

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

Направление подготовки
«01.03.02 Прикладная математика и информатика»

Отчет по лабораторной работе № 1
дисциплина "Математическая статистика"

Выполнил студент гр. 5030102/00101
Преподаватель:

Отто М.В.
Баженов А.Н.

Санкт-Петербург

2023

Содержание

1	Постановка задачи	4
2	Теория	5
2.1	Рассматриваемые распределения	5
2.2	Гистограмма	5
2.2.1	Построение гистограммы	5
2.3	Вариационный ряд	5
2.4	Выборочные числовые характеристики	6
2.4.1	Характеристики положения	6
2.4.2	Характеристики рассеяния	6
2.5	Боксплот Тьюки	6
2.5.1	Построение	6
2.6	Теоретическая вероятность выбросов	7
2.7	Эмпирическая функция распределения	7
2.7.1	Статический ряд	7
2.7.2	Эмпирическая функция распределения	7
2.7.3	Нахождение эмпирической функции распределения	8
2.8	Оценки плотности вероятности	8
2.8.1	Определение	8
2.8.2	Ядерные оценки	8
3	Результаты	10
3.1	Гистограммы и графики плотности распределения	10
3.2	Характеристики положения и рассеяния	11
3.3	Боксплот Тьюки	13
3.4	Доля выбросов	15
3.5	Теоретическая вероятность выбросов	15
3.6	Эмпирическая функция распределения	16
3.7	Ядерные оценки плотности распределения	18
4	Обсуждение	24
4.1	Гистограммы	24
4.2	Характеристики положения и рассеяния	24
4.3	Доля и теоретическая вероятность выбросов	24
4.4	Эмпирическая функция распределения. Ядерные оценки плотности	24
5	Литература	26

Список иллюстраций

1	Нормальное распределение	10
2	Распределение Коши	10
3	Распределение Пуассона	10
4	Равномерное распределение	11
5	Нормальное распределение	13
6	Распределение Коши	14
7	Распределение Пуассона	14
8	Равномерное распределение	14
9	Нормальное распределение	16
10	Распределение Коши	16
11	Распределение Пуассона	17
12	Равномерное распределение	17
13	Нормальное распределение, $n = 20$	18
14	Нормальное распределение, $n = 60$	18
15	Нормальное распределение, $n = 100$	19
16	Распределение Коши, $n = 20$	19
17	Распределение Коши, $n = 60$	20
18	Распределение Коши, $n = 100$	20
19	Распределение Пуассона, $n = 20$	21
20	Распределение Пуассона, $n = 60$	21
21	Распределение Пуассона, $n = 100$	22
22	Равномерное распределение, $n = 20$	22
23	Равномерное распределение, $n = 60$	23
24	Равномерное распределение, $n = 100$	23

1 Постановка задачи

Для 4 распределений:

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

1. Сгенерировать выборки размером 10, 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 характеристики положения данных: \bar{x} (выборочное среднее), $medx$ (выборочная медиана), z_R (полусумма экстремальных выборочных элементов), z_Q (полусумма квартилей), z_{tr} (усечённое среднее). Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

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

Вычислить оценку дисперсий по формуле:

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

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

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Пуассона

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

- Равномерное распределение

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

2.2 Гистограмма

2.2.1 Построение гистограммы

Множество значений, которое может принимать элемент выборки, разбивается на несколько интервалов. Чаще всего эти интервалы берут одинаковыми, но это не является строгим требованием. Эти интервалы откладываются на горизонтальной оси, затем над каждым рисуется прямоугольник. Если все интервалы были одинаковыми, то высота каждого прямоугольника пропорциональна числу элементов выборки, попадающих в соответствующий интервал. Если интервалы разные, то высота прямоугольника выбирается таким образом, чтобы его площадь была пропорциональна числу элементов выборки, которые попали в этот интервал.

2.3 Вариационный ряд

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

2.4 Выборочные числовые характеристики

2.4.1 Характеристики положения

- Выборочное среднее

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7)$$

- Выборочная медиана

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

- Полусумма экстремальных выборочных элементов

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

- Полусумма квартилей

Выборочная квартиль z_p порядка p определяется формулой:

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

Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \quad (11)$$

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

2.4.2 Характеристики рассеяния

Выборочная дисперсия

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

2.5 Боксплот Тьюки

2.5.1 Построение

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

«усов»:

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

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

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 :

$$\begin{cases} x < X_1^T \\ x > X_2^T \end{cases} \quad (15)$$

Теоретическая вероятность выбросов:

- для непрерывных распределений

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

- для дискретных распределений

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

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, \dots, z_k положенных в возрастающем порядке с указанием частот n_1, n_2, \dots, 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:

$$F_n^* = P^*(X < x) \quad (18)$$

2.7.3 Нахождение эмпирической функции распределения

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

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

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

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

Таблица 1: Таблица распределения

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

$$F_n^*(x) \approx F_X(x). \quad (20)$$

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 $f(x)$

$$\hat{f}(x) \approx f(x). \quad (21)$$

2.8.2 Ядерные оценки

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

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - x_i}{h_n}\right). \quad (22)$$

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

$$h_n \xrightarrow{n \rightarrow \infty} 0; \quad \frac{h_n}{n^{-1}} \xrightarrow{n \rightarrow \infty} \infty. \quad (23)$$

