

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ
ПЕТРА ВЕЛИКОГО

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ И МЕХАНИКИ
ВЫСШАЯ ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ФИЗИКИ

Математическая статистика
Отчёт по лабораторным работам №1-2

Выполнил:

Студент: Сачук Александр

Группа: 5030102/90201

Принял:

к. ф.-м. н., доцент

Баженов Андрей Николаевич

2022 г.

Содержание

1. Постановка задачи	5
2. Теория	6
2.1. Рассматриваемые распределения	6
2.2. Гистограммы	6
2.2.1. Определение и описание	6
2.2.2. Построение гистограммы	6
2.3. Вариационный ряд	7
2.4. Выборочные числовые характеристики	7
2.4.1. Характеристики положения	7
2.4.2. Характеристики рассеяния	7
2.5. Боксплот Тьюки	8
2.5.1. Построение	8
2.6. Теоретическая вероятность выбросов	8
2.7. Эмпирическая функция распределения	8
2.7.1. Статистический ряд	8
2.7.2. Эмпирическая функция распределения	8
2.7.3. Нахождение эмпирической функции распределения	9
2.8. Оценки плотности вероятности	9
2.8.1. Определение	9
2.8.2. Ядерные оценки	9
3. Результаты	11
3.1. Гистограммы	11
3.2. Характеристики положения и рассеяния	12
3.3. Боксплот Тьюки	14
3.4. Доля выбросов	16
3.5. Теоретическая вероятность выбросов	17
3.6. Эмпирическая функция распределения	17
3.7. Ядерные оценки плотности распределения	18
4. Реализация	23
5. Обсуждение	24
5.1. Гистограммы	24
5.2. Характеристики положения и рассеяния	24
5.3. Доля и теоретическая вероятность выбросов	24
5.4. Эмпирическая функция и ядерные оценки плотности распределения	25
6. Ссылки на библиотеки	26

7. Ссылки на репозиторий	27
------------------------------------	----

Список иллюстраций

1. Гистограмма и плотность вероятности для нормального распределения [N = 10, 100, 1000]	11
2. Гистограмма и плотность вероятности для распределения Коши [N = 10, 100, 1000]	11
3. Гистограмма и плотность вероятности для распределения Пуассона [N = 10, 100, 1000]	11
4. Гистограмма и плотность вероятности для равномерного распределения [N = 10, 100, 1000]	12
5. Нормальное распределение	14
6. Распределение Коши	15
7. Распределение Пуассона	15
8. Равномерное распределение	16
9. Нормальное распределение	17
10. Распределение Коши	17
11. Распределение Пуассона	18
12. Равномерное распределение	18
13. Нормальное распределение, $n = 20$	18
14. Нормальное распределение, $n = 60$	19
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, $n = 20$	20
20. Распределение Пуассона, $n = 60$	21
21. Распределение Пуассона, $n = 100$	21
22. Равномерное распределение, $n = 20$	21
23. Равномерное распределение, $n = 60$	22
24. Равномерное распределение, $n = 100$	22

Список таблиц

1. Таблица распределения	9
2. Нормальное распределение	12
3. Распределение Коши	13
4. Распределение Пуассона	13
5. Равномерное распределение	14
6. Доля выбросов	16
7. Теоретическая вероятность выбросов	17

1. Постановка задачи

Для четырех распределений:

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

Выполнить следующие задачи:

1. Сгенерировать выборки размером 10, 50 и 1000 элементов. Построить на одном рисунке гистограмму и график плотности распределения.
2. Сгенерировать выборки размером 10, 100 и 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных: \bar{x} , $medx$, z_R , z_Q , z_{tr} . Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

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

Вычислить оценку дисперсии 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

Плотности распределений:

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Пуассона

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

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}}, & x \in [-\sqrt{3}, \sqrt{3}] \\ 0, & x \notin [-\sqrt{3}, \sqrt{3}] \end{cases} \quad (6)$$

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

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

- 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 [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 :

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

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 [4, с. 38]

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

Правило Сильвермана [4, с. 44]

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

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

3. Результаты

3.1. Гистограммы

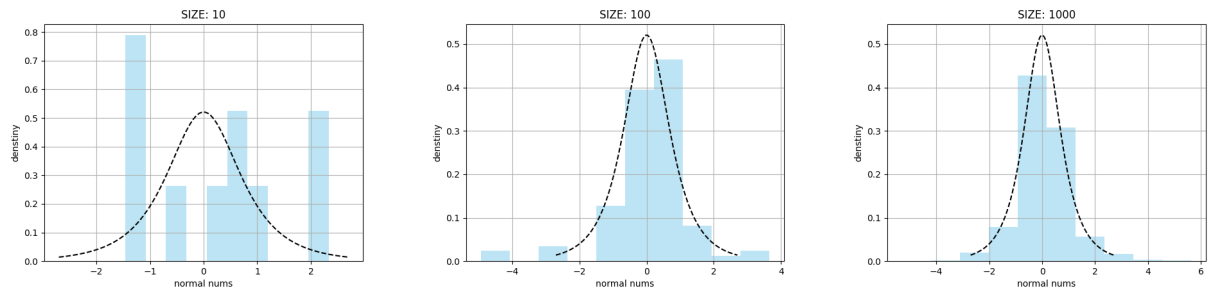


Рис. 1. Гистограмма и плотность вероятности для нормального распределения [$N = 10, 100, 1000$]

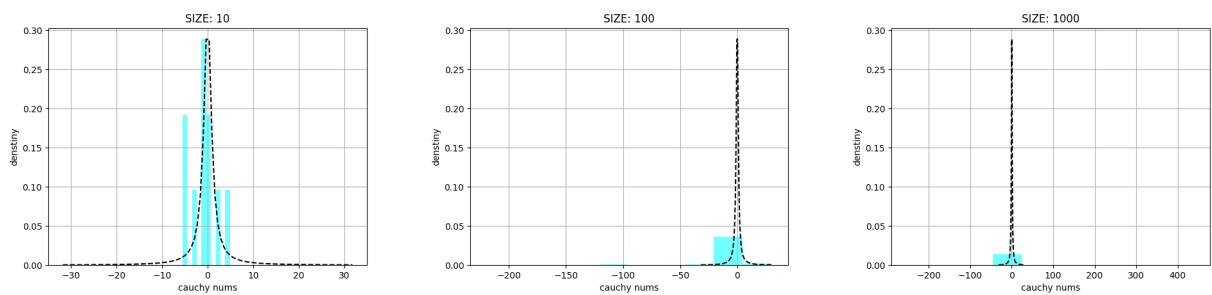


Рис. 2. Гистограмма и плотность вероятности для распределения Коши [$N = 10, 100, 1000$]

