Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт
Кафедра «Прикладная математика»

ЛАБОРАТОРНАЯ РАБОТА: 1-4
ПО ДИСЦИПЛИНЕ
«МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»

Выполнил студент
Келарев Михаил Алексеевич
группы 5030102/90101

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

Санкт-Петербург
2022

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Для 5 распределений:

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

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: $\overline{x}, medx, z_R, z_Q, z_{tr}$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

   $$E(z) = \overline{z}$$

   Вычислить оценку дисперсии по формуле:

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

   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 определения

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Лапласа

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

- Распределение Пуассона

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

- Равномерное распределение

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

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.

2.3 Вариационный ряд

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

2.4 Выборочные числовые характеристики

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

2.4.1 Характеристики положения

- Выборочное среднее
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
- Выборочная медиана
$$medx = \begin{cases} x_{(l+1)} & n = 2l + 1 \\ \frac{x_{(l)} + x_{(l+1)}}{2} & n = 2l \end{cases}$$
- Полусумма экстремальных выборочных элементов
$$z_R = \frac{x_{(1)} + x_{(n)}}{2}$$
- Полусумма квартилей
Выборочная квартиль z_p порядка p определяется формулой
$$z_p = \begin{cases} x_{([np]+1)} & np\text{–дробное} \\ x_{(np)} & np\text{–целое} \end{cases}$$
Полусумма квартилей
$$z_Q = \frac{z_{1/4} + z_{3/4}}{2}$$

- Усечённое среднее

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

2.4.2 Характеристики рассеяния

Выборочная дисперсия

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

2.5 Боксплот Тьюки

2.5.1 Определение

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

2.5.2 Описание

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

2.5.3 Построение

Границами ящика служат первый и третий квартили, линия в середине ящика — 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.

2.6 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:

$$\begin{bmatrix} x < X_1^T \\ x > X_2^T \end{bmatrix}$$

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

$$P_B^T = P(x < X_1^T) + P(x > x_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T))$$

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

2.7 Эмпирическая функция распределения

2.7.1 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.

2.7.2 Эмпирическая функция распределения

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

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

2.7.3 Нахождение э. ф. р.

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

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

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

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

Таблица 1: Таблица распределения

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

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

2.8 Оценки плотности вероятности

2.8.1 Определение

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

$$\widehat{f}(x) \approx f(x).$$

2.8.2 Ядерные оценки

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

$$\widehat{f_n}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right).$$

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

$$h_n \xrightarrow[n\to\infty]{} 0; \quad nh_n \xrightarrow[n\to\infty]{} \infty.$$

Такие оценки называются непрерывными ядерными.

Гауссово ядро:

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

Правило Сильвермана:

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

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

3 Реализация

Лабораторная работа выполнена на языке Python версии 3.10.
Использовались дополнительные библиотеки:

1. scipy

2. numpy

3. matplotlib

4. math

4 Результат

4.1 Гистограммы и графики плотности распределения

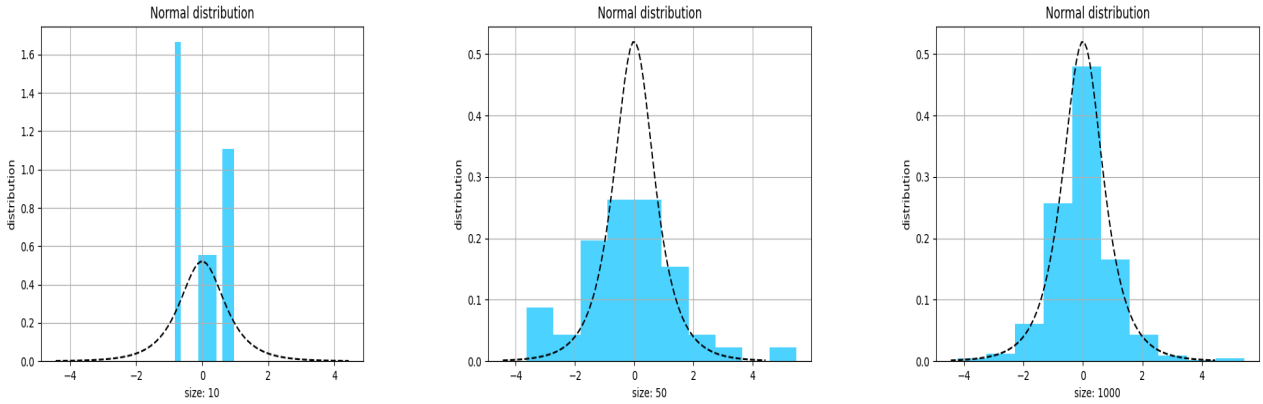

Рис. 1: Нормальное распределение

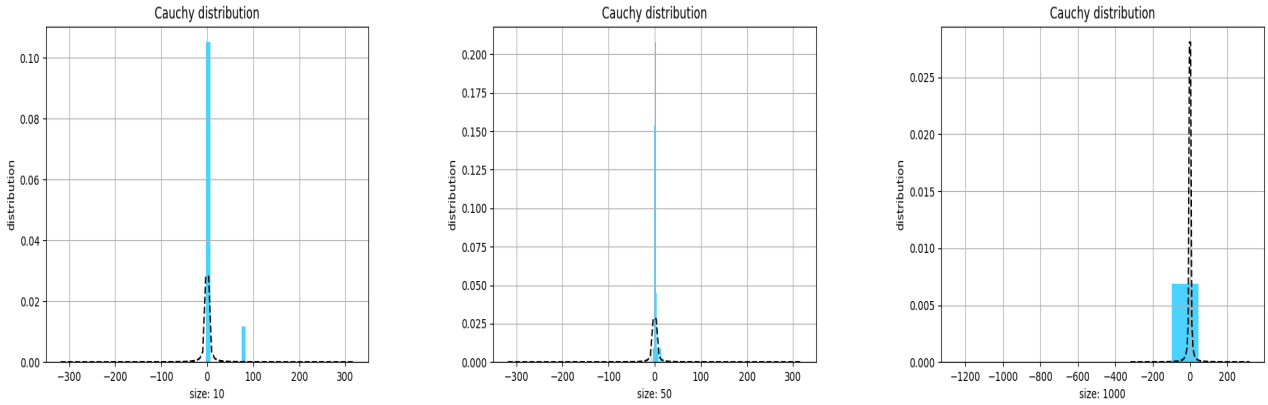

Рис. 2: Распределение Коши

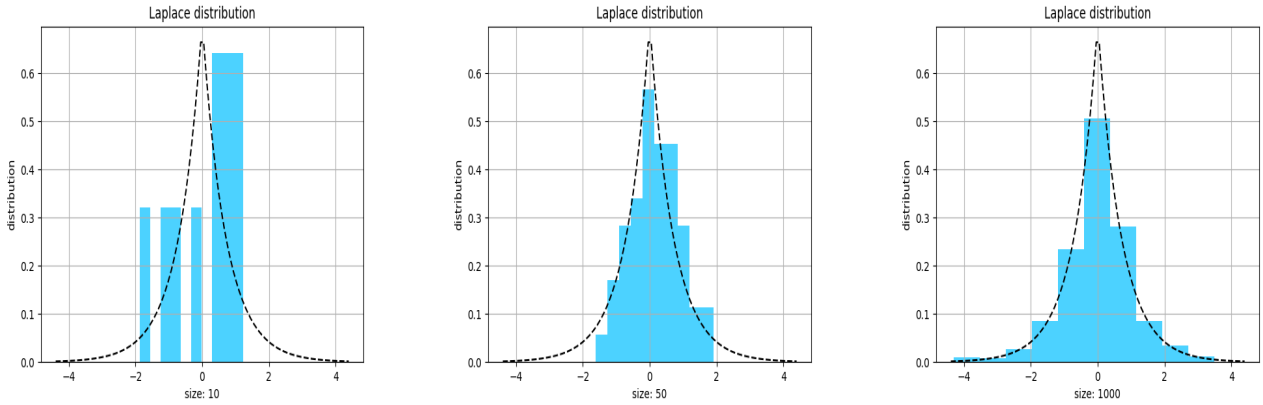

Рис. 3: Распределение Лапласа

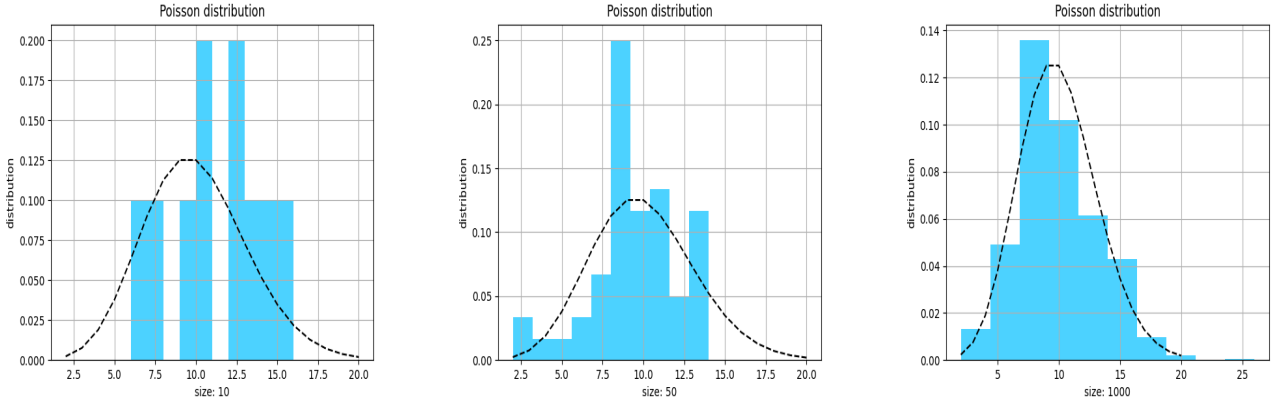

Рис. 4: Распределение Пауссона

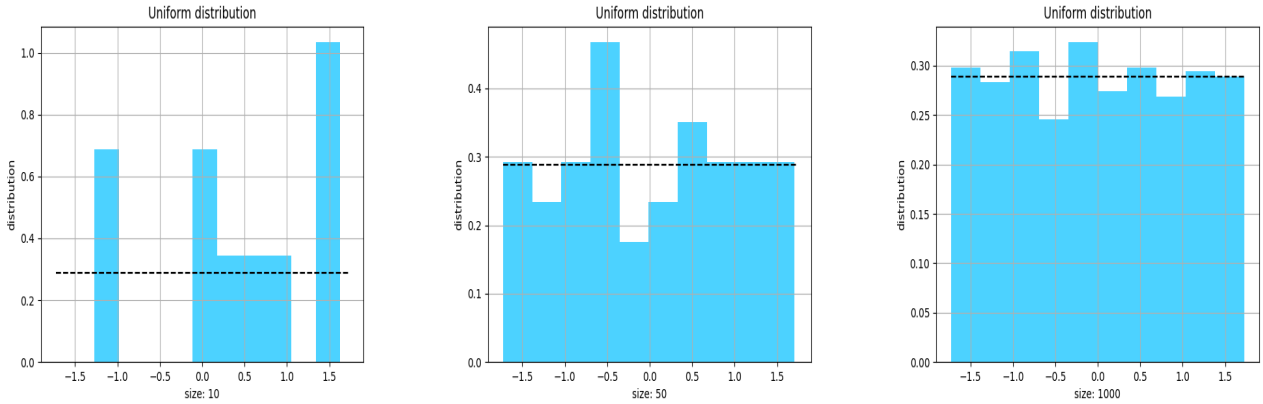

Рис. 5: Равномерное распределение

4.2 Характеристики положения и рассеяния

	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
n=10					
$E(z)$	0.00783	0.000649	0.011704	0.316865	0.279662
$D(z)$	0.101986	0.138258	0.182348	0.130745	0.118405
n=100					
$E(z)$	0.003355	0.005395	-0.002108	0.018576	0.032072
$D(z)$	0.009793	0.014229	0.090936	0.012374	0.011337
n=1000					
$E(z)$	-0.000548	-0.000252	-0.000612	0.001267	0.002887
$D(z)$	0.000986	0.001567	0.061974	0.001201	0.00119

Таблица 2: Нормальное распределение

	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
n=10					
$E(z)$	0.28875	-0.005228	1.474197	1.273873	0.737032
$D(z)$	93.628602	0.336607	2187.372982	15.339821	2.373734
n=100					
$E(z)$	-1.271446	-0.007581	-62.483751	0.016789	0.030006
$D(z)$	612.101245	0.02479	1508295.961798	0.053915	0.027198
n=1000					
$E(z)$	2.811126	0.001284	1382.966803	0.004563	0.005812
$D(z)$	3054.863093	0.002503	761373096.375691	0.005074	0.002654

Таблица 3: Распределение Коши

	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
n=10					
$E(z)$	0.009953	0.006635	0.010806	0.312161	0.246866
$D(z)$	0.098919	0.07654	0.406287	0.117389	0.082069
n=100					
$E(z)$	0.003592	0.001682	0.012879	0.018669	0.022897
$D(z)$	0.010065	0.00539	0.439571	0.009504	0.005988
n=1000					
$E(z)$	0.001955	0.000966	0.016544	0.004052	0.003654
$D(z)$	0.00102	0.000499	0.424402	0.001079	0.000609

Таблица 4: Распределение Лапласа

	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
n=10					
$E(z)$	9.9483	9.8055	10.2655	10.8815	10.7075
$D(z)$	0.903857	1.31842	1.86576	1.240708	1.10986
n=100					
$E(z)$	9.98972	9.836	10.947	9.9465	9.93686
$D(z)$	0.107242	0.224604	1.044691	0.167388	0.131238
n=1000					
$E(z)$	10.003314	9.9945	11.662	9.995	9.867934
$D(z)$	0.009953	0.00522	0.655256	0.002475	0.011043

Таблица 5: Распределение Пуасона

	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
n=10					
$E(z)$	-0.002391	-0.015133	0.003609	0.319112	0.310035
$D(z)$	0.107682	0.245241	0.044913	0.133545	0.165372
n=100					
$E(z)$	-0.001485	-0.002167	0.001143	0.012677	0.0321
$D(z)$	0.010127	0.02873	0.000586	0.015531	0.019983
n=1000					
$E(z)$	0.001444	0.001939	1.5e-05	0.003171	0.005647
$D(z)$	0.001067	0.003119	6e-06	0.001656	0.002112