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

Гауссово (нормальное) ядро

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

Правило Сильвермана

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

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

3 Результаты

3.1 Гистограммы и графики плотности распределения

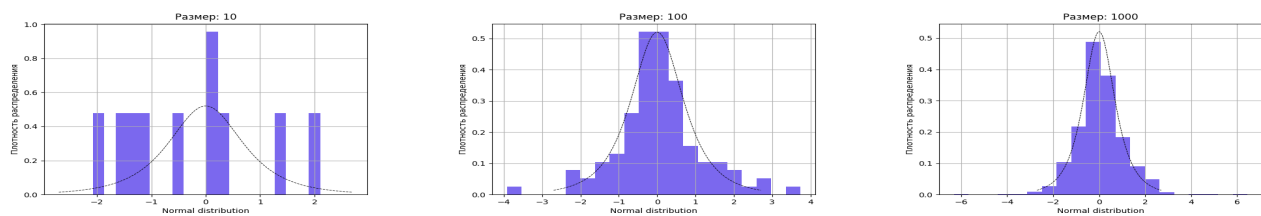


Рис. 1: Нормальное распределение

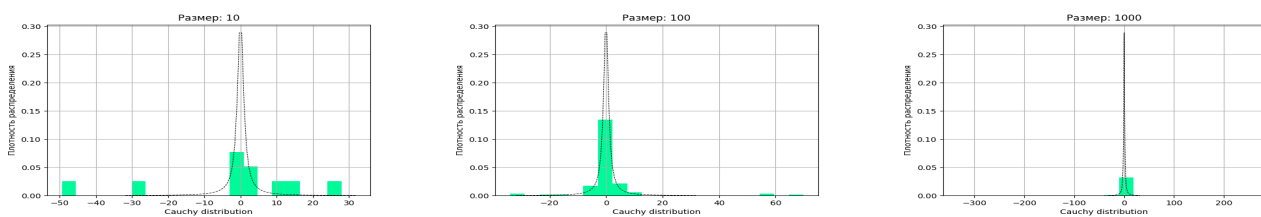


Рис. 2: Распределение Коши

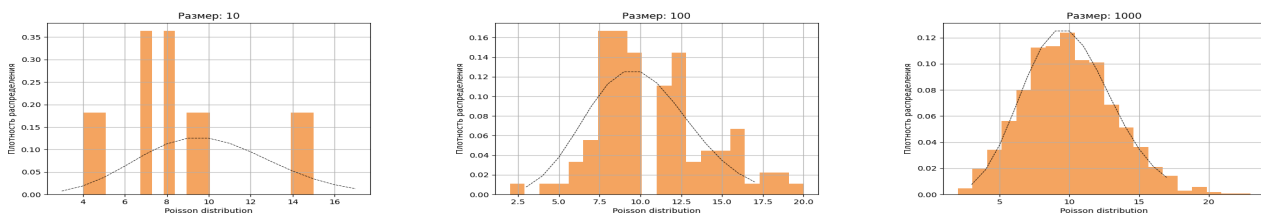


Рис. 3: Распределение Пуассона

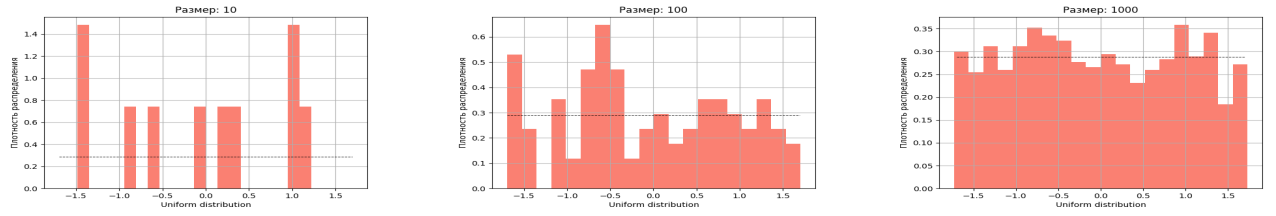


Рис. 4: Равномерное распределение

3.2 Характеристики положения и рассеяния

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Normal E(z) 10	-0.02	-0.03	-0.02	0.29	0.25
Normal D(z) 10	0.09	0.13	0.19	0.12	0.11
$E(z) \pm \sqrt{D(z)}$	[-0.3; 0.2]	[-0.4; 0.3]	[-0.4; 0.4]	[-0.1; 0.6]	[-0.1; 0.5]
$\hat{E}(z)$	0	0	0	0	0
Normal E(z) 100	-0.0001	0.003	0.01	0.01	0.02
Normal D(z) 100	0.009	0.01	0.09	0.01	0.01
$E(z) \pm \sqrt{D(z)}$	[-0.09; 0.09]	[-0.1; 0.1]	[-0.2; 0.3]	[-0.09; 0.1]	[-0.08; 0.1]
$\hat{E}(z)$	0	0	0	0	0
Normal E(z) 1000	0.001	0.0004	0.004	0.002	0.003
Normal D(z) 1000	0.0009	0.001	0.063	0.001	0.001
$E(z) \pm \sqrt{D(z)}$	[-0.03; 0.03]	[-0.03; 0.03]	[-0.2; 0.2]	[-0.03; 0.03]	[-0.03; 0.03]
$\hat{E}(z)$	0.0	0.0	0	0.0	0.0

Таблица 2: Нормальное распределение

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Cauchy E(z) 10	1.04	-0.005	5.3	1.1	0.6
Cauchy D(z) 10	1227.2	0.3	30466.1	5.07	1.1
E(z) $\pm\sqrt{D(z)}$	[-33.9; 36.07]	[-0.5; 0.5]	[-169.2; 179.8]	[-1.1; 3.3]	[-0.3; 1.7]
$\widehat{E}(z)$	-	0	-	-	-
Cauchy E(z) 100	1.6	0.005	85.05	0.03	0.04
Cauchy D(z) 100	2616.1	0.02	6492272.5	0.05	0.02
E(z) $\pm\sqrt{D(z)}$	[-49.4; 52.8]	[-0.1; 0.1]	[-2462.9; 2633.04]	[-0.2; 0.2]	[-0.1; 0.2]
$\widehat{E}(z)$	-	0	-	0	0
Cauchy E(z) 1000	-1.5	-0.0002	-734.3	0.003	0.004
Cauchy D(z) 1000	1880.6	0.002	468411808.4	0.004	0.002
E(z) $\pm\sqrt{D(z)}$	[-44.9; 41.8]	[-0.05; 0.04]	[-22377.2; 20908.4]	[-0.06; 0.07]	[-0.04; 0.05]
$\widehat{E}(z)$	-	0.0	-	0.0	0.0