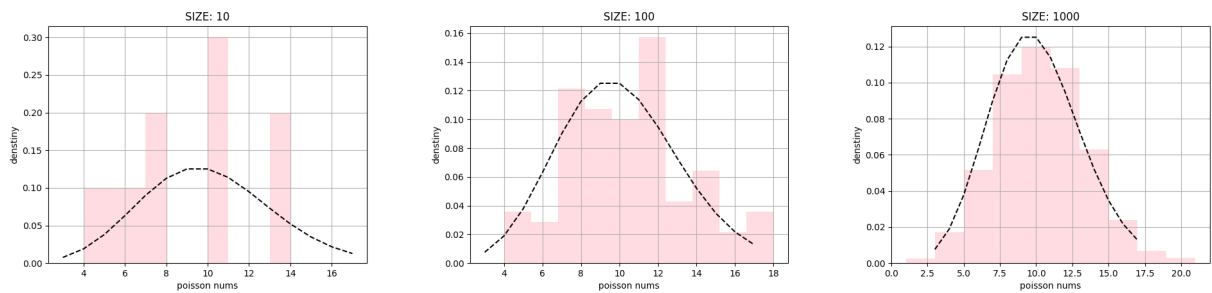


Рис. 3. Гистограмма и плотность вероятности для распределения Пуассона [$N = 10, 100, 1000$]

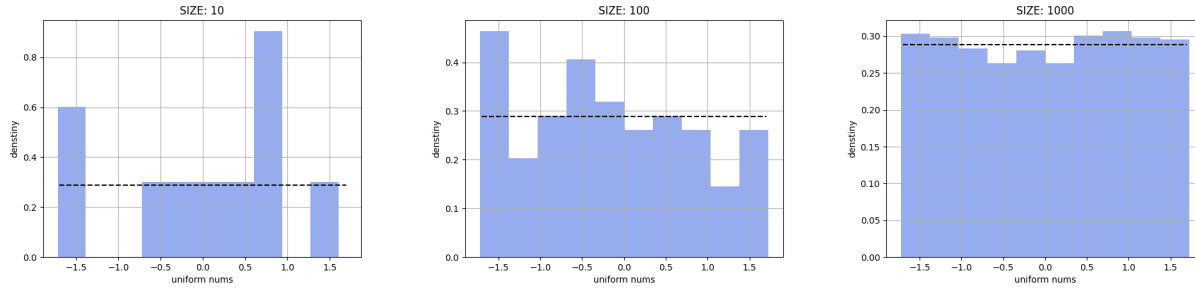


Рис. 4. Гистограмма и плотность вероятности для равномерного распределения $[N = 10, 100, 1000]$

3.2. 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 $k = 1$.

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Normal E(z) 10	0.011104	0.008388	0.007869	0.332015	0.291675
Normal D(z) 10	0.095662	0.129384	0.181631	0.126743	0.113728
$E(z) \pm \sqrt{D(z)}$	[-0.298189; 0.320397]	[-0.351312; 0.368088]	[-0.418313; 0.434051]	[-0.023995; 0.688025]	[-0.045561; 0.628911]
$\hat{E}(z)$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.3^{+0.3}_{-0.3}$	$0.3^{+0.3}_{-0.3}$
Normal E(z) 100	-0.002418	-0.007771	-0.010427	0.012414	0.021683
Normal D(z) 100	0.011254	0.017346	0.094536	0.013636	0.013397
$E(z) \pm \sqrt{D(z)}$	[-0.108503; 0.103667]	[-0.139475; 0.123933]	[-0.317894; 0.29704]	[-0.104359; 0.129187]	[-0.094062; 0.137428]
$\hat{E}(z)$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.3}_{-0.3}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$
Normal E(z) 1000	0.000378	-0.00027	-0.002986	0.001661	0.002754
Normal D(z) 1000	0.000998	0.001538	0.062421	0.001223	0.001163
$E(z) \pm \sqrt{D(z)}$	[-0.031213; 0.031969]	[-0.039487; 0.038947]	[-0.252828; 0.246856]	[-0.03331; 0.036632]	[-0.031349; 0.036857]
$\hat{E}(z)$	0.0	0.0	$0.0^{+0.2}_{-0.2}$	0.0	0.0

Таблица 2. Нормальное распределение

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Cauchy E(z) 10	9.529885	0.030773	47.393832	1.3084	0.782301
Cauchy D(z) 10	116013.195	0.305832	2898167.173	10.7123	1.88495
$E(z) \pm \sqrt{D(z)}$	[-331.077; 350.137]	[-0.522248; 0.583794]	[-1655.007; 1749.794]	[-1.96457; 4.581369]	[-0.590634; 2.155236]
$\widehat{E}(z)$	-	0.0	-	-	-
Cauchy E(z) 100	-0.502032	0.007131	-23.286978	0.043542	0.050521
Cauchy D(z) 100	1018.714	0.023291	2513115.181	0.051077	0.025707
$E(z) \pm \sqrt{D(z)}$	[-32.41934; 31.41528]	[-0.145483; 0.159745]	[-1608.568; 1561.993]	[-0.18246; 0.269544]	[-0.109813; 0.210855]
$\widehat{E}(z)$	-	0.0	-	-	-
Cauchy E(z) 1000	-1.856742	0.001006	-942.850162	0.005274	0.005073
Cauchy D(z) 1000	13277.33	0.002457	3317087798.66	0.004868	0.002564
$E(z) \pm \sqrt{D(z)}$	[-117.084; 113.371]	[-0.048562; 0.050574]	[-58537.015; 56651.314]	[-0.064497; 0.075045]	[-0.045563; 0.055709]
$\widehat{E}(z)$	-	0.0	-	-	-

