{-1}	10^{+1}_{-1}	11^{+1}_{-1}	10^{+1}_{-1}	10^{+1}_{-1}
$size = 1000$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	9.998205	9.991000	11.696000	9.995500	9.856278
$D(z)$	0.009659	0.008419	0.711084	0.002230	0.010905
$E(z) \pm \sqrt{D(z)}$	[9.899925; 10.096485]	[9.899245; 10.082755]	[10.852742; 12.539258]	[9.948277; 10.042723]	[9.751851; 9.960705]
$\widehat{E}(z)$	10^{+0}_{-0}	10^{+0}_{-0}	12^{+1}_{-1}	10^{+0}_{-0}	10^{+0}_{-0}

Таблица 5. Распределение Пуассона

uniform					
$size = 10$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.004121	0.013246	0.001783	-0.003022	0.007187
$D(z)$	0.106775	0.245253	0.045444	0.146939	0.175620
$E(z) \pm \sqrt{D(z)}$	[-0.322643; 0.330885]	[-0.481984; 0.508476]	[-0.211393; 0.214959]	[-0.386348; 0.380304]	[-0.411883; 0.426257]
$\widehat{E}(z)$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.5}_{-0.5}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$
$size = 100$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	-0.000908	0.003109	0.000780	0.012892	-0.000349
$D(z)$	0.010438	0.031048	0.000562	0.014659	0.020914
$E(z) \pm \sqrt{D(z)}$	[-0.103075; 0.101259]	[-0.173095; 0.179313]	[-0.022927; 0.024487]	[-0.108182; 0.133966]	[-0.144966; 0.144268]
$\widehat{E}(z)$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.2}_{-0.2}$
$size = 1000$	Mean	Median	z_R	z_Q	z_{tr}
$E(z)$	0.000354	-0.000030	0.000088	0.003005	0.000623
$D(z)$	0.000970	0.002928	0.000006	0.001492	0.001953
$E(z) \pm \sqrt{D(z)}$	[-0.030791; 0.031499]	[-0.054141; 0.054081]	[-0.002361; 0.002537]	[-0.035621; 0.041631]	[-0.043570; 0.044816]
$\widehat{E}(z)$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$

