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:
Фисюк Алексей Юрьевич

группа: 5030102/90201

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

Санкт-Петербург
2022

Оглавление

## 1. Постановка задачи

Для 5 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 $U(x, -\sqrt{3}, \sqrt{3})$

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: $x$, $med\ x$, $z\bar{}\ R$, $zQ$, $ztr$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:
   $E(z) = \overline{z}$ (1)
   Вычислить оценку дисперсии по формуле:
   $D(z) = \overline{z^2} - \bar{z}^2$ (2)
   Представить полученные данные в виде таблиц.

3. Сгенерировать выборки размером 20 и 100 элементов.
   Построить для них боксплот Тьюки. Для каждого распределения определить долю выбросов экспериментально (сгенерировав выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.

4. Сгенерировать выборки размером 20, 60 и 100 элементов.
   Построить на них эмпирические функции распределения и ядерные оценки плотности распределения на отрезке $[-4; 4]$ для непрерывных распределений и на отрезке $[6; 14]$ для распределения Пуассона.

# 2. Теория

## 2.1 Рассматриваемые распределения

Плотности:

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Лапласа

$$L(x, 0, \frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \quad (5)$$

- Распределение Пуассона

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \quad (6)$$

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при } |x| \leq \sqrt{3} \\ 0 & \text{при } |x| > \sqrt{3} \end{cases} \quad (7)$$

## 2.2 Гистограмма

Гистограмма в математической статистике — это функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него

Построение:

1) Множество значений, которые может принимать элемент выборки, разбивается на несколько, чаще всего равных, интервалов
2) Эти интервалы откладываются на горизонтальной оси
3) Над каждым из интервалов 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 Выборочные числовые характеристики

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

### 2.3.1 Характеристики положения

- Выборочное среднее

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

- Выборочная медиана

$$med\ x = \begin{cases} x_{l+1}, & \text{при n } = 2\text{l } + 1 \\ \frac{x_l + x_{l+1}}{2}, & \text{при n } = 2\text{l} \end{cases} \tag{9}$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_1 + x_n}{2} \tag{10}$$

- Полусумма квартилей

  Выборочная квартиль $z_p$ порядка $p$ определяется формулой

$$zp = \begin{cases} x_{[np]+1}, \text{при np дробном} \\ x_{[np]}, \text{при np целом} \end{cases} \tag{11}$$

- Полусумма квартилей

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

- Усечённое среднее

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

### 2.3.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\, n \tag{14}$$

## 2.4 Боксплот Тьюки

### 2.4.1 Построение

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

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

где $X_1$ — нижняя граница уса, $X_2$ — верхняя граница уса, $Q_1$ — первый квартиль, $Q_3$ — третий квартиль.

Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков

## 2.5 Теоретическая вероятность выбросов

Можно вычислить теоретические первый и третий квартили распределений – $Q_1$ и $Q_3$. По формуле (15) – теоретические нижнюю и верхнюю границы уса – $X_1$ и $X_2$.

Выбросы – величины $x$:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix} \quad (16)$$

Теоретическая вероятность выбросов:

- для непрерывных распределений
$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T))$$
- для дискретных распределений
$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T))$$

Выше $F(X) = P(x \le X)$ - функция распределения.

## 2.6 Эмпирическая функция распределения

Эмпирическая (выборочная) функция распределения (э. ф. р.) – относительная частота события $X < x$, полученная по данной выборке:

$$F^*(x) = P^*(X < x) \quad (19)$$

Для получения относительной частоты $P^*(X < x)$ просуммируем в статистическом ряде, построенном по данной выборке, все частоты $n_i$ , для которых элементы $z_i$ статистического ряда меньше $x$.

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

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

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

| $X^*$ | $z_1$ | $z_2$ | … | $z_k$ |
|---|---|---|---|---|
| $P$ | $\frac{n_1}{n}$ | $\frac{n_2}{n}$ | … | $\frac{n_k}{n}$ |

Таблица 1: Таблица распределения

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

$$F_n^*(x) \approx F_X(x) \quad (21)$$

## 2.8 Оценки плотности вероятности

Оценкой плотности вероятности $f(x)$ называется функция $\hat{f}(x)$, построенная на основе выборки, приближённо равная $f(x)$

$$\hat{f}(\mathrm{x}) \approx f(x) \tag{22}$$

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

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

Здесь функция $K(u)$ - ядро (ядерная функция), непрерывна и является плотностью вероятности, $x_1, \dots, x_n$ — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, обладающая свойствами:

$$h_n \underset{n\to\infty}{\longrightarrow} 0; \quad nh_n \underset{n\to\infty}{\longrightarrow} \infty \tag{24}$$

Такие оценки называются непрерывными ядерными

Гауссово (нормальное) ядро

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

Правило Сильвермана

$$h_n = 1.06\hat{\sigma}\, n^{-1/5}, \tag{26}$$

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

# 3. Реализация

Лабораторная работа выполнена на языке Python версии 3.9 в IDE Pycharm. Подключаемые библиотеки:

- numpy
- matplotlib
- scipy
- seaborn

Исходный код:

https://github.com/ayu-f/MathStat

# 4. Результаты

## 4.1 Гистограмма и график плотности распределения

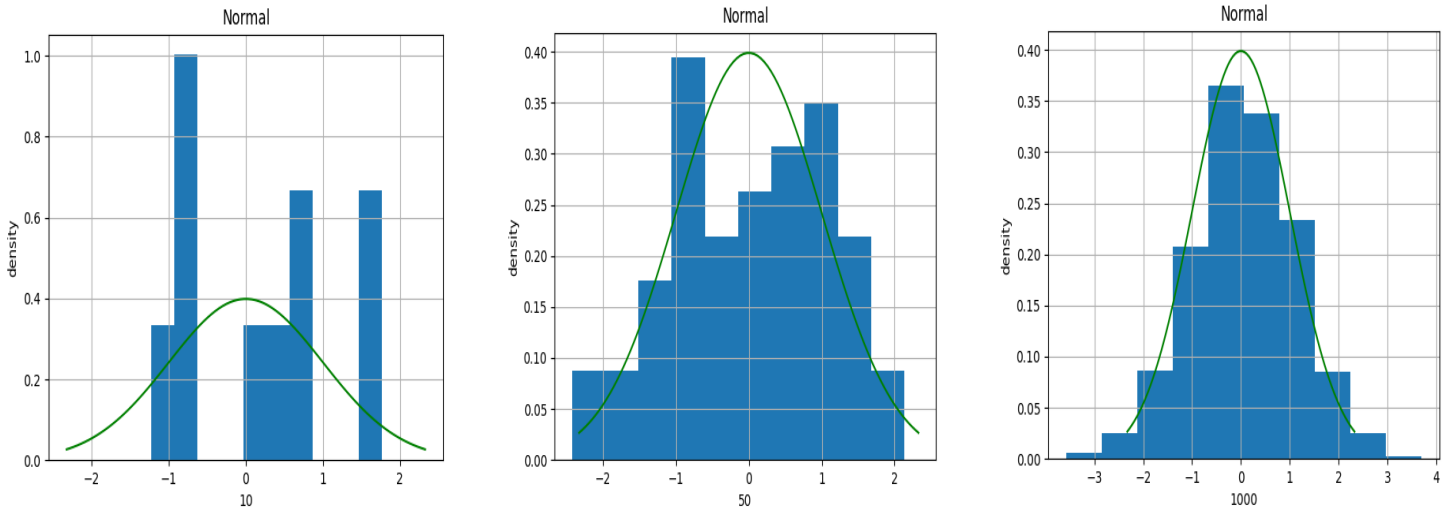


Рис. 1: Нормальное распределение

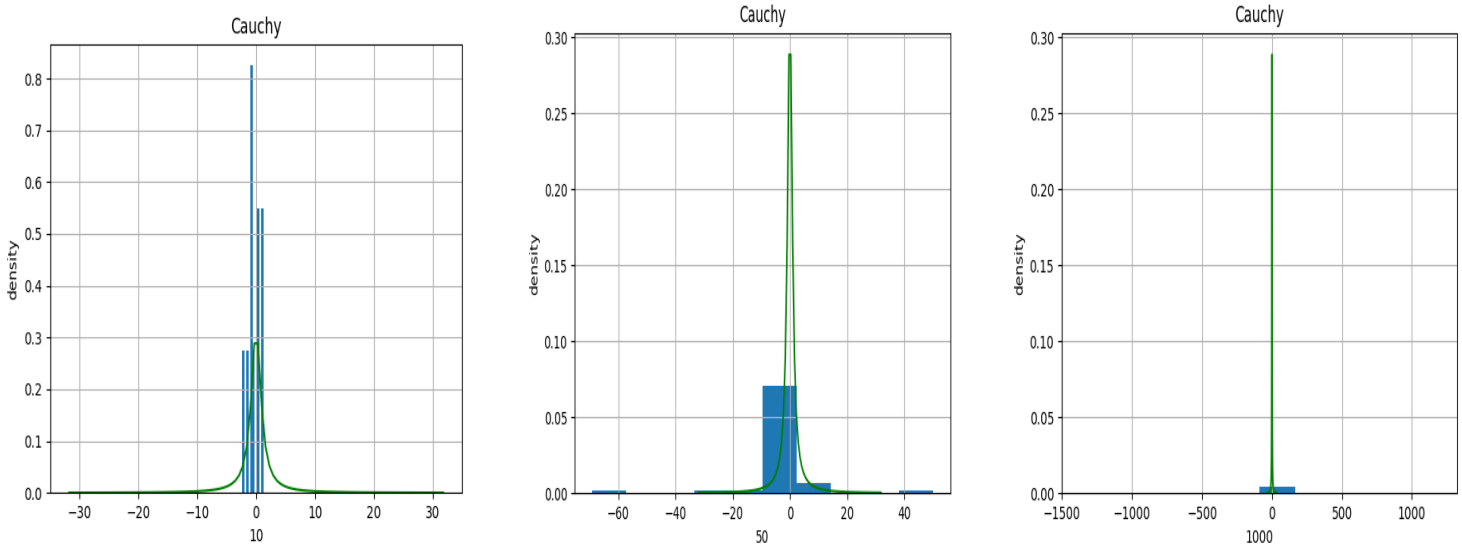


Рис. 2: Распределение Коши

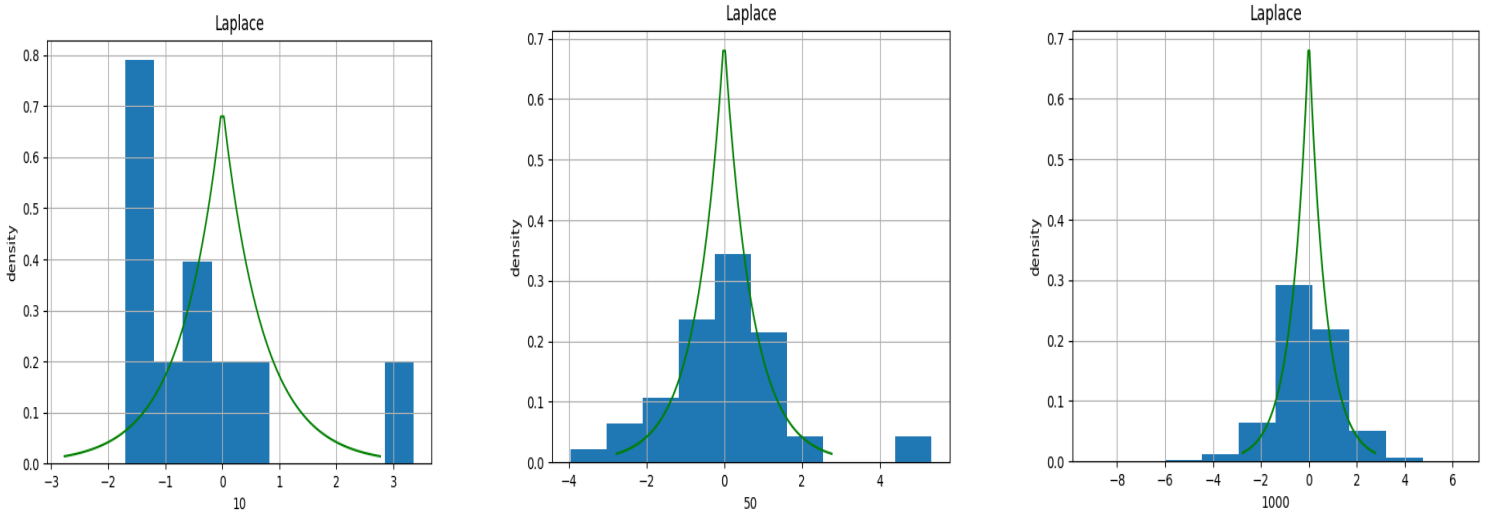


Рис. 3: Распределение Лапласа

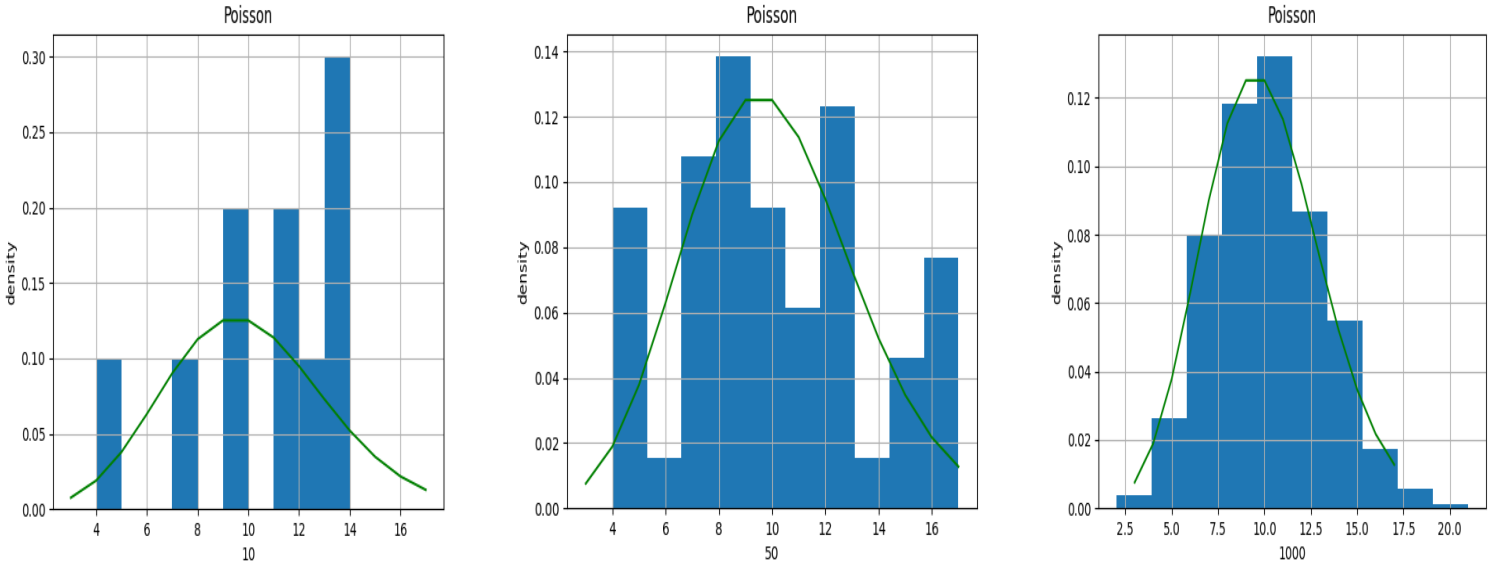


Рис. 4: Распределение Пуассона

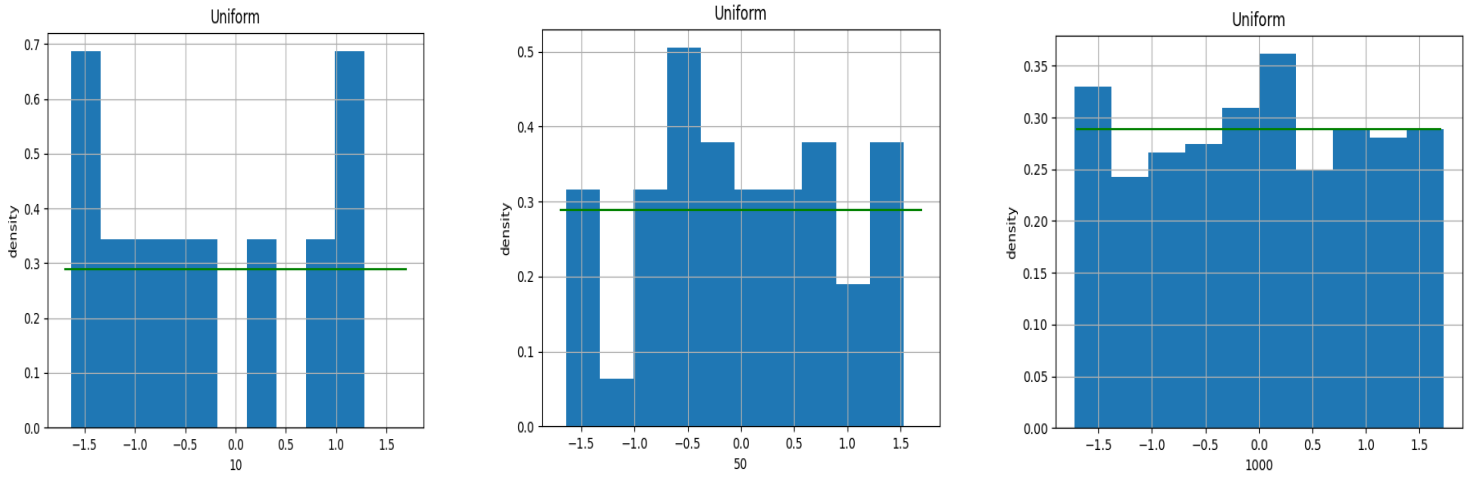


Рис. 5: Равномерное распределение

## 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

Округление сделано для $k = 1$

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 0.014389 | 0.012195 | 0.021778 | 0.322892 | 0.288555 |
| *D(z)* | 0.103622 | 0.146584 | 0.182411 | 0.129487 | 0.12315 |
| | | | | | |
| n = 100 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -0.005169 | -0.002888 | -0.005813 | 0.00934 | 0.020767 |
| *D(z)* | 0.010705 | 0.016466 | 0.089037 | 0.012262 | 0.012502 |
| | | | | | |
| n = 1000 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -0.000969 | -0.000121 | -0.011948 | -0.000157 | 0.00171 |
| *D(z)* | 0.000961 | 0.001511 | 0.057823 | 0.001234 | 0.001139 |