Таблица 3. Распределение Коши

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Poisson E(z) 10	9.9783	9.8555	10.249	10.913	10.765167
Poisson D(z) 10	1.049619	1.44237	1.823999	1.480431	1.31727
$E(z) \pm \sqrt{D(z)}$	[8.953791; 11.002809]	[8.654513; 11.056487]	[8.898445; 11.599555]	[9.69627; 12.12973]	[9.617443; 11.912891]
$\widehat{E}(z)$	10_{-1}^{+1}	10_{-1}^{+1}	10_{-2}^{+2}	11_{-1}^{+1}	11_{-1}^{+1}
Poisson E(z) 100	10.01941	9.848	11.0115	9.9795	9.9588
Poisson D(z) 100	0.1066	0.224396	1.015618	0.15533	0.132514
$E(z) \pm \sqrt{D(z)}$	[9.692913; 10.345907]	[9.374295; 10.321705]	[10.003721; 12.019279]	[9.585381; 10.373619]	[9.594775; 10.322825]
$\widehat{E}(z)$	$10_{-0.3}^{+0.3}$	$10_{-0.3}^{+0.3}$	11_{-1}^{+1}	10	10
Poisson E(z) 1000	10.001486	9.999	11.692	9.996	9.868494
Poisson D(z) 1000	0.009933	0.000999	0.667636	0.002484	0.010702
$E(z) \pm \sqrt{D(z)}$	[9.901822; 10.10115]	[9.967393; 10.030607]	[10.87491; 12.50909]	[9.94616; 10.04584]	[9.765044; 9.971944]
$\widehat{E}(z)$	$10_{-0.1}^{+0.1}$	10	$11.7_{-0.8}^{+0.8}$	10	10

Таблица 4. Распределение Пуассона

Characteristic	Mean	Median	z_R	z_Q	z_{tr}
Uniform $E(z)$ 10	0.022043	0.016613	0.015673	0.340336	0.33759
Uniform $D(z)$ 10	0.099705	0.231529	0.043987	0.122888	0.151837
$E(z) \pm \sqrt{D(z)}$	[-0.293718; 0.337804]	[-0.464562; 0.497788]	[-0.194058; 0.225404]	[-0.010218; 0.69089]	[-0.052073; 0.727253]
$\hat{E}(z)$	0.0	0.0	0.0	0.3	0.3
Uniform $E(z)$ 100	0.005039	0.0059	0.001225	0.022797	0.039371
Uniform $D(z)$ 100	0.010492	0.030199	0.000583	0.015156	0.020264
$E(z) \pm \sqrt{D(z)}$	[-0.097391; 0.107469]	[-0.167879; 0.179679]	[-0.02292; 0.02537]	[-0.100313; 0.145907]	[-0.102981; 0.181723]
$\hat{E}(z)$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.2}_{-0.2}$	0.0	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$
Uniform $E(z)$ 1000	0.00031	0.001299	-0.000103	0.002594	0.004221
Uniform $D(z)$ 1000	0.00103	0.002934	0.000005	0.001581	0.001991
$E(z) \pm \sqrt{D(z)}$	[-0.031784; 0.032404]	[-0.052867; 0.055465]	[-0.002339; 0.002133]	[-0.037168; 0.042356]	[-0.0404; 0.048842]
$\hat{E}(z)$	0.0	0.0	0.0	0.0	0.0