Таблица 3: Распределение Коши

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Poisson E(z) 10	9.9	9.8	10.2	10.9	10.7
Poisson D(z) 10	1.007	1.4	1.8	1.3	1.2
E(z) $\pm\sqrt{D(z)}$	[8.9; 10.9]	[8.6; 11.03]	[8.8; 11.6]	[9.7; 12.05]	[9.6; 11.8]
$\widehat{E}(z)$	10_{-1}^{+1}	10_{-1}^{+1}	10_{-1}^{+1}	11_{-1}^{+1}	11_{-1}^{+1}
Poisson E(z) 100	10.01	9.8	10.9	9.9	9.9
Poisson D(z) 100	0.1	0.1	0.9	0.1	0.1
E(z) $\pm\sqrt{D(z)}$	[9.6; 10.3]	[9.4; 10.2]	[9.9; 11.8]	[9.5; 10.3]	[9.6; 10.3]
$\widehat{E}(z)$	10_{-1}^{+1}	10_{-1}^{+1}	11_{-1}^{+1}	10	10
Poisson E(z) 1000	10.001	9.9	11.7	9.9	9.8
Poisson D(z) 1000	0.01	0.002	0.7	0.003	0.01
E(z) $\pm\sqrt{D(z)}$	[9.8; 10.1]	[9.9; 10.05]	[10.8; 12.5]	[9.9; 10.05]	[9.7; 9.9]
$\widehat{E}(z)$	10_{-1}^{+1}	10_{-1}^{+1}	11_{-1}^{+1}	10_{-1}^{+1}	$9_{-0.1}^{+0.1}$

Таблица 4: Распределение Пуассона

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Uniform $E(z)$ 10	0.02	0.03	0.01	0.3	0.3
Uniform $D(z)$ 10	0.09	0.2	0.04	0.1	0.1
$E(z) \pm \sqrt{D(z)}$	[-0.2; 0.3]	[-0.4; 0.5]	[-0.2; 0.2]	[-0.003; 0.6]	[-0.03; 0.7]
$\hat{E}(z)$	0	0	0	0	0
Uniform $E(z)$ 100	-0.004	-0.001	-0.0004	0.01	0.02
Uniform $D(z)$ 100	0.01	0.03	0.0005	0.01	0.02
$E(z) \pm \sqrt{D(z)}$	[-0.1; 0.09]	[-0.1; 0.1]	[-0.02; 0.02]	[-0.1; 0.1]	[-0.1; 0.1]
$\hat{E}(z)$	0.0	0	0.0	0	0
Uniform $E(z)$ 1000	0.0005	0.0006	-0.0006	0.01	0.03
Uniform $D(z)$ 1000	0.01	0.02	0.0006	0.01	0.02
$E(z) \pm \sqrt{D(z)}$	[-0.1; 0.1]	[-0.1; 0.1]	[-0.02; 0.02]	[-0.1; 0.1]	[-0.1; 0.1]
$\hat{E}(z)$	0.0	0.0	0.00	0.0	0.0

Таблица 5: Равномерное распределение

3.3 Боксплот Тьюки

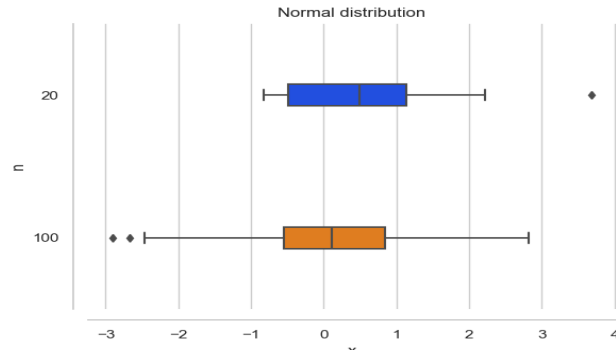


Рис. 5: Нормальное распределение

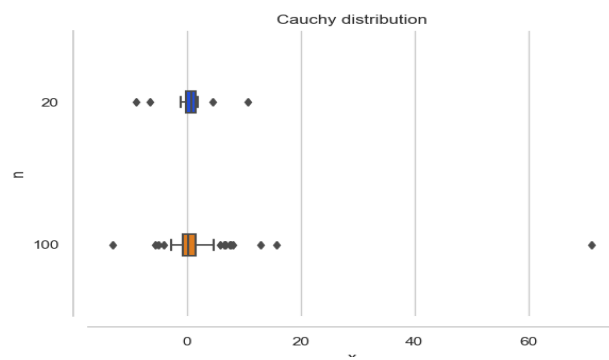


Рис. 6: Распределение Коши

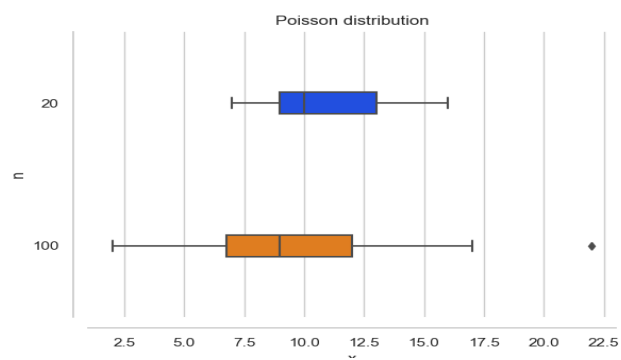


Рис. 7: Распределение Пуассона

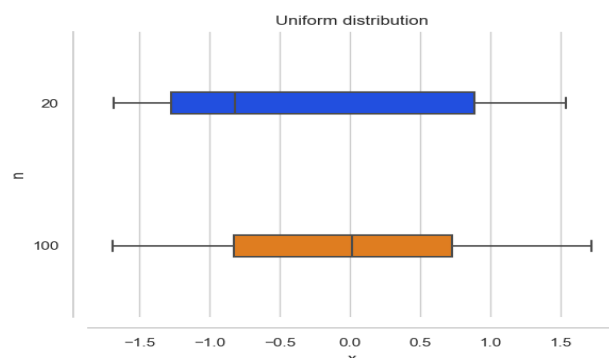


Рис. 8: Равномерное распределение

3.4 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	0.02305	0.007
Normal n=100	0.01495	0.007
Cauchy n=20	0.14755	0.156
Cauchy n=100	0.18427	0.156
Poisson n=20	0.0249	0.008
Poisson n=100	0.01578	0.008
Uniform n=20	0.00295	0
Uniform n=100	0.00059	0