Таблица 6: Нормальное распределение

4.3 Боксплот Тьюки

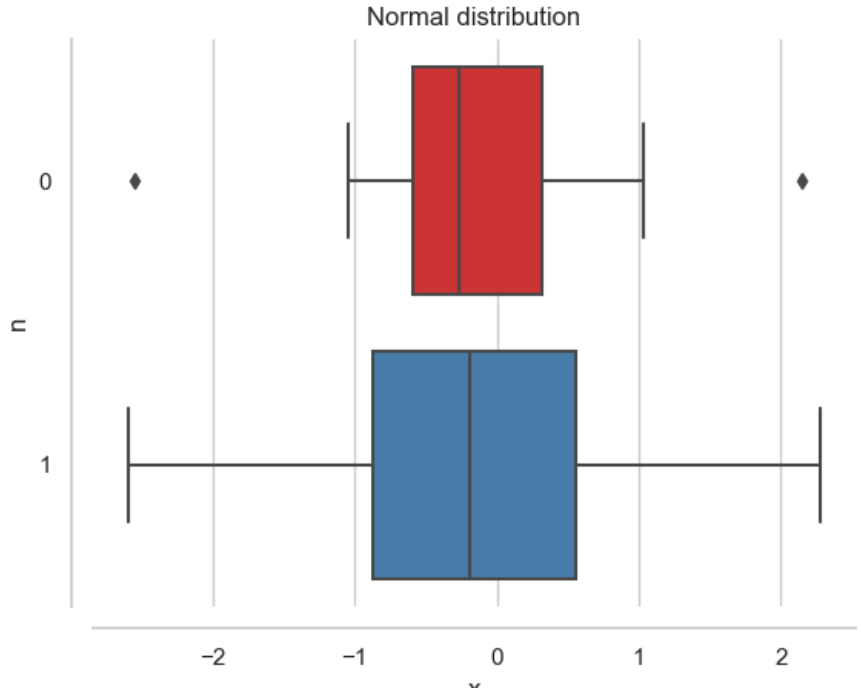

Рис. 6: Нормальное распределение

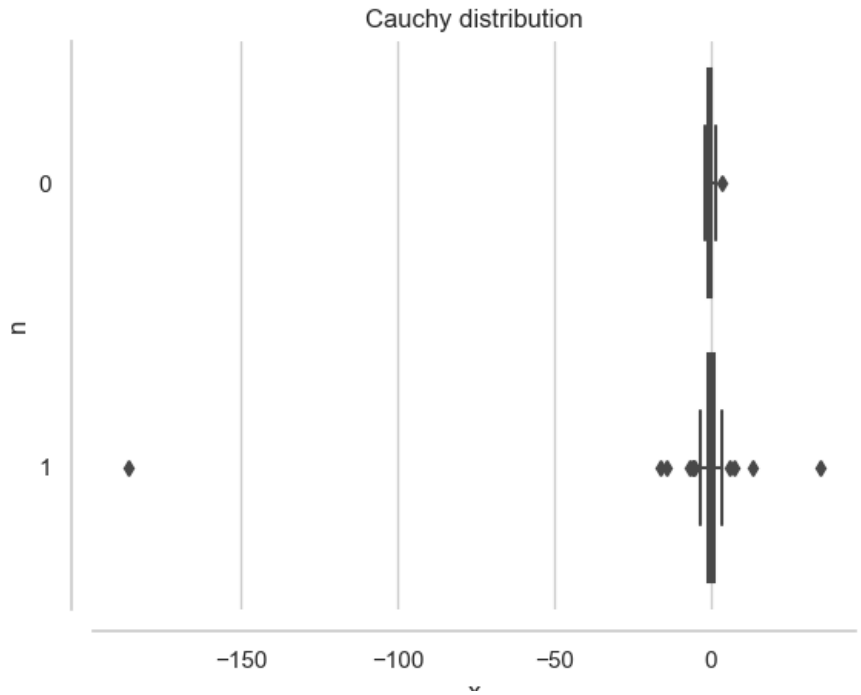

Рис. 7: Распределение Коши

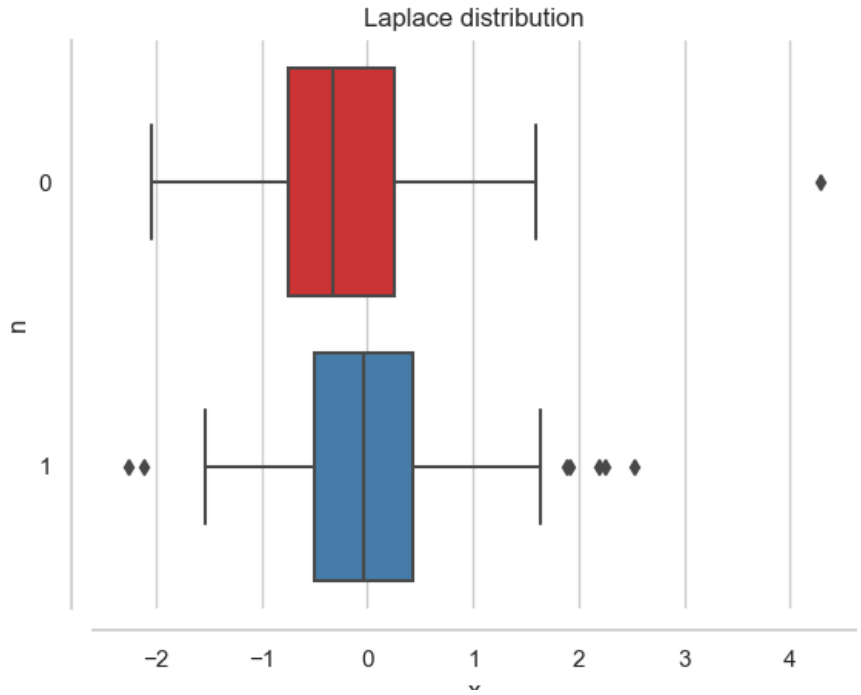

Рис. 8: Распределение Лапласа

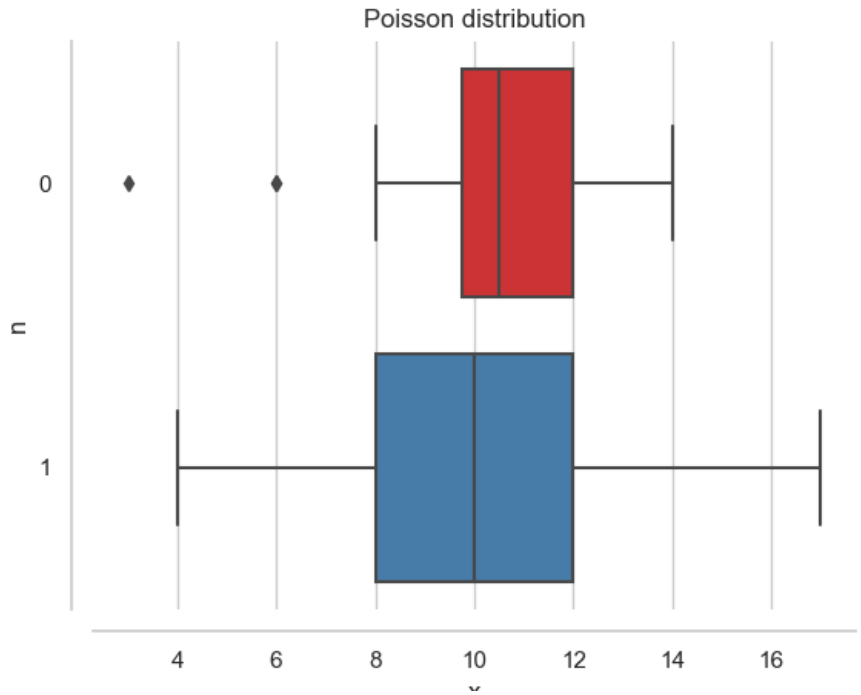

Рис. 9: Распределение Пуассона

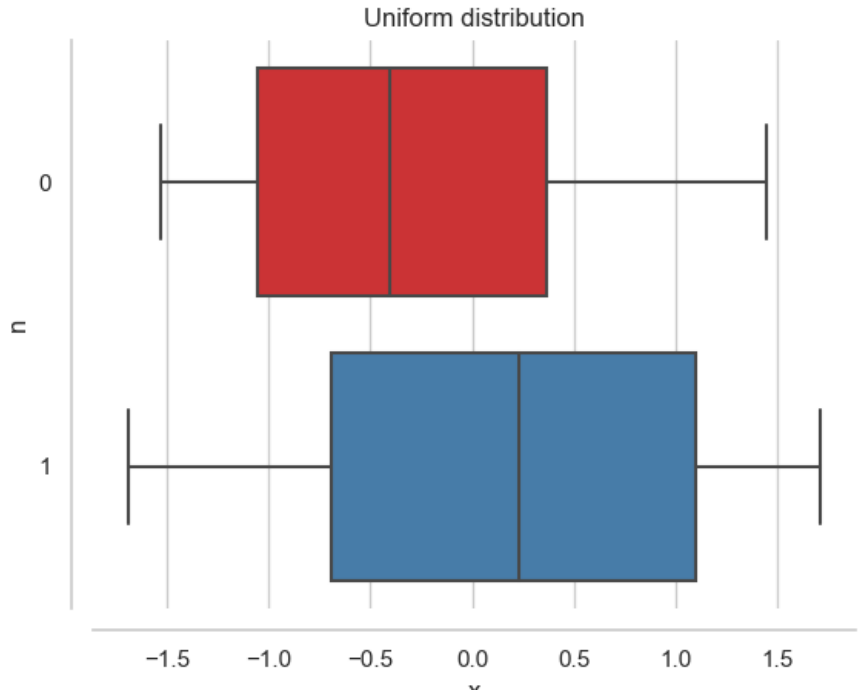

Рис. 10: Равномерное распределение

4.4 Доля выбросов

Выборка	Доля выбросов
Normal n = 20	0.0239
Normal n = 100	0.0099
Cauchy n = 20	0.1476
Cauchy n = 100	0.1541
Laplace n = 20	0.0771
Laplace n = 100	0.065
Poisson n = 20	0.022
Poisson n = 100	0.0111
Uniform n = 20	0.0026
Uniform n = 100	0