Таблица 5. Равномерное распределение

3.3. Боксплот Тьюки

- Нормальное распределение

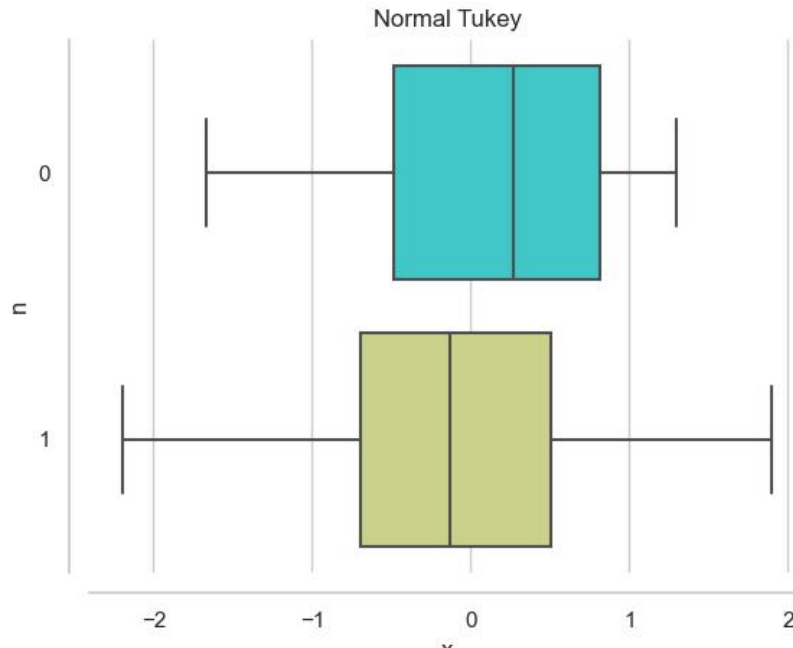


Рис. 5. Нормальное распределение

- Распределение Коши

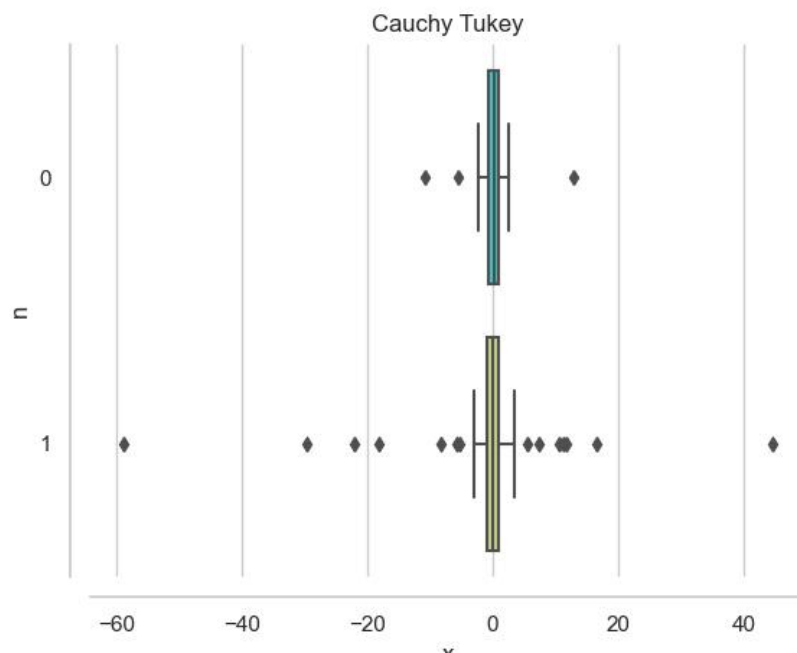


Рис. 6. Распределение Коши

- Распределение Пуассона

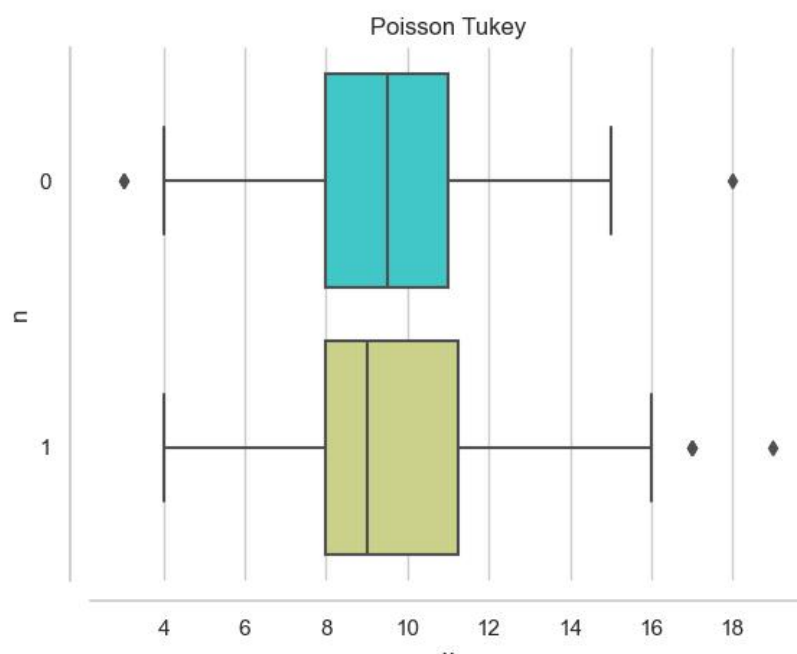


Рис. 7. Распределение Пуассона

- Равномерное распределение

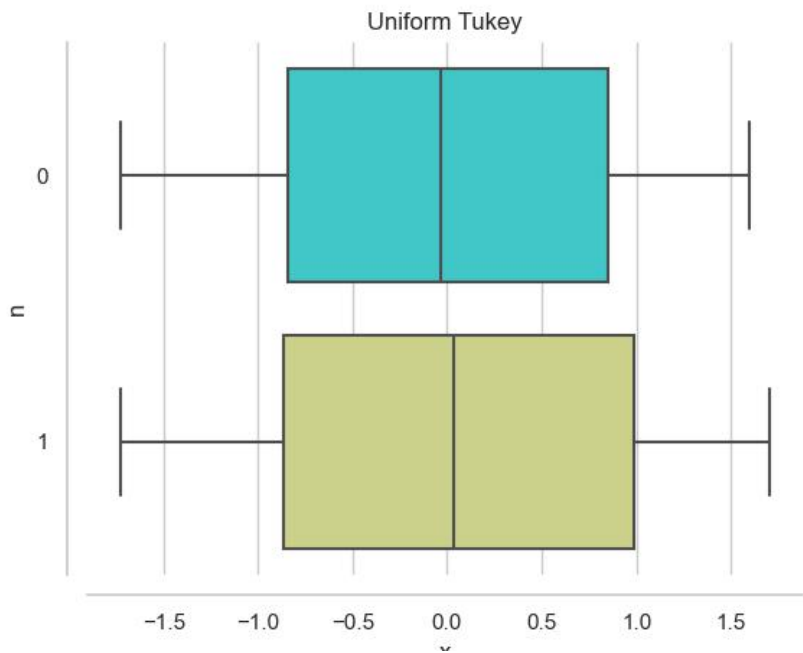


Рис. 8. Равномерное распределения

3.4. 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	0.02
Cauchy n = 20	0.15
Cauchy n = 100	0.18
Poisson n = 20	0.02
Poisson n = 100	0.01
Uniform n = 20	0
Uniform n = 100	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