Таблица 2: Нормальное распределение

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 3.538687 | 0.005747 | 17.621658 | 1.220385 | 0.7136 |
| *D(z)* | 22604.624487 | 0.313332 | 564514.92 | 13.407058 | 2.239713 |
| | | | | | |
| n = 100 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -1.4071 | -0.00721 | -68.423193 | 0.026804 | 0.034252 |
| *D(z)* | 3165.113041 | 0.023949 | 7499545.88 | 0.048978 | 0.024529 |
| | | | | | |
| n = 1000 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -1.761315 | -0.00033 | -887.914482 | -0.000279 | 0.002199 |
| *D(z)* | 2093.687604 | 0.002336 | 518897246 | 0.004768 | 0.002418 |

Таблица 3: Распределение Коши

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -0.010182 | -0.001877 | -0.027369 | 0.402402 | 0.320974 |
| *D(z)* | 0.192717 | 0.139431 | 0.804849 | 0.222492 | 0.159666 |
| | | | | | |
| n = 100 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |

| | | | | | |
|---|---|---|---|---|---|
| *E(z)* | -0.00007 | 0.001586 | -0.050559 | 0.019895 | 0.027567 |
| *D(z)* | 0.019491 | 0.010984 | 0.833799 | 0.01897 | 0.012073 |
| | | | | | |
| n = 1000 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 0.002939 | 0.001933 | -0.016726 | 0.005174 | 0.005352 |
| *D(z)* | 0.001939 | 0.00113 | 0.741087 | 0.001963 | 0.001295 |

Таблица 4: Распределение Лапласа

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 10.0131 | 9.891 | 10.3065 | 10.941 | 10.803 |
| *D(z)* | 1.079498 | 1.480119 | 1.954308 | 1.399019 | 1.278635 |
| | | | | | |
| n = 100 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 10.00841 | 9.8425 | 10.922 | 9.9715 | 9.95278 |
| *D(z)* | 0.095403 | 0.214944 | 1.052916 | 0.150438 | 0.118814 |
| | | | | | |
| n = 1000 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 9.999174 | 9.9955 | 11.6815 | 9.9905 | 9.866174 |
| *D(z)* | 0.009634 | 0.00423 | 0.647808 | 0.00516 | 0.011437 |

Таблица 5: Распределение Пуассона

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -0.000015 | 0.001647 | -0.000806 | 0.317138 | 0.317472 |
| *D(z)* | 0.098668 | 0.227861 | 0.04336 | 0.127716 | 0.151463 |
| | | | | | |
| n = 100 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | 0.005672 | 0.00801 | 0.00102 | 0.024088 | 0.042511 |
| *D(z)* | 0.010117 | 0.030217 | 0.000519 | 0.014795 | 0.020165 |
| | | | | | |
| n = 1000 | | | | | |
| | $\bar{x}$ | *med x* | $z_R$ | $z_Q$ | $z_{tr}$ |
| *E(z)* | -0.002387 | -0.003544 | 0.000052 | -0.00159 | 0.000402 |
| *D(z)* | 0.001023 | 0.003124 | 0.0000007 | 0.001528 | 0.002028 |

