Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

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

Выполнил студент:
Самутичев Евгений Романович
группа: 3630102/70201

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

Санкт-Петербург
2020 г.

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Для каждого из 5 распределений:

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

выполнить следующее:

1. сгенерировать массив случайных данных (выборку) размера: 10, 50, 1000 и построить графики плотности вероятности (функции вероятности для распределения Пуассона как дискретного).

2. выборку размера: 10, 100, 1000 - сгенерировать 1000 раз, для каждой генерации произвести вычисления выборочных характеристик $\bar{x}, \operatorname{med} x, z_R, z_Q, z_{tr}$ для всех генераций в рамках одного размера выборки получить значения среднего характеристик положения:
$$E(z) = \bar{z} \tag{1}$$
и дисперсию:
$$D(z) = \bar{z^2} - \bar{z}^2 \tag{2}$$
Представить полученные данные в виде таблиц.

3. сгенерировать выборки размера 20 и 100, построить боксплот Тьюки. Определить долю выбросов экспериментально (сгенерировав выборку каждого размера 1000 раз) 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/функции распределения на отрезке $[-4, 4]$ для непрерывных распределений и на отрезке $[6, 14]$ для распределения Пуассона.

2 Теория

2.1 Распределения

Пусть задано вероятностное пространство $(\Omega, \mathcal{F}, \mathbf{P})$, на котором определена *случайная величина* $\xi : \Omega \to \mathbb{R}$ т.е. функция $\xi(\omega)$ такая что $\xi^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}(\mathbb{R})$. Она индуцирует вероятностную меру на $\mathbb{R}$ как $\mathbf{P}_\xi(B) = \mathbf{P}(\xi^{-1}(B))$ которая и носит название *распределения вероятностей* случайной величины [1].

Функция $F_\xi(x) = \mathbf{P}_\xi(-\infty, x], x \in \mathbb{R}$ называется *функцией распределения* случайной величины ξ. Случайная величина может быть:

1. *дискретной*, если распределение представимо в виде $\mathbf{P}_\xi(B) = \sum\limits_{k:x_k \in B} p(x_k)$, где $p(x_k) = \mathbf{P}_\xi\{x_k\}$ для конечного $\{x_1, ..., x_n\}$ или счетного $\{x_1, ..., x_k, ...\}$ подмножества вещественных чисел. В этом случае функция $p(x_k)$ называется таблицей распределения.
2. *непрерывной*, если $F(x)$ непрерывна
3. *абсолютно непрерывной*, если существует такая неотрицательная функция $f_\xi(x)$ называемая *плотностью вероятности*, что $F(x) = \int\limits_{-\infty}^{x} f(y)dy$

В работе рассматриваются следующие распределения:

1. *Нормальное* $N(x, 0, 1)$ - абсолютно непрерывное, задается плотностью
$$f_N(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3}$$
2. *Коши* $C(x, 0, 1)$ - абсолютно непрерывное, задается плотностью
$$f_C(x) = \frac{1}{\pi(x^2 + 1)} \tag{4}$$
3. *Лапласа* $L(x, 0, \frac{1}{\sqrt{2}})$ - абсолютно непрерывное, задается плотностью
$$f_L(x) = \frac{1}{2\sqrt{2}} e^{-\frac{1}{\sqrt{2}}|x|} \tag{5}$$
4. *Пуассона* $P(k, 10)$ - дискретное, задается на $\{1, 2, ..., k, ...\}$ как
$$p(k) = \frac{10^k}{k!} e^{-10} \tag{6}$$
5. *Равномерное* $U(x, -\sqrt{3}, \sqrt{3})$ - абсолютно непрерывное, задается плотностью
$$f_U(x) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{если } x \in [-\sqrt{3}, \sqrt{3}] \\ 0 & \text{иначе} \end{cases} \tag{7}$$

2.2 Гистограмма

Все приведенные распределения характеризуются таблицей (для дискретных) или плотностью (для абсолютно непрерывных). Эмпирическим аналогом таблицы или плотности является *гистограмма* [2]. Гистрограмма строится по группированным данным. Предполагаемую область значений случайной величины ξ делят на некоторое количество интервалов:

Пусть $A_1, ..., A_k$ - интервалы на прямой. Обозначим $\nu_j, j \in \{1, ..., k\}$ - число элементов выборки, попавших в интервал A_j. Размер выборки в этих обозначениях равен $n = \sum\limits_{j=1}^{k} \nu_j$. На каждом из интервалов строят прямоугольник, площадь которого пропорциональна ν_j, общая площадь всех прямоугольников должна равняться единице (нормировка гистограммы), поэтому высота каждого определяется как $f_j = \frac{\nu_j}{nl_j}$. Полученная фигура из объединения прямоугольников и называется гистограммой.

2.3 Вариационный ряд

Если элементы выборки $x_1, ..., x_n$ упорядочить по возрастанию на каждом элементарном исходе (рассматриваем их как случайные величины), получится новый набор случайных величин, называемый *вариационным рядом*:

$$x_{(1)} \leq ... \leq x_{(n)}$$

Элемент $x_{(k)}$ называется *k-ой порядковой статистикой* [1] .

2.4 Выборочные характеристики

При работе с выборкой нам неизвестно распределение по которому она получена, а значит и соответствующие характеристики распределения. Однако, существуют оценки - т.н. *выборочные характеристики*:

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

- Выборочная медиана
$$\operatorname{med} x = \begin{cases} x_{(k+1)} & \text{при } n = 2k+1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{при } n = 2k \end{cases} \tag{9}$$

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

- Выборочный квантиль уровня α
$$z_\alpha = \frac{x_{(\lfloor q \rfloor + 1)} + x_{(\lceil q \rceil + 1)}}{2}, \text{где } q = (n-1)\alpha \tag{11}$$

  формула, используемая в **NumPy**, в этом случае $z_0 = \min\limits_{i=1,...,n} x_{(i)}, z_1 = \max\limits_{i=1,...,n} x_{(i)}$, $z_{0.5} = \operatorname{med} x$

[1] [2] стр. 10

- Полусумма квантилей

$$z_Q = \frac{z_{0.25} + z_{0.75}}{2} \tag{12}$$

- Усеченное среднее

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

Выборочные характеристики как борелевские функции от случайных величин (выборки) также являются случайными величинами, поэтому в работе и производится усреднение их значений для 1000 генераций и вычисление дисперсии.

2.5 Выбросы

2.5.1 Определение

Результат измерения, выделяющийся из выборки называется *выбросом*. Простейший критерий основан на межквартильном расстоянии, выбросами считаются элементы выборки лежащие вне диапазона $[X_1, X_2]$:

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

, где LQ, UQ - выборочные нижний и верхний квартили.

Теоретическая вероятность выбросов для непрерывных распределений:

$$P_{outlier} = P(x < X_1) + P(x > X_2) = F(X_1) + (1 - F(X_2)) \tag{15}$$

, а для дискретных с учетом возможного скачка

$$P_{outlier} = F(X_1) - (F(X_1+) - F(X_1)) + (1 - F(X_2)) \tag{16}$$

2.5.2 Доля выбросов

Проведем следующий эксперимент $1, ..., i, ..., N$ раз: сгенерируем выборку размера n и подсчитаем число выбросов k_i, используя определение (1), но с выборочными квартилями. Тогда доля выбросов в i-м эксперименте:

$$P_i = \frac{k_i}{n} \tag{17}$$

Собственно *долей выбросов* будем называть величину

$$P = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{18}$$

, с дисперсией

$$D = \frac{1}{N} \sum_{i=1}^{N} P_i^2 - P^2 \tag{19}$$

2.6 Боксплот Тьюки

2.6.1 Описание

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей: в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. [3]

2.6.2 Построение

Границами ящика служат LQ и UQ, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов): X_1 и X_2 (1).

2.7 Эмпирическая функция распределения

Эмпирической функцией распределения, построенной по выборке $(x_1, ..., x_n)$ объема n называется случайная функция $F_n^* : \mathbb{R} \times \Omega \to [0, 1]$, которая имеет вид

$$F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i < y) \tag{20}$$

где I - индикатор события $x_i < y$ [2]

2.8 Ядерная оценка плотности распределения

Пусть $(x_1, ..., x_n)$ - выборка полученная по распределению с некоторой плотностью f, требуется оценить функцию f. *Ядерным оценщиком плотности* называется [4]

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{21}$$

где K - т.н. *ядро* (некоторая неотрицательная функция), $h > 0$ - сглаживающий параметр, именуемый *шириной полосы*.