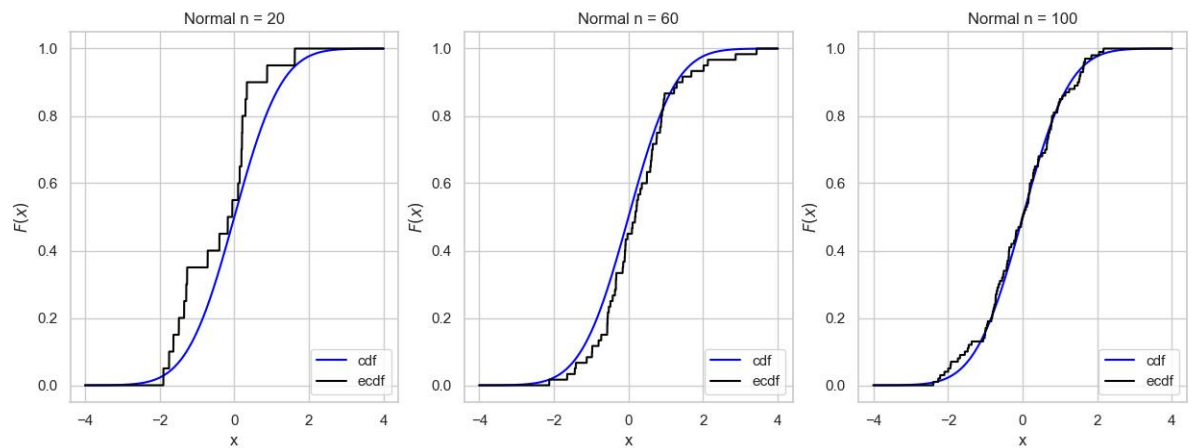


Рис. 9. Нормальное распределение

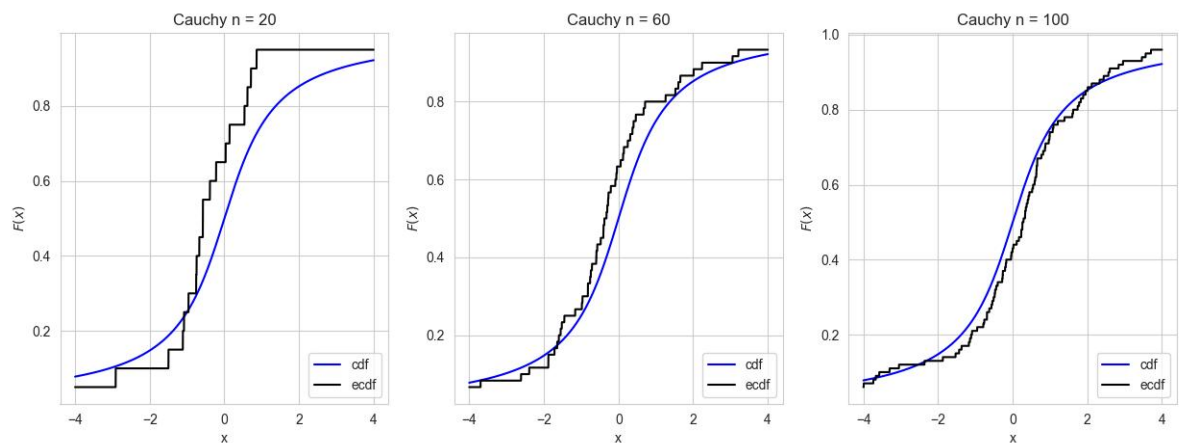


Рис. 10. Распределение Коши

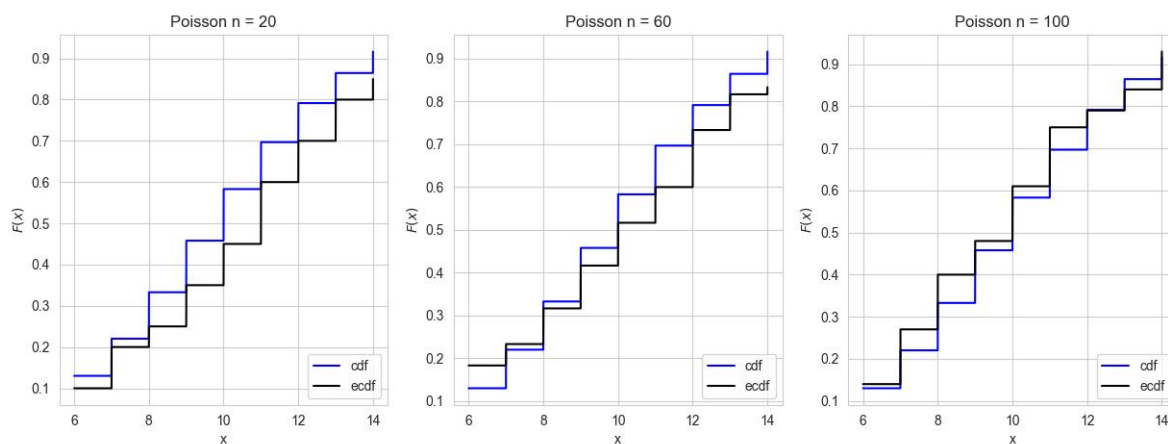


Рис. 11. Распределение Пуассона

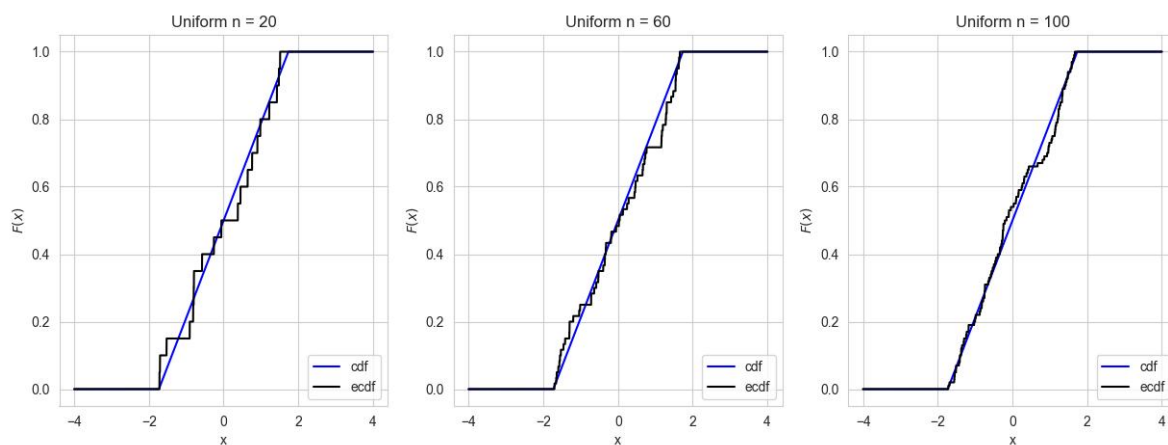
