

Санкт-Петербургский политехнический университет Петра Великого
Физико Механический Институт
Высшая школа прикладной математики и вычислительной физики

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

по дисциплине
«Математическая статистика»

Выполнила студент
группы 5030102/90101

Кузин Иван Никитович

Проверил
Доцент, к.ф.-м.н.

Баженов Александр Николаевич

Санкт-Петербург
2022

СОДЕРЖАНИЕ

| | |
|---|----------|
| СПИСОК ИЛЛЮСТРАЦИЙ | 3 |
| 1 Постановка задачи | 4 |
| 2 Теория | 4 |
| 2.1 Распределения | 4 |
| 2.2 Гистограмма | 4 |
| 2.2.1 Определение | 4 |
| 2.2.2 Графическое описание | 5 |
| 2.2.3 Использование | 5 |
| 3 Программная реализация | 5 |
| 4 Результаты | 6 |
| 4.1 Гистограммы и графики плотности распределения | 6 |
| 5 Обсуждение | 8 |
| 6 Приложение | 8 |

СПИСОК ИЛЛЮСТРАЦИЙ

| | | |
|---|---|---|
| 1 | Нормальное распределение (1) | 6 |
| 2 | Распределение Коши (2) | 6 |
| 3 | Распределение Лапласа (3) | 7 |
| 4 | Распределение Пуассона (4) | 7 |
| 5 | Равномерное распределение (5) | 7 |

1 Постановка задачи

Для 5 распределений:

1. $N(x, 0, 1)$ – нормальное распределение
2. $C(x, 0, 1)$ – распределение Коши
3. $L(x, 0, \frac{1}{\sqrt{2}})$ – распределение Лапласа
4. $P(k, 10)$ – распределение Пуассона
5. $U(x, -\sqrt{3}, \sqrt{3})$ – равномерное распределение

Сгенерировать выборки размером 10, 50 и 1000 элементов.

Построить на одном рисунке гистограмму и график плотности распределения.

2 Теория

2.1 Распределения

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Лапласа

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

- Распределение Пуассона

$$P(k, 10) = \frac{10^k}{k!} e^{-10} \quad (4)$$

- Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при } |x| \leq \sqrt{3} \\ 0 & \text{при } |x| > \sqrt{3} \end{cases} \quad (5)$$

2.2 Гистограмма

2.2.1 Определение

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

2.2.2 Графическое описание

Графически гистограмма строится следующим образом. Сначала 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.

2.2.3 Использование

Гистограммы применяются в основном для визуализации данных на начальном этапе статистической обработки.

Построение гистограмм используется для получения эмпирической оценки плотности распределения случайной величины. Для построения гистограммы наблюдаемый диапазон изменения случайной величины разбивается на несколько интервалов и подсчитывается доля от всех измерений, попавшая в каждый из интервалов. Величина каждой доли, отнесенная к величине интервала, принимается в качестве оценки значения плотности распределения на соответствующем интервале.

3 Программная реализация

Лабораторная работа выполнена на языке Python версии 3.9 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. `scipy`
2. `numpy`
3. `matplotlib`
4. `math`

В приложении находится ссылка на GitHub репозиторий с исходным кодом.

4 Результаты

4.1 Гистограммы и графики плотности распределения

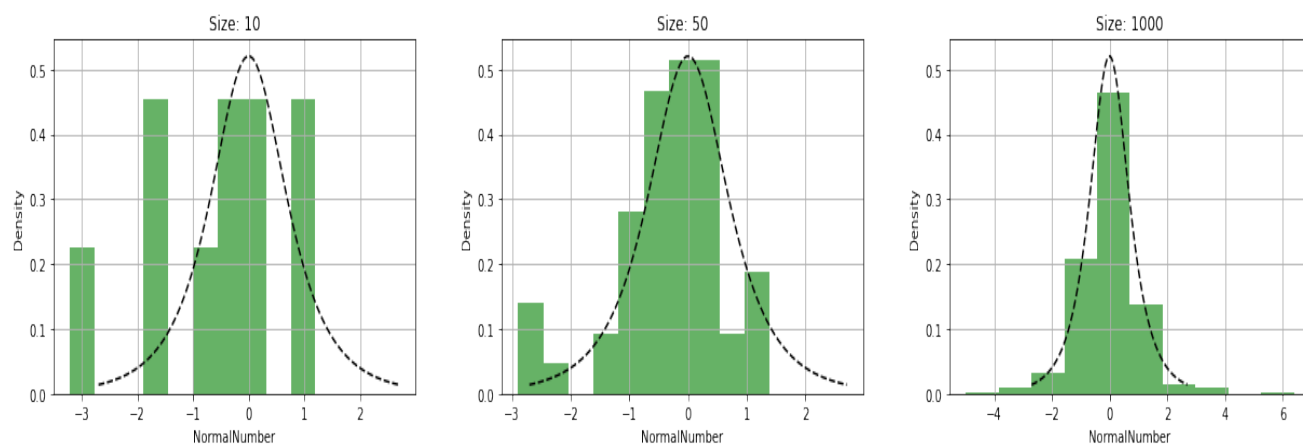


Рис. 1: Нормальное распределение (1)

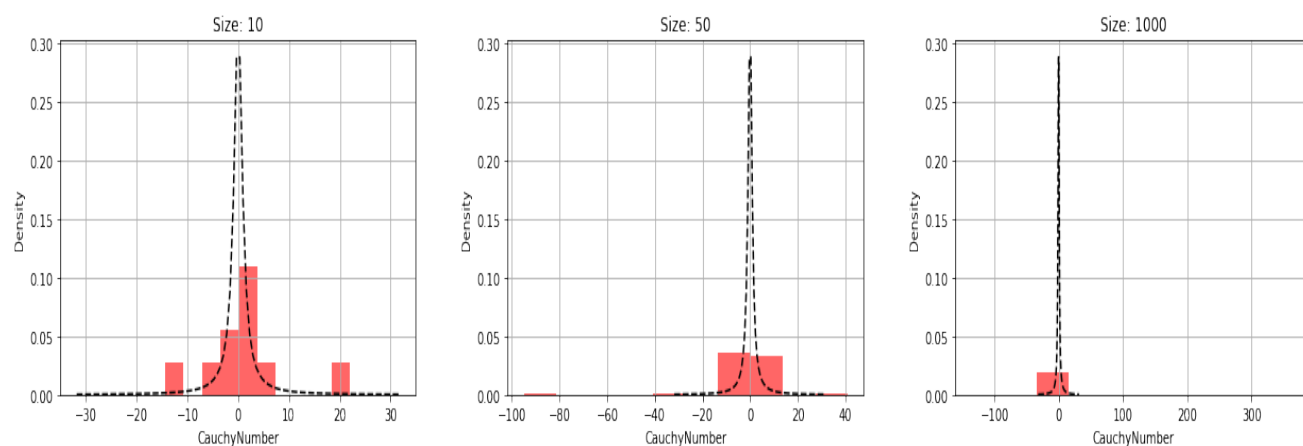


Рис. 2: Распределение Коши (2)

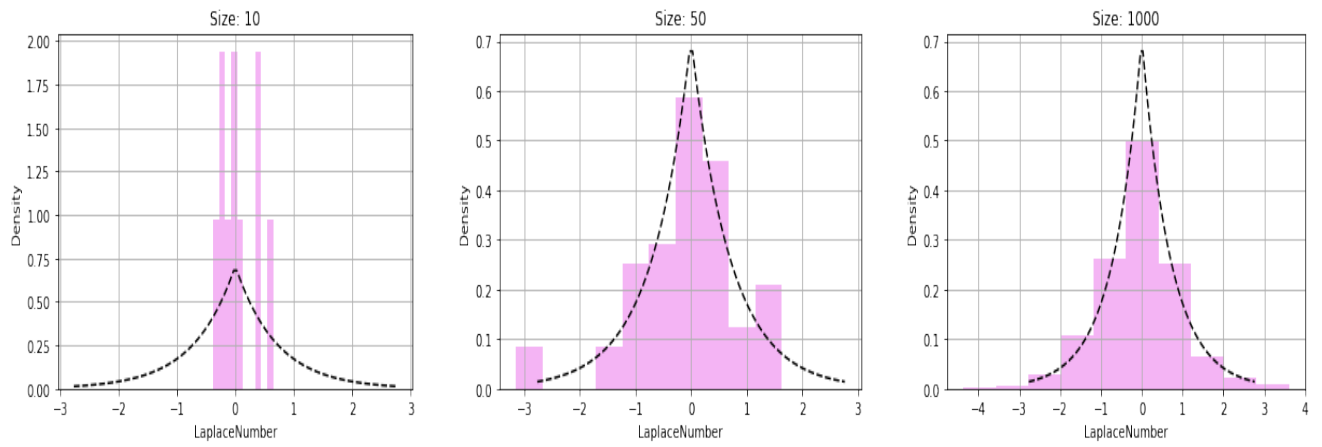


Рис. 3: Распределение Лапласа (3)

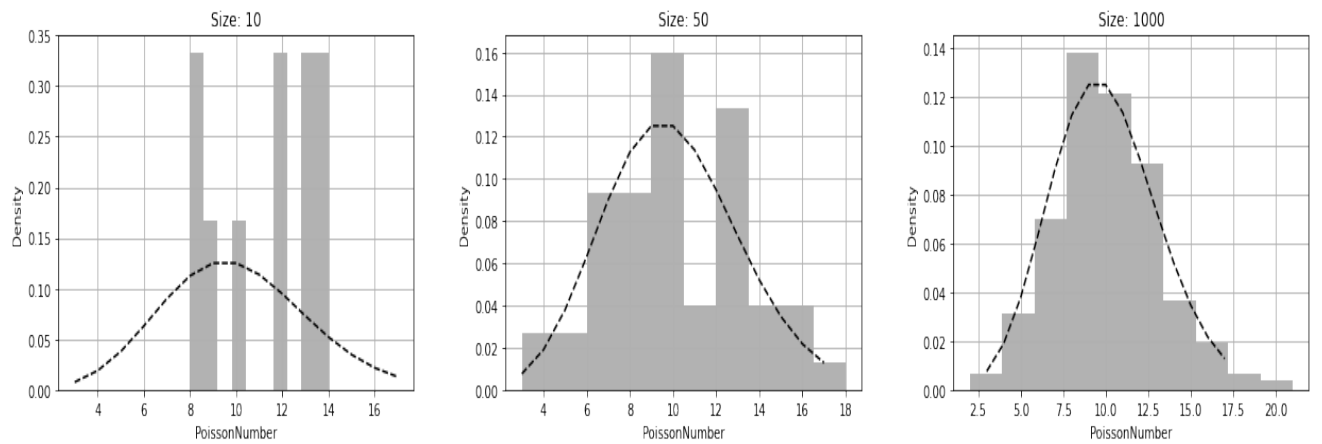


Рис. 4: Распределение Пуассона (4)