Таблица 6: Равномерное распределение

## 4.3 Боксплот Тьюки

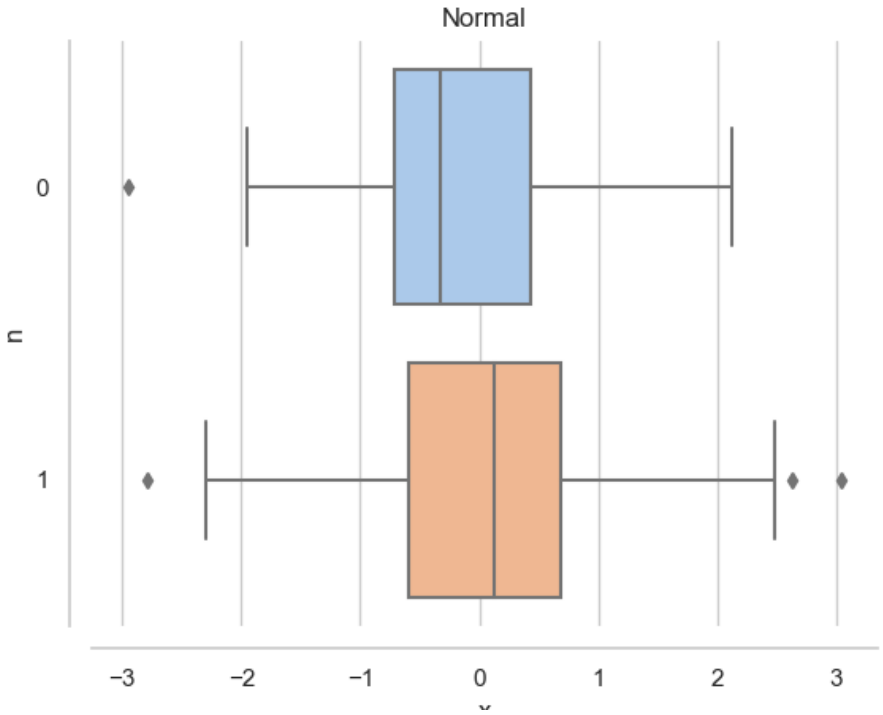


Рис. 6: Нормальное распределение

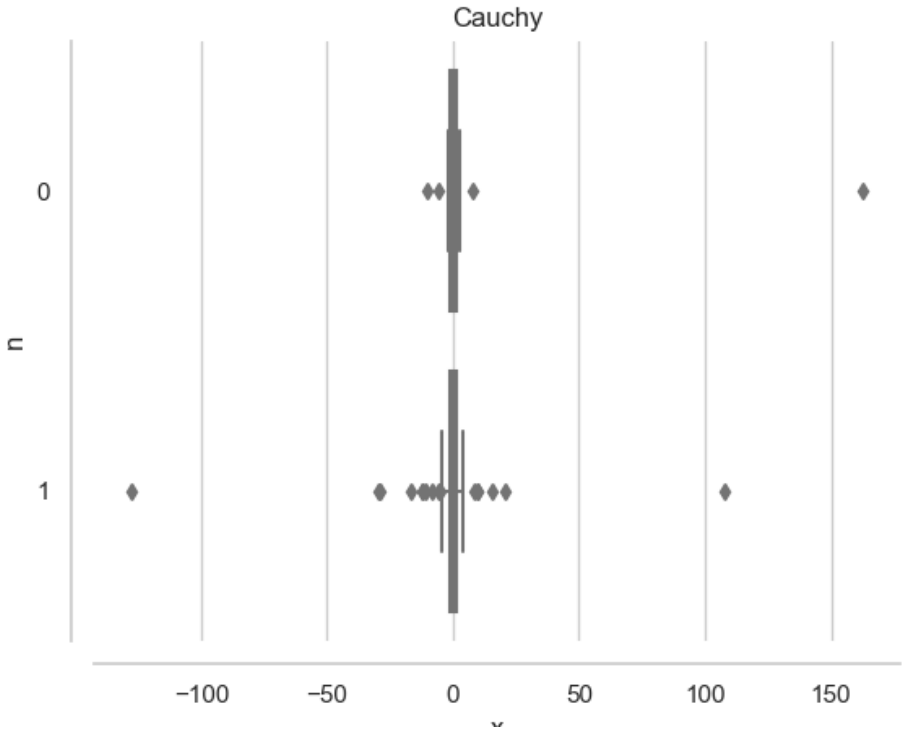


Рис. 7: Распределение Коши

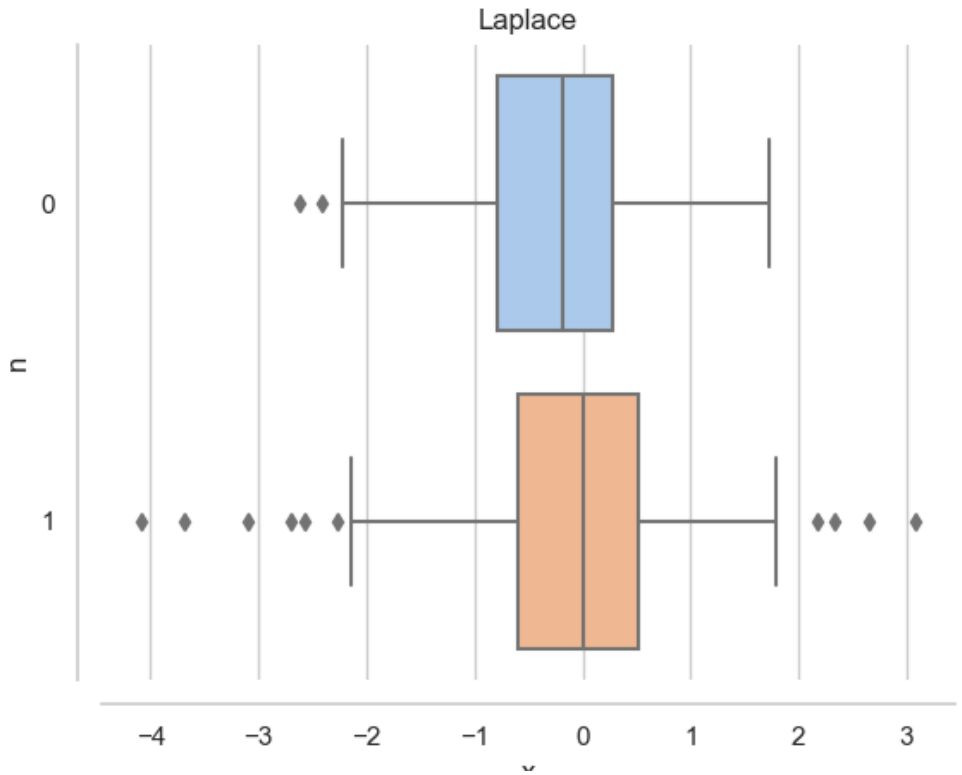


Рис. 8: Распределение Лапласа

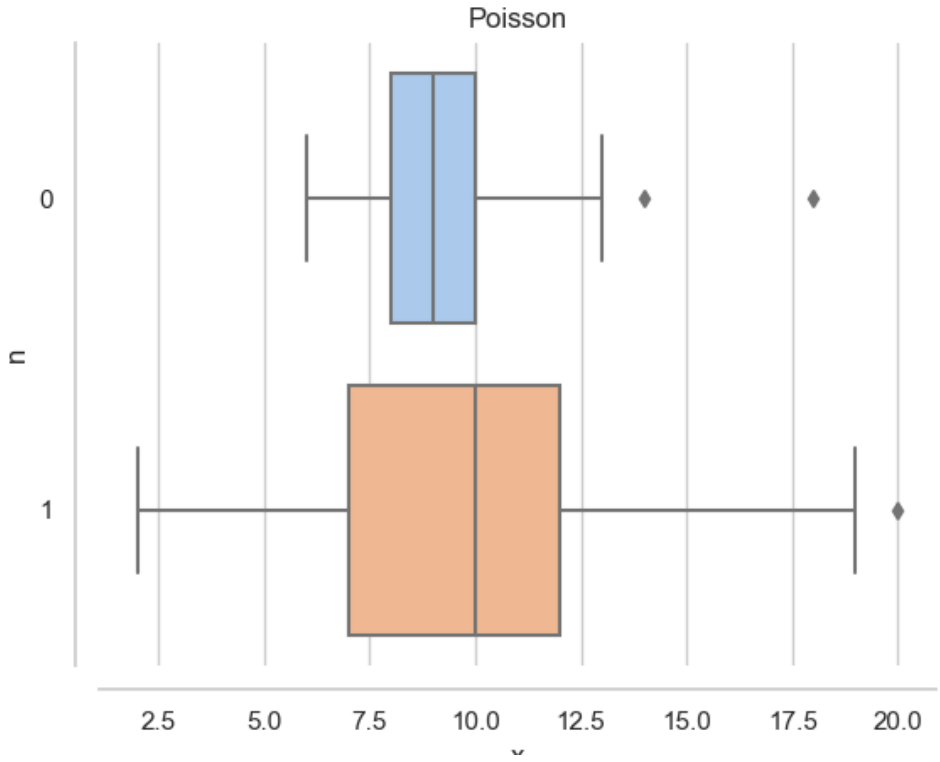


Рис. 9: Распределение Пуассона

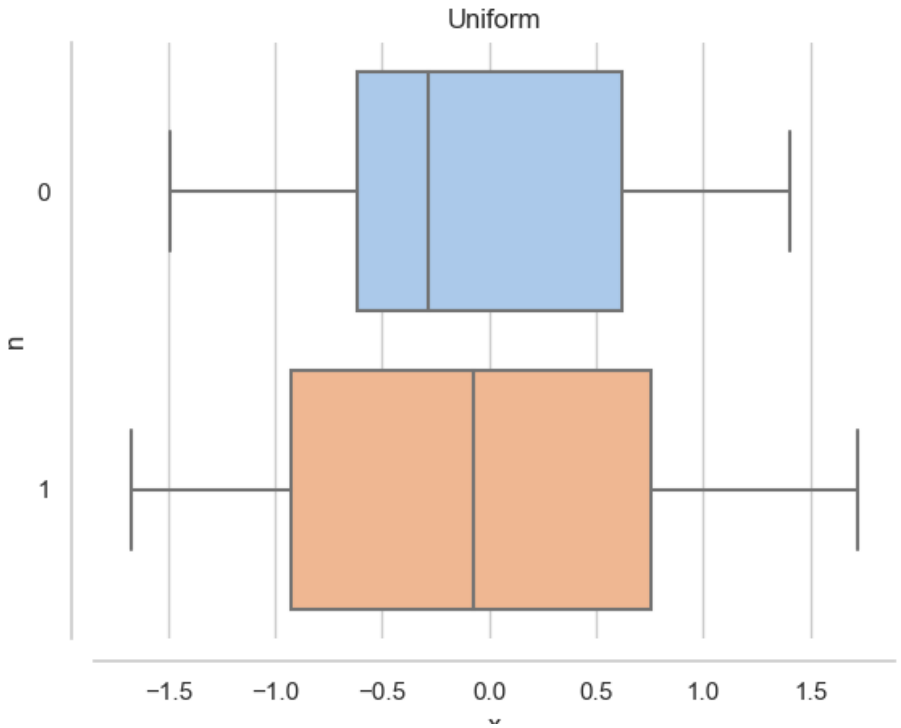


Рис. 10: Равномерное распределение

## 4.4 Доля выбросов

Выборка случайна, поэтому в качестве оценки рассеяния можно взять дисперсию пуассоновского потока: $D_n \approx \sqrt{n}$

Доля $pn = \frac{D_n}{n} = \frac{1}{\sqrt{n}}$

Для $n = 20$: $pn = \frac{1}{\sqrt{20}}$ – примерно 0.2 или 20%

Для $n = 100$: $pn = \frac{1}{\sqrt{100}}$ – 0.1 или 10%

Исходя из этого можно решить, сколько знаков оставлять в доле выбросов.

| Выборка | Доля выбросов |
|---|---|
| Normal n = 20 | 0.022 |
| Normal n = 100 | 0.009 |
| Cauchy n = 20 | 0.151 |
| Cauchy n = 100 | 0.155 |
| Laplace n = 20 | 0.069 |
| Laplace n = 100 | 0.065 |
| Poisson n = 20 | 0.024 |
| Poisson n = 100 | 0.010 |
| Uniform n = 20 | 0.002 |
| Uniform n = 100 | 0.0 |

Таблица 7: Доля выбросов

## 4.5 Теоретическая вероятность выбросов

| Распределение | $Q_1^T$ | $Q_3^T$ | $X_1^T$ | $X_2^T$ | $P_B^T$ |
|---|---|---|---|---|---|
| Нормальное распределение | -0.674 | 0.674 | -2.698 | 2.698 | 0.007 |
| Распределение Коши | -1 | 1 | -4 | 4 | 0.156 |
| Распределение Лапласа | -0.490 | 0.490 | -1.961 | 1.961 | 0.063 |
| Распределение Пуассона | 8 | 12 | 2 | 18 | 0.008 |
| Равномерное распределение | -0.866 | 0.866 | -3.464 | 3.464 | 0 |