Таблица 6: Доля выбросов

3.5 Теоретическая вероятность выбросов

Распределение	Q_1^T	Q_3^T	X_1^T	X_2^T	P_B^T
Нормальное распределение	-0.674	0.674	-2.698	2.698	0.007
Распределение Коши	-1	1	-4	4	0.156
Распределение Пуассона	8	12	2	18	0.008
Равномерное распределение	-0.866	0.866	-3.464	3.464	0

Таблица 7: Теоретическая вероятность выбросов

3.6 Эмпирическая функция распределения

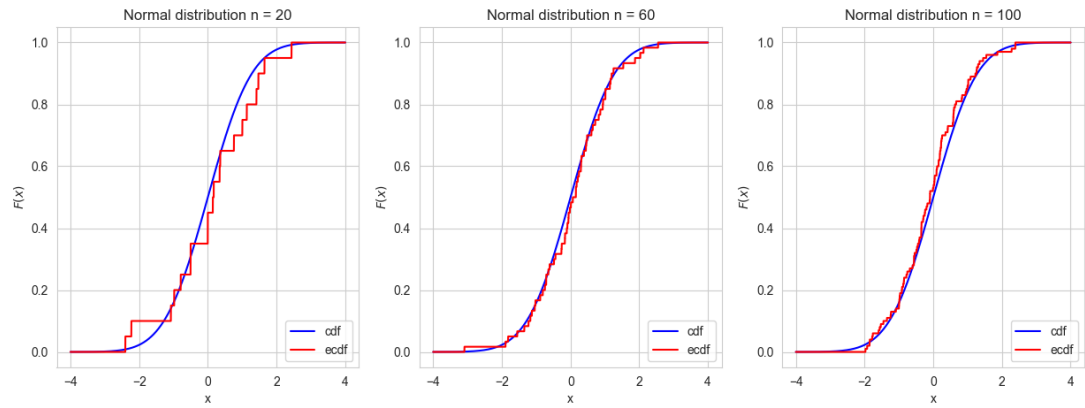


Рис. 9: Нормальное распределение

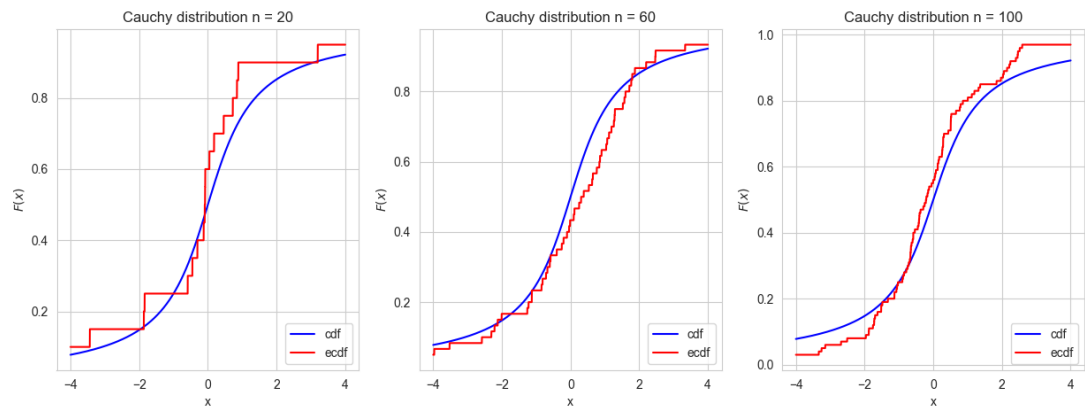


Рис. 10: Распределение Коши

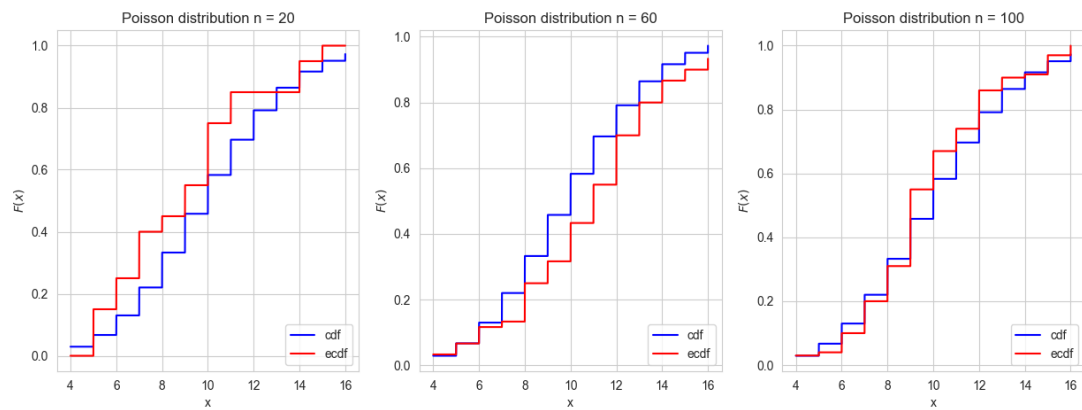


Рис. 11: Распределение Пуассона

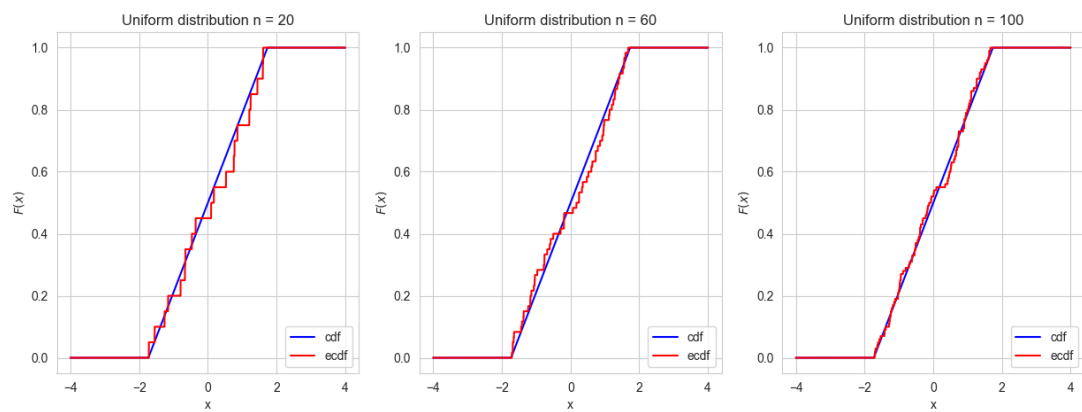


Рис. 12: Равномерное распределение

3.7 Ядерные оценки плотности распределения

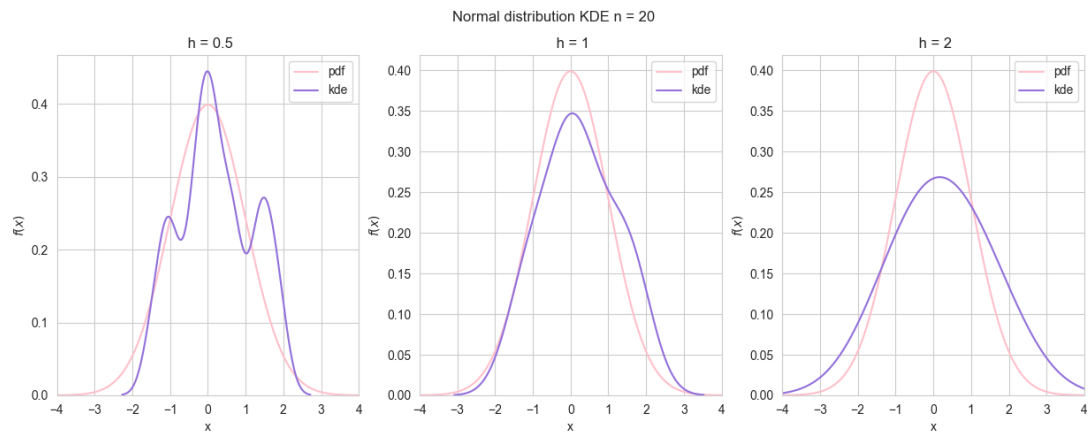


Рис. 13: Нормальное распределение, $n = 20$

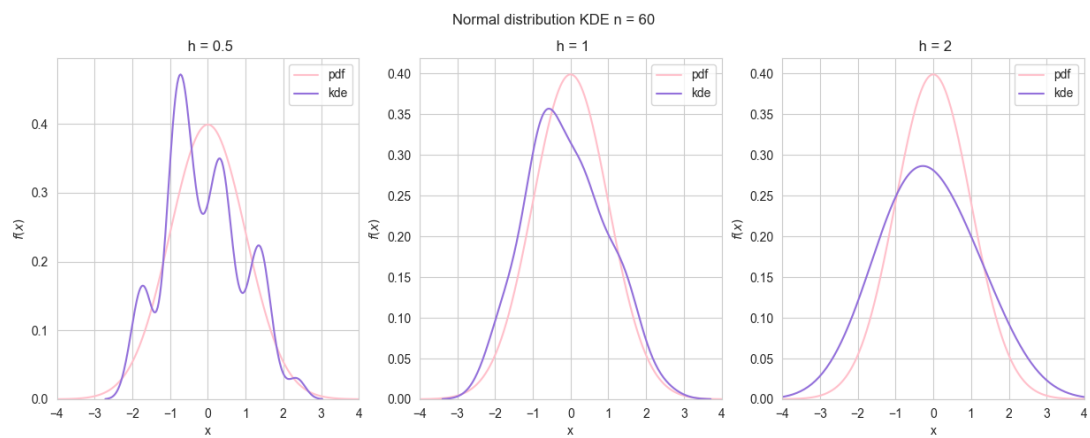


Рис. 14: Нормальное распределение, $n = 60$

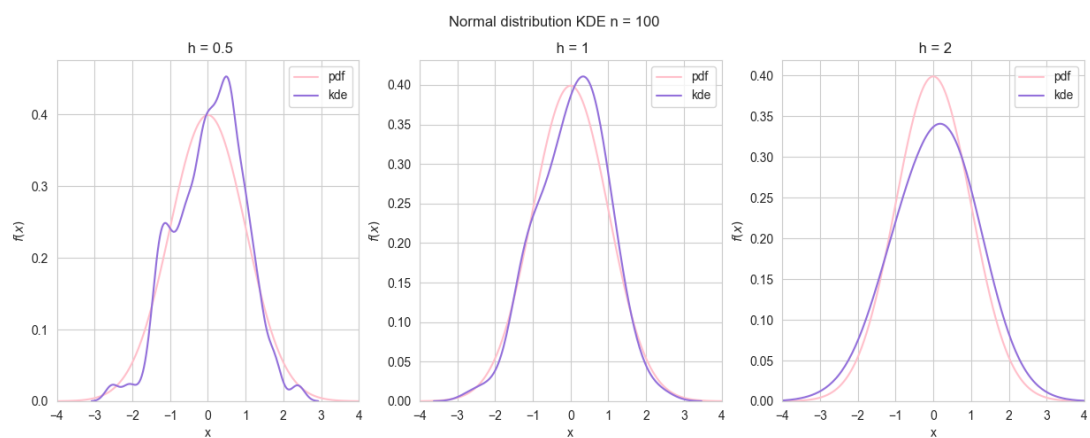