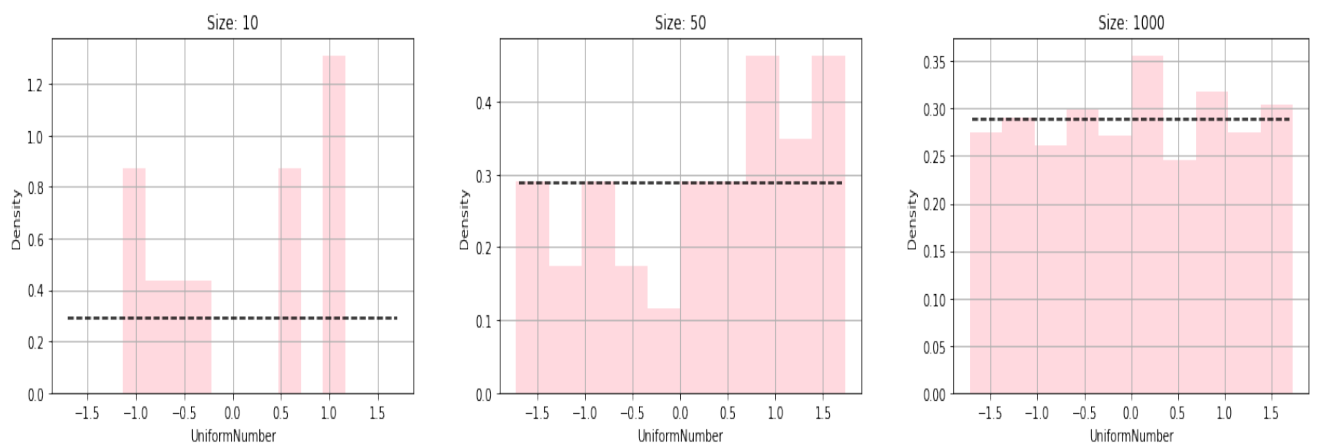


Рис. 5: Равномерное распределение (5)

5 Обсуждение

По результатам проделанной работы можем сделать вывод о том, что чем больше выборка для каждого из распределений, тем ближе ее гистограмма к графику плотности вероятности того закона, по которому распределены величины сгенерированной выборки. Чем меньше выборка, тем менее она показательна - тем хуже по ней определяется характер распределения величины.

Визуально очень трудно отличить гистограммы друг от друга, тем более при маленьких выборках. При выборке из 10 элементов вид гистограммы много сильно отличается от плотности распределения. Чем больше выборка, тем точнее становится гистограмма. На выборке из 1000 элементов можем отличить и распознать с большей вероятностью равномерное распределение (все прямоугольники примерно на одном уровне), а также распределение Пуассона (оно визуально шире чем распределение Лапласа и нормальное). Однако отличить между собой распределение Лапласа и нормальное тяжело. Так как визуально гистограммы получились похожи друг на друга и без подписей отличить их почти не представляется возможным.

Также можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Из полученных графиков можно увидеть, что только при распределении Пуассона на выборке из 1000 элементов, максимум графика плотности вероятности совпал с максимумом гистограммы. Также наблюдаются всплески гистограмм, что наиболее хорошо прослеживается на распределении Коши.

6 Приложение

Код программы GitHub URL:

<https://github.com/workivan/mat-ver-stat/>

Санкт-Петербургский политехнический университет Петра Великого
Физико Механический институт
Высшая школа прикладной математики и вычислительной физики

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №2

по дисциплине
«Математическая статистика»

Выполнила студент
группы 5030102/90101

Кузин Иван Никитович

Проверил
Доцент, к.ф.-м.н.

Баженов Александр Николаевич

Санкт-Петербург
2022

СОДЕРЖАНИЕ

| | |
|--|----------|
| СПИСОК ТАБЛИЦ | 3 |
| 1 Постановка задачи | 4 |
| 2 Теория | 4 |
| 2.1 Распределения | 4 |
| 2.2 Вариационный ряд | 4 |
| 2.3 Выборочные числовые характеристики | 5 |
| 2.3.1 Характеристики положения | 5 |
| 2.3.2 Характеристики рассеяния | 5 |
| 3 Программная реализация | 5 |
| 4 Результаты | 6 |
| 4.1 Характеристики положения и рассеяния | 6 |
| 5 Обсуждение | 9 |
| 6 Приложение | 9 |

СПИСОК ТАБЛИЦ

| | | |
|---|---|---|
| 1 | Нормальное распределение (3) | 6 |
| 2 | Распределение Коши (4) | 7 |
| 3 | Распределение Лапласа (5) | 7 |
| 4 | Распределение Пуассона (6) | 8 |
| 5 | Равномерное распределение (7) | 8 |

1 Постановка задачи

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.

2 Теория

2.1 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 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 Вариационный ряд

Вариационным рядом называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются. Запись вариационного ряда: $x_{(1)}, x_{(2)}, \dots$, Элементы вариационного ряда $x_{(i)} (i = 1, 2, \dots, n)$ называются порядковыми статистиками.

2.3 Выборочные числовые характеристики

С помощью выборки образуются её числовые характеристики. Это числовые характеристики дискретной случайной величины X^* , принимающей выборочные значения $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

2.3.1 Характеристики положения

- Выборочное среднее

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (8)$$

- Выборочная медиана

$$medx = \begin{cases} x_{(l+1)} & n = 2l + 1 \\ \frac{x_{(l)} + x_{(l+1)}}{2} & n = 2l \end{cases} \quad (9)$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \quad (10)$$

- Полусумма квартилей

Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & np\text{—дробное} \\ x_{(np)} & np\text{—целое} \end{cases} \quad (11)$$

Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \quad (12)$$

- Усечённое среднее

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

2.3.2 Характеристики рассеяния

Выборочная дисперсия

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

3 Программная реализация

Лабораторная работа выполнена на языке Python версии 3.7 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. scipy
2. numpy

В приложении находится ссылка на GitHub репозиторий с исходным кодом.

4 Результаты

4.1 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 | -0.017468 | -0.019852 | -0.025488 | 0.300647 | 0.260236 |
| Normal D(z) 10 | 0.099305 | 0.145161 | 0.182558 | 0.119238 | 0.11468 |
| $E(z) \pm \sqrt{D(z)}$ | [0.332595; 0.297659] | [0.400852; 0.361148] | [0.452756; 0.40178] | [0.044662; 0.645956] | [0.078408; 0.59888] |
| $\hat{E}(z)$ | 0 | 0 | 0 | 0 | 0 |
| Normal E(z) 100 | -0.000866 | 0.002153 | -0.004746 | 0.013516 | 0.027086 |
| Normal D(z) 100 | 0.010386 | 0.01633 | 0.085266 | 0.012678 | 0.012396 |
| $E(z) \pm \sqrt{D(z)}$ | [0.102778; 0.101046] | [0.125636; 0.129942] | [0.296749; 0.287257] | [0.099081; 0.126113] | [0.084251; 0.138423] |
| $\hat{E}(z)$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Normal E(z) 1000 | 0.000153 | 0.000089 | 0.007519 | 0.001509 | 0.002836 |
| Normal D(z) 1000 | 0.000981 | 0.001642 | 0.059635 | 0.001228 | 0.001195 |
| $E(z) \pm \sqrt{D(z)}$ | [0.031168; 0.031474] | [0.040433; 0.040611] | [0.236684; 0.251722] | [0.033534; 0.036552] | [0.031733; 0.037405] |
| $\hat{E}(z)$ | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 |

Таблица 1: Нормальное распределение (3)

| Characteristic | Mean | Median | z_R | z_Q | z_{tr} |
|------------------------|---------------------------|-------------------------|-------------------------------|-------------------------|-------------------------|
| Cauchy E(z) 10 | 0.562654 | -0.006675 | 2.898445 | 1.104381 | 0.671502 |
| Cauchy D(z) 10 | 523.363485 | 0.295683 | 12824.001857 | 4.591925 | 1.076515 |
| $E(z) \pm \sqrt{D(z)}$ | [22.314485; 23.439793] | [0.550442; 0.537092] | [110.344665; 116.141555] | [1.038497; 3.247259] | [0.36605; 1.709054] |
| $\hat{E}(z)$ | - | 0 | - | - | - |
| Cauchy E(z) 100 | -0.7079 | -0.001573 | -34.396555 | 0.03319 | 0.038699 |
| Cauchy D(z) 100 | 398.608006 | 0.027015 | 945555.067 | 0.057985 | 0.027952 |
| $E(z) \pm \sqrt{D(z)}$ | [20.67307; 19.25727] | [0.165935; 0.162789] | [1006.793114; 938.000004] | [0.207611; 0.273991] | [0.12849; 0.205888] |
| $\hat{E}(z)$ | - | 0 | - | 0 | 0 |
| Cauchy E(z) 1000 | -0.510318 | 0.000322 | -260.66918 | 0.004265 | 0.003941 |
| Cauchy D(z) 1000 | 290.388668 | 0.002172 | 69909833.801229 | 0.00484 | 0.002387 |
| $E(z) \pm \sqrt{D(z)}$ | [17.551112; 16.530476] | [0.046283; 0.046927] | [8621.879248; 8100.540888] | [0.065305; 0.073835] | [0.044916; 0.052798] |
| $\hat{E}(z)$ | - | 0.0 | - | 0.0 | 0.0 |