Таблица 8: Теоретическая вероятность выбросов

## 4.6 Эмпирическая функция распределения

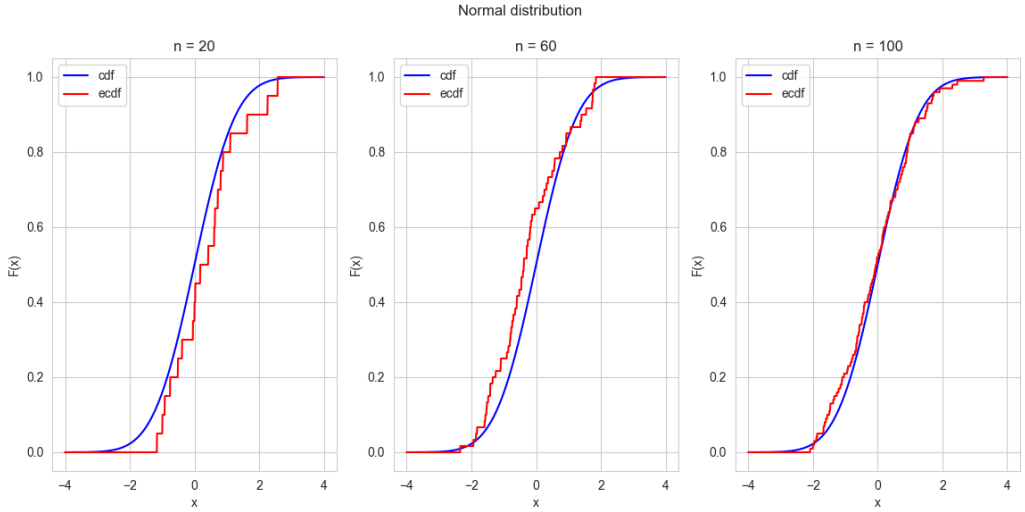


Рис. 11: Нормальное распределение

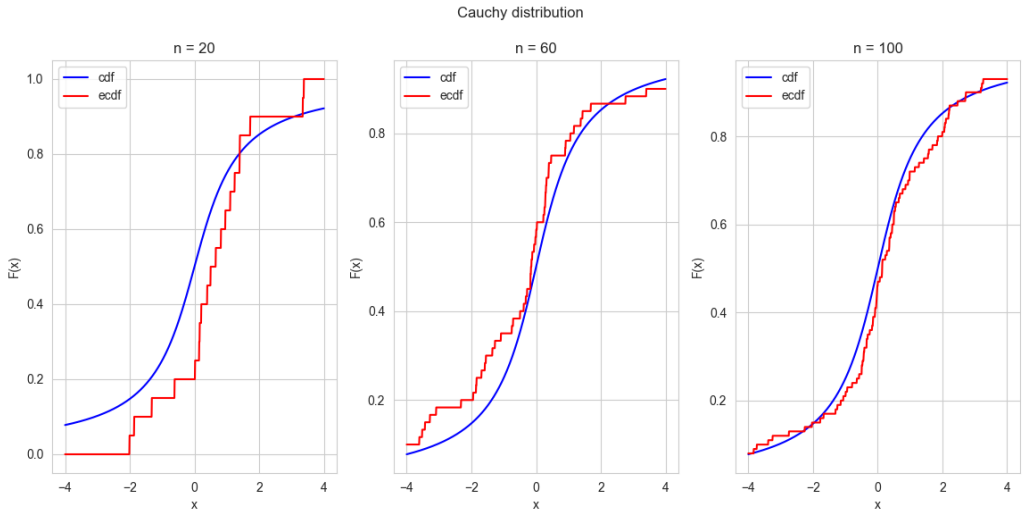


Рис. 12: Распределение Коши

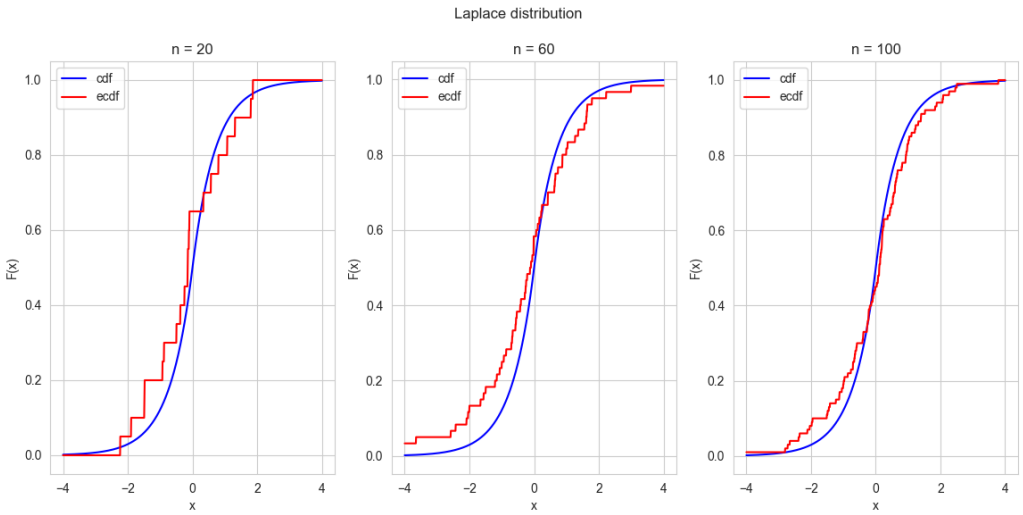


Рис. 13: Распределение Лапласа

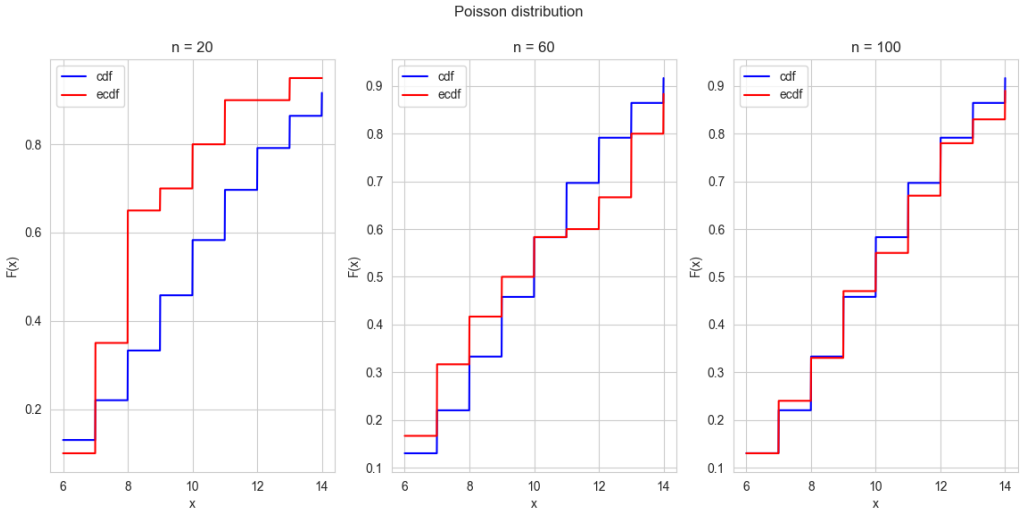


Рис. 14: Распределение Пуассона

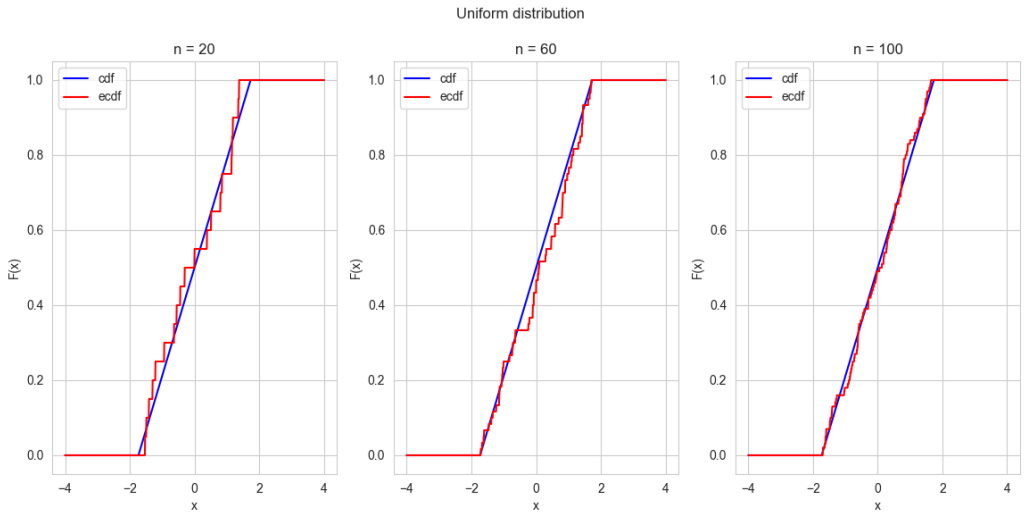


Рис. 15: Равномерное распределение

## 4.7 Ядерные оценки плотности распределения

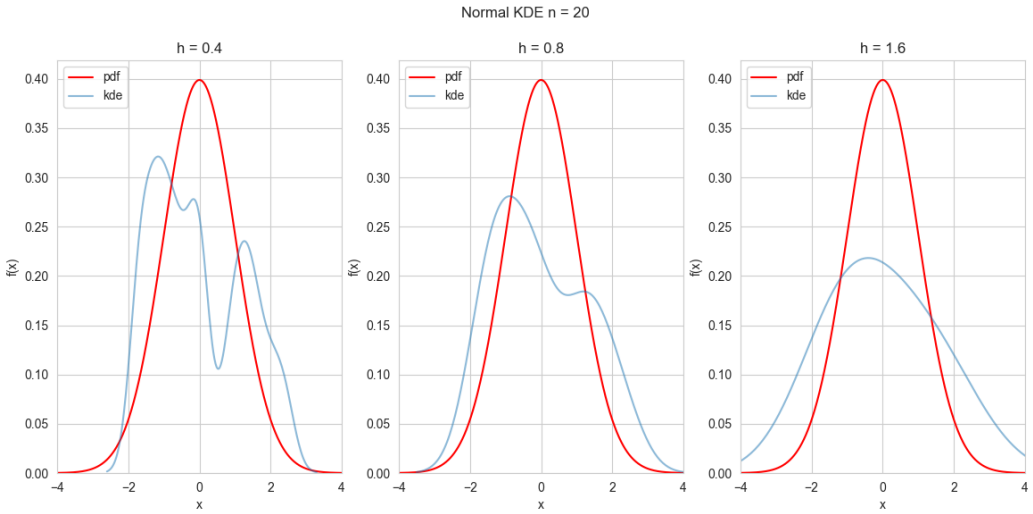


Рис. 16: Нормальное распределение, n = 20

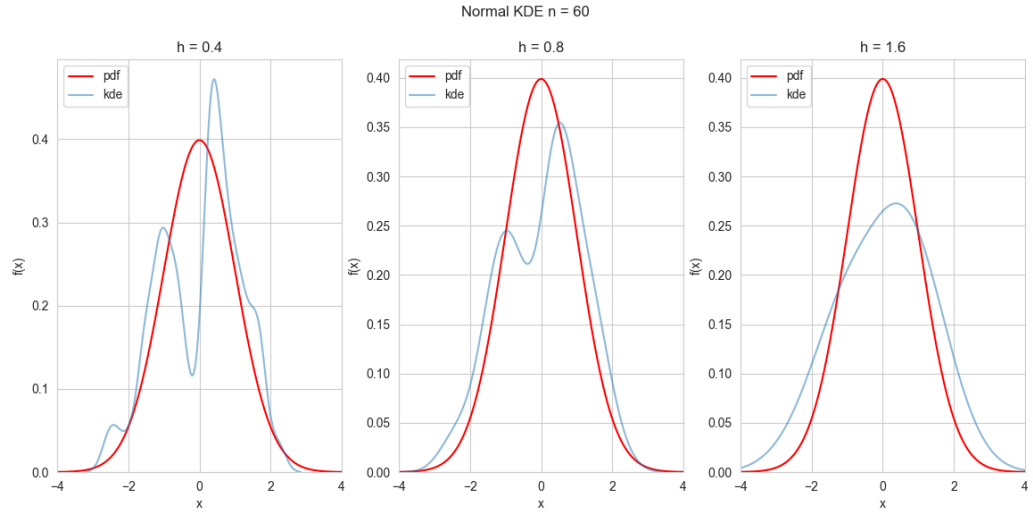


Рис. 17: Нормальное распределение, n = 60

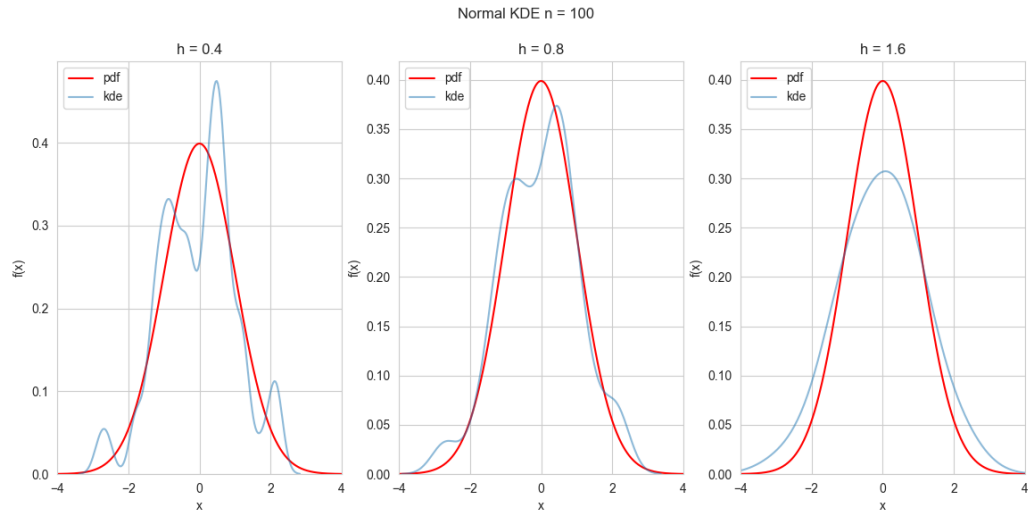


Рис. 18: Нормальное распределение, n = 100

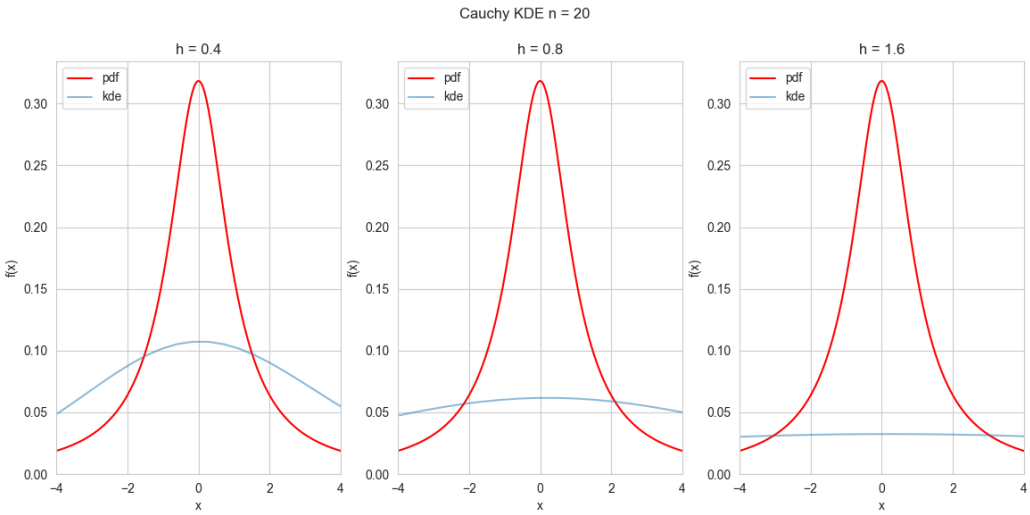


Рис. 19: Распределение Коши n = 20

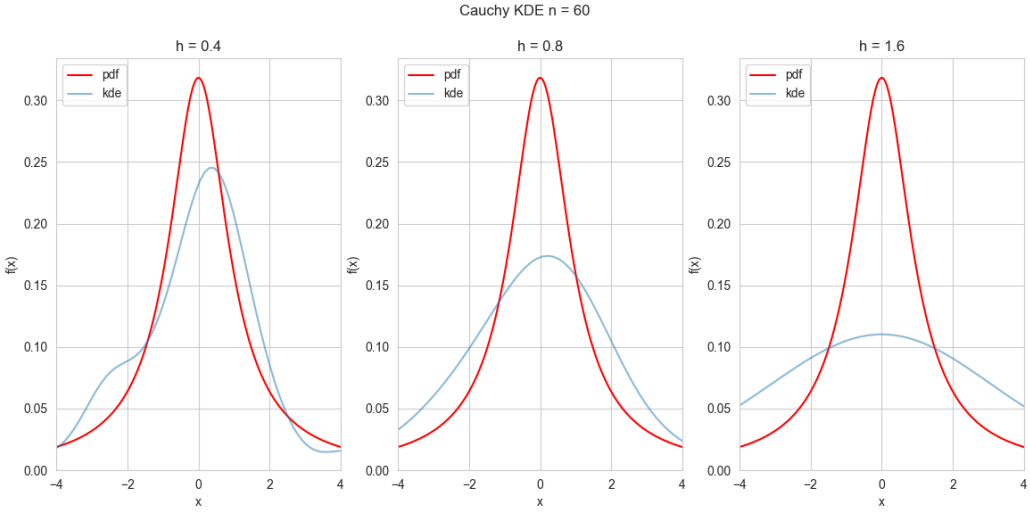


Рис. 20: Распределение Коши n = 60

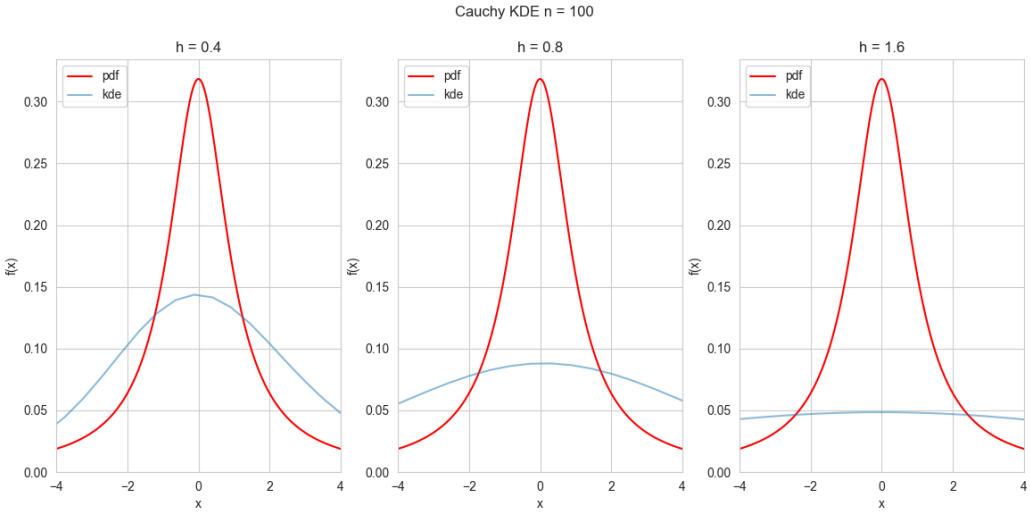


Рис. 21: Распределение Коши n = 100

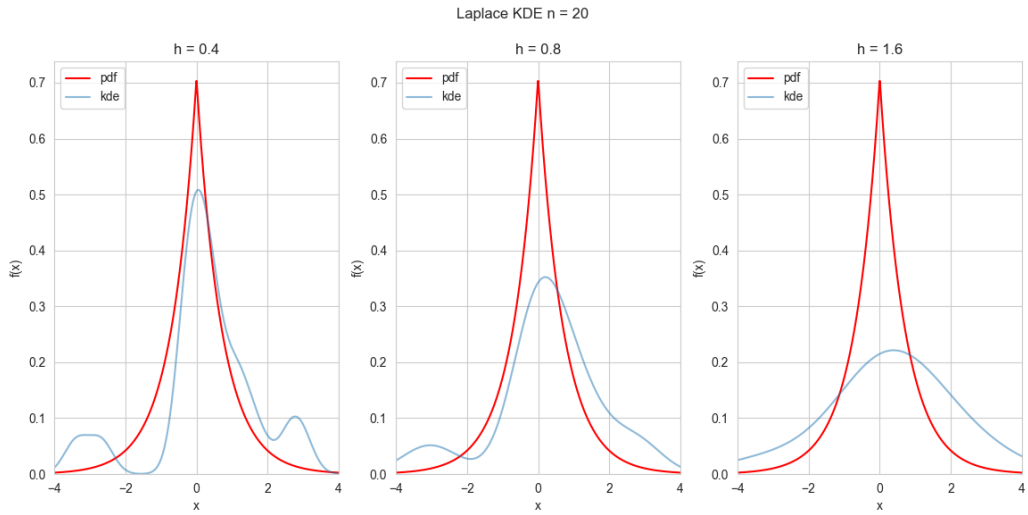


Рис. 22: Распределение Лапласа n = 20

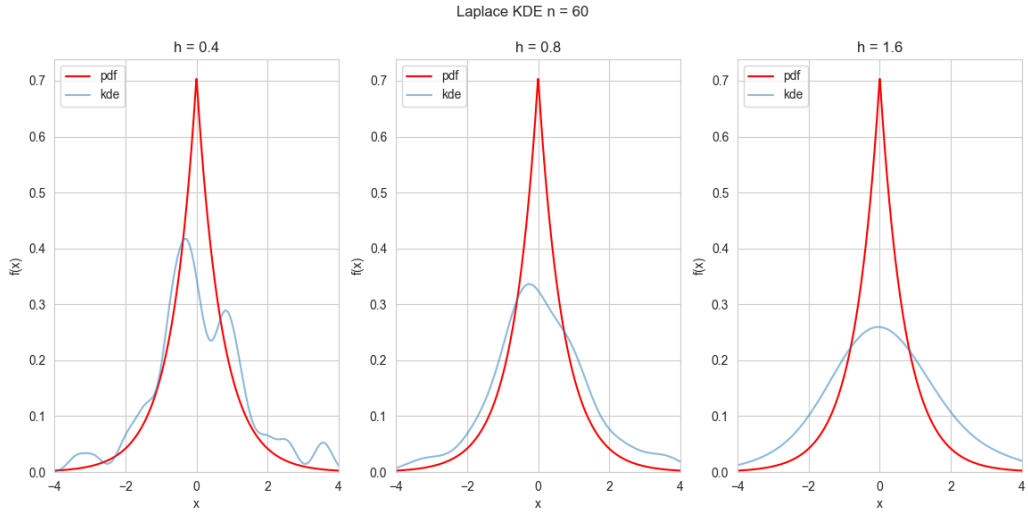


Рис. 23: Распределение Лапласа n = 60