Таблица 6. Равномерное распределение

4.3. Боксплот Тьюки

- Нормальное распределение

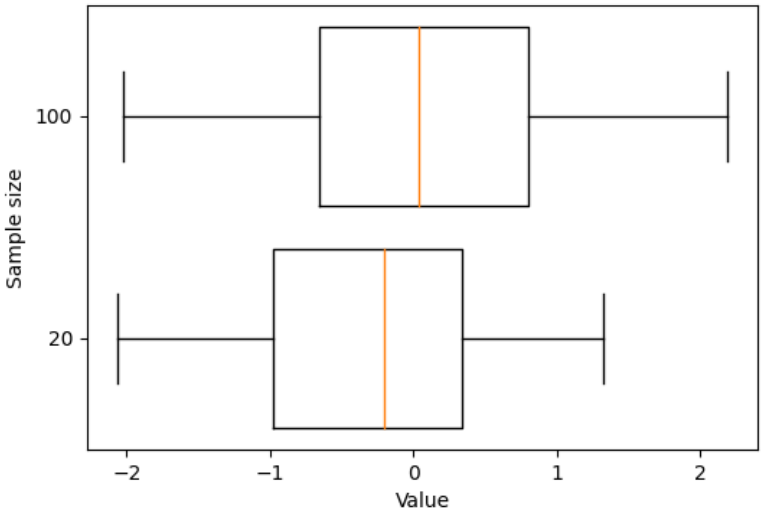

Рис. 6. Боксплот Тьюки Нормальное распределение

- Распределение Коши

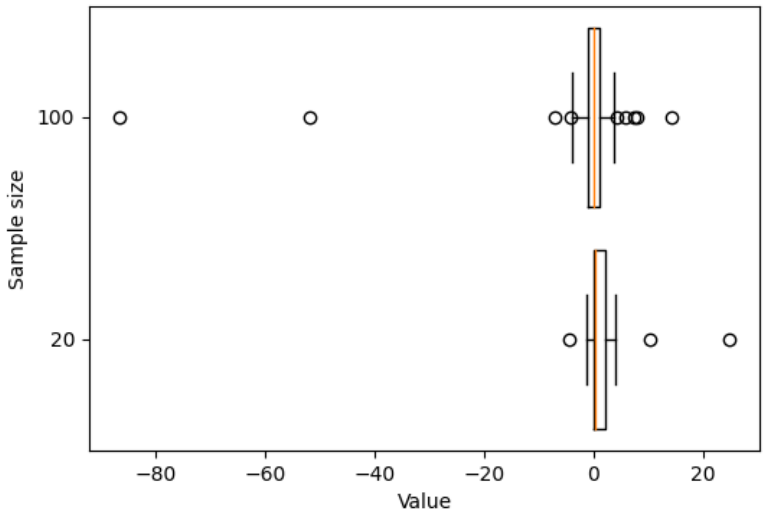

Рис. 7. Боксплот Тьюки Распределение Коши

- Распределение Лапласа

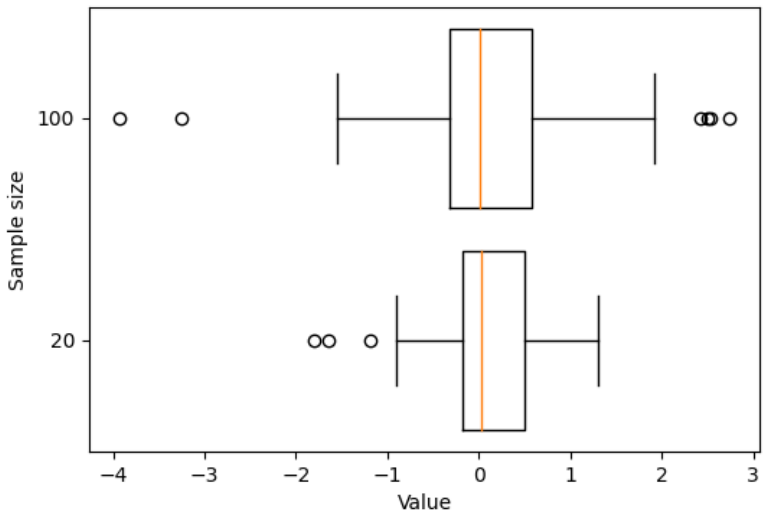

Рис. 8. Боксплот Тьюки Распределение Лапласа

- Распределение Пуассона

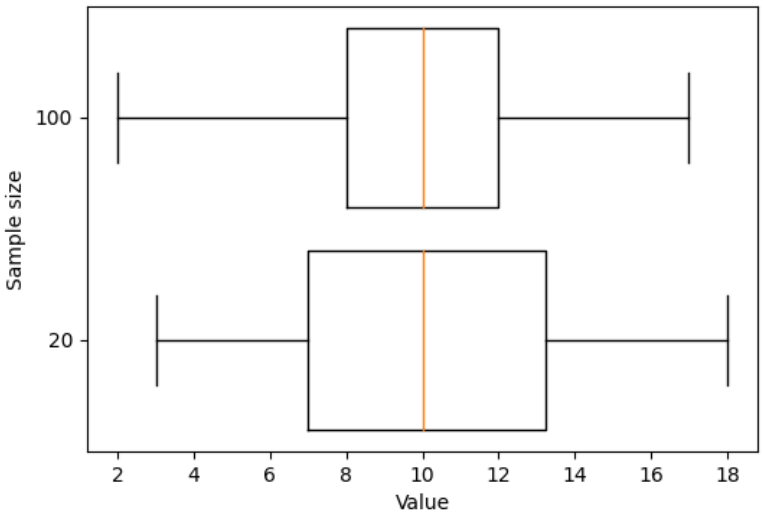

Рис. 9. Боксплот Тьюки Распределение Пуассона

- Равномерное распределение

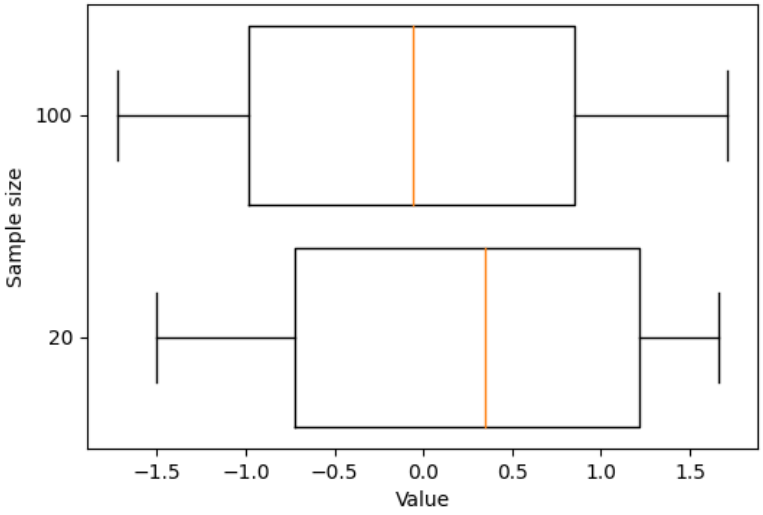

Рис. 10. Боксплот Тьюки Равномерное распределение

4.4. Доля выбросов

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

Выборка	Доля выбросов
Normal n = 20	0.02
Normal n = 100	0.01
Cauchy n = 20	0.15
Cauchy n = 100	0.15
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 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