Таблица 2: Распределение Коши (4)

| Characteristic | Mean | Median | z_R | z_Q | z_{tr} |
|------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Laplace E(z) 10 | -0.006405 | -0.006188 | -0.005512 | 0.293618 | 0.229066 |
| Laplace D(z) 10 | 0.010247 | 0.006208 | 0.425481 | 0.010082 | 0.006497 |
| $E(z) \pm \sqrt{D(z)}$ | [0.031154; 0.030002] | [0.022723; 0.021909] | [0.651449; 0.626013] | [0.029889; 0.032657] | [0.022585; 0.025539] |
| $\hat{E}(z)$ | 0 | 0 | 0 | 0 | 0 |
| Laplace E(z) 100 | 0.002803 | 0.001905 | 0.023725 | 0.017159 | 0.021071 |
| Laplace D(z) 100 | 0.057048 | 0.041891 | 0.492694 | 0.492542 | 0.095815 |
| $E(z) \pm \sqrt{D(z)}$ | [-0.226576; 0.251118] | [-0.199293; 0.210051] | [-0.680069; 0.723773] | [-0.679985; 0.723641] | [-0.294498; 0.324580] |
| $\hat{E}(z)$ | 0.0 | 0.0 | 0 | 0 | 0.0 |
| Laplace E(z) 1000 | -0.000576 | -0.000407 | -0.012718 | 0.001384 | 0.001477 |
| Laplace D(z) 1000 | 0.000935 | 0.000498 | 0.407977 | 0.000978 | 0.000579 |
| $E(z) \pm \sqrt{D(z)}$ | [0.022723; 0.021909] | [0.651449; 0.626013] | [0.029889; 0.032657] | [0.022585; 0.025539] | [0.022652; 0.023563] |
| $\hat{E}(z)$ | 0.00 | 0.00 | 0 | 0.00 | 0.00 |

Таблица 3: Распределение Лапласа (5)

| Characteristic | Mean | Median | z_R | z_Q | z_{tr} |
|------------------------|--------------------------|--------------------------|---------------------------|--------------------------|--------------------------|
| Poisson E(z) 10 | 10.0015 | 9.87 | 10.265 | 10.9455 | 10.786167 |
| Poisson D(z) 10 | 1.101808 | 1.5811 | 1.936775 | 1.46278 | 1.362525 |
| $E(z) \pm \sqrt{D(z)}$ | [0.031154; 0.030002] | [0.022723; 0.021909] | [0.651449; 0.626013] | [0.029889; 0.032657] | [0.022585; 0.025539] |
| $\widehat{E}(z)$ | 10_{-1}^{+1} | 10_{-1}^{+1} | 10_{-2}^{+2} | 10_{-2}^{+2} | 10_{-1}^{+1} |
| Poisson E(z) 100 | 9.9932 | 9.844 | 10.9475 | 9.959 | 9.93526 |
| Poisson D(z) 100 | 0.10332 | 0.204664 | 0.997494 | 0.159819 | 0.122566 |
| $E(z) \pm \sqrt{D(z)}$ | [9.671766; 10.314634] | [9.391602; 10.296398] | [9.948754; 11.946246] | [9.559226; 10.358774] | [9.585166; 10.285354] |
| $\widehat{E}(z)$ | 10_{-1}^{+1} | 10_{-1}^{+1} | 10_{-2}^{+2} | 10_{-2}^{+2} | 10_{-1}^{+1} |
| Poisson E(z) 1000 | 10.000822 | 9.997 | 11.671 | 9.9955 | 9.86806 |
| Poisson D(z) 1000 | 0.00993 | 0.002991 | 0.746759 | 0.00223 | 0.01131 |
| $E(z) \pm \sqrt{D(z)}$ | [9.901173; 10.100471] | [9.94231; 10.05169] | [10.806848; 12.535152] | [9.948277; 10.042723] | [9.761712; 9.974408] |
| $\widehat{E}(z)$ | 10_{-1}^{+1} | 10_{-1}^{+1} | 10_{-2}^{+2} | 10_{-2}^{+2} | 10_{-1}^{+1} |

Таблица 4: Распределение Пуассона (6)

| Characteristic | Mean | Median | z_R | z_Q | z_{tr} |
|------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| Uniform E(z) 10 | -0.007779 | -0.011541 | -0.00537 | 0.316201 | 0.304339 |
| Uniform D(z) 10 | 0.104723 | 0.233967 | 0.047982 | 0.130148 | 0.159694 |
| $E(z) \pm \sqrt{D(z)}$ | [0.331388; 0.31583] | [0.495242; 0.47216] | [0.224418; 0.213678] | [0.044559; 0.676961] | [0.095278; 0.703956] |
| $\widehat{E}(z)$ | 0 | 0 | 0.0 | 0 | 0 |
| Uniform E(z) 100 | -0.003973 | -0.004643 | 0.001269 | 0.012524 | 0.02923 |
| Uniform D(z) 100 | 0.009763 | 0.029286 | 0.000551 | 0.014414 | 0.019828 |
| $E(z) \pm \sqrt{D(z)}$ | [0.102781; 0.094835] | [0.175775; 0.166489] | [0.022204; 0.024742] | [0.107534; 0.132582] | [0.111582; 0.170042] |
| $\widehat{E}(z)$ | 0.00 | 0.0 | 0.00 | 0.0 | 0.0 |
| Uniform E(z) 1000 | -0.001949 | -0.002912 | - 0.00696 | -0.000109 | 0.000962 |
| Uniform D(z) 1000 | 0.001006 | 0.00297 | 0.00006 | 0.001499 | 0.001982 |
| $E(z) \pm \sqrt{D(z)}$ | [0.033667; 0.029769] | [0.05741; 0.051586] | [0.002545; 0.002353] | [0.038826; 0.038608] | [0.043558; 0.045482] |
| $\widehat{E}(z)$ | 0.00 | 0.00 | 0.0 | 0.00 | 0.0 |

Таблица 5: Равномерное распределение (7)

5 Обсуждение

Исходя из данных, приведенных в таблицах, можно судить о том, что дисперсия характеристик рассеяния для распределения Коши является некой аномалией: значения слишком большие даже при увеличении размера выборки - понятно, что это результат выбросов, которые мы могли наблюдать в результатах предыдущего задания.

6 Приложение

Код программы GitHub URL:

<https://github.com/workivan/mat-ver-stat>

Санкт-Петербургский политехнический университет Петра Великого
Физико Механический институт
Высшая школа прикладной математики и вычислительной физики

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №3

по дисциплине
«Математическая статистика»

Выполнила студент
группы 5030102/90101

Кузин Иван Никитович

Проверил
Доцент, к.ф.-м.н.

Баженов Александр Николаевич

Санкт-Петербург
2022

Содержание

| | |
|--|----------|
| Список иллюстраций | 3 |
| 1 Постановка задачи | 4 |
| 2 Теория | 4 |
| 2.1 Боксплот Тьюки | 4 |
| 2.1.1 Определение | 4 |
| 2.1.2 Описание | 4 |
| 2.1.3 Построение | 4 |
| 2.2 Теоретическая вероятность выбросов | 5 |
| 3 Программная реализация | 5 |
| 4 Результаты | 6 |
| 4.1 Боксплот Тьюки | 6 |
| 4.2 Доля выбросов | 8 |
| 4.3 Теоретическая вероятность выбросов | 9 |
| 5 Обсуждение | 9 |
| 6 Приложение | 9 |

Список иллюстраций

| | | |
|---|-------------------------------------|---|
| 1 | Нормальное распределение | 6 |
| 2 | распределение Коши | 6 |
| 3 | распределение Лапласа | 7 |
| 4 | распределение Пуассона | 7 |
| 5 | равномерное распределение | 8 |

1 Постановка задачи

Для 5 распределений:

1. $N(x, 0, 1)$ – нормальное распределение
2. $C(x, 0, 1)$ – распределение Коши
3. $L(x, 0, \frac{1}{\sqrt{2}})$ – распределение Лапласа
4. $P(k, 10)$ – распределение Пуассона
5. $U(x, -\sqrt{3}, \sqrt{3})$ – равномерное распределение

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.

2 Теория

2.1 Боксплот Тьюки

2.1.1 Определение

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей

2.1.2 Описание

Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили и выбросы. Несколько таких ящичков можно нарисовать бок о бок, чтобы визуальнo сравнить одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящичка позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

2.1.3 Построение

Границами ящичка служат первый и третий квартили, линия в середине ящичка — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину «усов» определяют разность первого квартиля и полутора межквартильных расстояний и сумма третьего квартиля и полутора межквартильных расстояний. Формула имеет вид

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1) \quad (1)$$

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.

2.2 Теоретическая вероятность выбросов

Встроенными средствами языка программирования Python в среде разработки PyCharm можно вычислить теоретические первый и третий квартили распределений (Q_1^T и Q_3^T соответственно). По формуле (1) можно вычислить теоретические нижнюю и верхнюю границы уса (X_1^T и X_2^T соответственно). Выбросами считаются величины x , такие что:

$$\begin{cases} x < X_1^T \\ x > X_2^T \end{cases} \quad (2)$$

Теоретическая вероятность выбросов для непрерывных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T)) \quad (3)$$

где $F(X) = P(x \leq X)$ - функция распределения. Теоретическая вероятность выбросов для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T)) \quad (4)$$

где $F(X) = P(x \leq X)$ - функция распределения

3 Программная реализация

Лабораторная работа выполнена на языке Python версии 3.7 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. `scipy` - статические распределения и функции
2. `seaborn` - построение графиков, визуализация
3. `matplotlib` - построение графиков
4. `math` - использование математических функций

В приложении находится ссылка на GitHub репозиторий с исходным кодом.