Как правило используется нормальное (или гауссово) ядро, в силу его удобных математических свойств:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{22}$$

В случае если используется гауссово ядро и оцениваемая плотность является гауссовой, оптимальный выбор для h определяется т.н. *правилом Сильвермана* [4]:

$$h_n = \left(\frac{4s_n^5}{3n}\right)^{\frac{1}{5}} \approx 1.06 s_n n^{-\frac{1}{5}} \tag{23}$$

где s_n - выборочное среднеквадратичное отклонение (корень из выборочной дисперсии)

3 Реализация

Работа выполнена с использованием языка **Python** в интегрированной среде разработки **PyCharm**, были задействованы библиотеки:

- **NumPy** - работа с массивами данных, построение вариационного ряда и вычисления характеристик, вычисление квартилей для дальнейшего подсчета выбросов
- **SciPy** - модуль **stats** для генерации данных по распределениям, вычисления ядерной оценки плотности
- **Matplotlib** - отрисовка гистограмм и графиков, построение боксплотов

4 Результаты

4.1 Гистограммы и графики

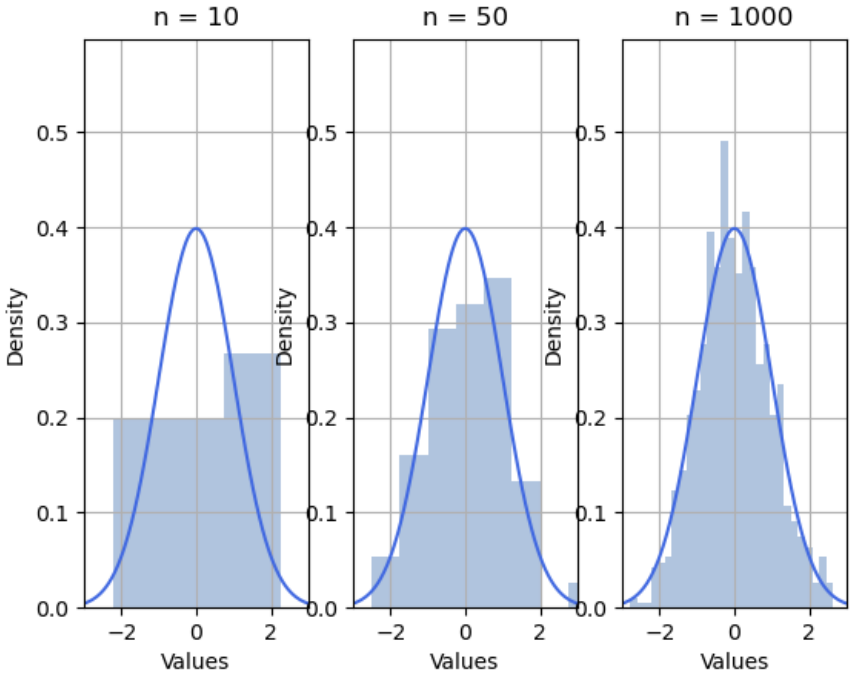

Рис. 1: Нормальное распределение

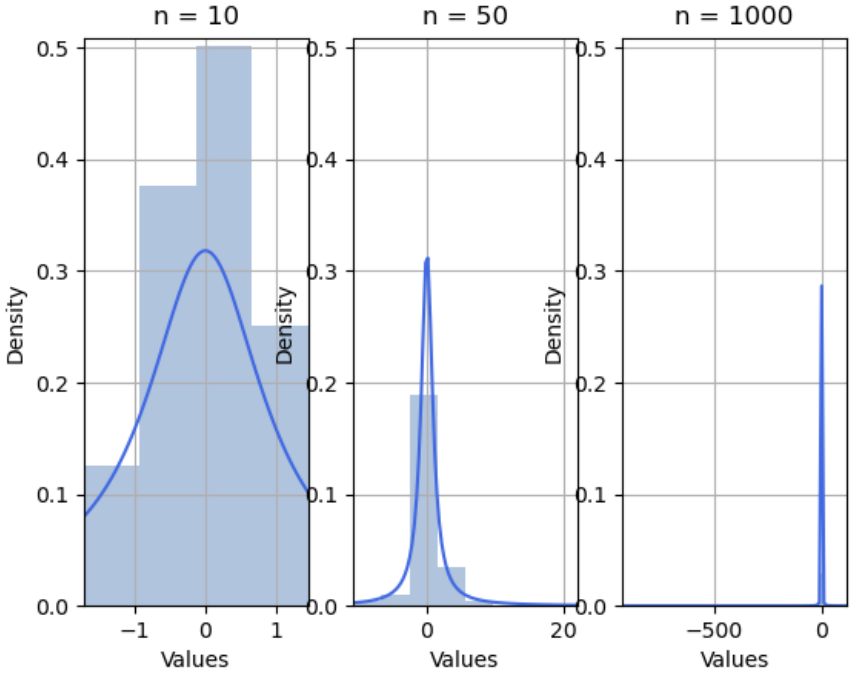

Рис. 2: Распределение Коши

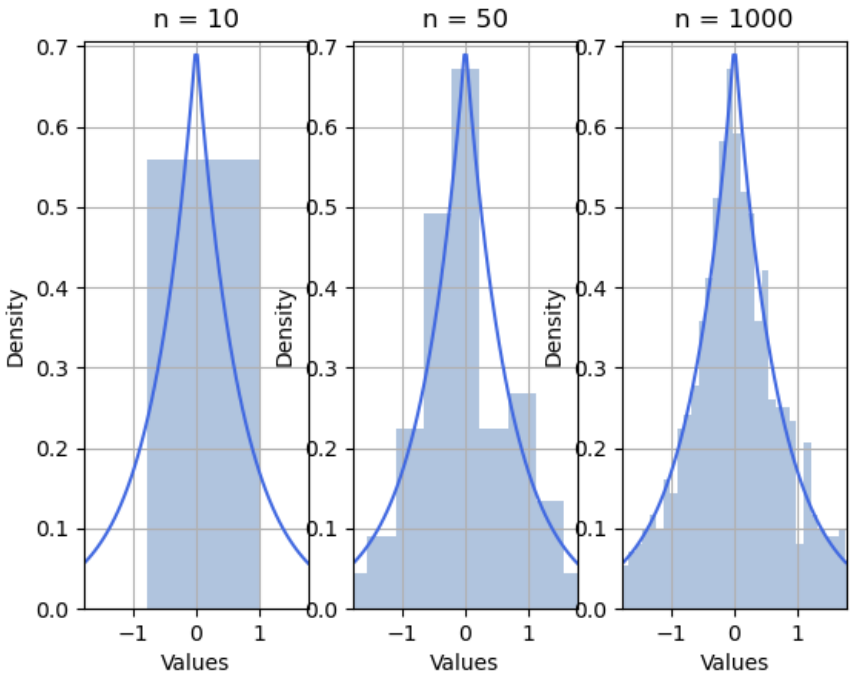

Рис. 3: Распределение Лапласа

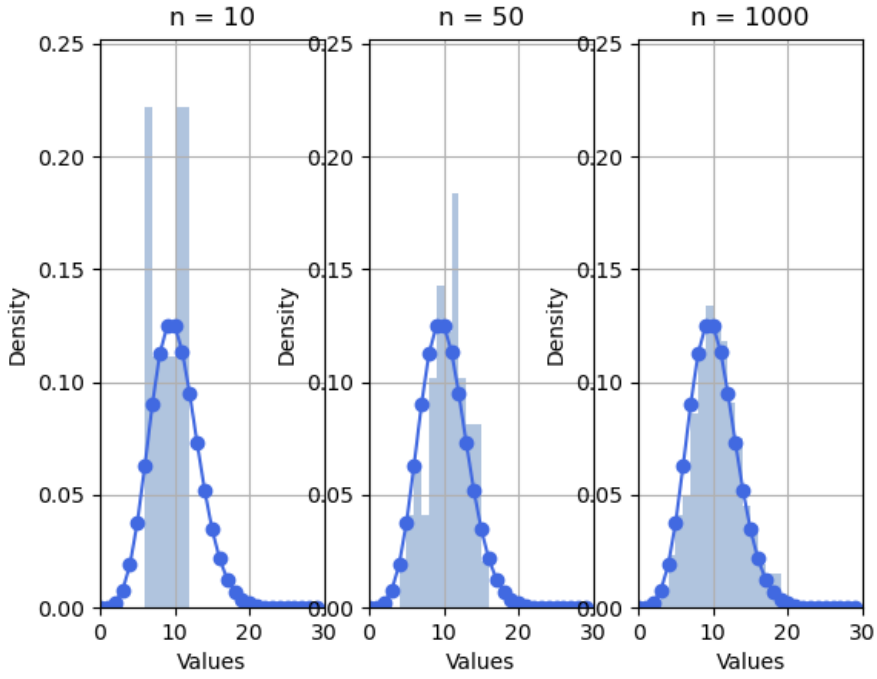

Рис. 4: Распределение Пуассона

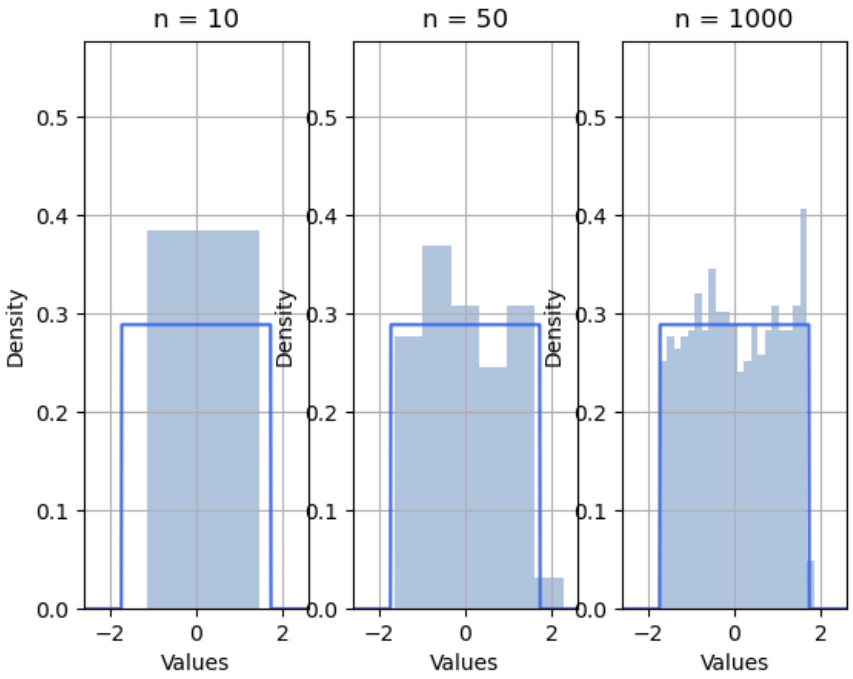

Рис. 5: Равномерное распределение

4.2 Выборочные характеристики

	$\bar{x}$ (8)	med x (9)	z_R (10)	z_Q (12)	z_{tr} (13)
$n = 10$					
$E(z)$	-0.0	-0.0	-0.0	-0.0	-0.1
$D(z)$	0.094467	0.130519	0.187448	0.105385	0.069492
$n = 100$					
$E(z)$	0.0	0.0	0.0	0.01	-0.01