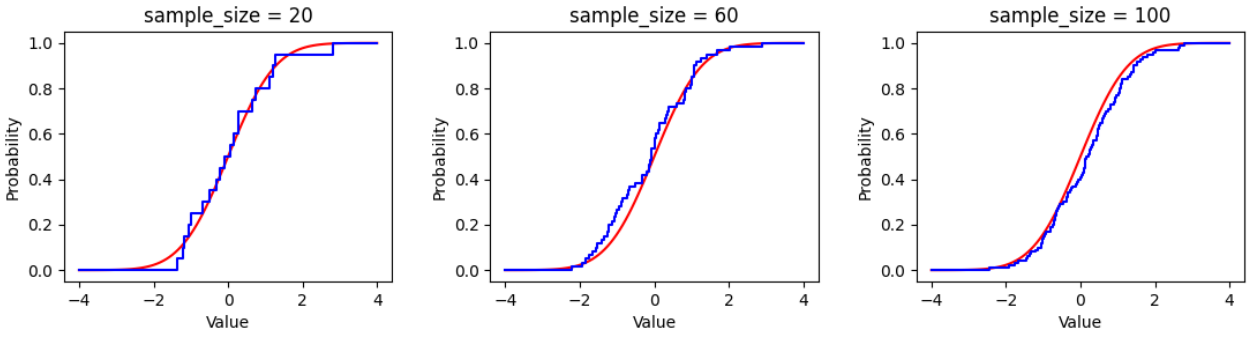

Рис. 11. Нормальное распределение, Эмпирическая функция распределения

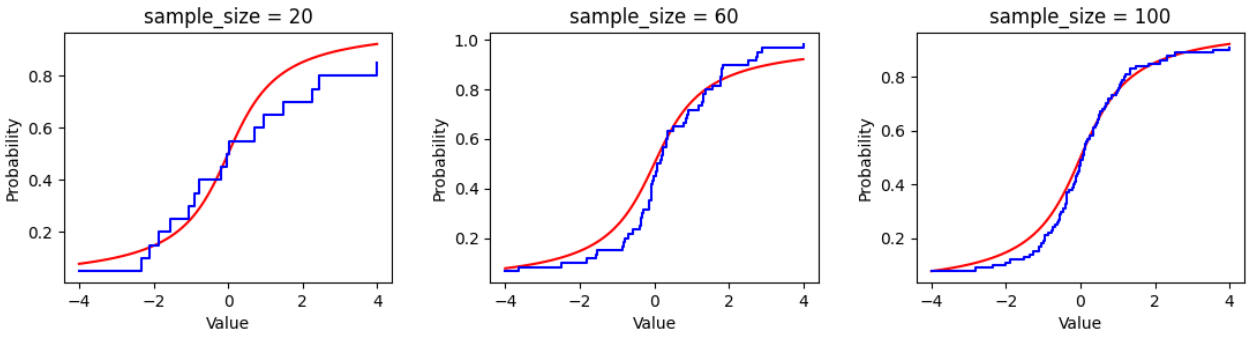

Рис. 12. Распределение Коши, Эмпирическая функция распределения

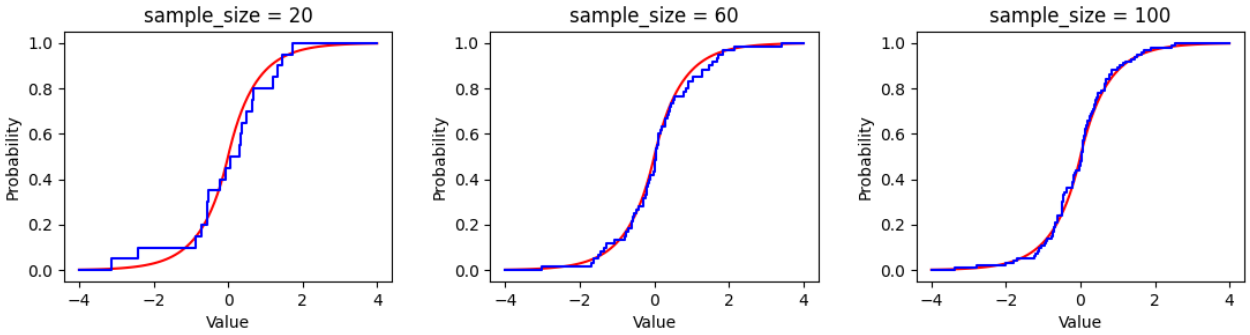

Рис. 13. Распределение Лапласа, Эмпирическая функция распределения

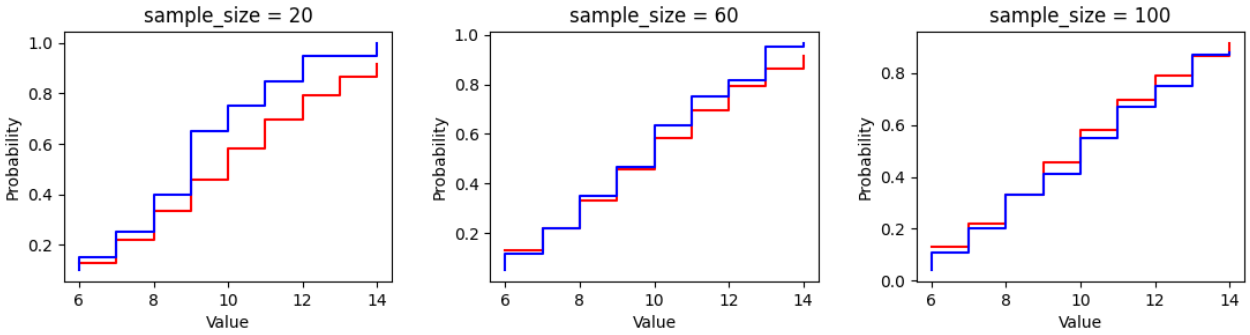

Рис. 14. Распределение Пуассона, Эмпирическая функция распределения

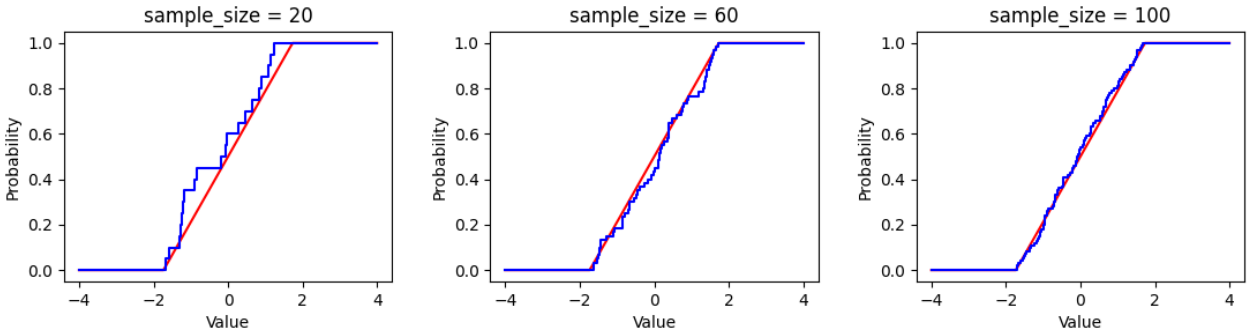

Рис. 15. Равномерное распределение, Эмпирическая функция распределения

4.7. Ядерные оценки плотности распределения

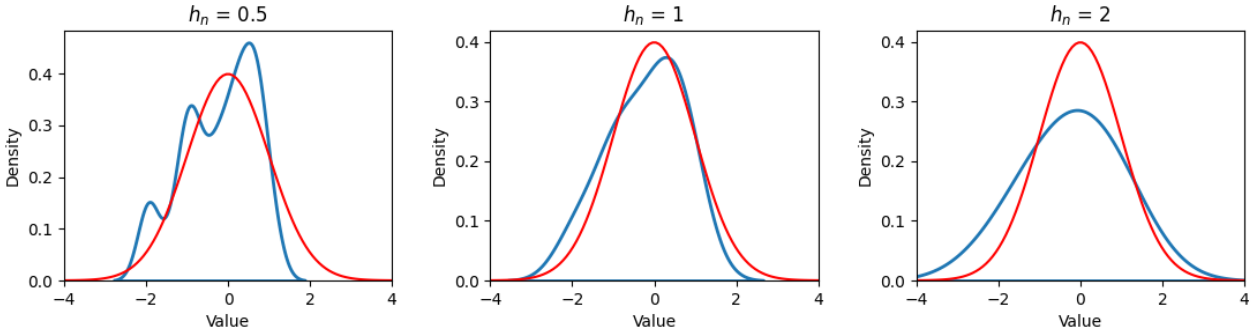

Рис. 16. Нормальное распределение, $n = 20$