4 Результаты

4.1 Боксплот Тьюки

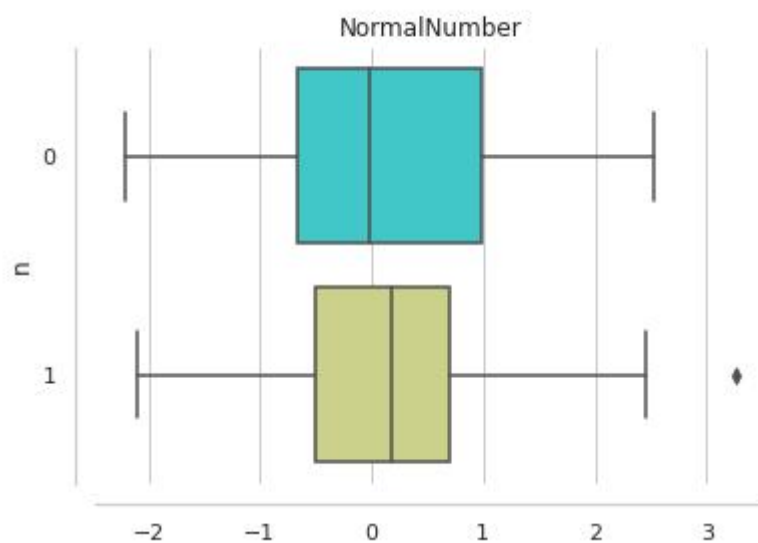


Рис. 1: Нормальное распределение

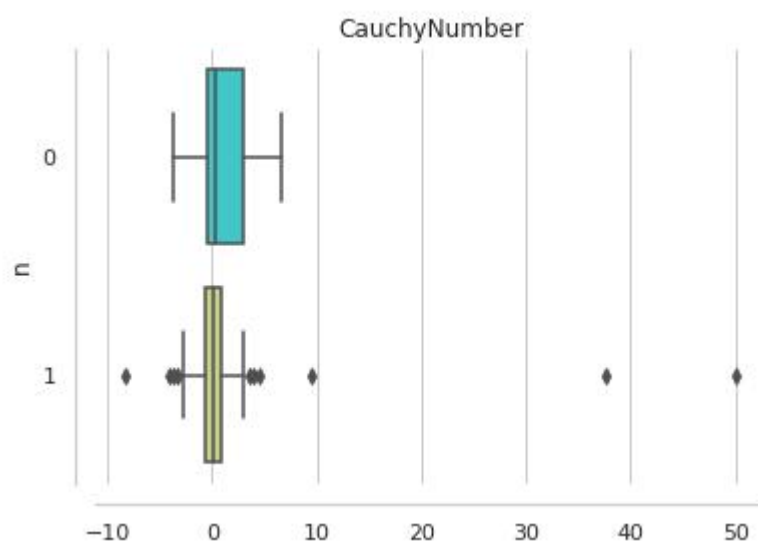


Рис. 2: распределение Коши

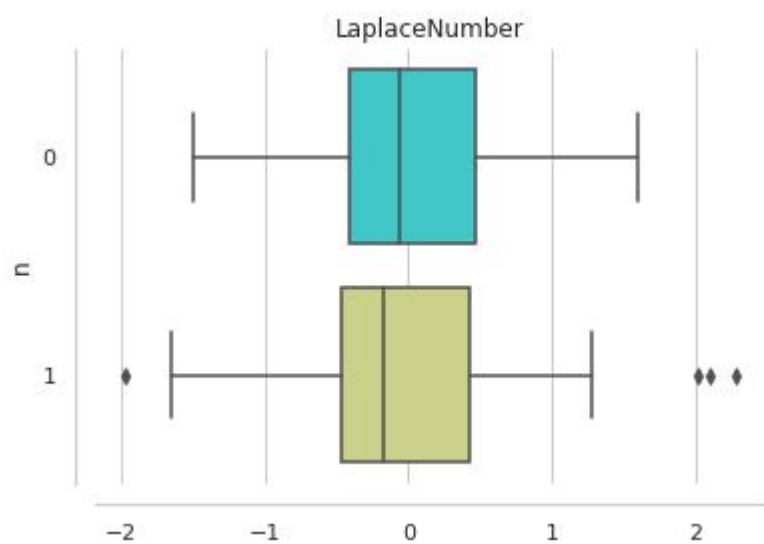


Рис. 3: распределение Лапласа

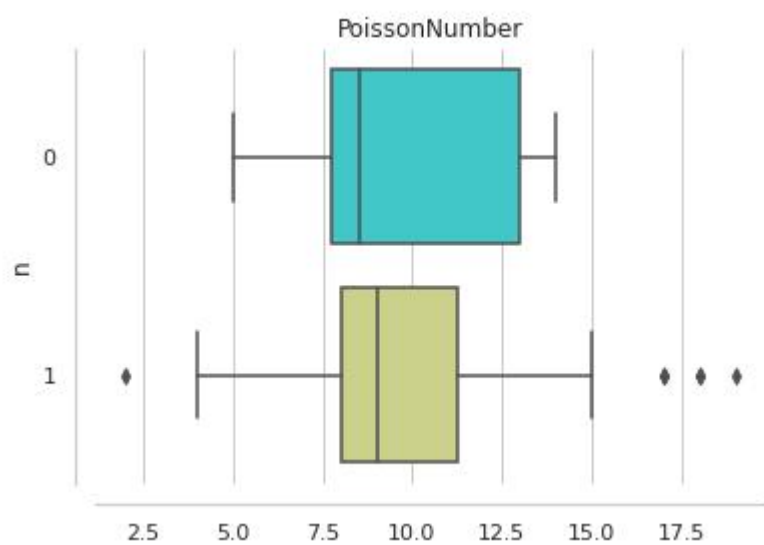


Рис. 4: распределение Пуассона

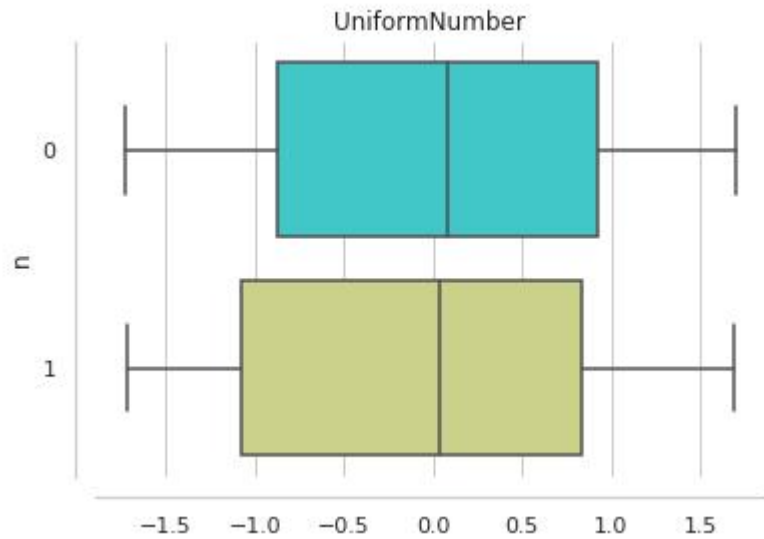


Рис. 5: равномерное распределение

4.2 Доля выбросов

Округление доли выбросов:

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

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

Для $n = 20$: $p_n = \frac{1}{\sqrt{20}}$ - 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 | 0.024 | 0.007 |
| Normal n=100 | 0.014 | 0.007 |
| Cauchy n=20 | 0.151 | 0.156 |
| Cauchy n=100 | 0.185 | 0.156 |
| Laplace n=20 | 0.075 | 0.063 |
| Laplace n=100 | 0.081 | 0.063 |
| Poisson n=20 | 0.024 | 0.008 |
| Poisson n=100 | 0.016 | 0.008 |
| Uniform n=20 | 0.002 | 0 |
| Uniform n=100 | 0 | 0 |

Таблица 1: Доля выбросов

4.3 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 2: Теоретическая вероятность выбросов

5 Обсуждение

По данным, приведенным в таблице, можно сказать, что чем больше выборка (в нашем случае для 100 элементов), тем ближе доля выбросов будет к теоретической оценке. Снова доля выбросов для распределения Коши значительно выше, чем для остальных распределений. Для распределений: нормального, Лапласа и Пуассона погрешность при большой выборке составила не более 2 процентов. При увеличении выборки равномерное распределение показывает стремительный рост к теоретической оценке - выбросы практически не наблюдаются.

Ящички с «усами» в удобной форме показывает многие важные характеристики выборки, такие как медиана, первый и третий квартили и другие. Исходя из которых можно делать выводы касательно природы входных данных, распределений.

6 Приложение

Код программы GitHub URL:

<https://github.com/workivan/mat-ver-stat.git>

Санкт-Петербургский политехнический университет Петра Великого
Физико Механический институт
Высшая школа прикладной математики и вычислительной физики

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №4

по дисциплине
«Математическая статистика»

Выполнила студент
группы 5030102/90101

Кузин Иван Никитович

Проверил
Доцент, к.ф.-м.н.

Баженов Александр Николаевич

Санкт-Петербург
2022

СОДЕРЖАНИЕ

| | |
|--|-----------|
| СПИСОК ИЛЛЮСТРАЦИЙ | 3 |
| 1 Постановка задачи | 4 |
| 2 Теория | 4 |
| 2.1 Эмпирическая функция распределения | 4 |
| 2.1.1 Статистический ряд | 4 |
| 2.1.2 Эмпирическая функция распределения | 4 |
| 2.1.3 Нахождение э. ф. р. | 4 |
| 2.2 Оценки плотности вероятности | 5 |
| 2.2.1 Определение | 5 |
| 2.2.2 Ядерные оценки | 5 |
| 3 Программная реализация | 5 |
| 4 Результаты | 6 |
| 4.1 Эмпирическая функция распределения | 6 |
| 4.2 Ядерные оценки плотности распределения | 8 |
| 5 Обсуждение | 15 |
| 6 Приложение | 16 |

СПИСОК ИЛЛЮСТРАЦИЙ

| | | |
|----|--|----|
| 1 | нормальное распределение | 6 |
| 2 | распределение Коши | 6 |
| 3 | распределение Лапласа | 7 |
| 4 | распределение Пуассона | 7 |
| 5 | равномерное распределение | 8 |
| 6 | нормальное распределение, $n = 20$ | 8 |
| 7 | нормальное распределение, $n = 60$ | 9 |
| 8 | нормальное распределение, $n = 100$ | 9 |
| 9 | распределение Коши, $n = 20$ | 10 |
| 10 | распределение Коши, $n = 60$ | 10 |
| 11 | распределение Коши, $n = 100$ | 11 |
| 12 | распределение Лапласа, $n = 20$ | 11 |
| 13 | распределение Лапласа, $n = 60$ | 12 |
| 14 | распределение Лапласа, $n = 100$ | 12 |
| 15 | распределение Пуассона, $n = 20$ | 13 |
| 16 | распределение Пуассона, $n = 60$ | 13 |
| 17 | распределение Пуассона, $n = 100$ | 14 |
| 18 | равномерное распределение, $n = 20$ | 14 |
| 19 | равномерное распределение, $n = 60$ | 15 |
| 20 | равномерное распределение, $n = 100$ | 15 |

1 Постановка задачи

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 Эмпирическая функция распределения

2.1.1 Статистический ряд

Статистическим рядом назовем совокупность, состоящую из последовательности $\{z_i\}_{i=1}^k$ попарно различных элементов выборки, расположенных по возрастанию, и последовательности $\{n_i\}_{i=1}^k$ частот, с которыми эти элементы содержатся в выборке.