$D(z)$	0.009651	0.015116	0.091466	0.011977	0.011309
$n = 1000$					
$E(z)$	-0.0	0.001	0.0	-0.0	-0.001
$D(z)$	0.001049	0.00153	0.062347	0.001313	0.001193

Таблица 1: Нормальное распределение

	$\bar{x}$	med x	z_R	z_Q	z_{tr}
$n = 10$					
$E(z)$	-0.0	-0.0	-1.0	-0.0	-0.2
$D(z)$	490.607293	0.332166	11914.643438	1.132547	0.210165
$n = 100$					
$E(z)$	2.0	0.0	79.0	0.0	-0.01
$D(z)$	3581.697241	0.026278	8922533.221739	0.04842	0.025421
$n = 1000$					
$E(z)$	1.0	0.0	642.0	0.0	-0.0
$D(z)$	2223.205146	0.00256	547218440.611133	0.004966	0.002619

Таблица 2: Распределение Коши

	$\bar{x}$	med x	z_R	z_Q	z_{tr}
$n = 10$					
$E(z)$	0.0	0.0	0.0	0.0	-0.1
$D(z)$	0.098335	0.072722	0.382323	0.089897	0.041919
$n = 100$					
$E(z)$	-0.0	0.0	-0.0	-0.0	-0.01
$D(z)$	0.009719	0.00561	0.433362	0.009396	0.005806
$n = 1000$					
$E(z)$	0.0	0.001	-0.0	0.001	-0.0
$D(z)$	0.000918	0.000483	0.441511	0.000944	0.000561

Таблица 3: Распределение Лапласа

	$\bar{x}$	med x	z_R	z_Q	z_{tr}
$n = 10$					
$E(z)$	10.0	10.0	10.0	10.0	7.0
$D(z)$	0.955624	1.3806	1.744716	1.128935	0.704541
$n = 100$					
$E(z)$	10.0	9.9	11.0	9.9	9.6
$D(z)$	0.097956	0.194391	0.997104	0.14328	0.110723
$n = 1000$					
$E(z)$	10.0	10.0	12.0	9.994	9.84
$D(z)$	0.010385	0.003484	0.6581	0.002748	0.011585

Таблица 4: Распределение Пуассона

	$\bar{x}$	med x	z_R	z_Q	z_{tr}
$n = 10$					
$E(z)$	0.0	0.0	-0.0	0.0	-0.1
$D(z)$	0.10033	0.234165	0.043909	0.136123	0.119729
$n = 100$					
$E(z)$	0.0	-0.0	0.001	0.0	-0.02
$D(z)$	0.009457	0.028559	0.00059	0.014028	0.018067
$n = 1000$					
$E(z)$	-0.001	-0.002	4e-05	-0.001	-0.003
$D(z)$	0.00102	0.003073	6e-06	0.001465	0.002005