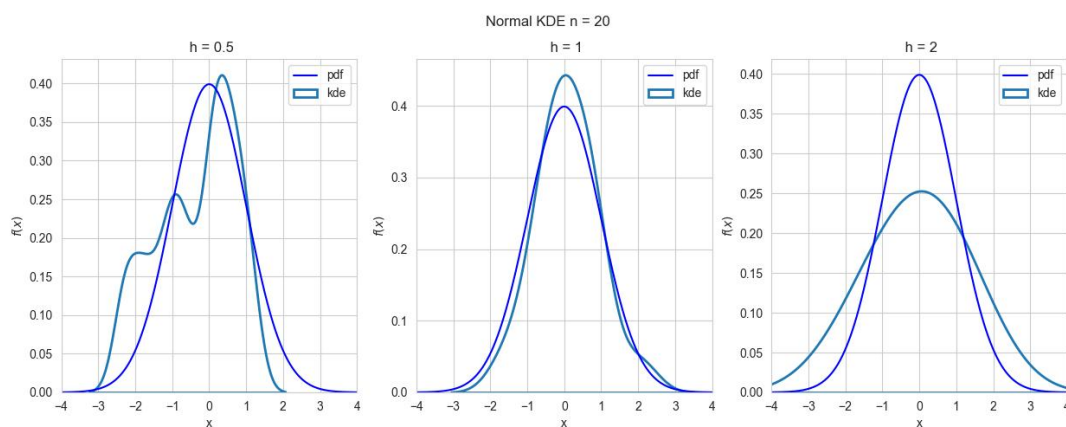
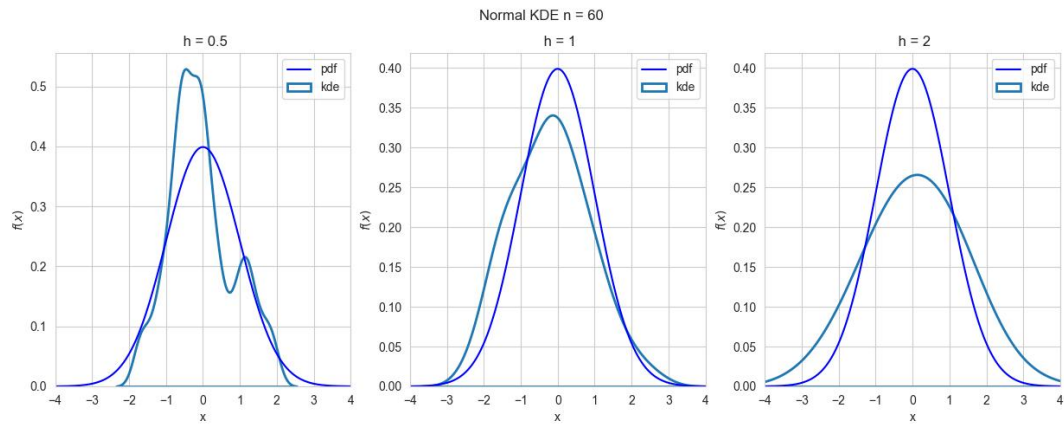
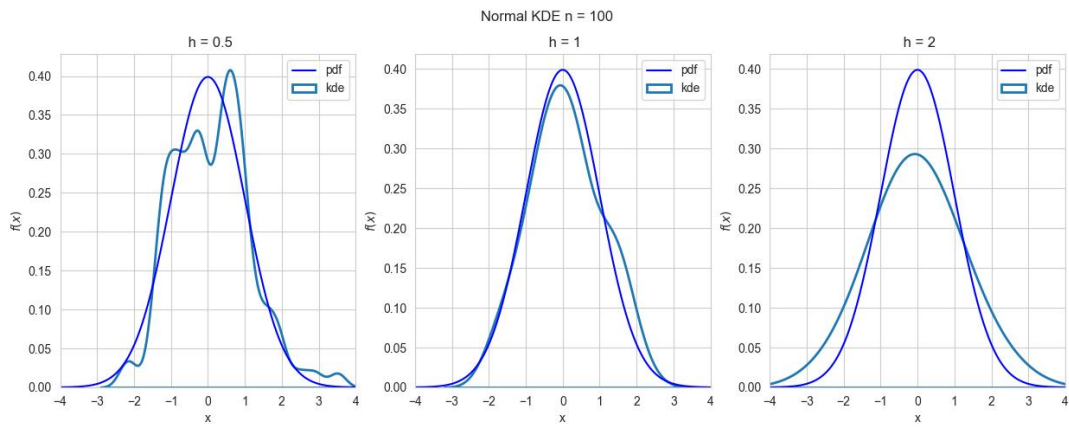
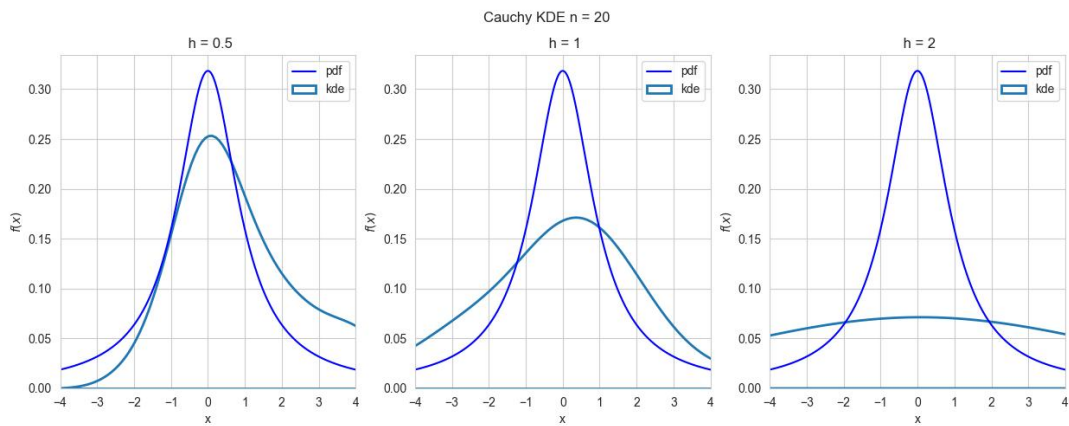
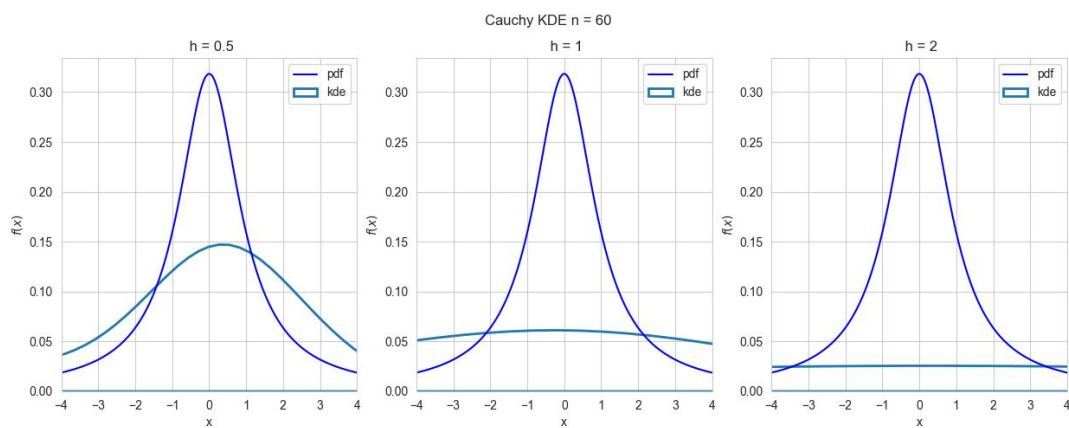
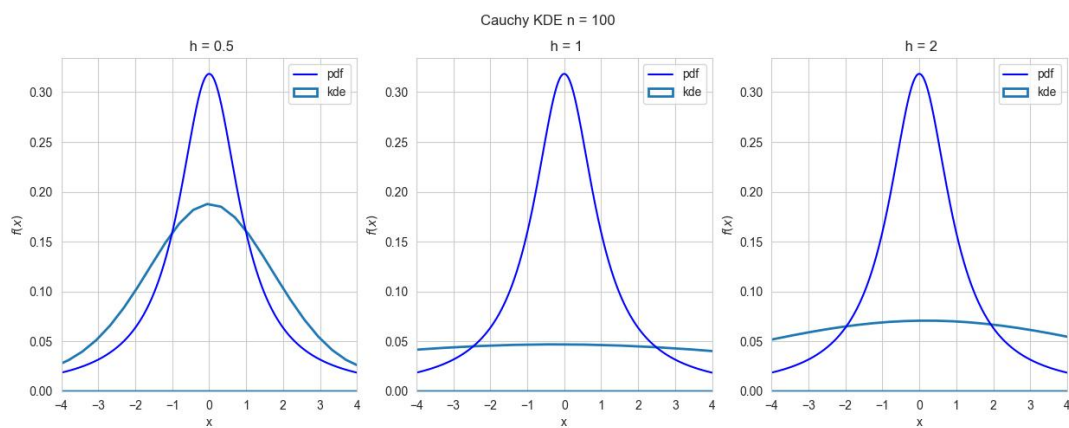
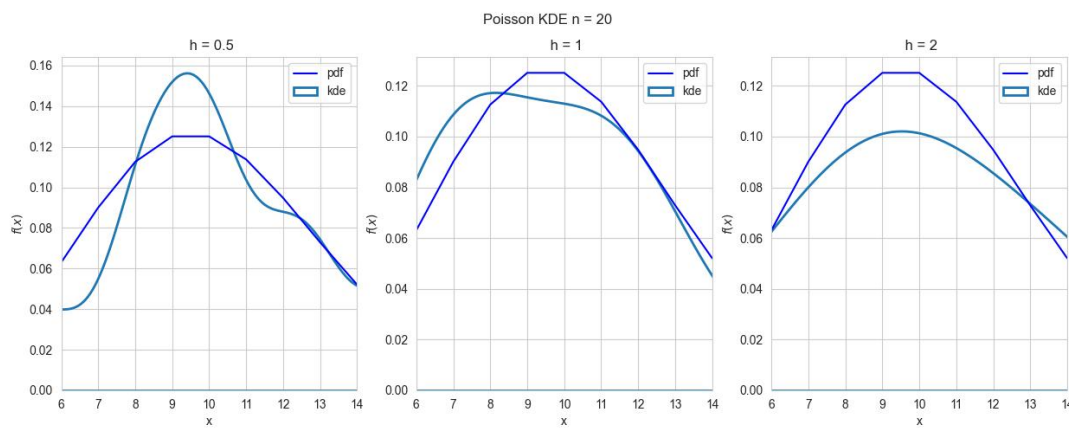