2.1.2 Эмпирическая функция распределения

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

$$F_n^*(x) = P^*(X < x). \quad (1)$$

2.1.3 Нахождение э. ф. р.

$$F_n^*(x) = \frac{1}{n} \sum_{z_i < x} n_i. \quad (2)$$

$F_n^*(x)$ — 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

$$F_n^*(x) \approx F_X(x). \quad (3)$$

2.2 Оценки плотности вероятности

2.2.1 Определение

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

$$\hat{f}(x) \approx f(x). \quad (4)$$

2.2.2 Ядерные оценки

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

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - x_i}{h_n}\right). \quad (5)$$

$K(u)$ - ядро, т. е. непрерывная функция, являющаяся плотностью вероятности, x_1, \dots, x_n - элементы выборки, а $\{h_n\}_{n \in \mathbb{N}}$ - последовательность элементов из \mathbb{R}_+ такая, что

$$h_n \xrightarrow{n \rightarrow \infty} 0; \quad nh_n \xrightarrow{n \rightarrow \infty} \infty. \quad (6)$$

Такие оценки называются непрерывными ядерными.

Гауссово ядро:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \quad (7)$$

Правило Сильвермана:

$$h_n = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5}, \quad (8)$$

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

3 Программная реализация

Лабораторная работа выполнена на языке Python версии 3.7 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. `scipy` (генерация выборок)
2. `statsmodels` (построение эмпирических функций распределения)
3. `matplotlib`, `seaborn` (визуализация)
4. `numpy` (вычисление ряда числовых характеристик)

В приложении находится ссылка на GitHub репозиторий с исходным кодом.