Таблица 5: Равномерное распределение

4.3 Боксплоты

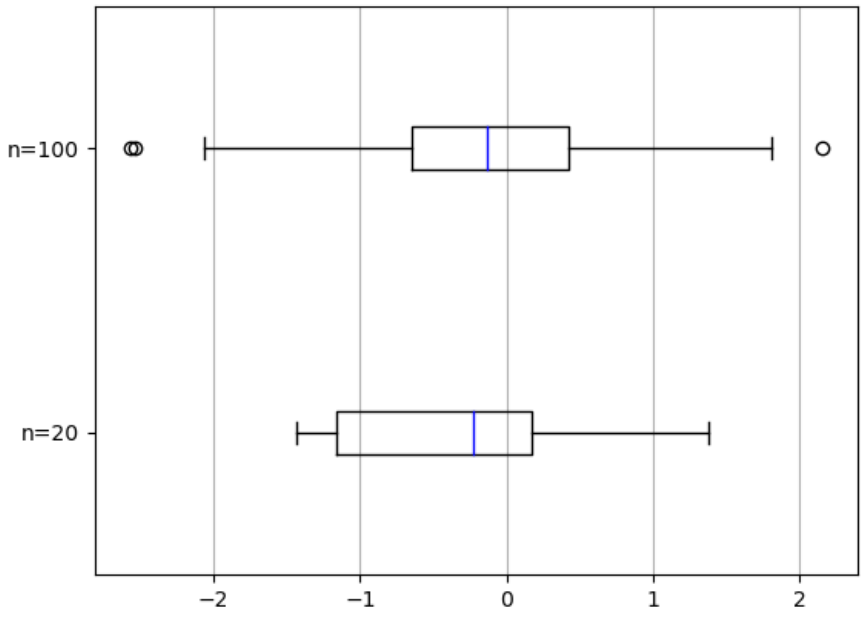

Рис. 6: Нормальное распределение

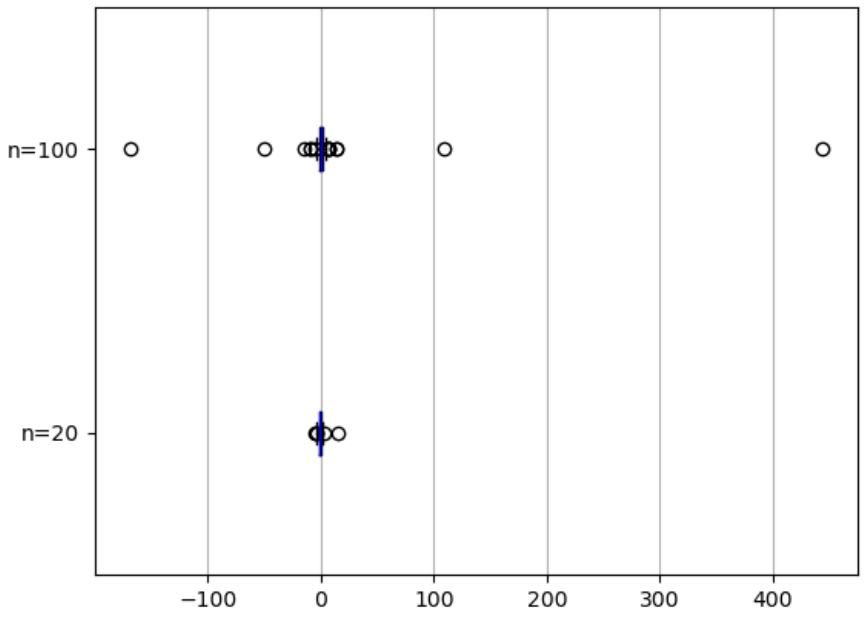

Рис. 7: Распределение Коши

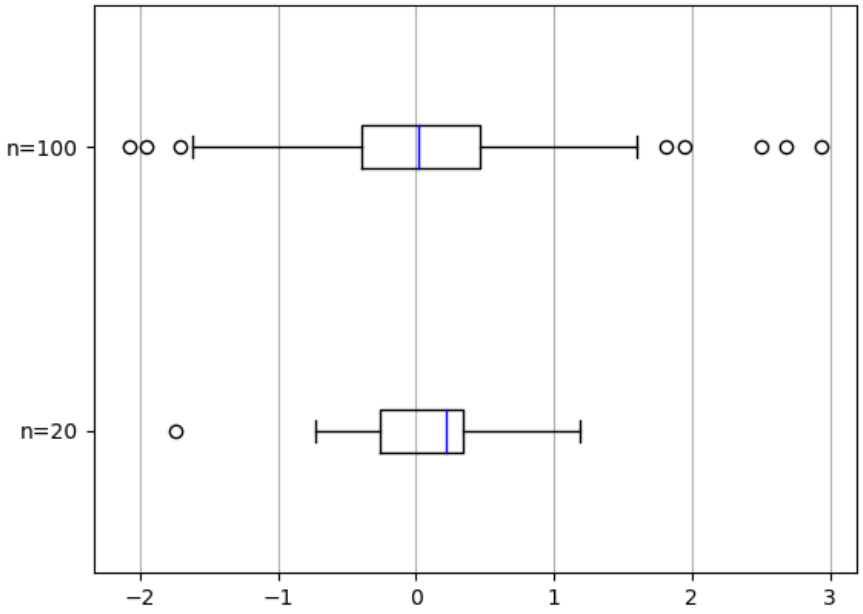

Рис. 8: Распределение Лапласа

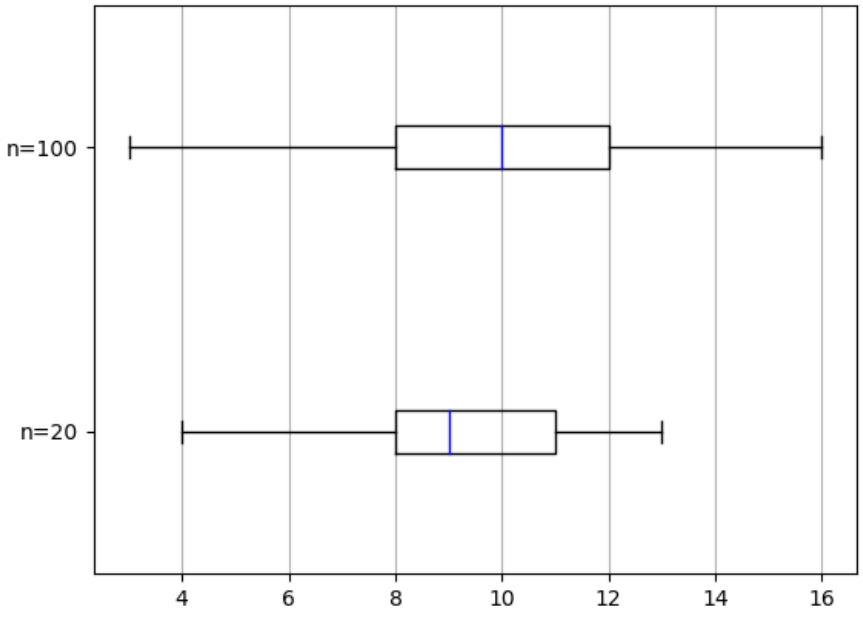

Рис. 9: Распределение Пуассона

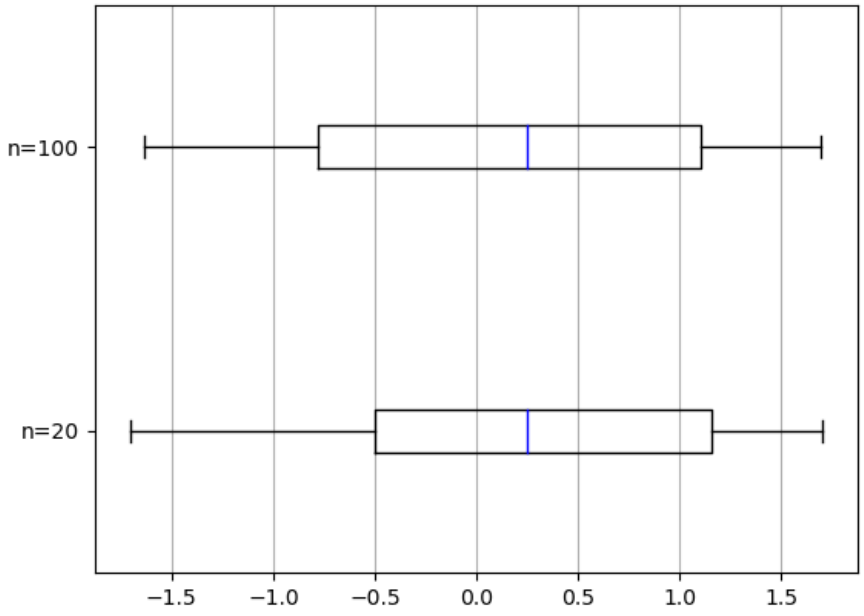

Рис. 10: Равномерное распределение

4.4 Теоретическая вероятность выбросов

Подсчитана для каждого распределения при помощи модуля stats библиотеки SciPy (см. Реализация):

Распределение	normal	cauchy	laplace	poisson	uniform
$P_{outlier}$ (15), (16)	0.007	0.156	0.0625	0.008	0.0

Таблица 6: Теоретическая вероятность выбросов

4.5 Доля выбросов

Распределение	normal	cauchy	laplace	poisson	uniform
$n = 20$					
$P(18)$	0.025	0.147	0.070	0.022	0.0023
$D(19)$	0.002085	0.005248	0.004219	0.001801	0.0002
$n = 100$					
P	0.0105	0.156	0.0658	0.0108	0.0
D	0.000185	0.001068	0.0009	0.000236	0.0