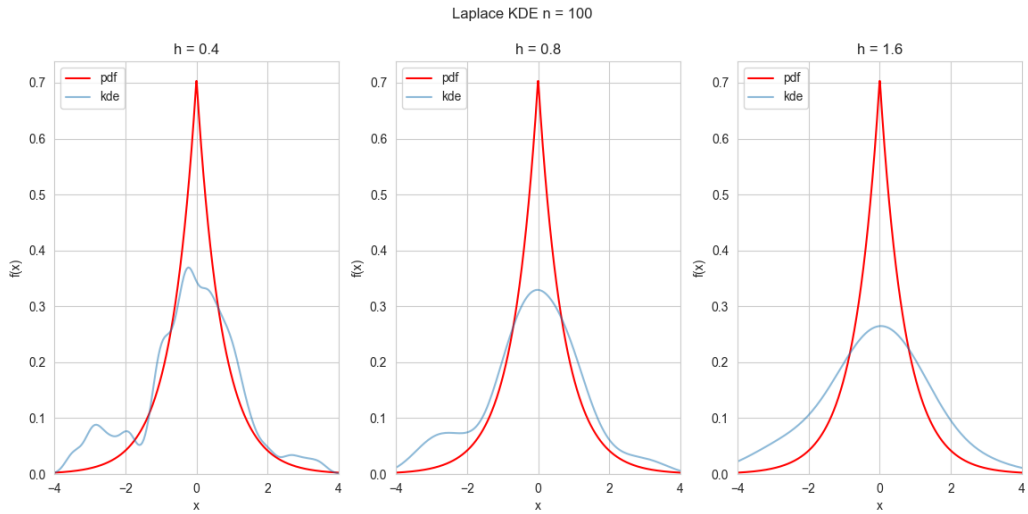


Рис. 24: Распределение Лапласа n = 100

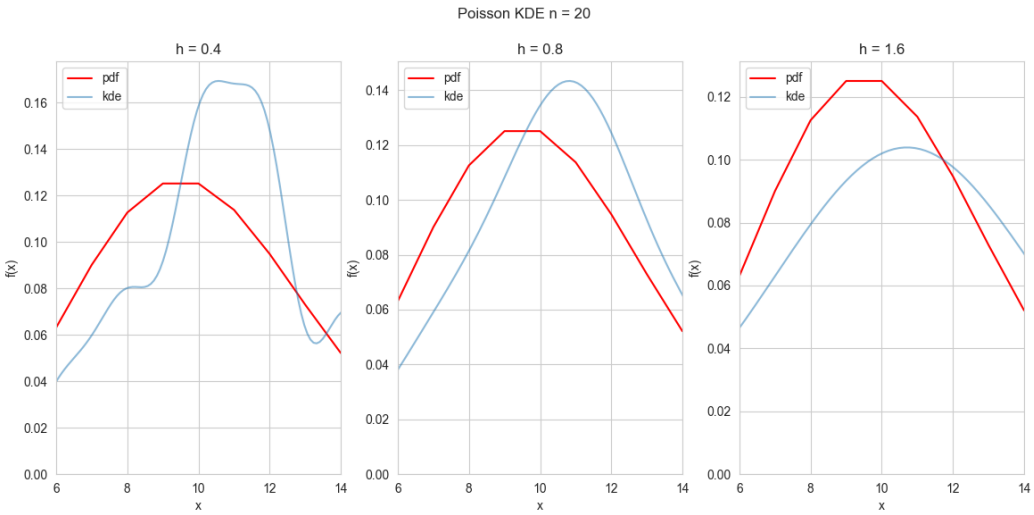


Рис. 25: Распределение Пуассона n = 20

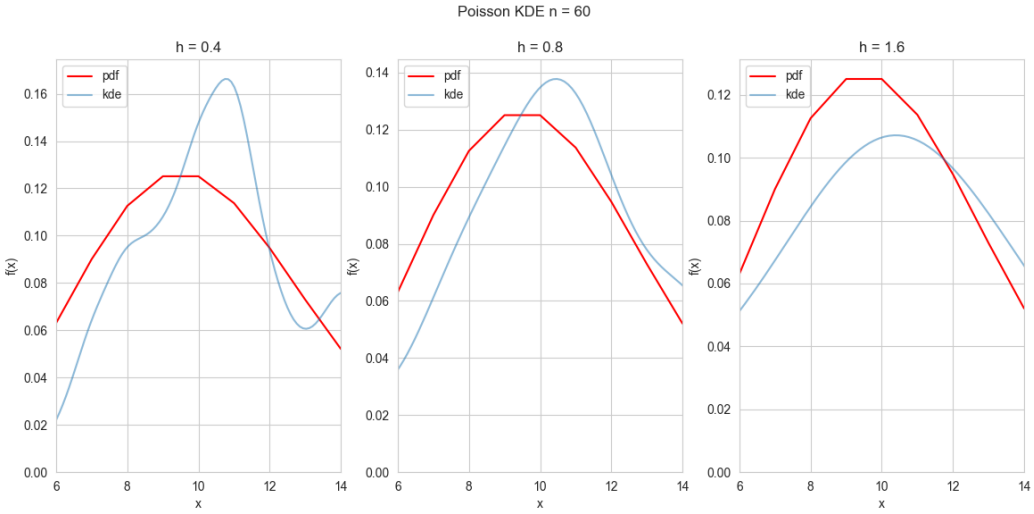


Рис. 26: Распределение Пуассона n = 60

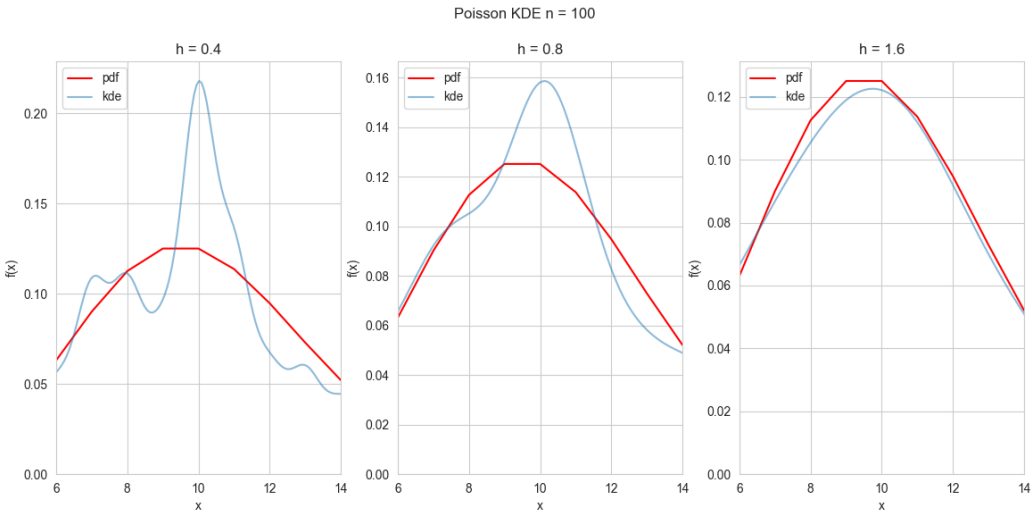


Рис. 27: Распределение Пуассона n = 100

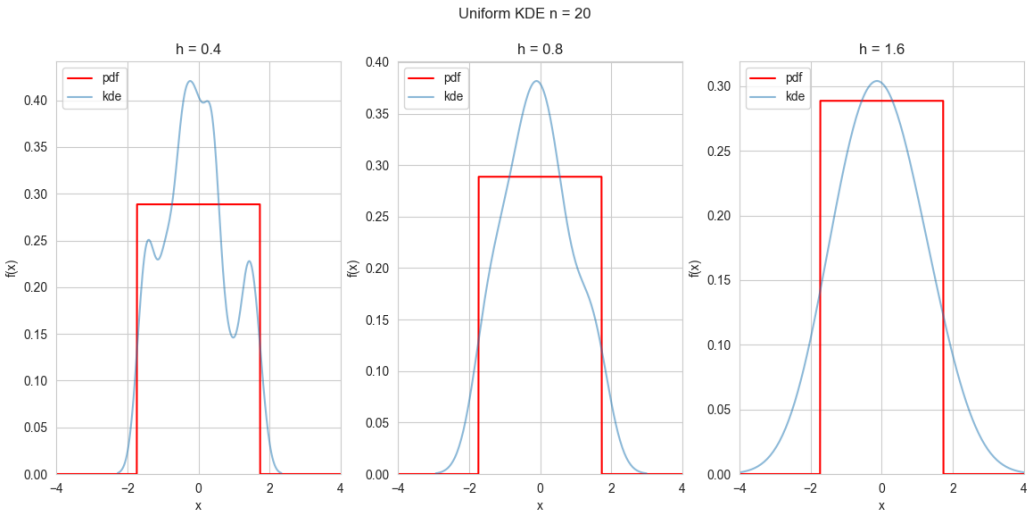


Рис. 28: Равномерное распределение n = 20

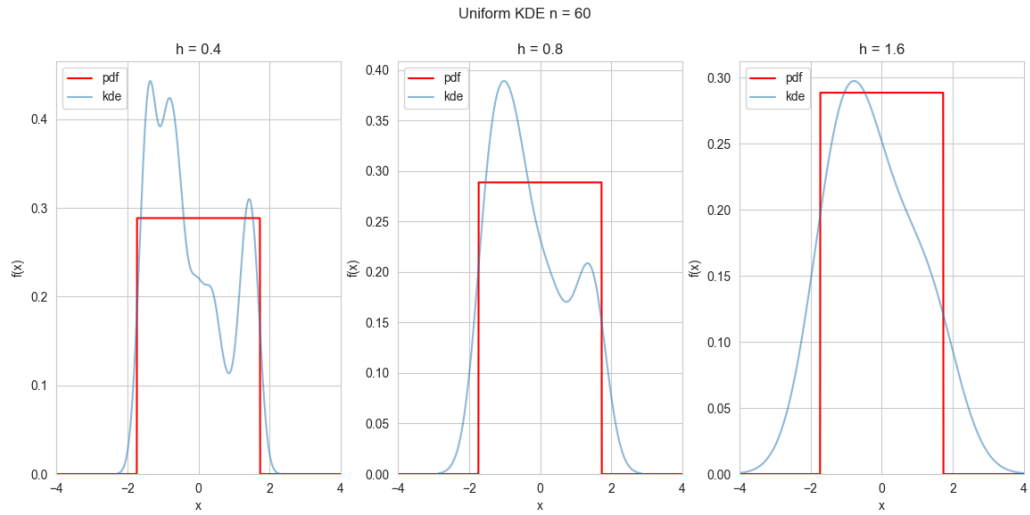


Рис. 29: Равномерное распределение n = 60

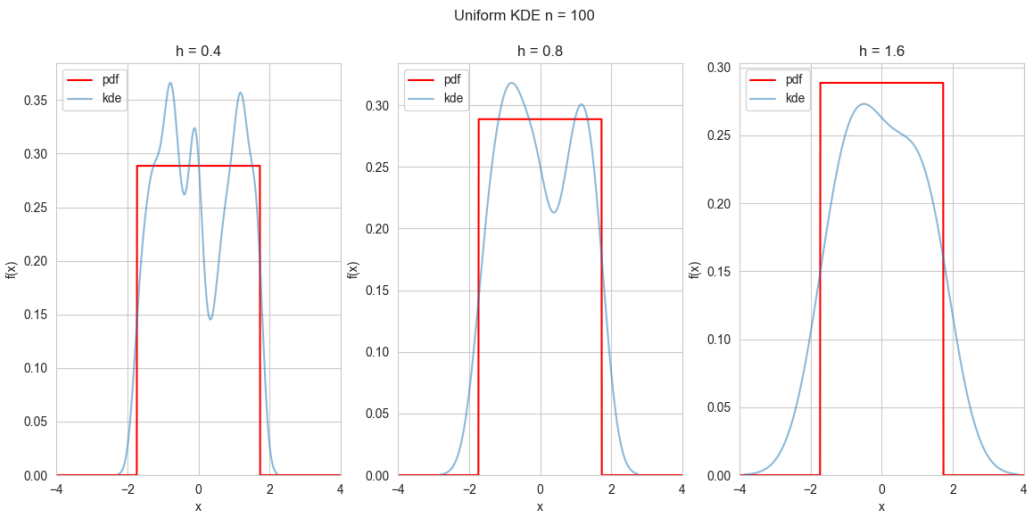


Рис. 30: Равномерное распределение n = 100

# 5. Обсуждение

## 5.1 Гистограмма и график плотности распределения

По результатам проделанной работы видно, что чем увеличение выборки из распределения приближает гистограмму к графику плотности распределения. Следовательно, чем меньше выборка, тем менее она показательна - тем хуже по ней определяется характер распределения величины. Кроме того, можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Всплески на гистограммах относительно плотности также уменьшаются при увеличении выборки.