4 Результаты

4.1 Эмпирическая функция распределения

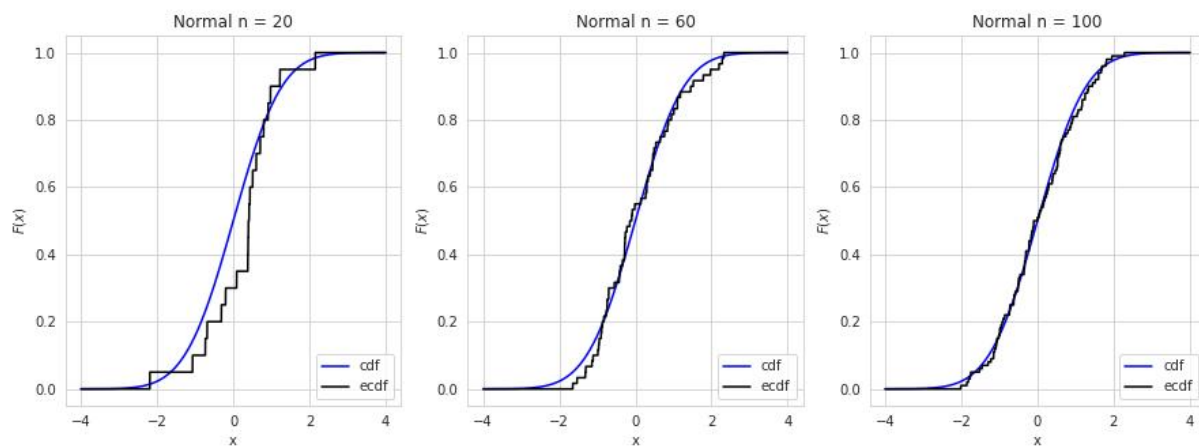


Рис. 1: нормальное распределение

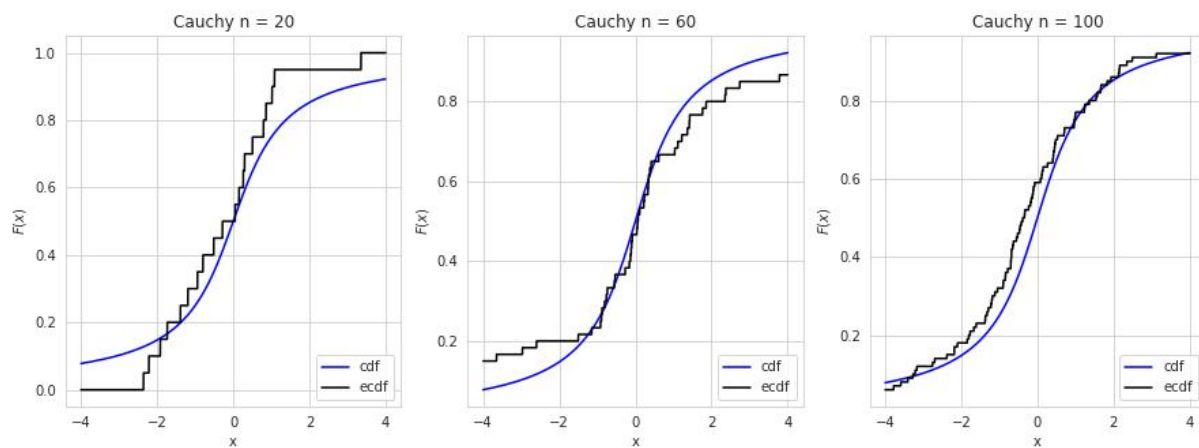


Рис. 2: распределение Коши

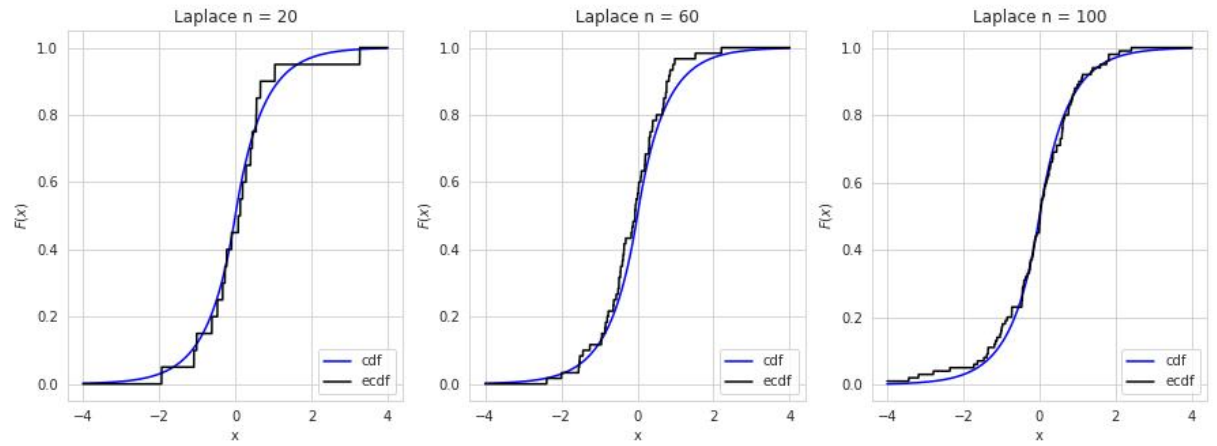


Рис. 3: распределение Лапласа

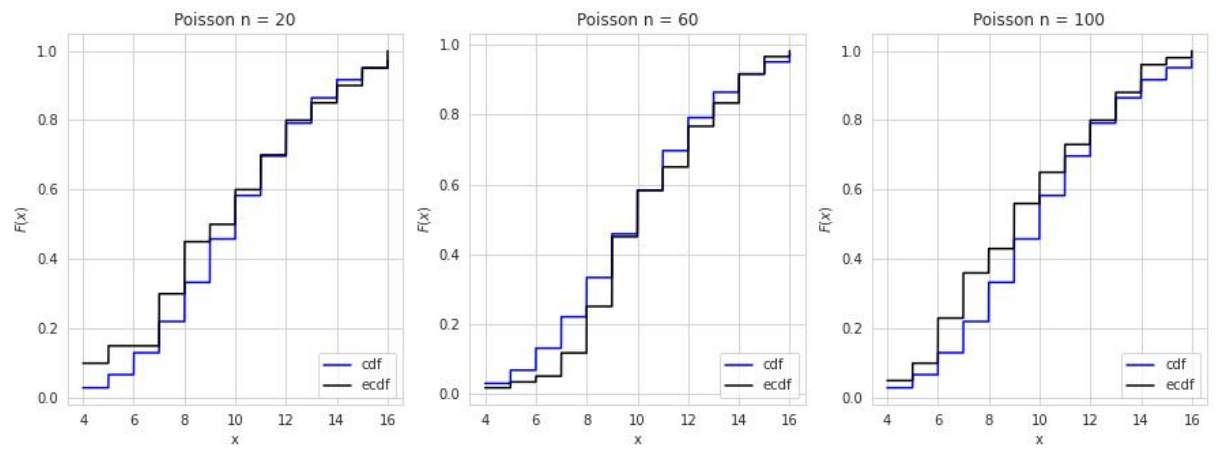


Рис. 4: распределение Пуассона

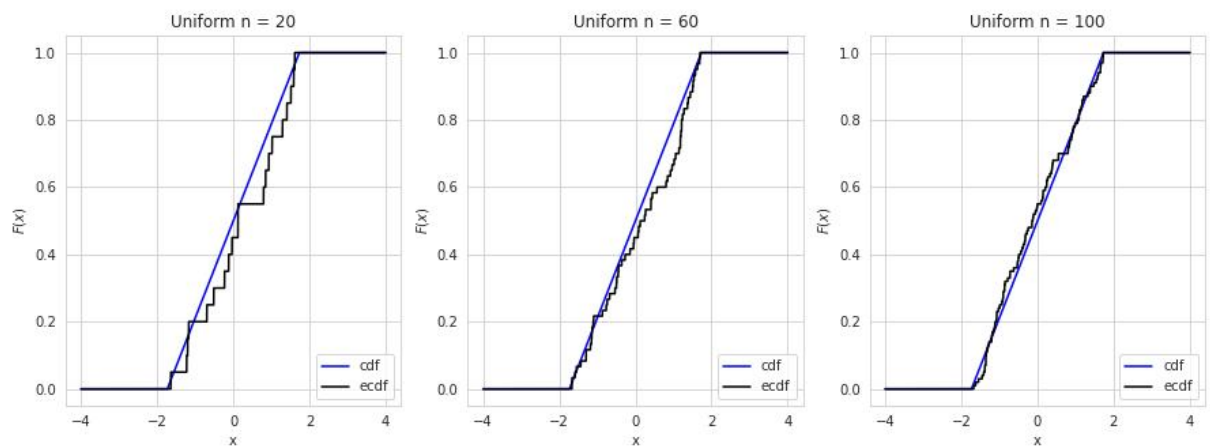


Рис. 5: равномерное распределение

4.2 Ядерные оценки плотности распределения

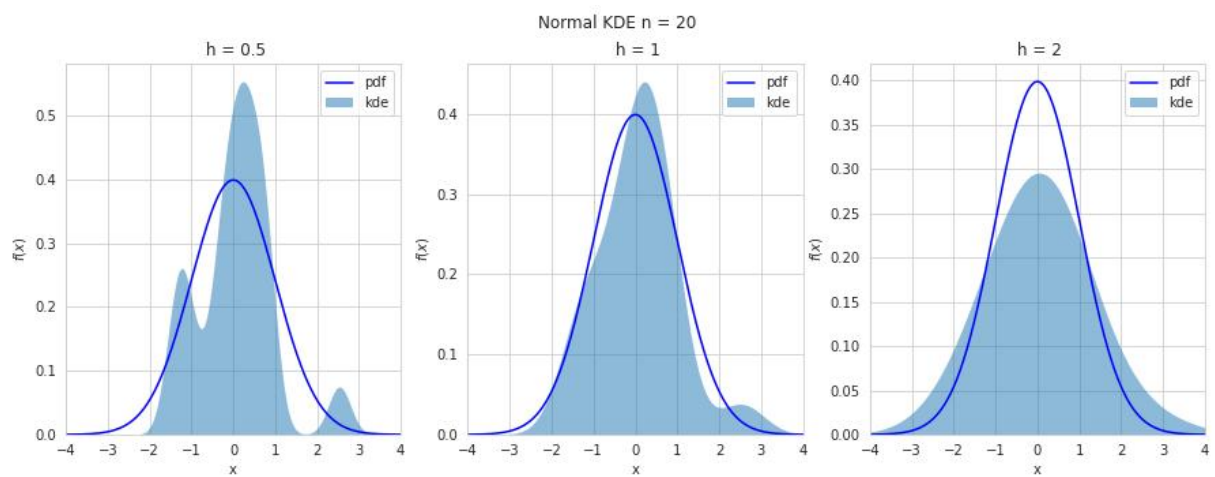


Рис. 6: нормальное распределение, $n = 20$

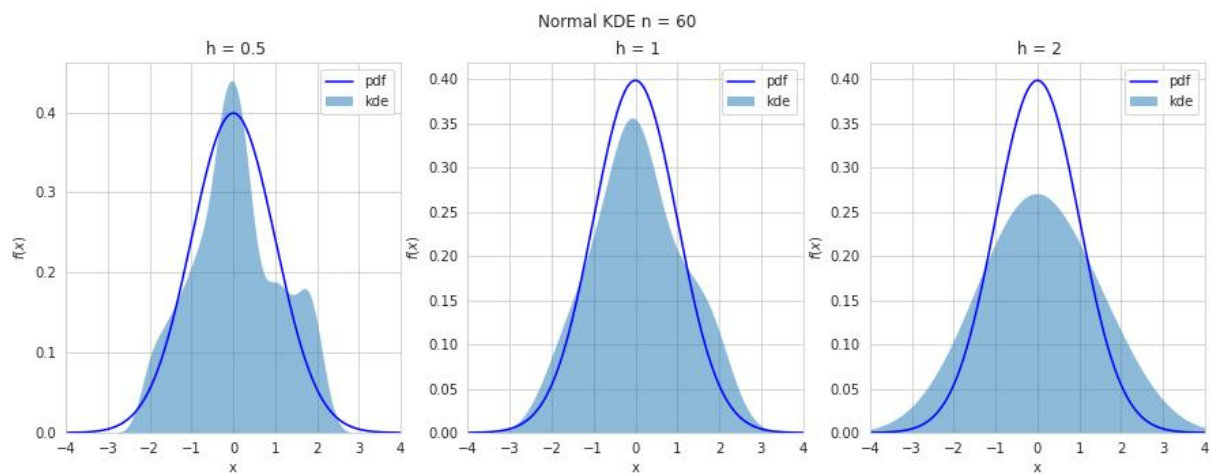


Рис. 7: нормальное распределение, $n = 60$

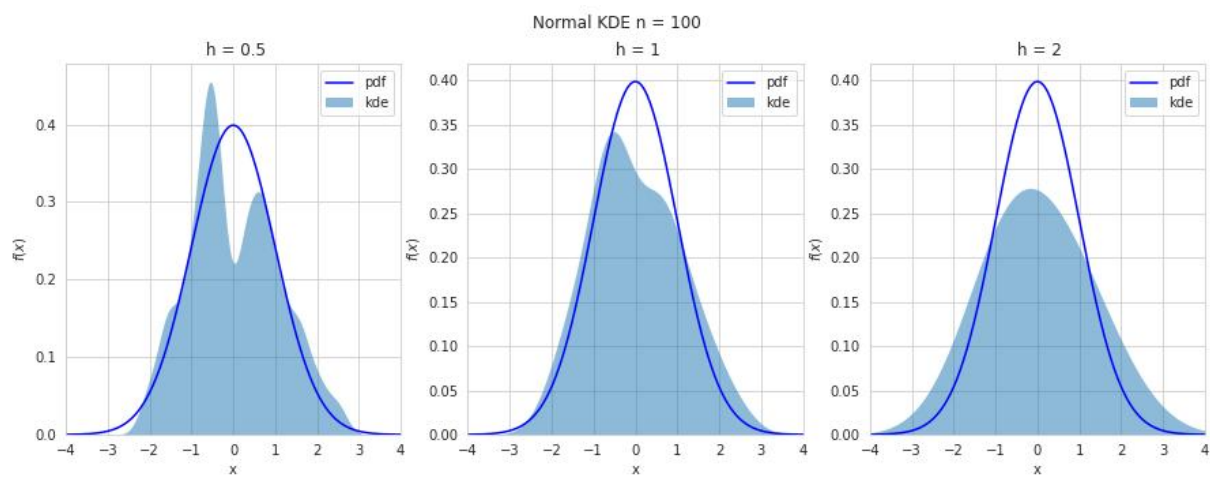


Рис. 8: нормальное распределение, $n = 100$

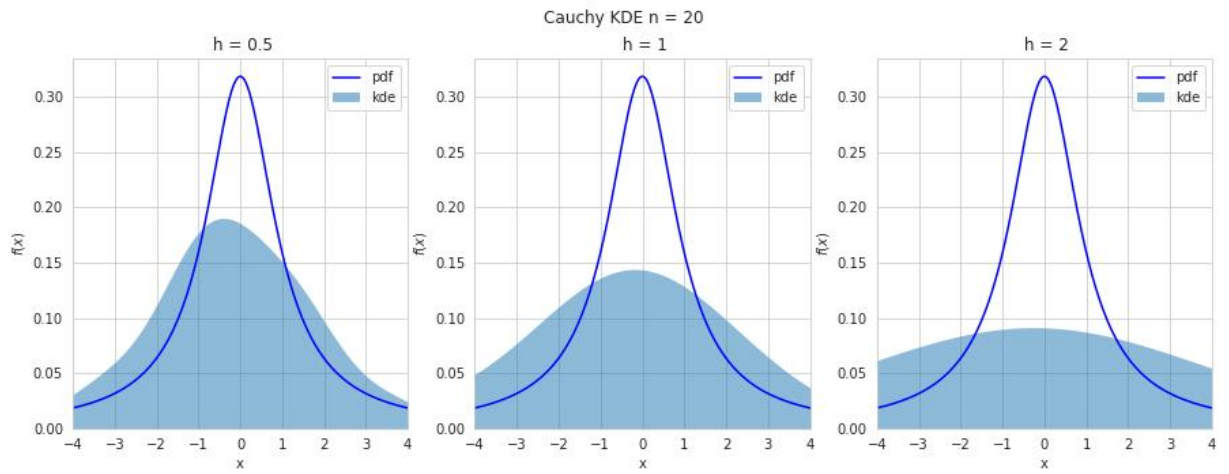