Рис. 15: Нормальное распределение, $n = 100$

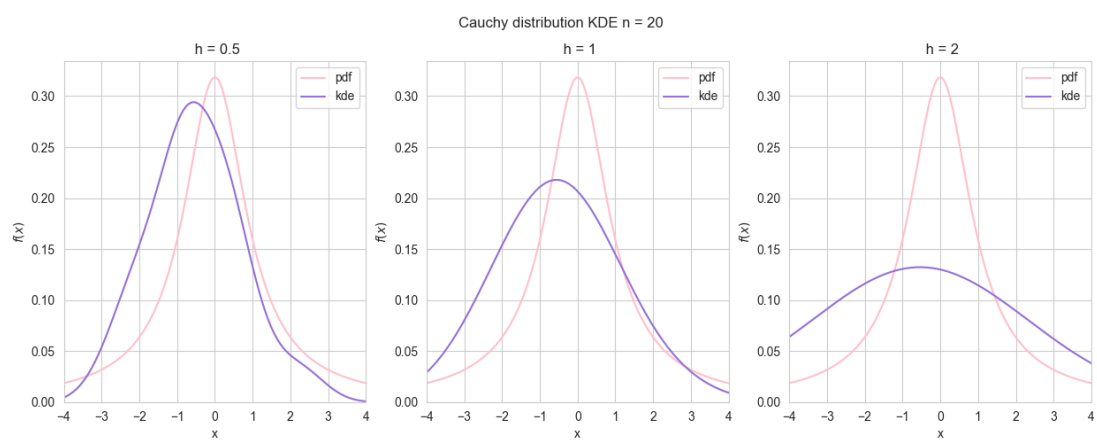


Рис. 16: Распределение Коши, $n = 20$

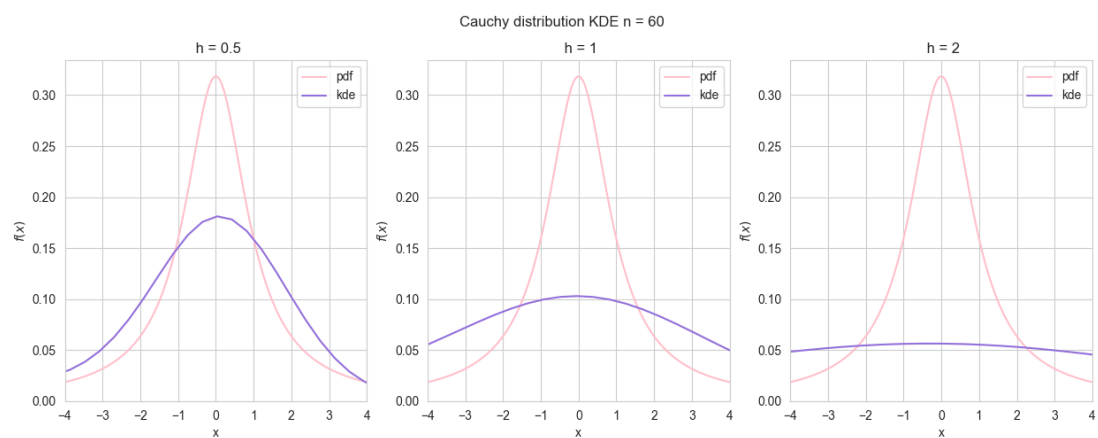


Рис. 17: Распределение Коши, $n = 60$

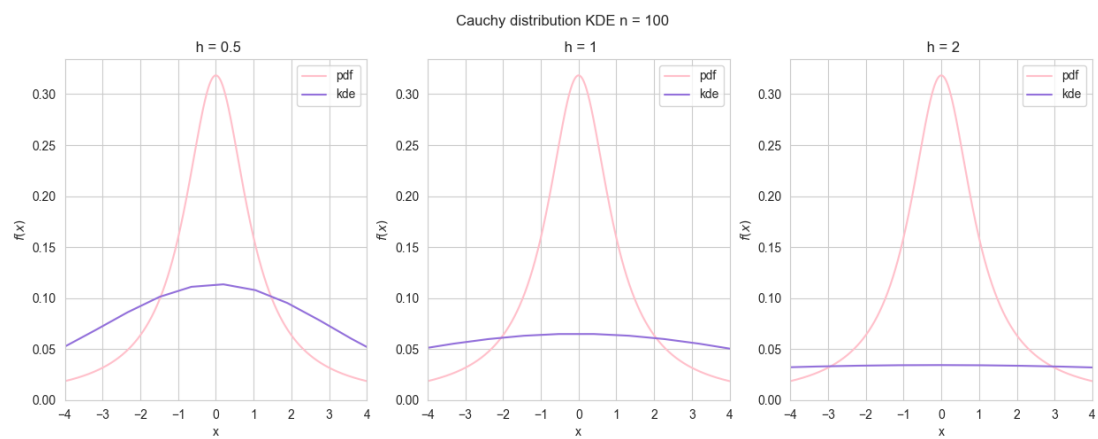


Рис. 18: Распределение Коши, $n = 100$

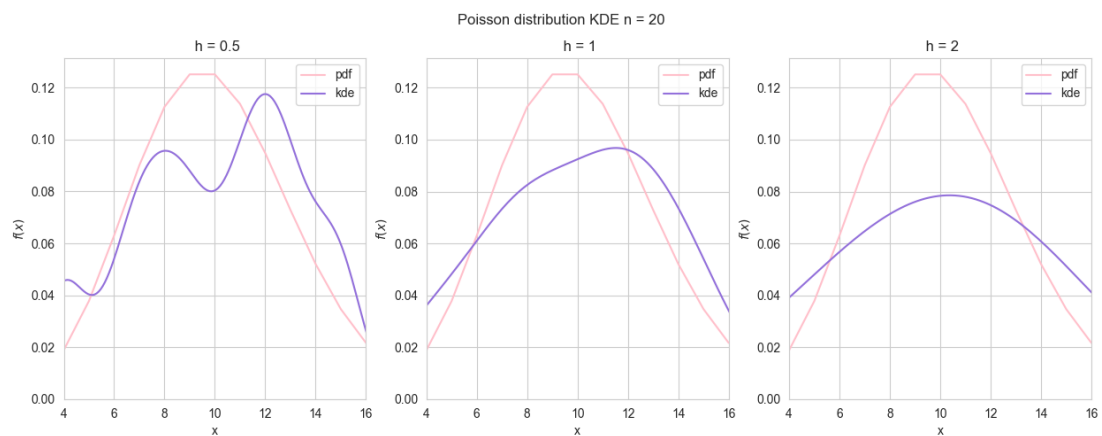


Рис. 19: Распределение Пуассона, $n = 20$

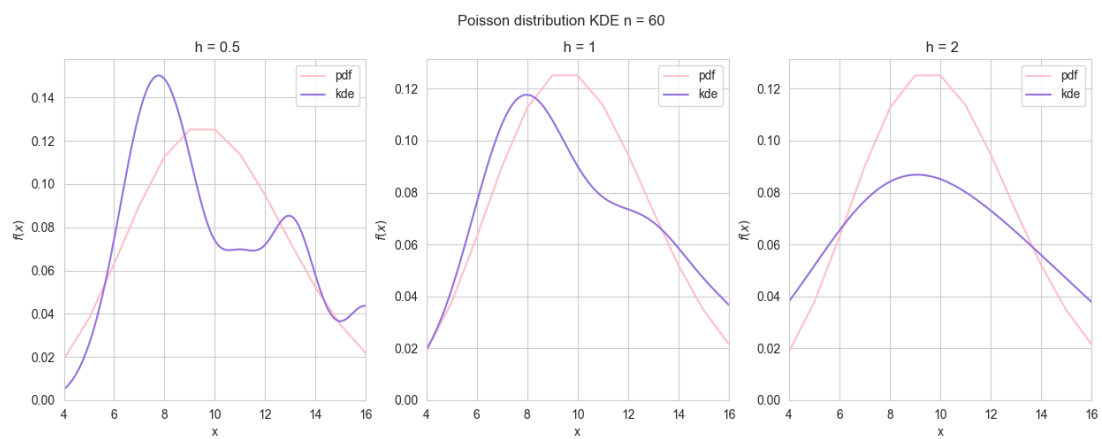


Рис. 20: Распределение Пуассона, $n = 60$

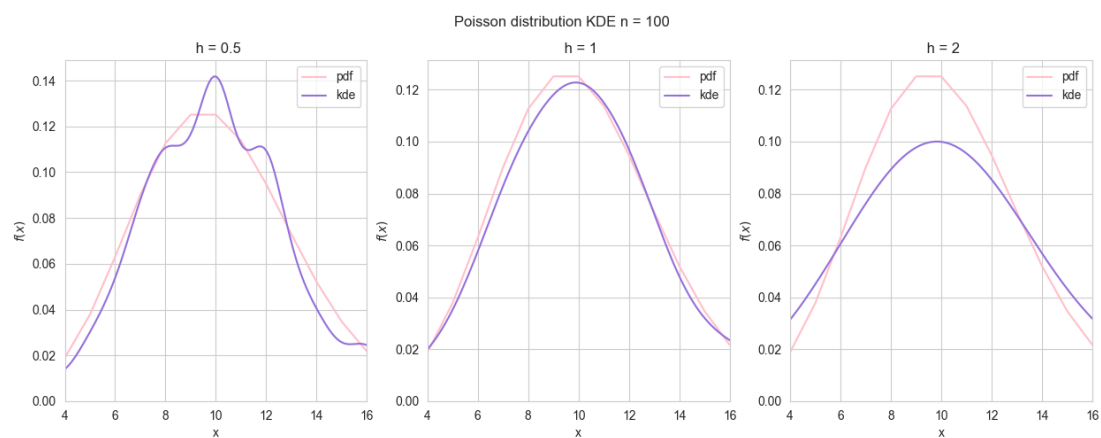


Рис. 21: Распределение Пуассона, $n = 100$

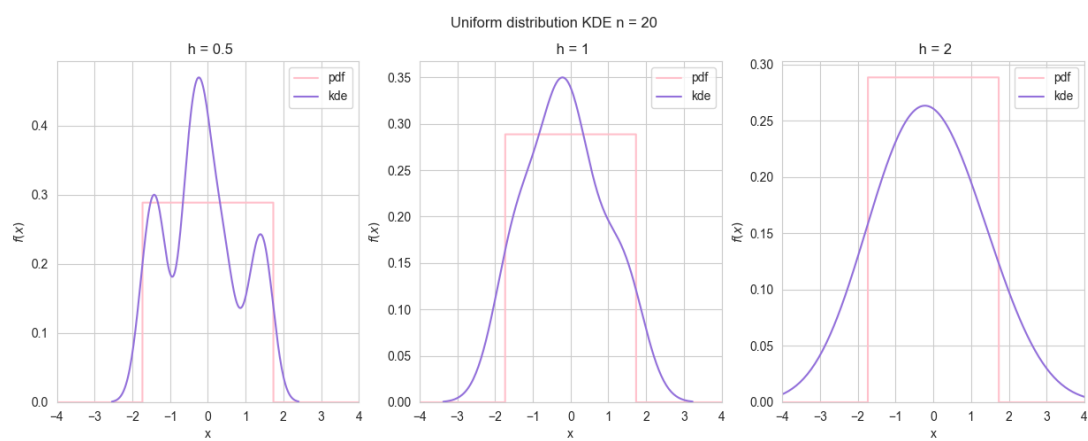


Рис. 22: Равномерное распределение, $n = 20$

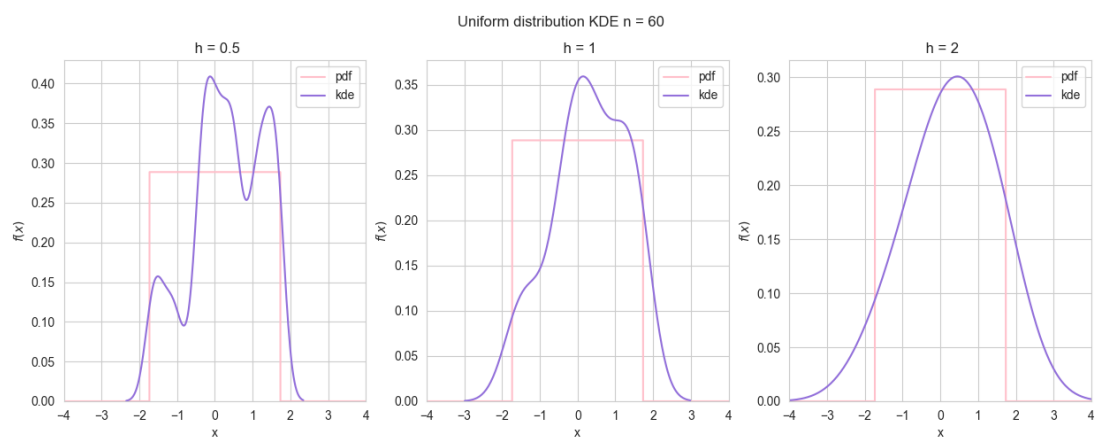


Рис. 23: Равномерное распределение, $n = 60$

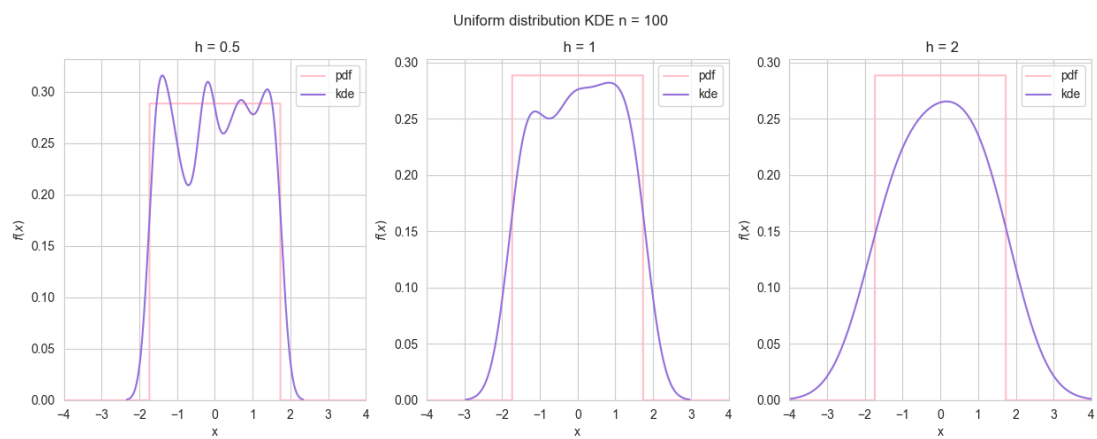


Рис. 24: Равномерное распределение, $n = 100$

4 Обсуждение

4.1 Гистограммы