Таблица 7: Доля выбросов

4.6 Эмпирические функции распределения

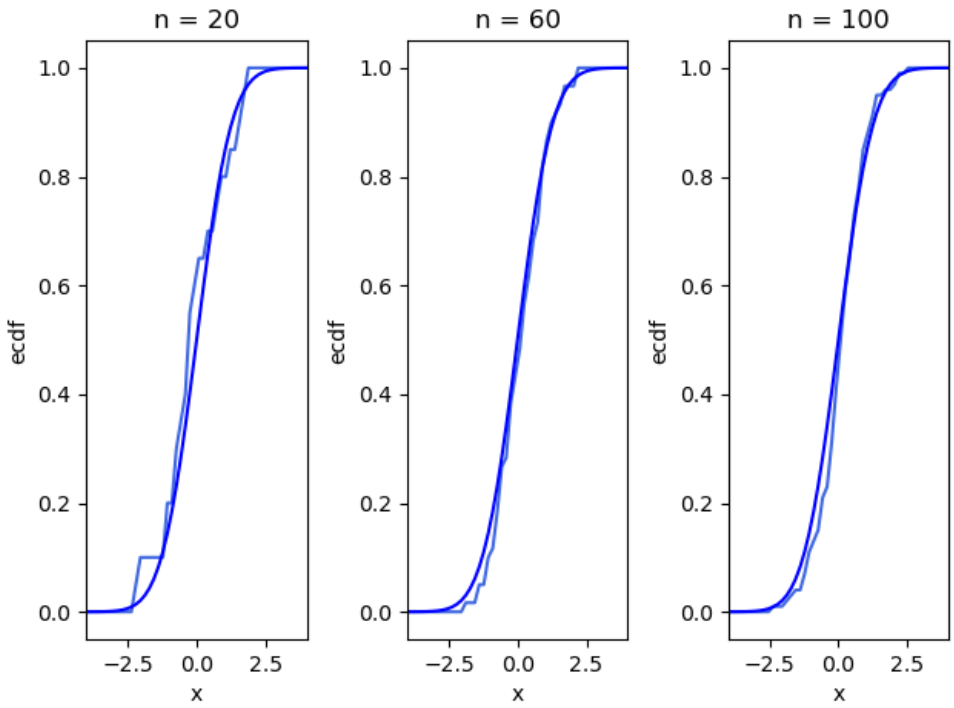

Рис. 11: Нормальное распределение

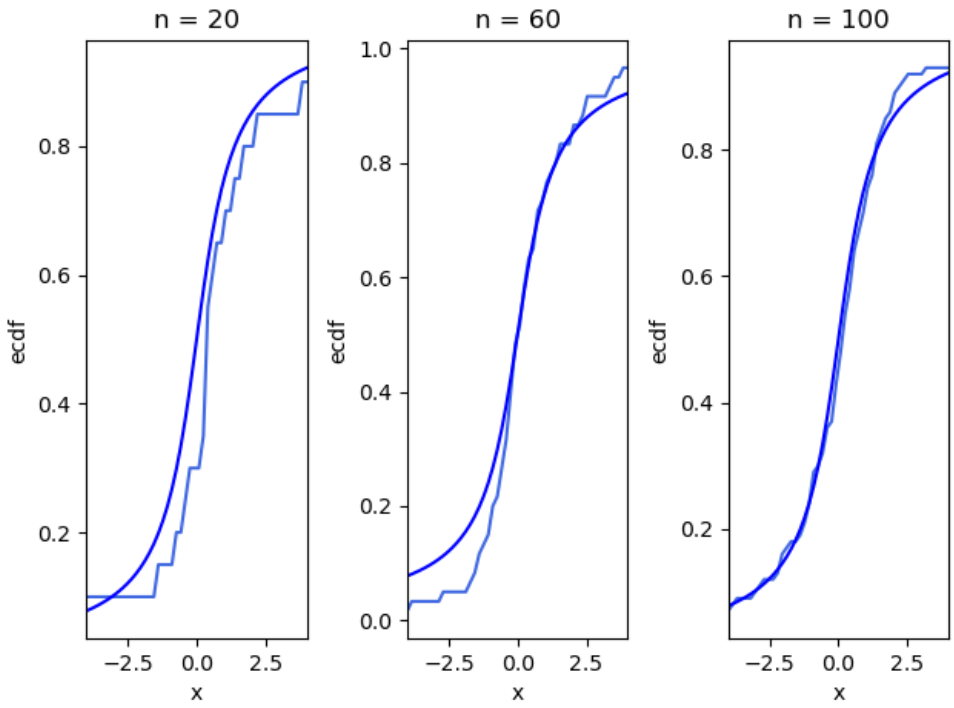

Рис. 12: Распределение Коши

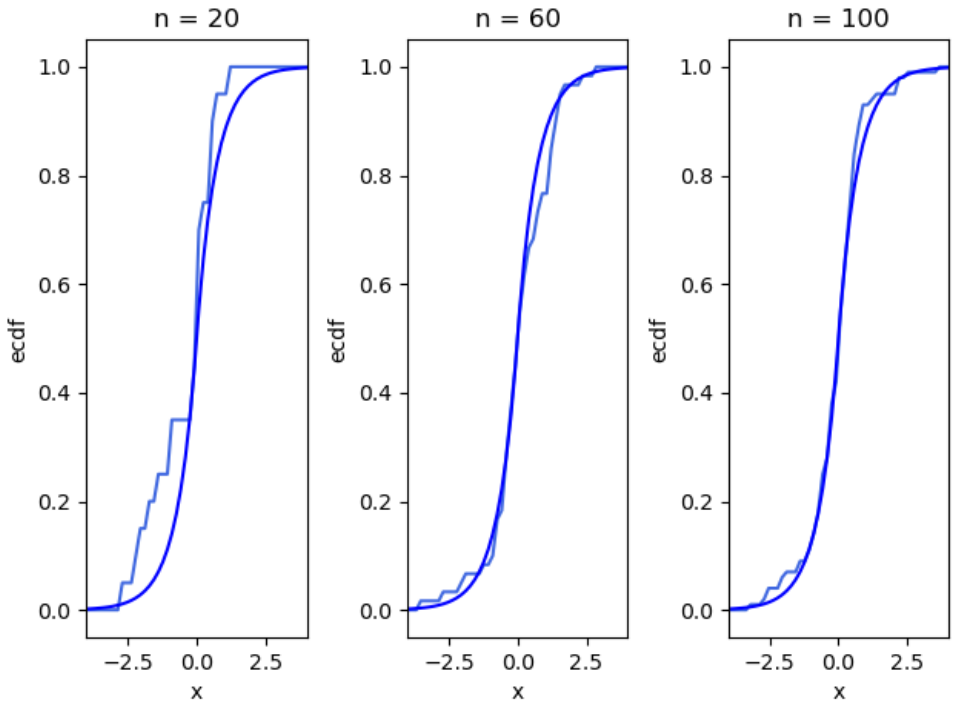

Рис. 13: Распределение Лапласа

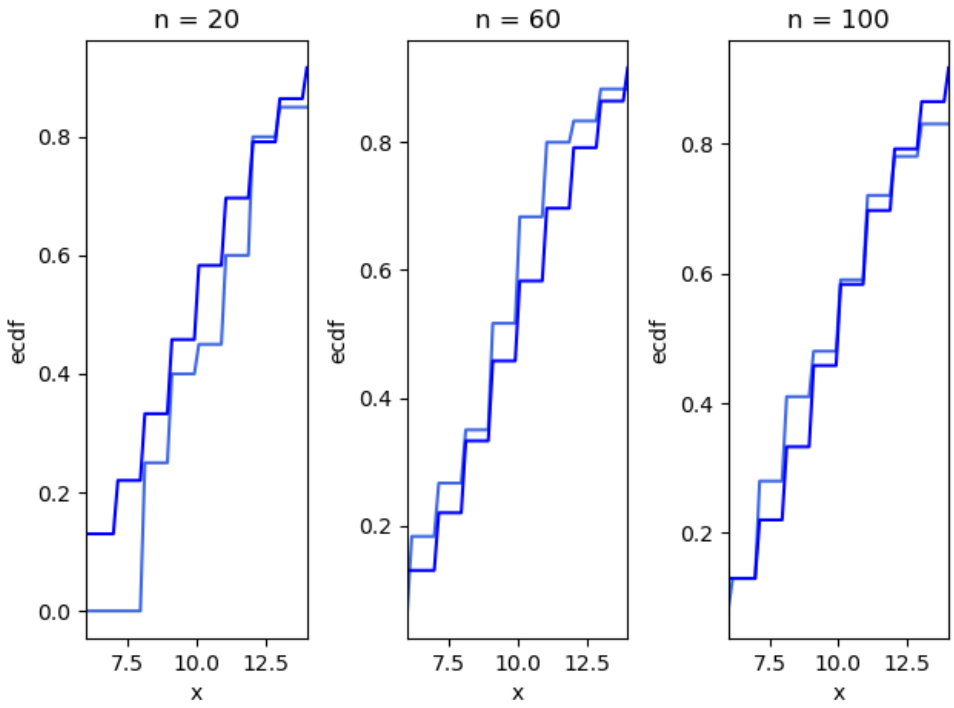

Рис. 14: Распределение Пуассона

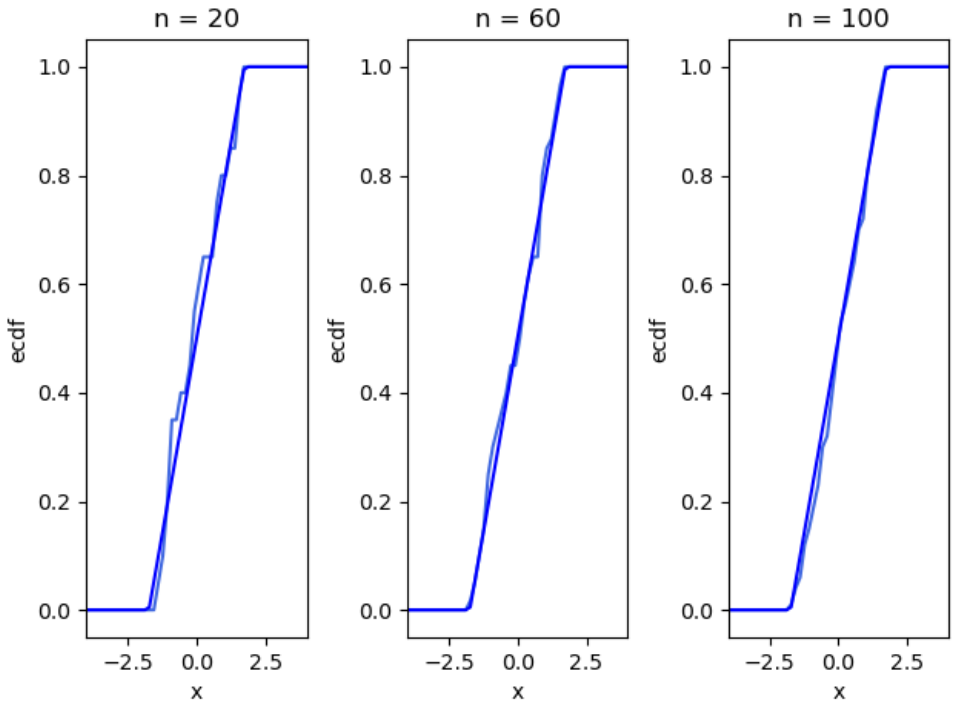

Рис. 15: Равномерное распределение

4.7 Ядерные оценки плотности распределения

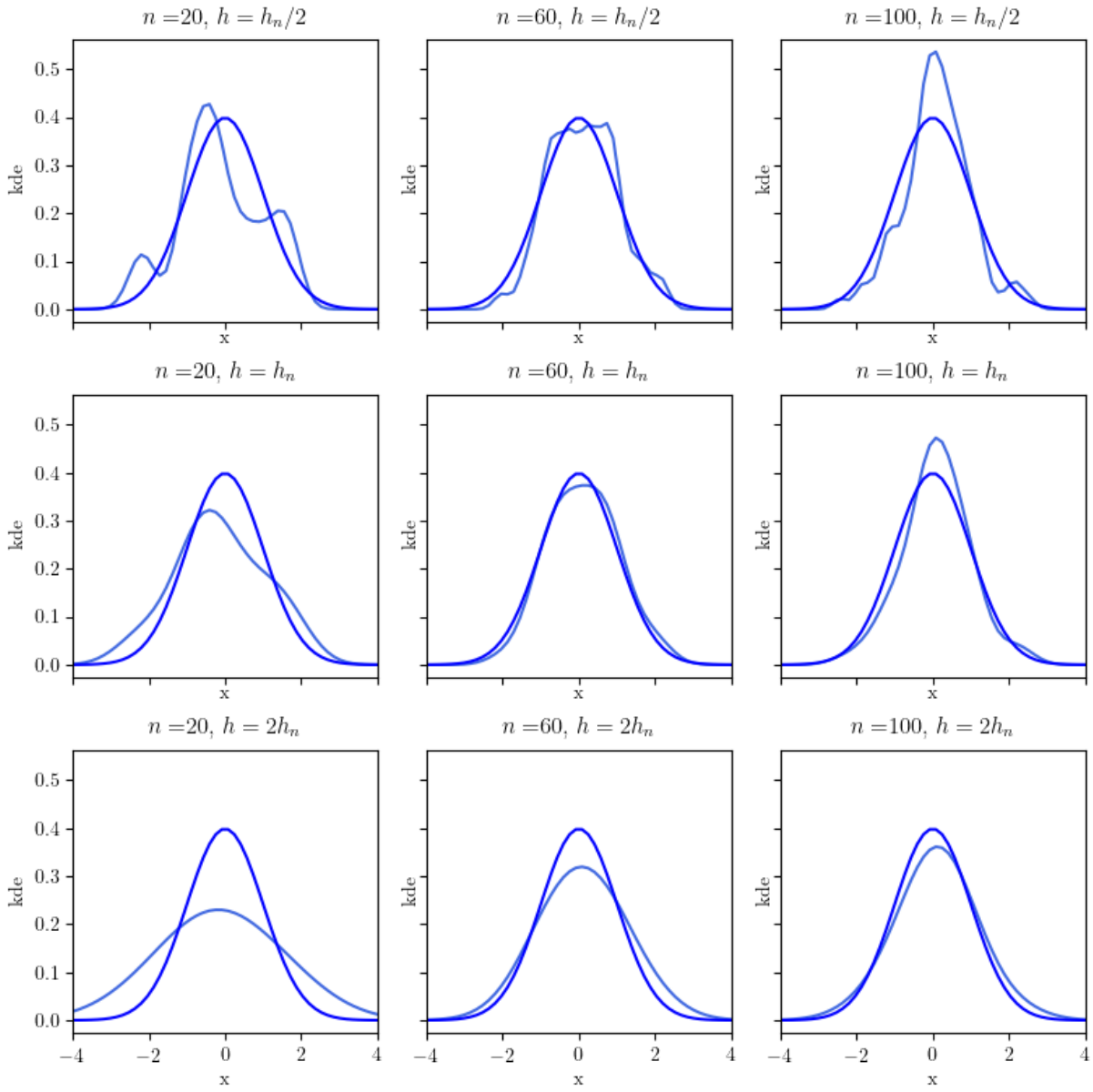

Рис. 16: Нормальное распределение

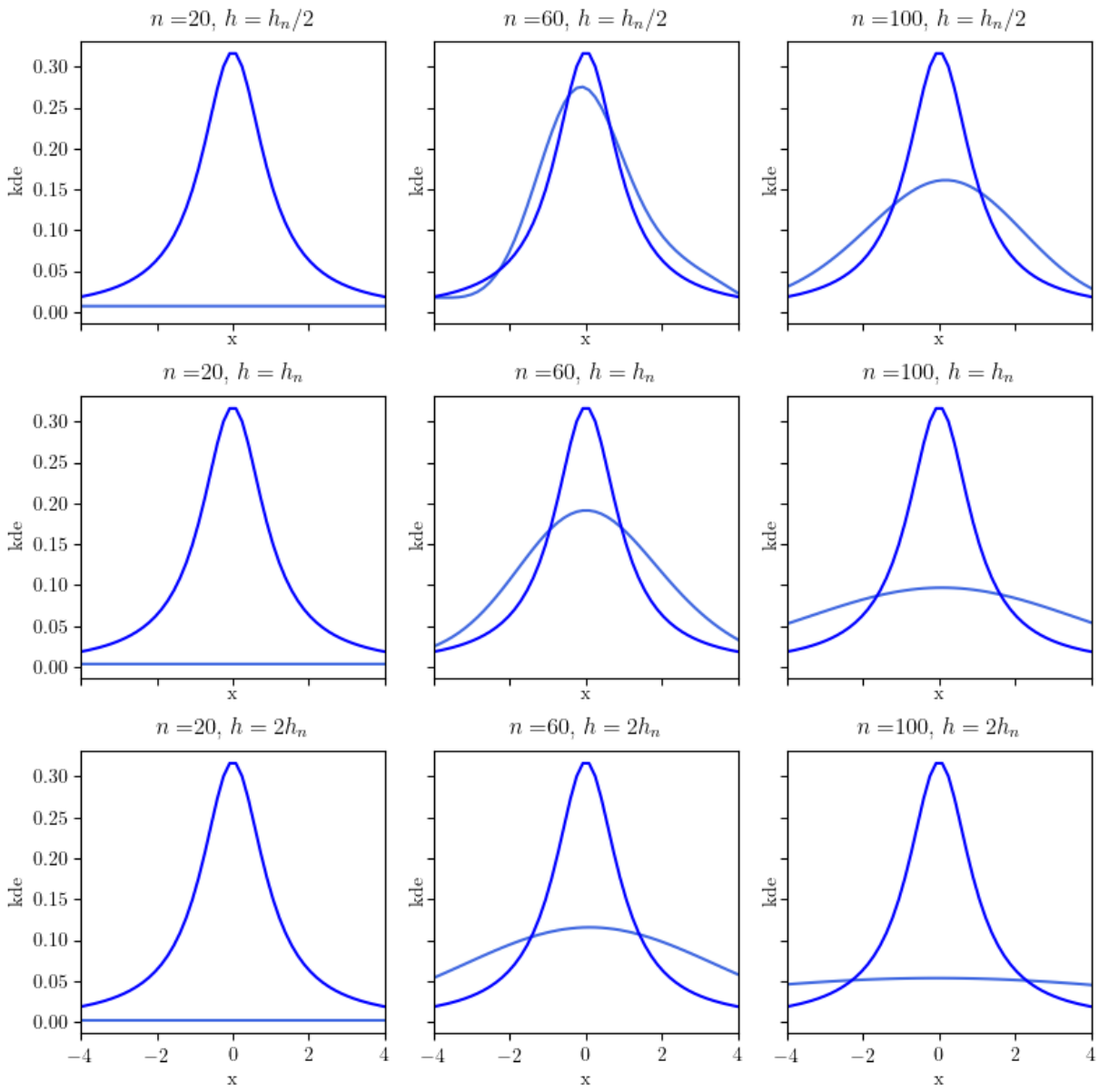

Рис. 17: Распределение Коши

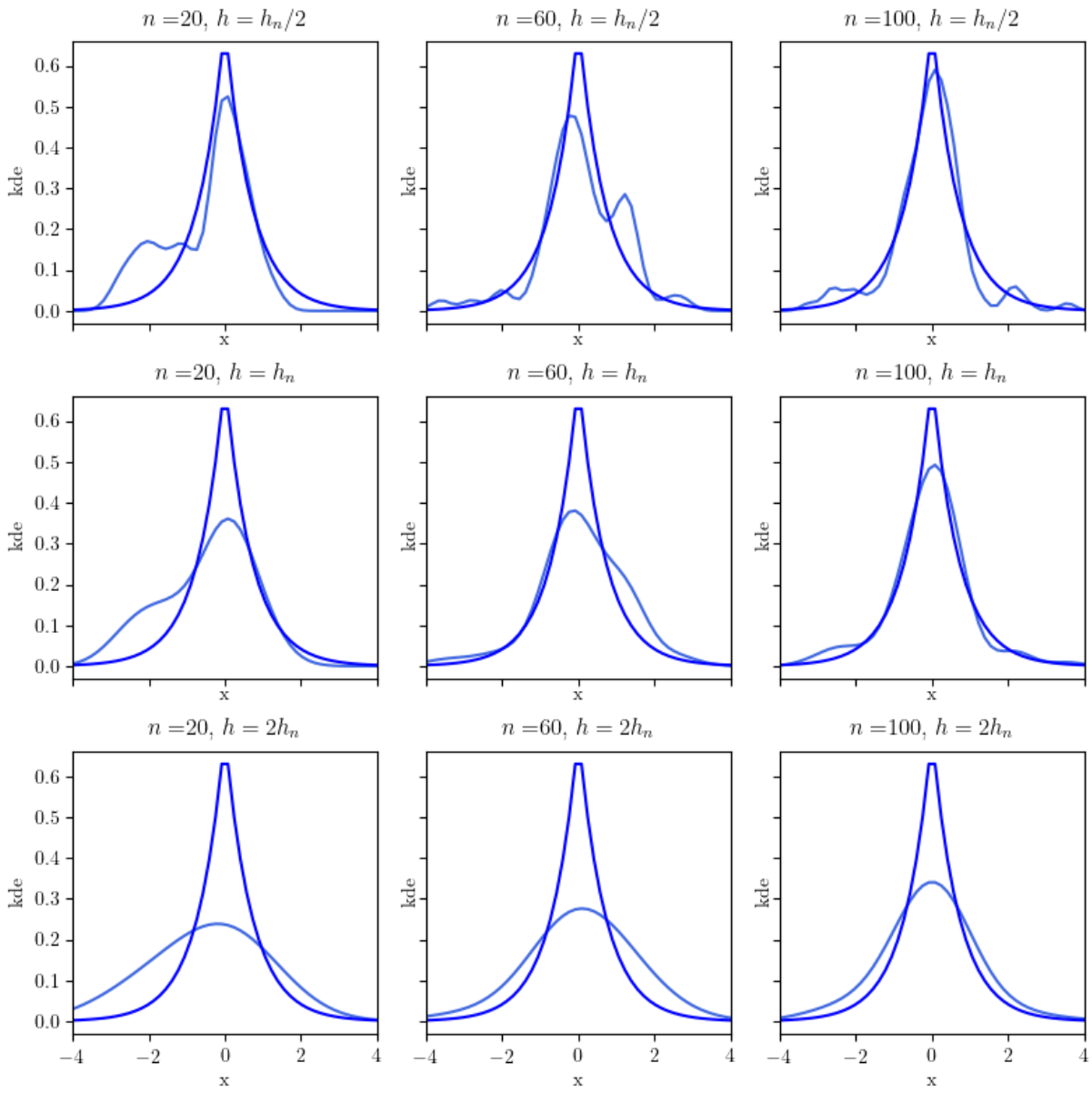

Рис. 18: Распределение Лапласа

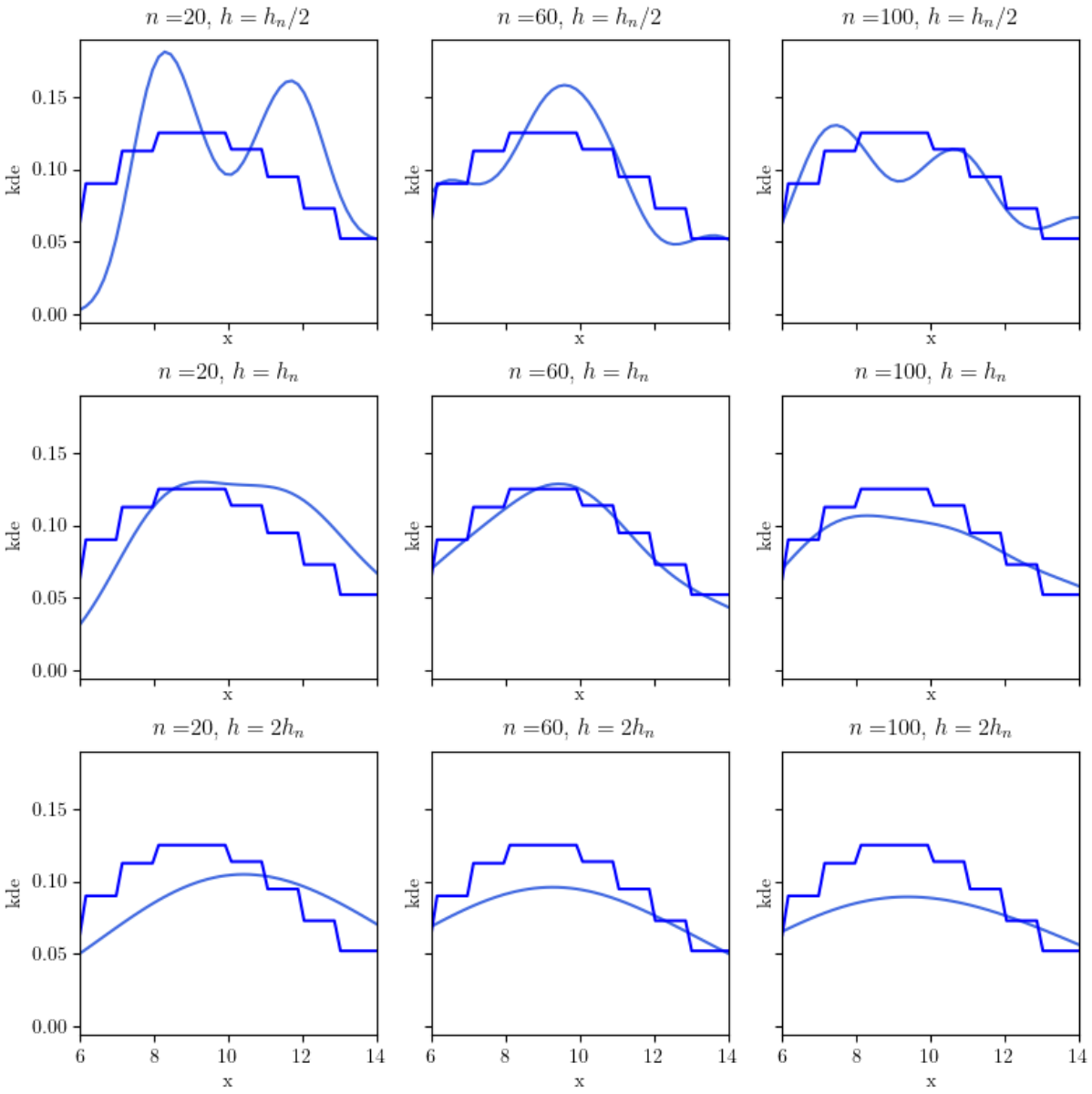

Рис. 19: Распределение Пуассона

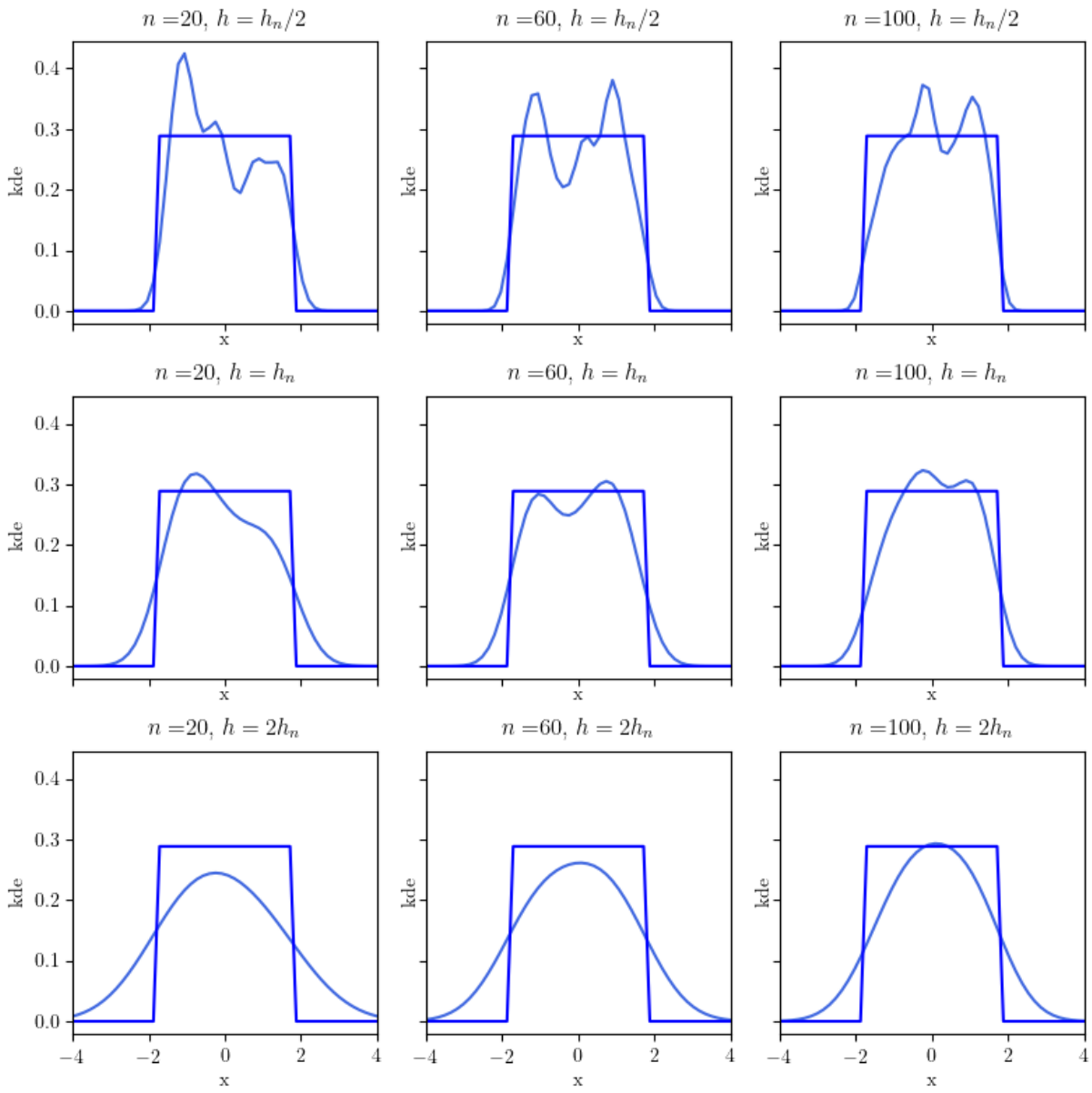

Рис. 20: Равномерное распределение

5 Обсуждение

5.1 Гистограммы и графики

Проведенный эксперимент подтверждает **утверждение**: *пусть плотность распределения по которому построена выборка является непрерывной функцией. Если число интервалов гистограммы $k(n)$ стремится к бесконечности таким образом что $\lim\limits_{n\to\infty}\frac{k(n)}{n}=0$, то имеет место сходимость по вероятности гистограммы к плотности.* [2] Действительно мы взяли $k(n)=\lceil\sqrt{n}\rceil$ и очевидно условие утверждения в таком случае выполнено, при этом гистограмма при увеличении n заполняет площадь под графиком плотности (кусочно-линейной функции вероятности для распределения Пуассона), а это и означает сходимость по вероятности.