Рис. 9: распределение Коши, $n = 20$

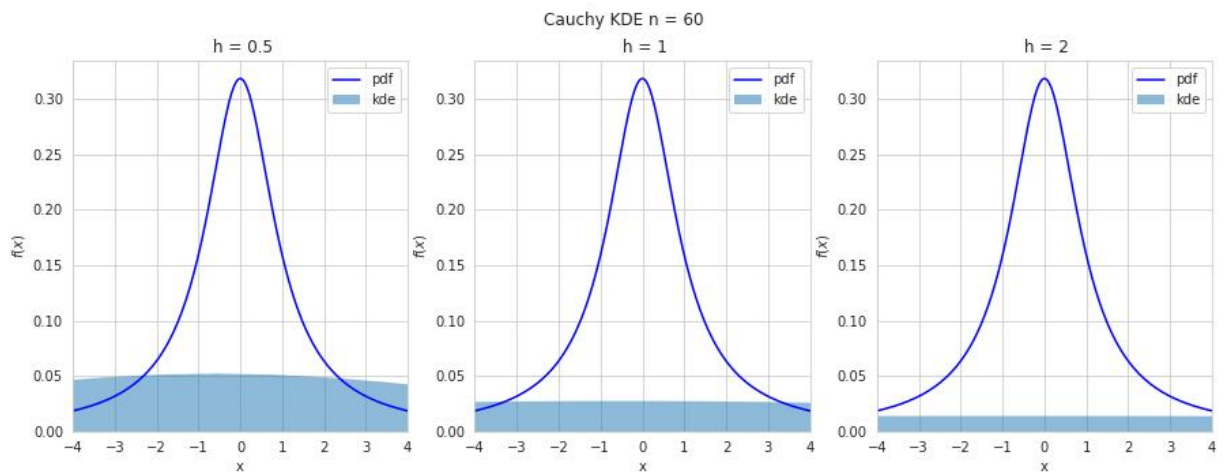


Рис. 10: распределение Коши, $n = 60$

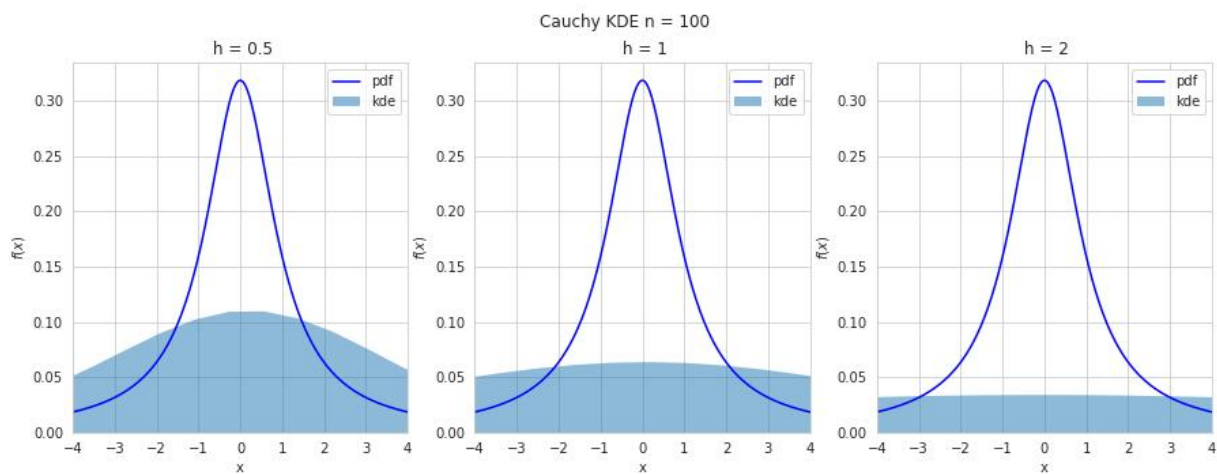


Рис. 11: распределение Коши, $n = 100$

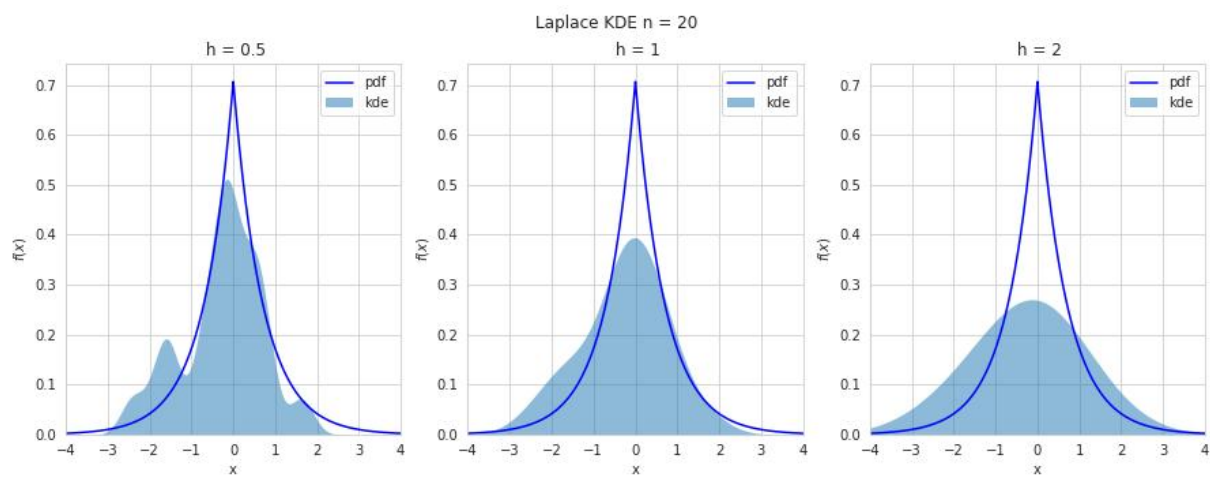


Рис. 12: распределение Лапласа, $n = 20$

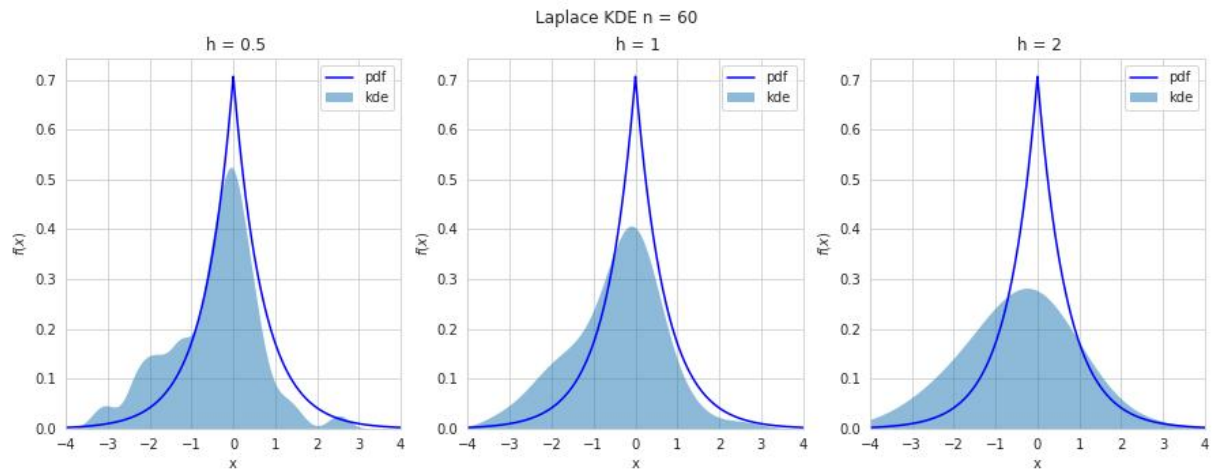


Рис. 13: распределение Лапласа, $n = 60$

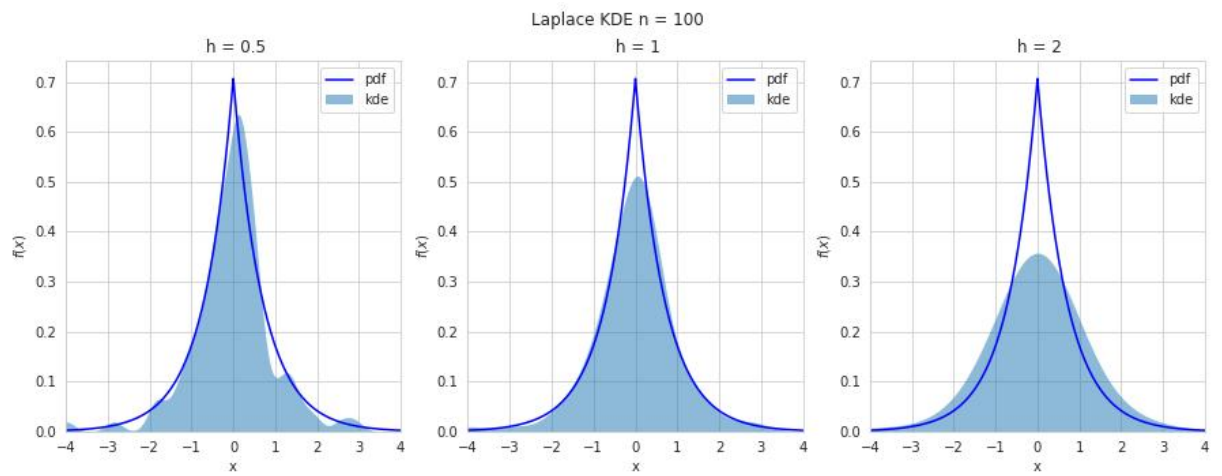


Рис. 14: распределение Лапласа, $n = 100$

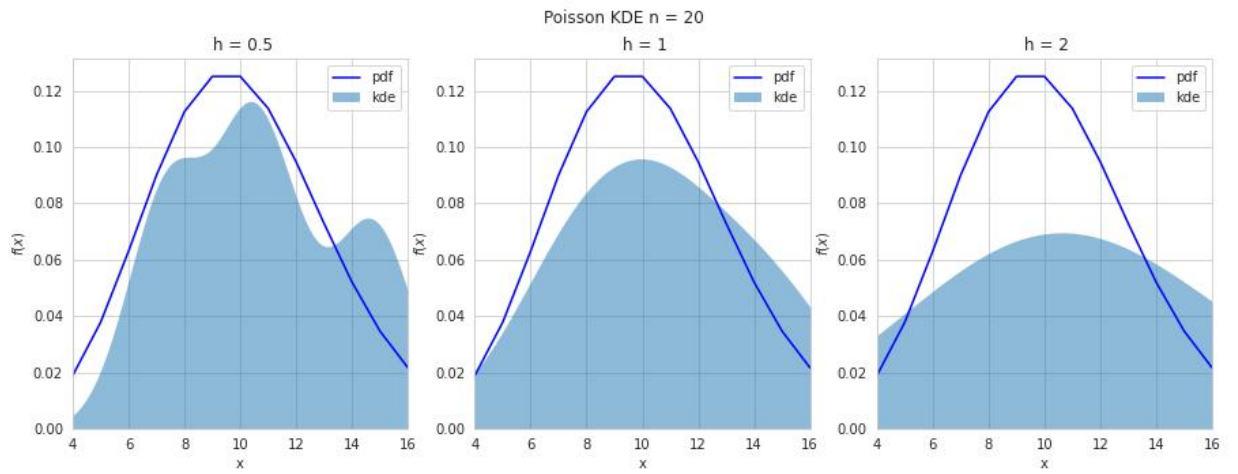


Рис. 15: распределение Пуассона, $n = 20$

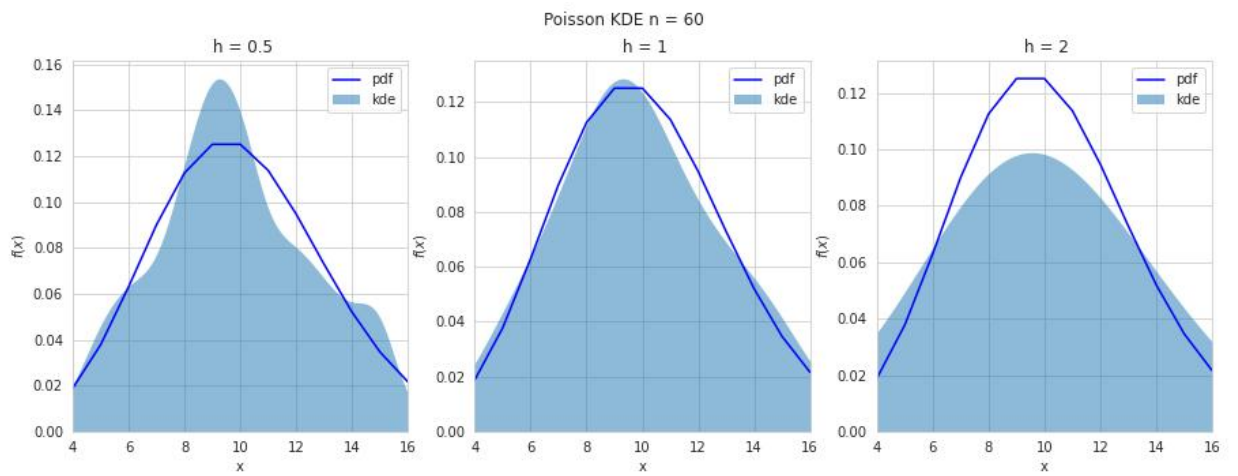


Рис. 16: распределение Пуассона, $n = 60$

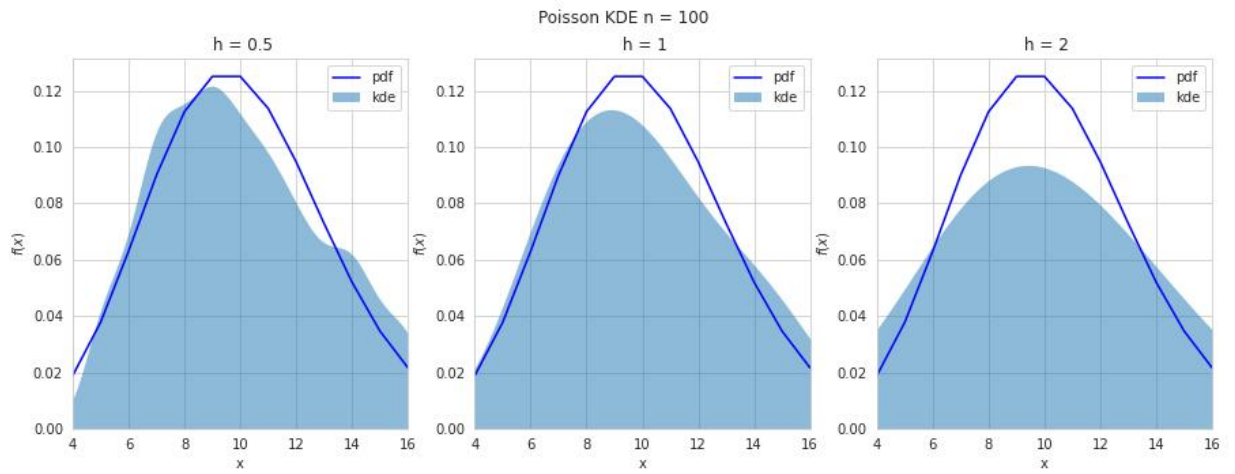


Рис. 17: распределение Пуассона, $n = 100$

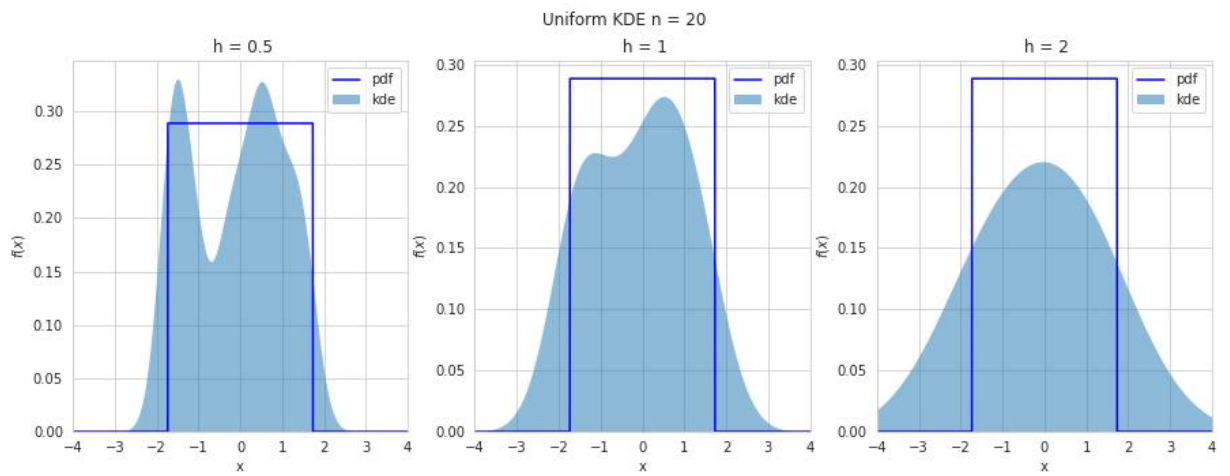
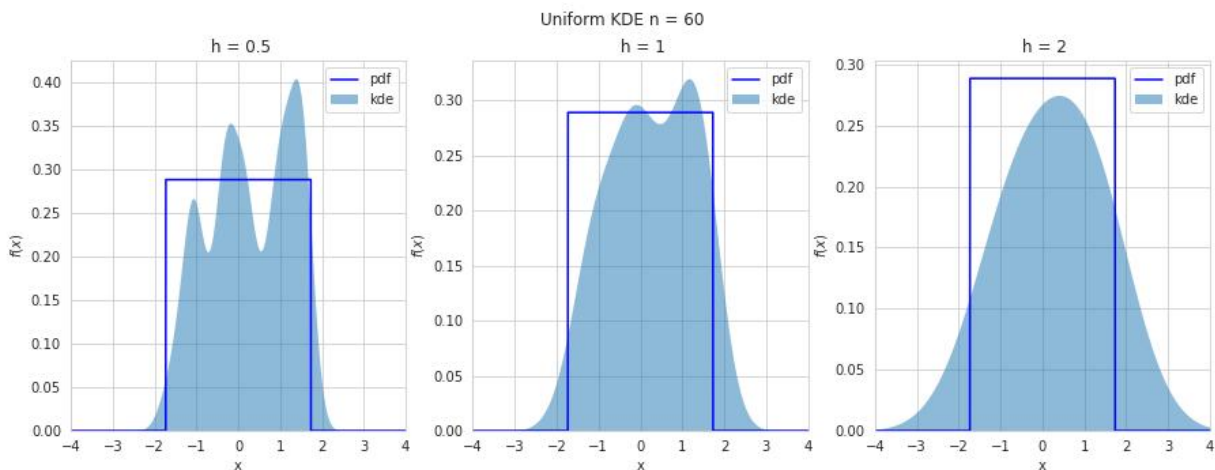