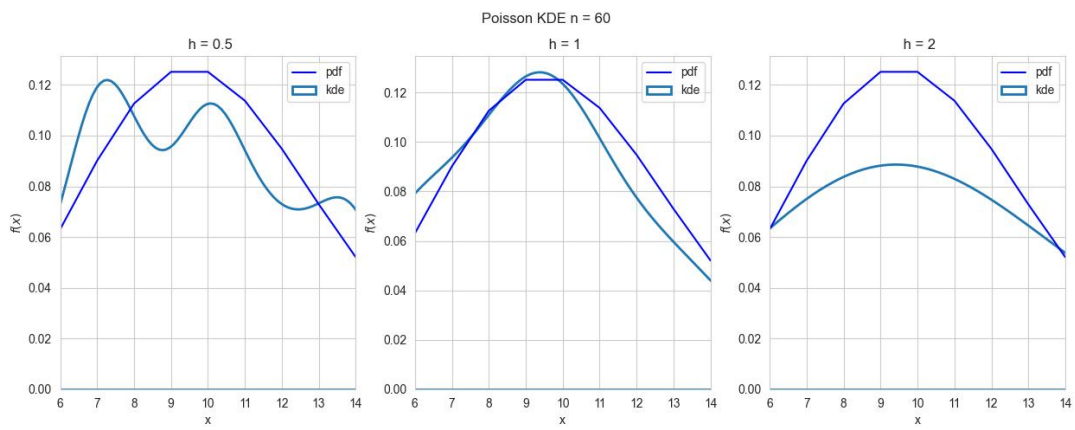
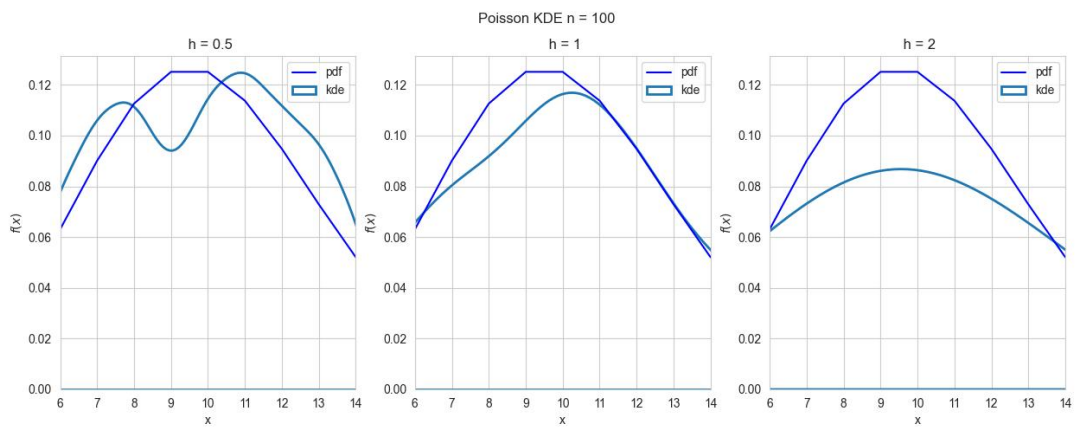
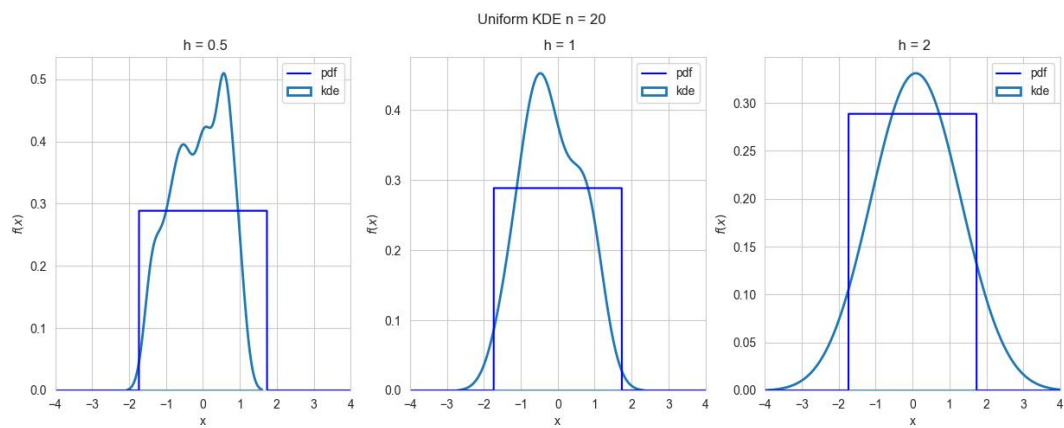
Рис. 12. Равномерное распределение