Таблица 7: Практическая доля выбросов

4.5 Теоретическая вероятность выбросов

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

Таблица 8: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

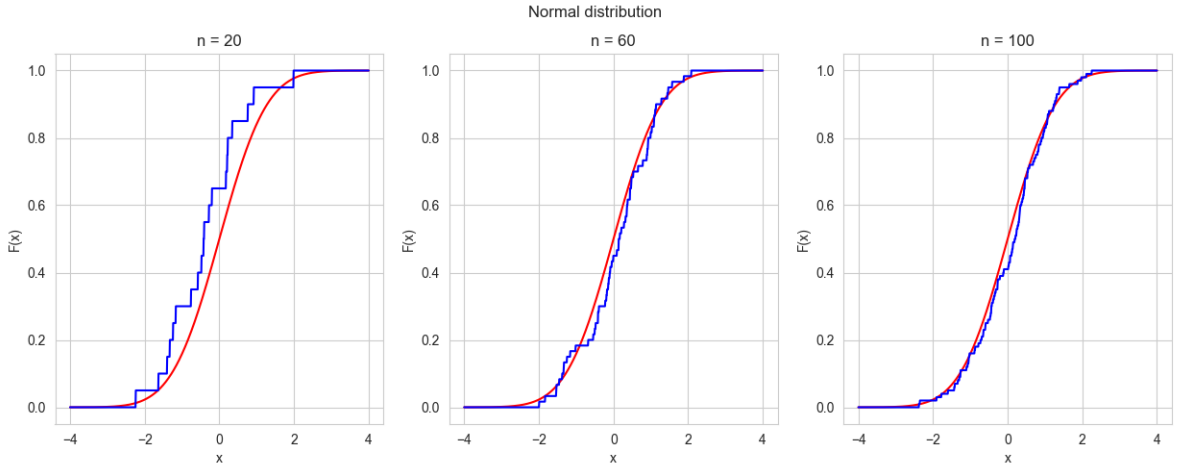

Рис. 11: Нормальное распределение

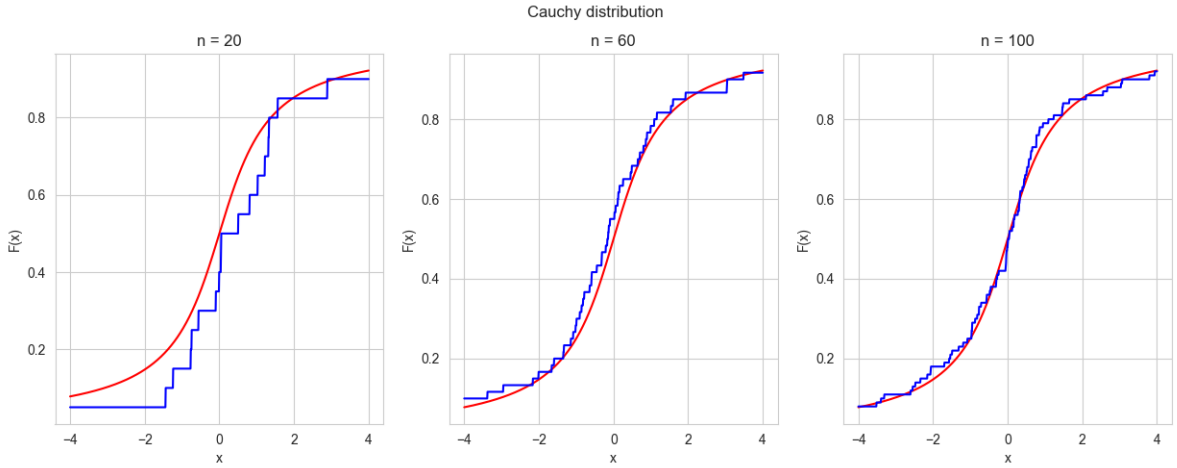

Рис. 12: Распределение Коши

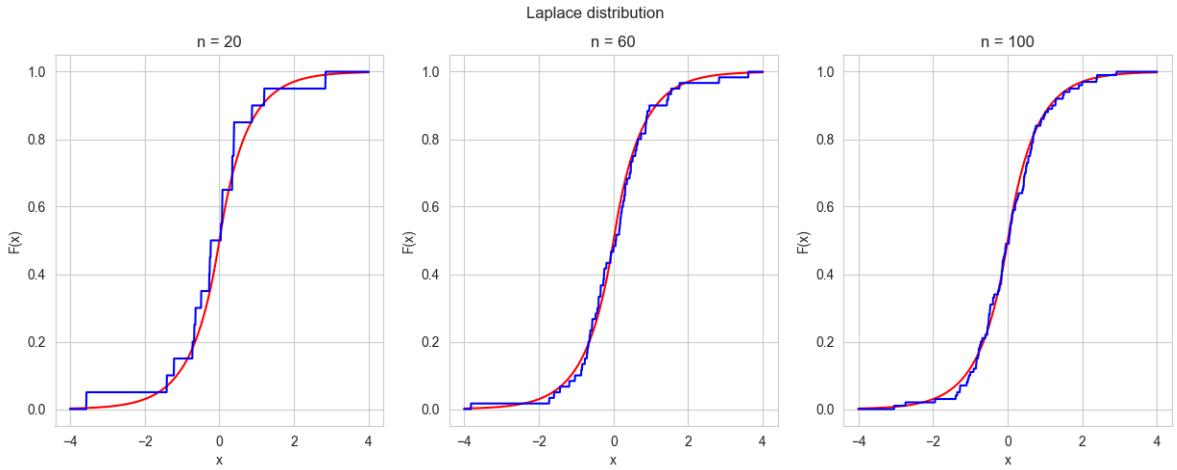

Рис. 13: Распределение Лапласа

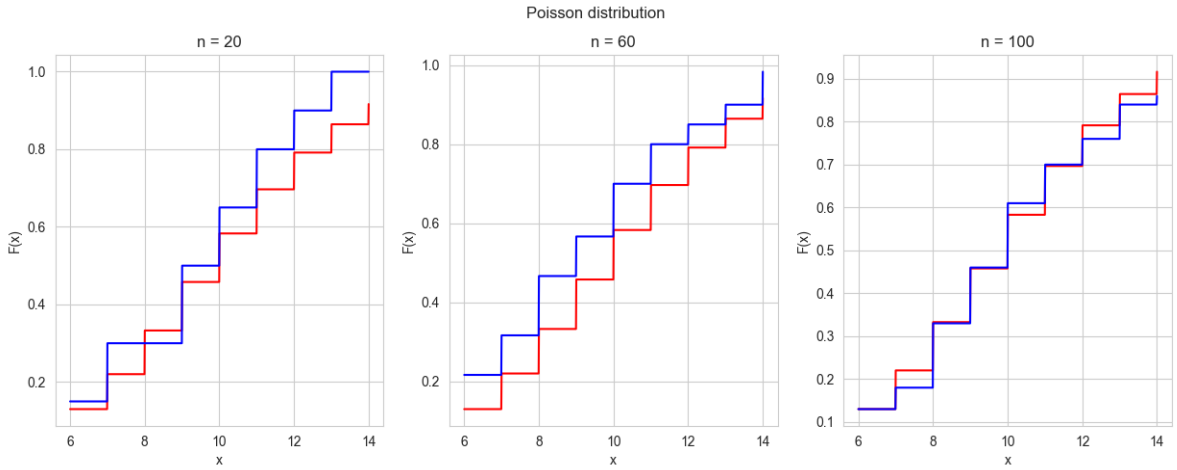

Рис. 14: Распределение Пуассона

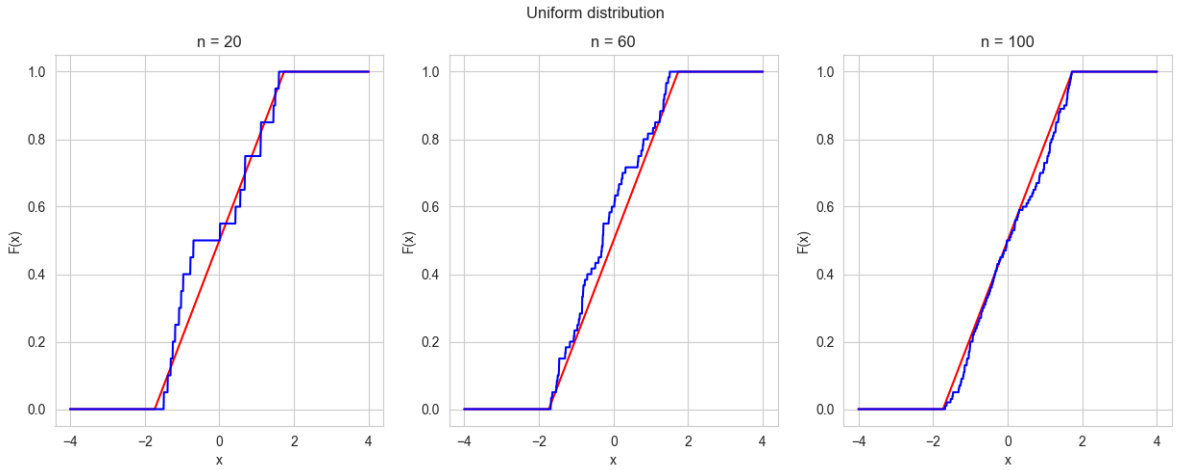

Рис. 15: Равномерное распределение