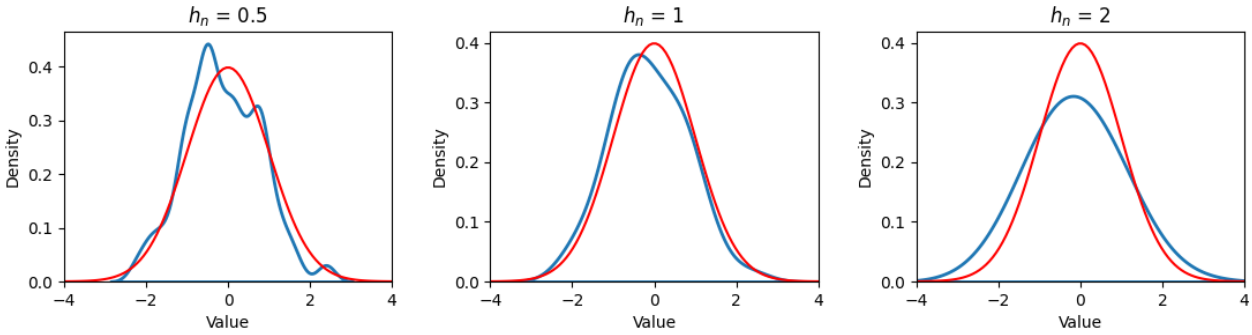

Рис. 17. Нормальное распределение, $n = 60$

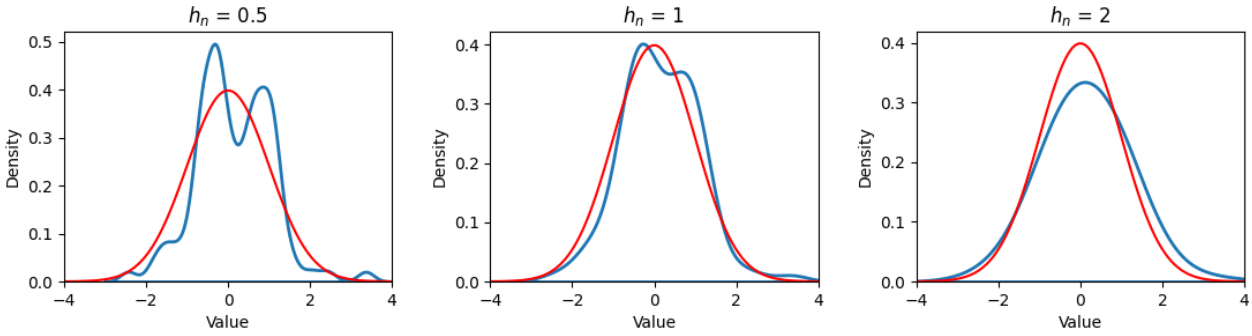

Рис. 18. Нормальное распределение, $n = 100$

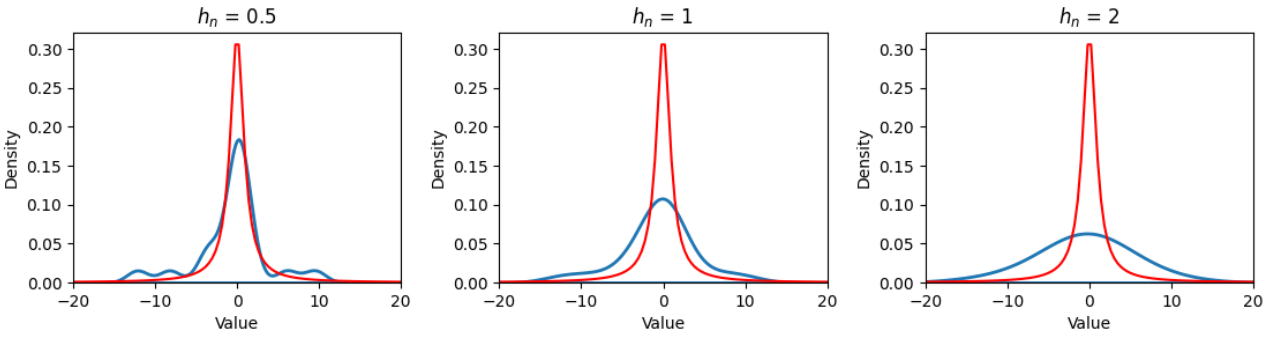

Рис. 19. Распределение Коши, $n = 20$

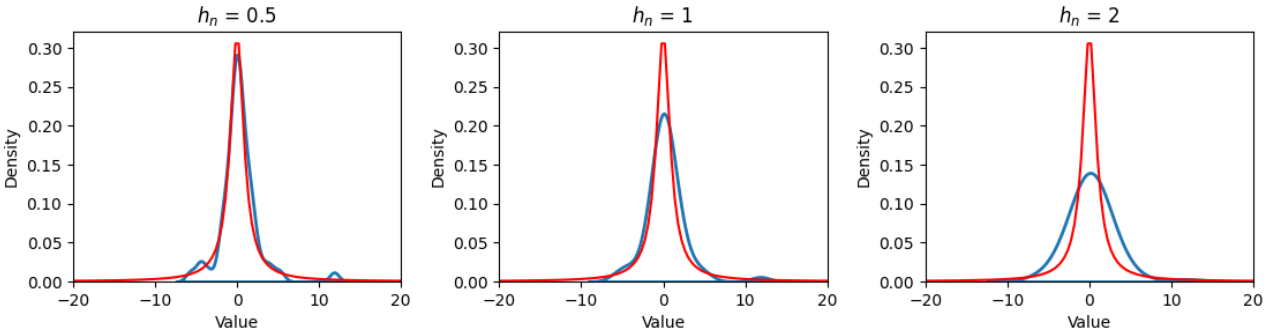

Рис. 20. Распределение Коши, $n = 60$

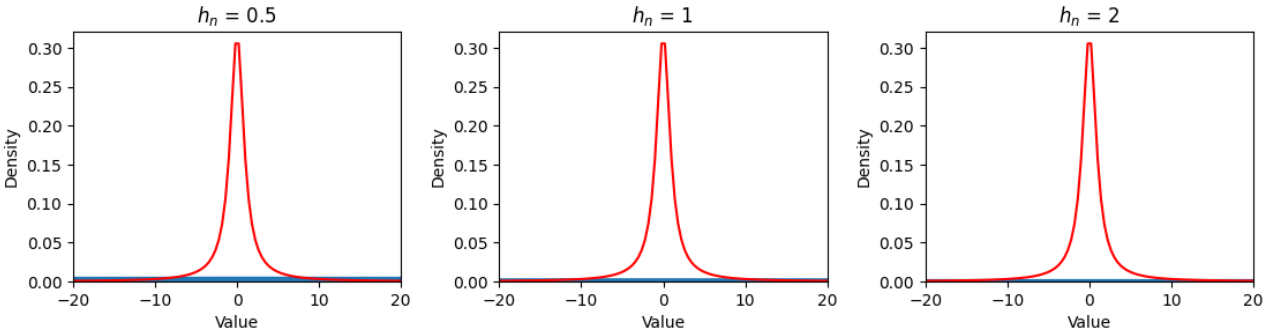

Рис. 21. Распределение Коши, $n = 100$

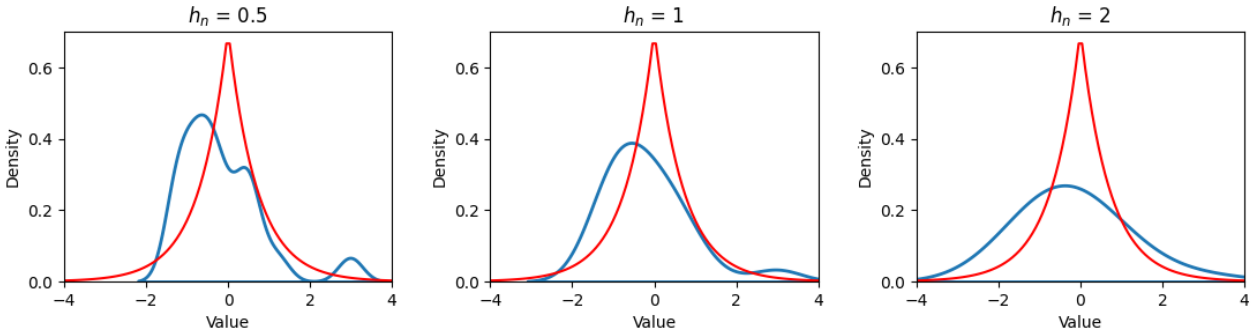

Рис. 22. Распределение Лапласа, $n = 20$

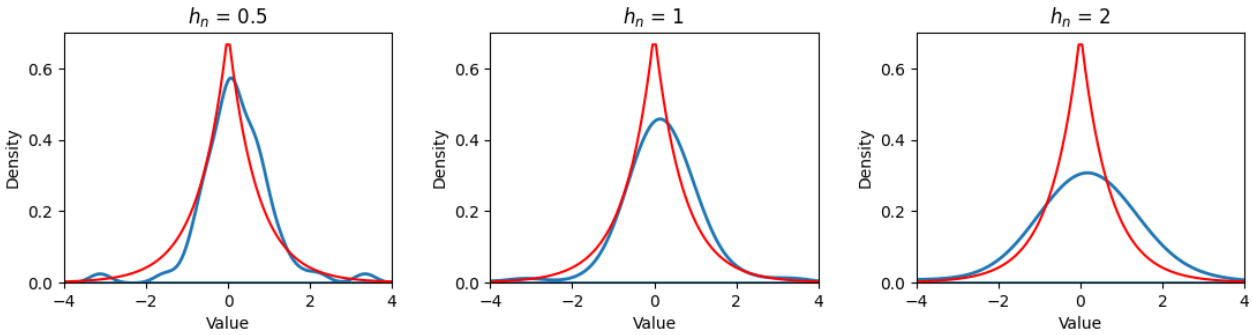

Рис. 23. Распределение Лапласа, $n = 60$

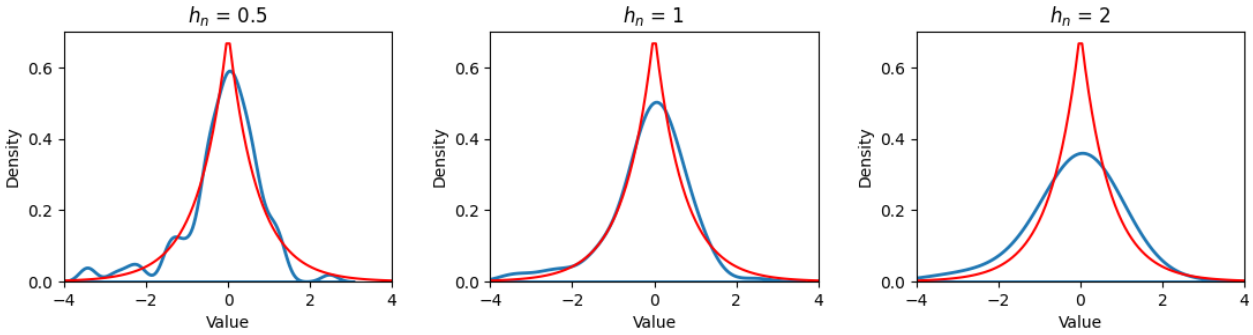

Рис. 24. Распределение Лапласа, $n = 100$

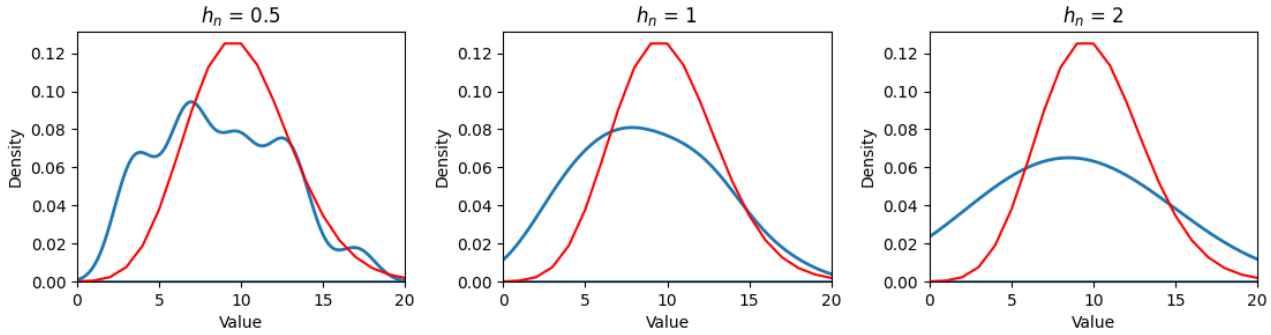