Из результатов работы можно сделать вывод, что чем больше выборка для каждого из данных распределений, тем ближе ее гистограмма к графику плотности вероятности того закона, по которому распределены величины сгенерированной выборки. Чем меньше выборка, тем менее она показательна, т.е. тем хуже по ней определяется характер распределения величины.

Из дополнительных наблюдений можно выделить, что почти нигде не совпали максимумы гистограмм и плотностей распределения. Также наблюдаются всплески гистограмм, что наиболее хорошо прослеживается на распределении Коши.

4.2 Характеристики положения и рассеяния

Исходя из данных, приведенных в таблицах, можно выделить, что у распределения Коши значения дисперсии характеристик рассеяния слишком большие даже при увеличении размера выборки, что является следствием выбросов, наблюдаемых на графиках предыдущего задания.

4.3 Доля и теоретическая вероятность выбросов

Из данных в таблице, можно сказать, что чем больше выборка, тем ближе доля выбросов будет к теоретической оценке. Снова для распределения Коши доля выбросов значительно выше, чем для остальных распределений. Для равномерного распределения выбросов мы не получили, равномерное распределение же повторило теоретическую оценку

4.4 Эмпирическая функция распределения. Ядерные оценки плотности

Можем наблюдать на иллюстрациях с эмпирической функцией распределения, что ступенчатая эмпирическая функция распределения тем лучше приближает функцию распределения реальной выборки, чем мощнее эта выборка. Заметим так же, что для распределения Пуассона и равномерного распределения отклонение функций друг от друга наибольшее.

Рисунки, посвященные ядерным оценкам, иллюстрируют сближение ядерной оценки и функции плотности вероятности для всех h с ростом размера выборки. Для распределения Пуассона наиболее ярко видно, как сглаживает отклонения увеличение параметра сглаживания h .

В зависимости от особенностей распределений для их описания лучше подходят разные параметры h в ядерной оценке: для равномерного распределения и распределения Пуассона лучше подойдет параметр $h = 2h_n$, для нормального и Коши $h = h_n$. Такие значения дают вид ядерной оценки наиболее близкий к плотности, характерной данным распределениям.

Также можно увидеть, что чем больше коэффициент при параметре сглаживания, тем меньше изменений знака производной у аппроксимирующей функции, вплоть до того, что при $h = 2h_n$ функция становится унимодальной на рассматриваемом промежутке. Также видно, что при $h = h_n/2$ по полученным приближениям становится сложно сказать плотность вероятности какого распределения они должны повторять, так как они очень похожи между собой.

5 Литература

Язык программирования *Python* 3.10

Подключенные библиотеки для *Python*: *math*, *matplotlib*, *numpy*, *seaborn*, *scipy*

Ссылка на Github: <https://github.com/OttoMarina/MathStat/tree/main/Lab1-4>