4.7 Ядерные оценки плотности распределения

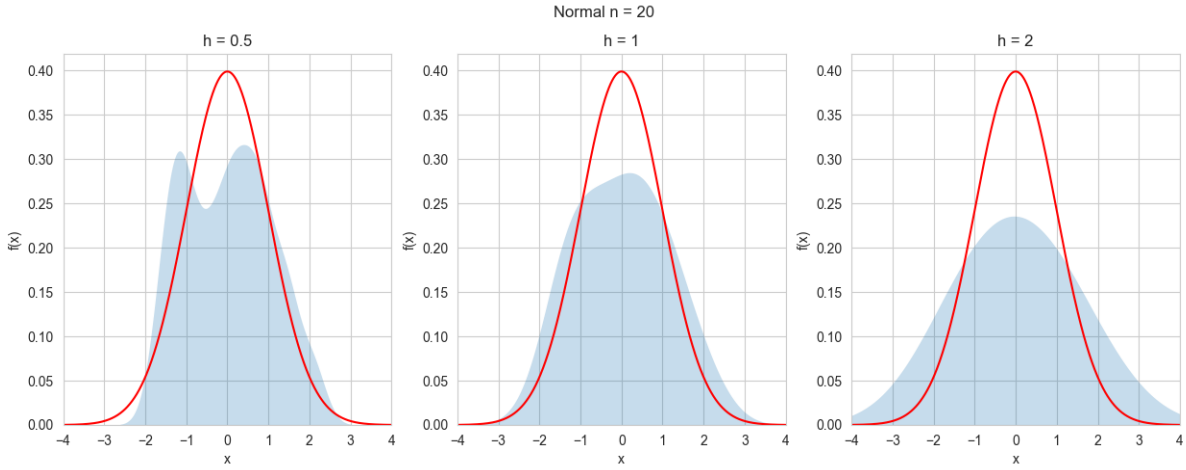

Рис. 16: Нормальное распределение, n = 20

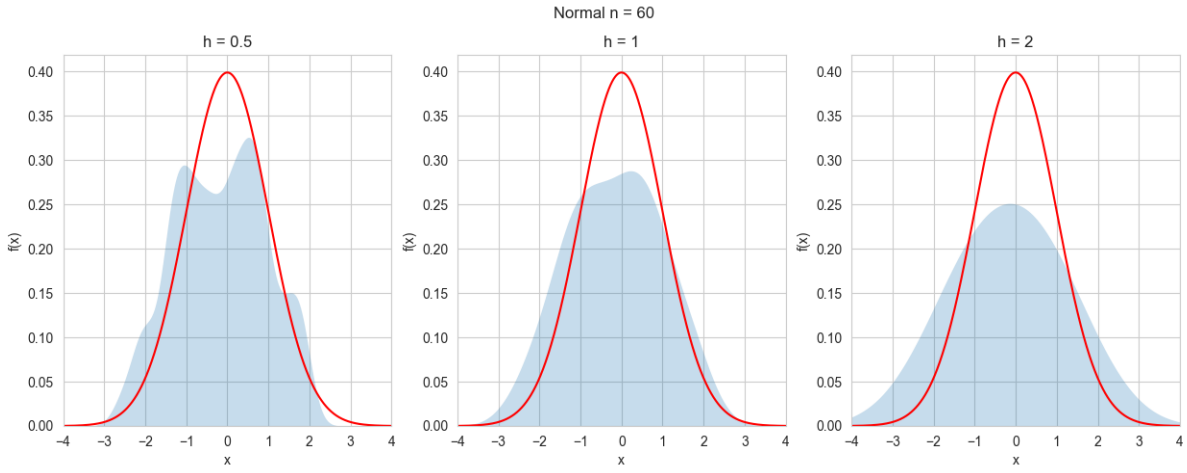

Рис. 17: Нормальное распределение, n = 60

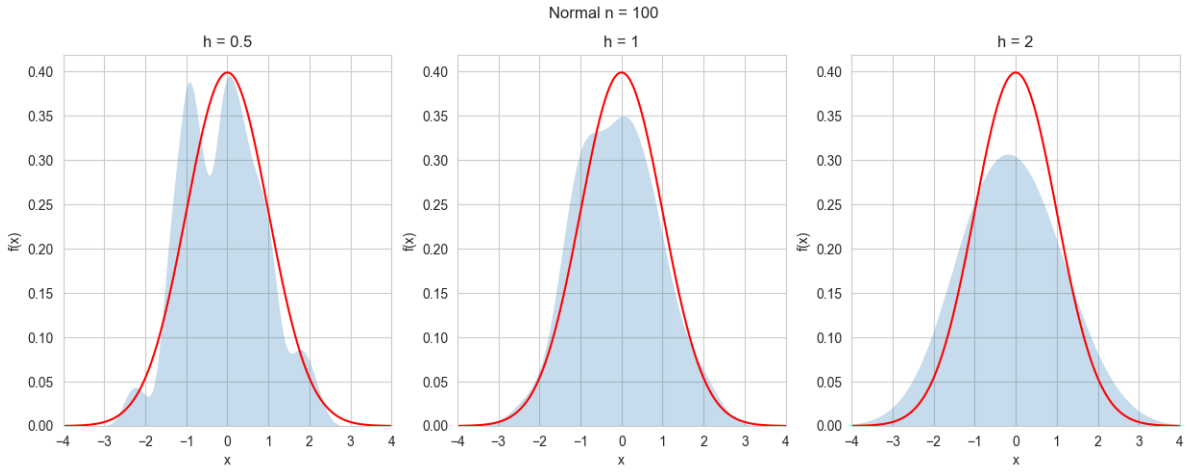

Рис. 18: Нормальное распределение, n = 100

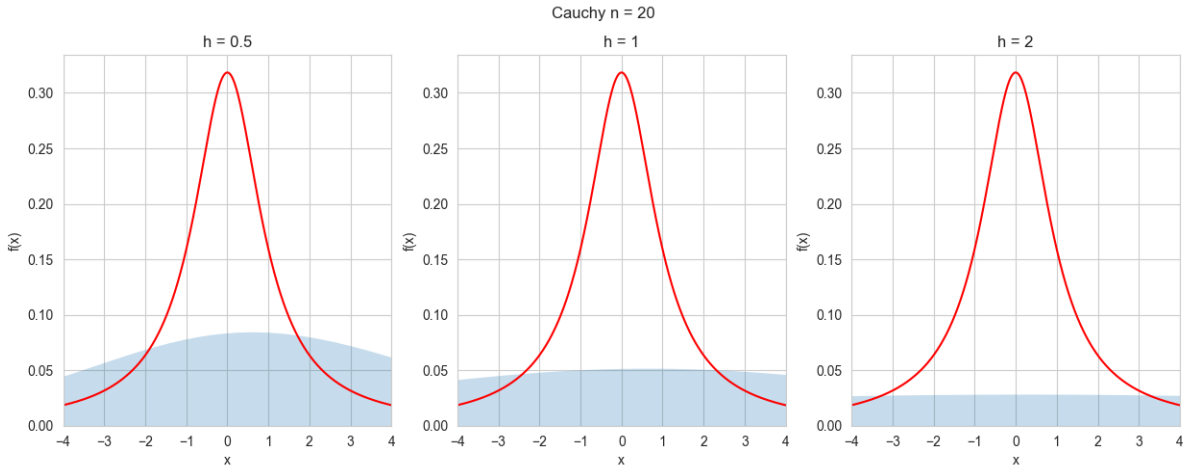

Рис. 19: Распределение Коши, n = 20

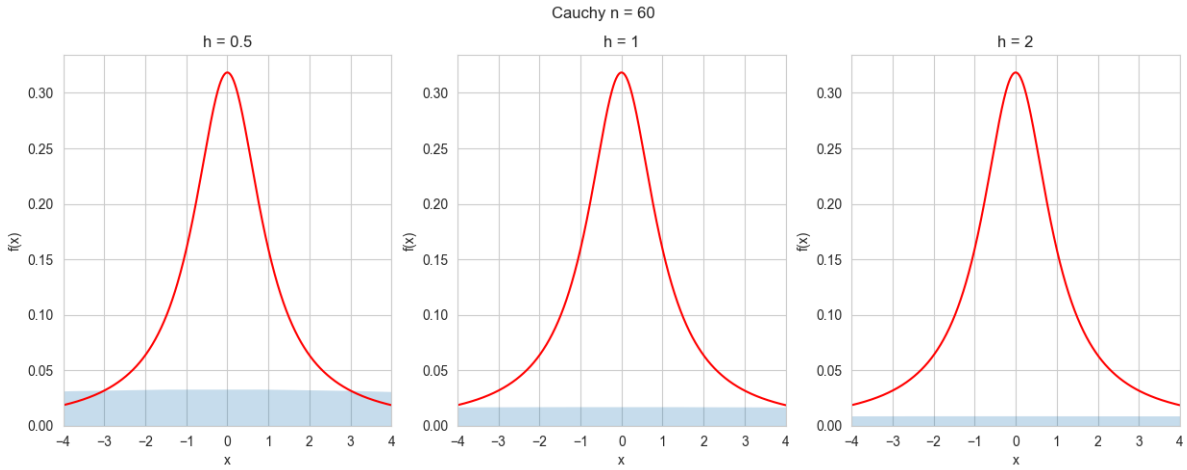

Рис. 20: Распределение Коши, n = 60

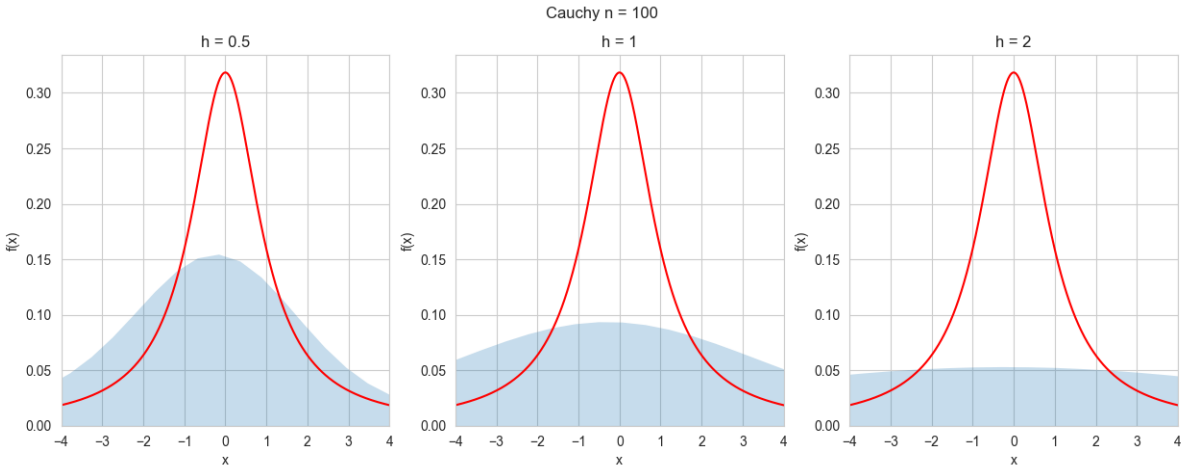

Рис. 21: Распределение Коши, n = 100

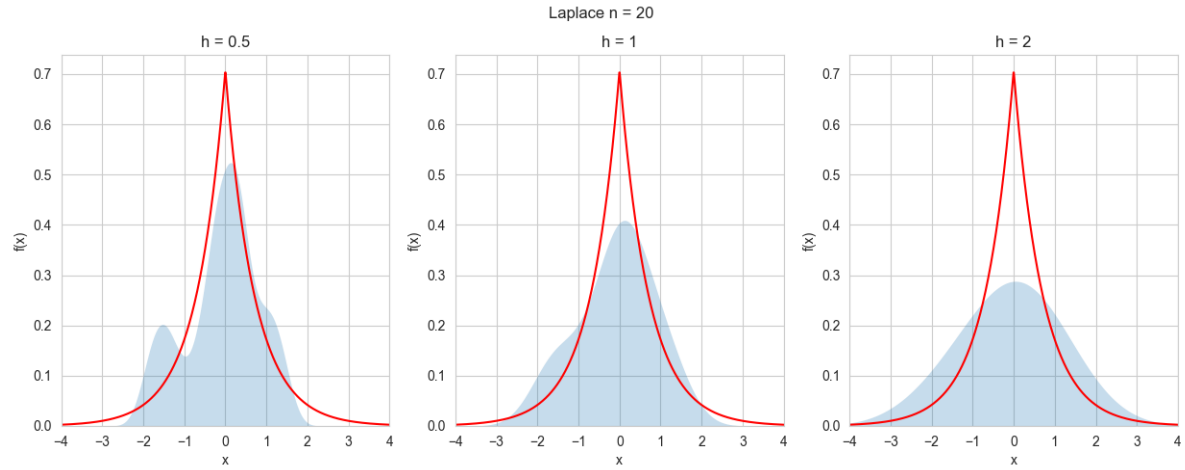