5.2 Математическое ожидание и медиана

Для каждого из указанных в постановке задачи распределений, приведем теоретические значения математического ожидания и медианы:

- $N(x,0,1):\mathbf{E}=0,\text{med}=0$
- $C(x,0,1):\mathbf{E}-\text{не определено},\text{med}=0$
- $L(x,0,\frac{1}{\sqrt{2}}):\mathbf{E}=0,\text{med}=0$
- $P(k,10):\mathbf{E}=10,\text{med}=10$
- $U(x,-\sqrt{3},\sqrt{3}):\mathbf{E}=0,\text{med}=0$

Как известно, *выборочное среднее является несмещенной и состоятельной оценкой для математического ожидания*[2] Это объясняет то что для всех распределений кроме распределения Коши - выборочное среднее при росте n стремится к математическому ожиданию, для распределения Коши последовательность вычислений не демонстрирует никакой сходимости (см. таблицу 2), поскольку у него отсутствует математическое ожидание. В тоже время медиана имеется у всех распределений и к ней сходится выборочная медиана.

5.3 Полусуммы: z_R и z_Q

Полусумма квартилей z_Q и экстремальных выборочных элементов z_R оценивают центр симметрии распределения, из таблиц наблюдается что z_Q ближе к медиане и последовательность вычислений $E(z)$ для z_Q при увеличении n сходится, в тоже время последовательность значений $E(z)$ для z_R расходится при распределении Коши. Таким образом оценка через полусумму квартилей лучше, хотя и требует больше вычислений.