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, $n = 20$

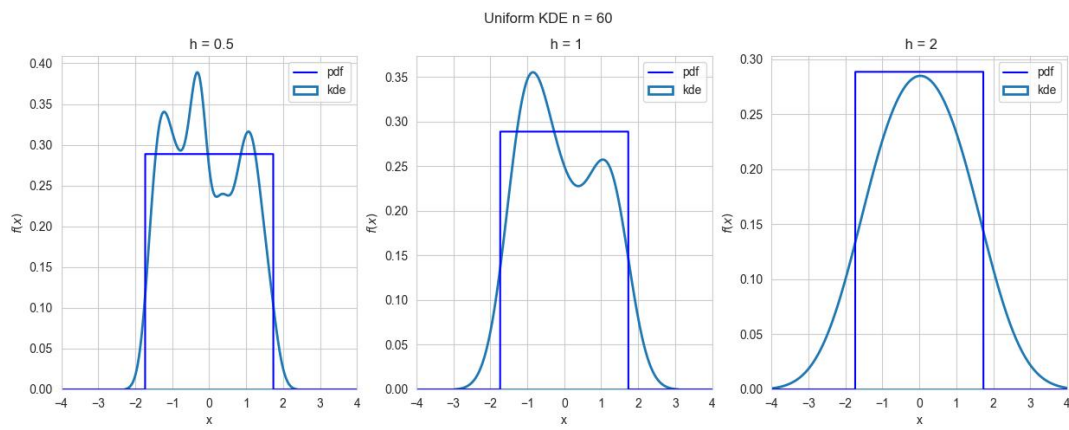


Рис. 23. Равномерное распределение, $n = 60$

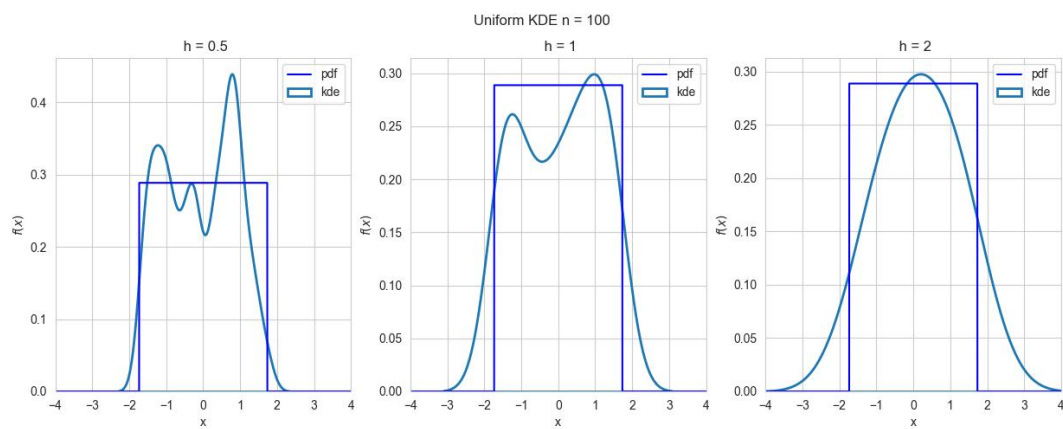


Рис. 24. Равномерное распределение, $n = 100$

4. Реализация

Данная лабораторная работа была выполнена с использованием языка программирования Python 3.10 в среде разработки Visual Studio Code с использованием следующих библиотек:

- `scipy` версии 1.8.0
- `numpy` версии 1.22.0
- `matplotlib` версии 3.5.1
- `seaborn` версии 0.11.2

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 обратное: 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 и ядерные оценки плотности распределения

Можем наблюдать на иллюстрациях с э. ф. р., что ступенчатая эмпирическая функция распределения тем лучше приближает функцию распределения реальной выборки, чем мощнее эта выборка. Заметим так же, что для распределения Пуассона и 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.

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/AS2/Mathematical-statistics> - GitHub репозиторий