Рис. 22: Распределение Лапласа, n = 20

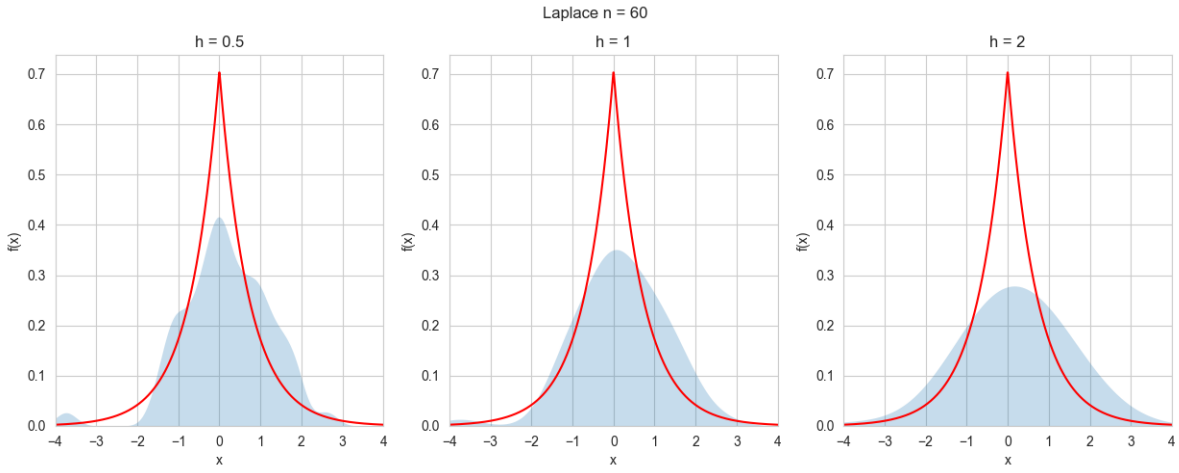

Рис. 23: Распределение Лапласа, n = 60

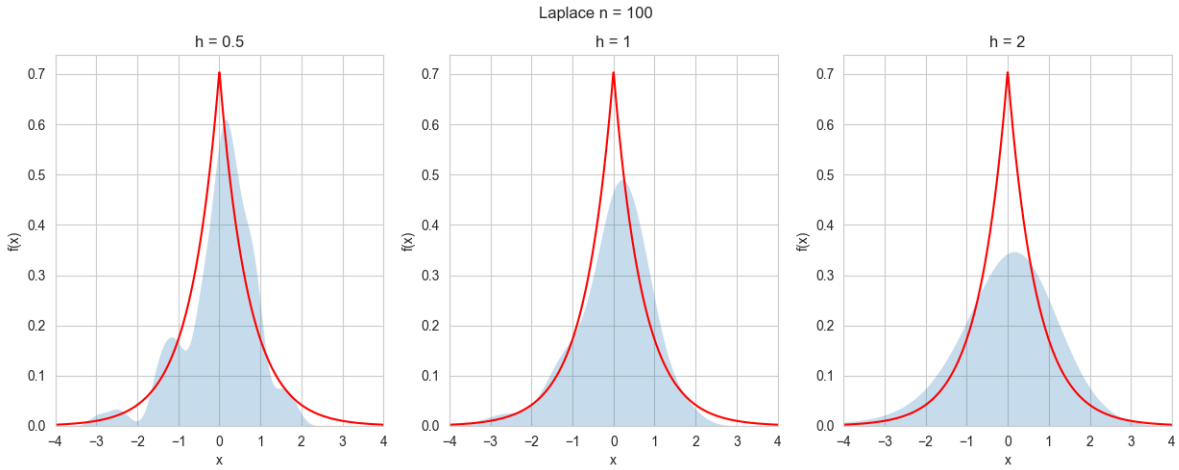

Рис. 24: Распределение Лапласа, n = 100

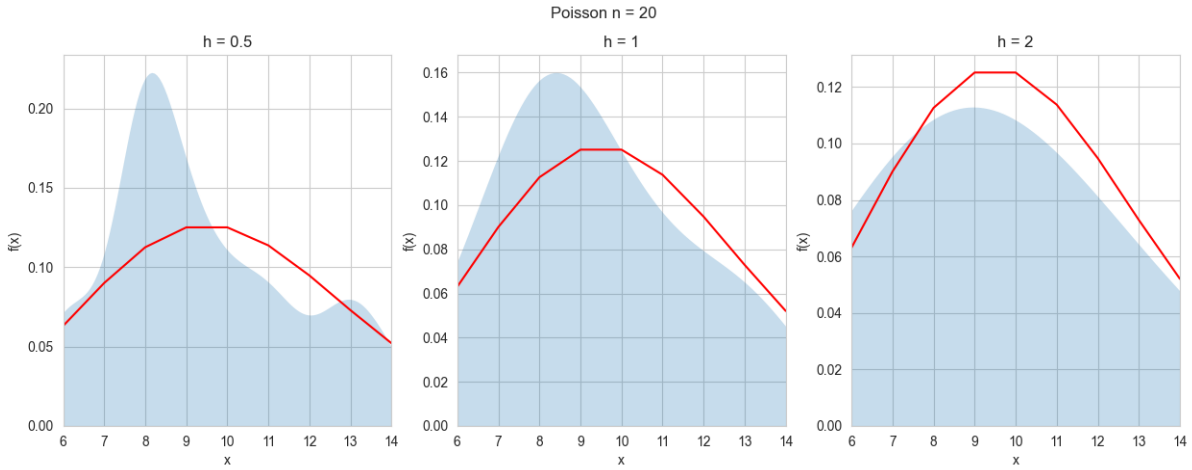

Рис. 25: Распределение Пуассона, n = 20

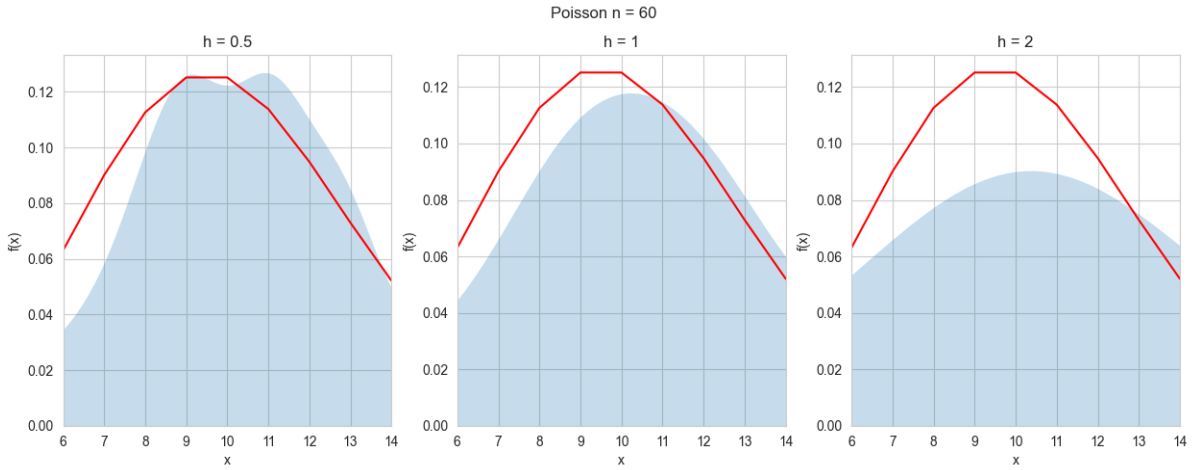

Рис. 26: Распределение Пуассона, n = 60

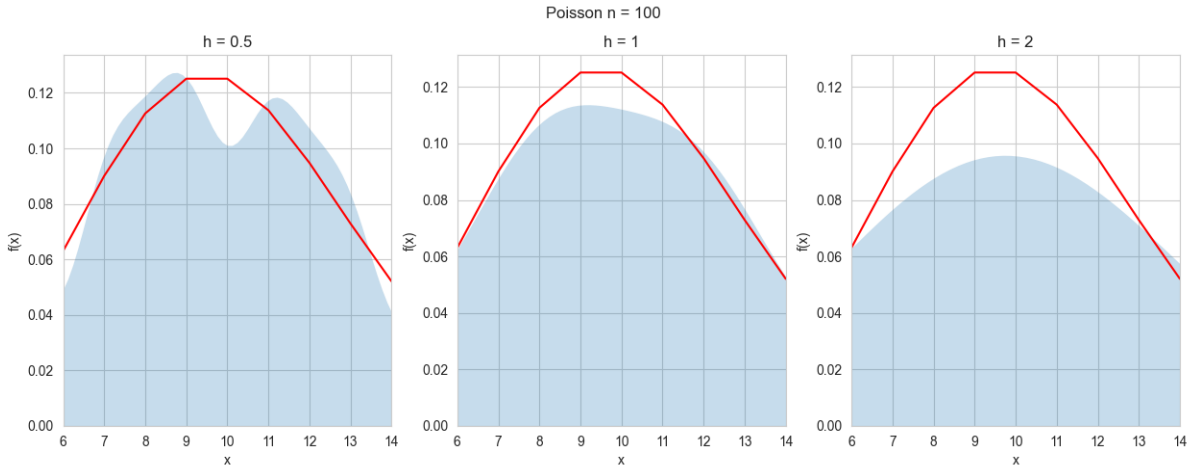

Рис. 27: Распределение Пуассона, n = 100

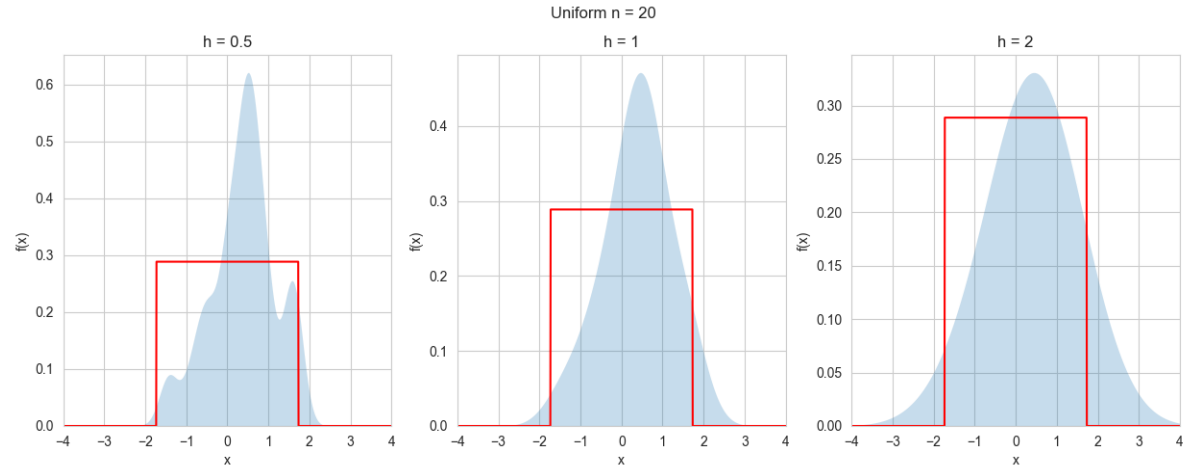

Рис. 28: Равномерное распределение, n = 20

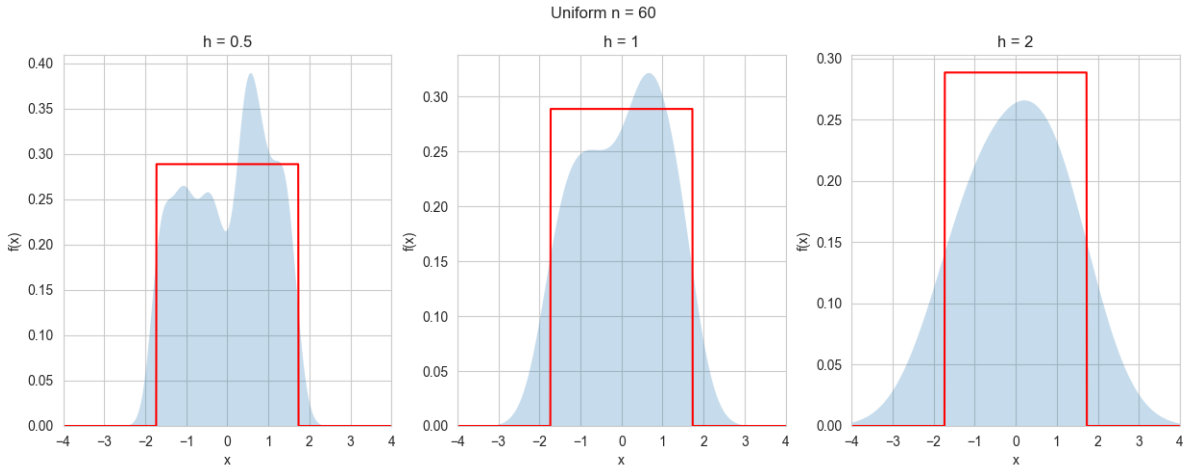

Рис. 29: Равномерное распределение, n = 60