Рис. 25. Распределение Пуассона, $n = 20$

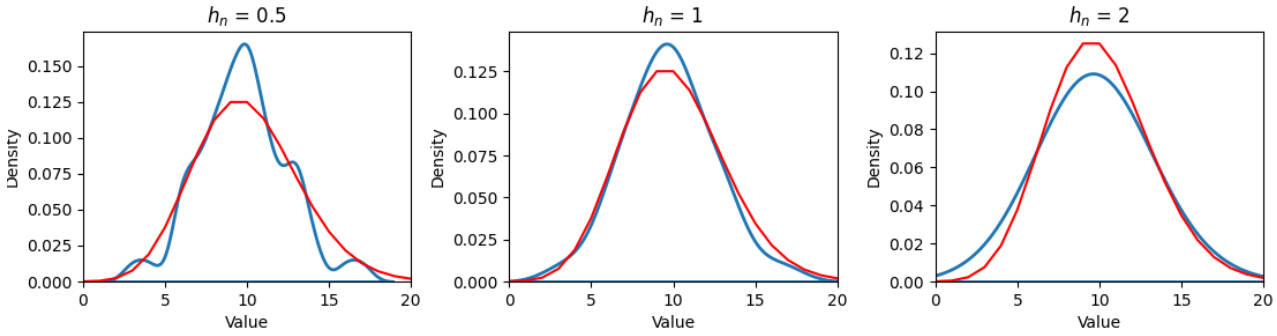

Рис. 26. Распределение Пуассона, $n = 60$

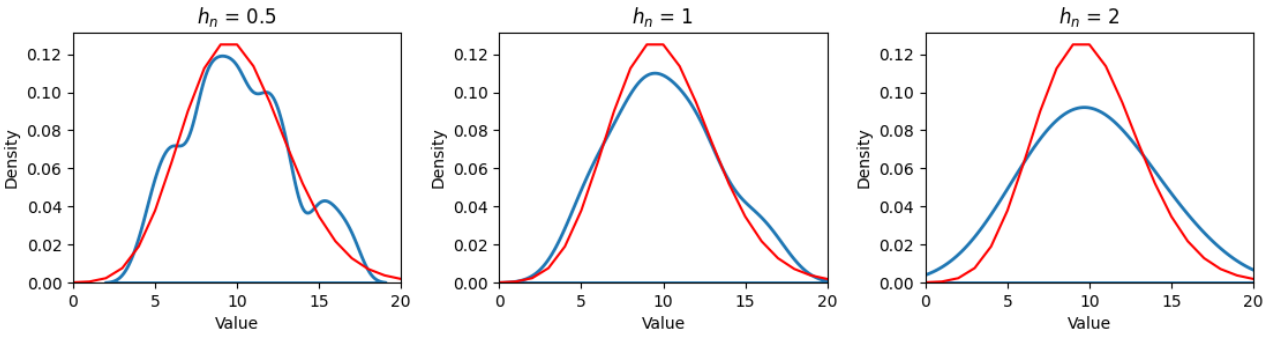

Рис. 27. Распределение Пуассона, $n = 100$

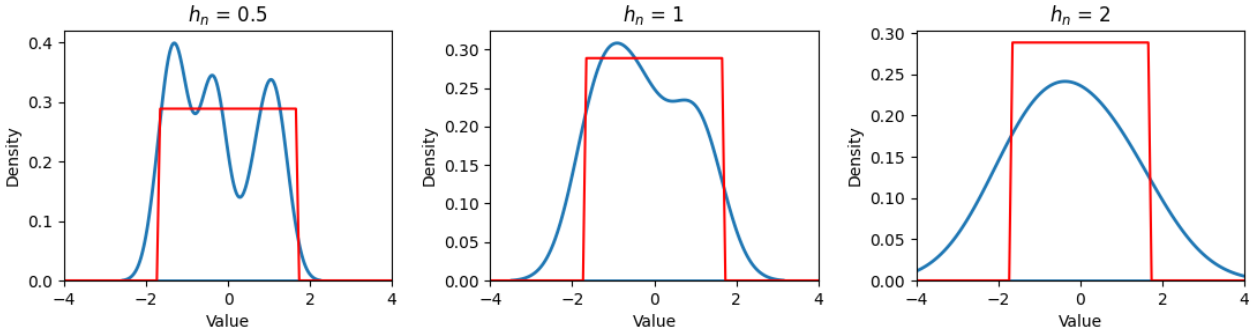

Рис. 28. Равномерное распределение, $n = 20$

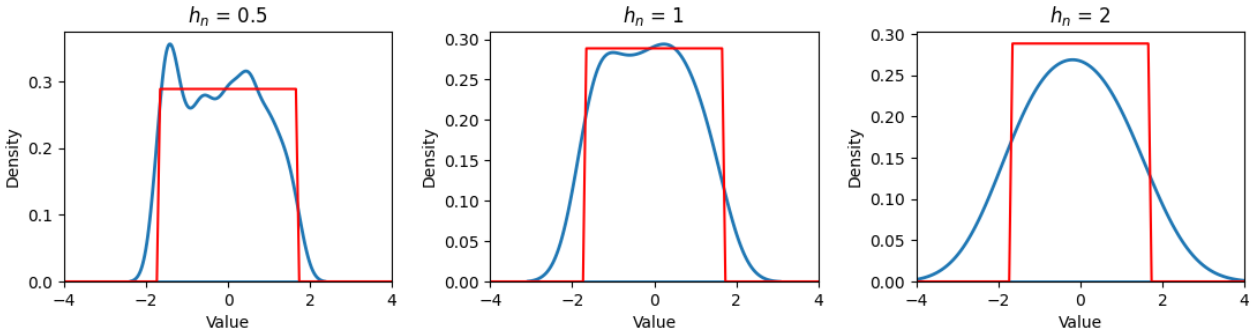

Рис. 29. Равномерное распределение, $n = 60$

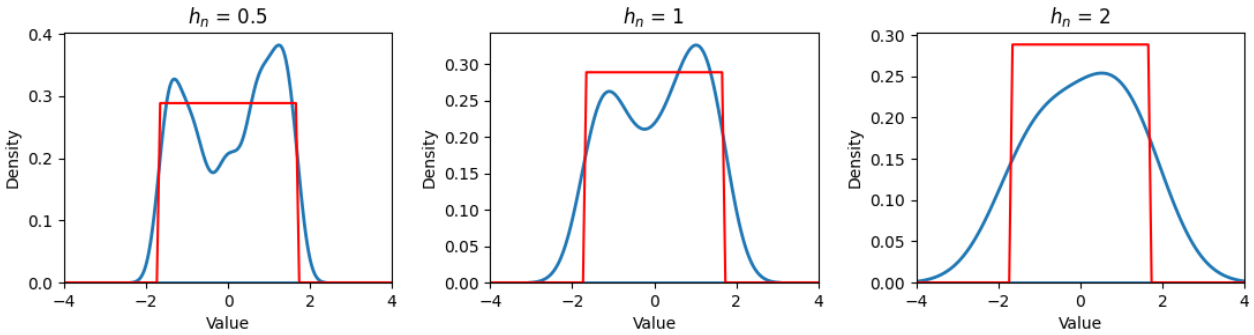

Рис. 30. Равномерное распределение, $n = 100$

5. Обсуждение

5.1. Гистограммы

Полученные результаты работы говорят о том, что при увеличении размеров выборок, гистограммы все ближе к графику плотности вероятности того закона, по которому были