5.4 Упорядочение характеристик

Для $n=1000$ приведем упорядочение характеристик положения по каждому распределению:

- $N(x,0,1):z_{tr}<z_Q\le\bar{x}\le z_R<\text{med }x$
- $C(x,0,1):z_{tr}\le z_Q\le\text{med }x<\bar{x}<z_R$

[2] [2] стр. 17

- $L(x, 0, \frac{1}{\sqrt{2}}) : z_{tr} \leq z_R \leq \bar{x} < \text{med } x \leq z_Q$
- $P(k, 10) : z_{tr} < z_Q < \text{med } x \leq \bar{x} < z_R$
- $U(x, -\sqrt{3}, \sqrt{3}) : z_{tr} < \text{med } x < z_Q \leq \bar{x} < z_R$

5.5 Выбросы

Из полученных таблиц видно что доля выбросов близка к теоретической. Наибольшая при этом у распределения Коши, что также видно по боксплоту Тьюки (рис. 7). Вторая по величине у распределения Лапласа. Для остальных выборок доля выбросов не превосходит 95%, а значит можно считать что они соответствуют гипотетическим распределениям.

5.6 Эмпирическая функция распределения

Существует **теорема** [2]: *Пусть* $(x_1, ..., x_n)$ *- выборка из распределения с некоторой функцией распределения* F *и пусть* F_n^* *- эмпирическая функция распределения построенная по этой выборке. Тогда* $F_n^*(y) \xrightarrow{p} F(y), \forall y \in \mathbb{R}$ Полученные графики подтверждают данный теоретический факт, с ростом n эмпирическая функцяи распределения все ближе к истинной.