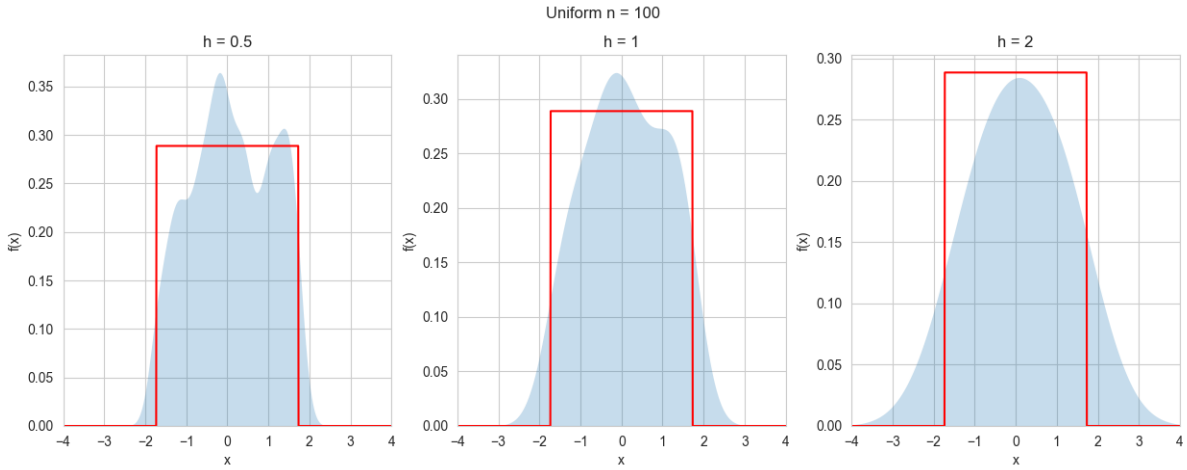

Рис. 30: Равномерное распределение, n = 100

5 Обсуждение

5.1 Гистограмма и график плотности распределения

По результатам проделанной работы 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. Также можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Также наблюдаются всплески гистограмм, что наиболее хорошо прослеживается на распределении Коши

5.2 Характеристики положения и рассеяния

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.

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 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 равномерного и пуассоновского распределений оптимальным значением параметра является $h = 2h_n$, для распределений Лапласа - $h = h_n/2$, а 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 $\hat{h_n}$, 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 $h = 2h_n$ 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/Fourroubles/Math-Statistic

7 Список литературы

1. Histogram. URL: https://en.wikipedia.org/wiki/Histogram
2. Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.
3. Box plot. URL: https://en.wikipedia.org/wiki/Box_plot