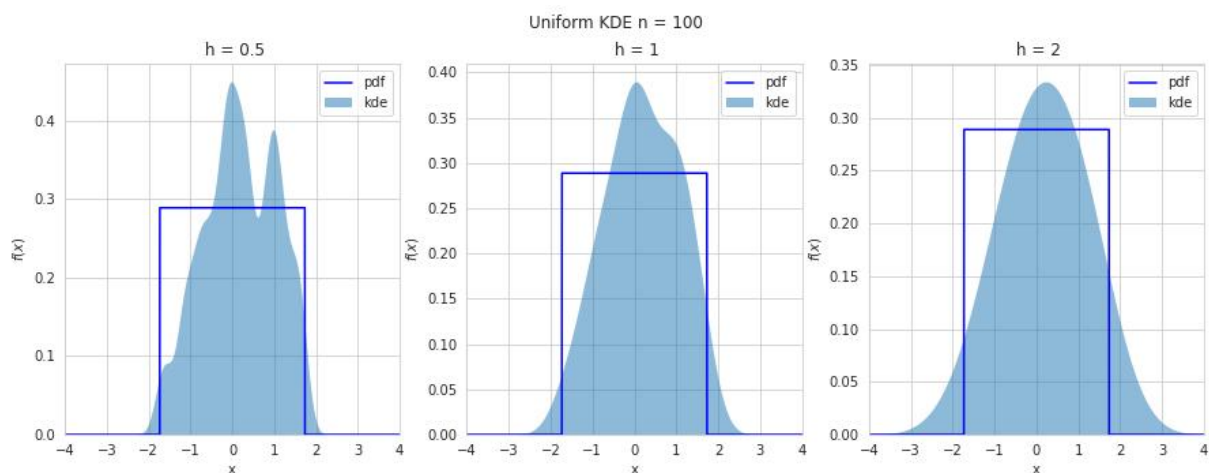


Рис. 18: равномерное распределение, $n = 20$

Рис. 19: равномерное распределение, $n = 60$ Рис. 20: равномерное распределение, $n = 100$

5 Обсуждение

Можем наблюдать на иллюстрациях с эмпирическими функциями, 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 нормального, равномерного и пуассоновского распределений оптимальным значением параметра является $h = h_n$, а для распределений Коши и Лапласа - $h = h_n/2$.

Также можно увидеть, что чем больше коэффициент при параметре сглаживания \hat{h}_n , тем меньше изменений знака производной у аппроксимирующей функции, вплоть до того, что при $h = h_n/2$ 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 Приложение

Код программы GitHub URL:

<https://github.com/workivan/mat-ver-stat>