5.7 Ядерная оценка плотности распределения

Для нормального распределения наилучшие результаты показал выбор h по правилу Сильвермана, что обосновано теоретически т.к. он оптимален в некотором смысле (см. Теория), как и для распределения Пуассона. Для распределения Лапласа хорошие результаты в приближении плотности распределения имеем как при h_n, так и при $0.5h_n$. Плотность равномерного распределения аппроксимируется неудачно т.к. оно далеко от гауссова, как и распределение Коши.

6 Приложения

1. Исходный код лабораторной 1 https://github.com/zhenyatos/statlabs/tree/master/Lab1
2. Исходный код лабораторной 2 https://github.com/zhenyatos/statlabs/tree/master/Lab2
3. Исходный код лабораторной 3 https://github.com/zhenyatos/statlabs/tree/master/Lab3
4. Исходный код лабораторной 4 https://github.com/zhenyatos/statlabs/tree/master/Lab4

Список литературы

[1] А. Н. Ширяев, *Вероятность-1*. Изд. МЦНМО, Москва, 2017. 551 стр.

[2] Н. И. Чернова, *Математическая статистика: Учеб. пособие*. Новосиб. гос. ун-т. Новосибирск, 2007. 148 стр.

[3] Ящик с усами // Википедия. [2020—2020]. Дата обновления: 12.01.2020. URL: https://ru.wikipedia.org/?oldid=104502300 (дата обращения: 12.01.2020)

[4] Ядерная оценка плотности // Википедия. [2020—2020]. Дата обновления: 05.01.2020. URL: https://ru.wikipedia.org/?oldid=104368872 (дата обращения: 05.01.2020).