сгенерированы элементы выборок. Верно и обратно: чем меньше выборка, тем хуже по ней можно определить закон, по которой эта выборка генерировалась.

Также одним из ключевых выводов является тот факт, что по маленькому размеру выборки (n = 10) очень трудно отличить гистограммы, а, следовательно, и определить закон, по которой генерировалась выборка. Действительно, гистограмма выборки, построенной по распределению Пуассона при n = 10, могла бы с тем же успехом описывать график равномерного распределения (если не учитывать один единственный всплеск гистограммы, который вообще мог остаться незамеченным при более широких интервалах боксов гистограммы).

При выборках n = 1000 видно, что гистограммы уже достаточно неплохо приближаются к графикам плотностей соответствующих законов распределения: в равномерном распределении отклонения гистограммы от графика незначительны, а в нормальном распределении уже наблюдаются «хвосты», которые позволяют отличить треугольное распределение от нормального.

5.2. Характеристики положения и рассеяния

Проанализировав полученные результаты из таблиц, можно сразу же заметить, что дисперсия распределения Коши аномально зашкаливает, даже при увеличении числа элементов выборки. Такая дисперсия является следствием тех выбросов, которые мы наблюдали в предыдущем задании.

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 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 равномерного распределения отклонение функций друг от друга наибольшее.

5.5. Ядерные оценки плотности распределения

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 - $h = h_n/2$, а 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 $h = h_n/2$ по полученным приближениям становится сложно сказать плотность вероятности какого распределения они должны повторять, так как они очень похожи между собой.

6. Ссылки на библиотеки

1. `https://scipy.org/` - SciPy
2. `https://numpy.org/` - NumPy
3. `https://numpy.org/` - Matplotlib
4. `https://seaborn.pydata.org/` - Seaborn

7. Ссылки на репозиторий

`https://github.com/maloxit/matstat` - GitHub репозиторий