Формы гистограмм при небольших выборках весьма похожи друг на друга, они имеют более общий вид и меньше похожи на график плотности распределения.

5.2 Характеристики положения и рассеяния

Приведенные данные показывают, что большие выборки лучше уточняют значение характеристик случайной величины.
Для равномерного, нормального распределений и распределения Лапласа характеристики положения близки к нулю. У распределения Коши появляются аномально большие числа, обусловленные бесконечной дисперсией случайной величины. Также это можно было увидеть на количестве выбросов из пункта 4.1.
В распределении Пуассона среднее значение E(z) близко к 10 при любой выборке – это параметр данного распределения.

5.3 Доля и теоретическая вероятность выбросов

Боксплот Тьюки позволяет наглядно оценить характеристики распределений. Видно, как медиана равномерного, нормального и распределения Лапласа приближается к нулю, что было установлено аналитически выше. Кроме того, остальные характеристики также проще и удобнее оценивать.

Полученные данные в таблице показывают, что для всех распределений чем больше выборка, тем ближе найденная доля выбросов к теоретической. В распределении Коши продолжается наблюдаться большая доля выбросов, а у равномерного распределения она почти отсутствует.

5.4 Эмпирическая функция и ядерные оценки плотности распределения

По приведенным графикам эмпирический функций можно сделать вывод о том, что с увеличением мощности выборки эмпирическая функция становится ближе к эталонной. Из всех распределений больше всего разница между эмпирическими функциями наблюдается у распределения Пуассона.

По графикам ядерных оценок видно, что при увеличении размера выборки ядерные оценки сближаются к функциям плотности вероятности для всех h. Однако оптимальный параметр h отличается в зависимости от распределения. Например, для распределения Лапласа и Коши больше всего подходит параметр h=$h_n$/2. Для распределения Пуассона и равномерного оптимальнее h=2$h_n$. Кроме того, при увеличении параметра h уменьшается верхнее значение точки графика ядерной функции.
Поэтому можно сделать вывод о том, что параметр сглаживание стоит выбирать с умом для каждого отдельного распределения. Чрезмерное сглаживание может привести к потери особенности  распределения, в то же время недостаточно сглаженная кривая может создать ложные особенности случайного распределения.