Санкт-Петербургский политехнический университет Петра Великого
Физико-Механический институт
«Высшая школа прикладной математики»

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1-4

по дисциплине «Математическая статистика»

Выполнил студент:
Ярмак Дмитрий Юрьевич
группа: 3630102/90101

Проверил:

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

Санкт-Петербург 2022 г.

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Для четырех распределений:

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

1. Сгенерировать выборки по 10, 50, 1000 элементов. Построить на одном рисунке гистаграмму и график плотности распределения.

2. Сгенерировать выборки по 10, 100, 1000 элементов. Для каждой выборки вычислить следующие статические 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 их квадратов:

$$E(z) = \overline{z} \tag{1}$$

Вычислить оценку дисперсии по формуле:

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

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 (сгенерировать выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.

4. Сгенерировать выборки размера 20, 60, 100 элементов. Построить на них эмперические 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

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

- Равномерное распределение

$$U(x,-\sqrt{3},\sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & , |x| \le \sqrt{3} \\ 0 & , |x| > \sqrt{3} \end{cases} \tag{4}$$

- Распределение Коши

$$C(x,0,1) = \frac{1}{\pi} \frac{1}{x^2+1} \tag{5}$$

- Распределение Пуассона

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

2.2 Гистограмма

Гистограмма в математической статистике - это функция, приближающая плотность вероятности некоторого распределения, поостроенная на основе выборки из нее.

Графически гистограммы строятся следующим образом. Сначала множество значений, которые 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.

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

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

2.4 Выборочные числовые характеристики

2.4.1 Характеристики положения

- Выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

- Выборочная медиана

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

- Полусумма экстремальных выборочных элементов

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

- Полусумма квартилей

  Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{, при } np \text{ дробном} \\ x_{(np)} & \text{, при } np \text{ целом} \end{cases} \tag{10}$$

  Полусумма квартилей

$$z_Q = \frac{z_{\frac{1}{4}} + z_{\frac{3}{4}}}{2} \tag{11}$$

- Усеченное среднее

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

2.4.2 Характеристики рассеяния

Выборочная дисперсия

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

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":

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

где X_1 - нижняя граница уса, X_2 - верхняя граница уса, Q_1 - первый квартиль, Q_3 - третий квартиль. Данные, выходящие за границу усов (выбросы), отображаются на графике в виде маленьких кружков.

2.6 Теоретическая вероятность выбросов

Можно вычислять теоретические первый и третий квартили распределений Q_1^T и Q_3^T . По ф-ле (14) - теоретические нижнюю и верхнюю границы уса X_1^T и X_2^T. Выбросы - величины x:

$$\begin{cases} x < X_1^T \\ x > X_2^T \end{cases} \tag{15}$$

Теоретическая вероятность выбросов:

- для непрерывных распределений

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

- для дискретных распределений

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

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

2.7 Эмперическая функция распределения

2.7.1 Статистический ряд

Статистический ряд - последовательность различных элементов выборки $z_1, z_2, ..., z_k$, расположенных в возрастающем порядке с указанием частот $n_1, n_2, ..., n_k$, с которыми эти элементы содержатся в выборке. Обычно описываются 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:

$$F_n^*(x) = P^*(X < x) \tag{18}$$

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

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

$F^*(x)$ - 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. приближенным значением, генеральной функции распределения

$$F_n^*(x) \approx F_X(x) \tag{20}$$

2.8 Оценка плотности вероятности

2.8.1 Определение

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

$$\hat{f}(x) \approx f(x) \tag{21}$$

2.8.2 Ядерные оценки

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

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

Здесь функция K(u), называемая ядерной (ядром), непрерывна и является плотностью вероятности, $x_1, ..., x_n$ - элементы выборки, h_n - любая последовательность положительных чисел, обладающая свойтсвами

$$\lim_{n->\infty} h_n = 0; \lim_{n->\infty} \frac{h_n}{n^{-1}} = \infty \tag{23}$$

Такие оценки называют непрерывными ядерными.

Гауссово (нормальное) ядро

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

Правило Сильвермана

$$h_n = 1.06\hat{\sigma} n^{-\frac{1}{5}} \tag{25}$$

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

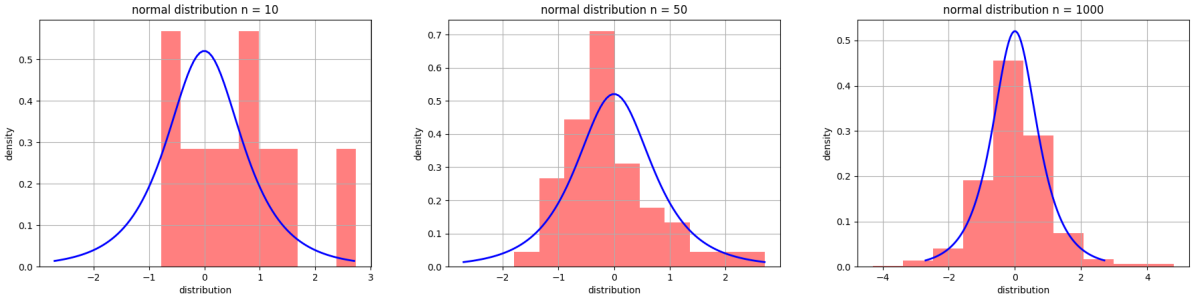
3 Реализация

Лабораторная работа выполнена при помощи языка программирования Phyton и библиотек numpy, matplotlib, scipy и т.п. в среде программирования PyCharm.

4 Результаты

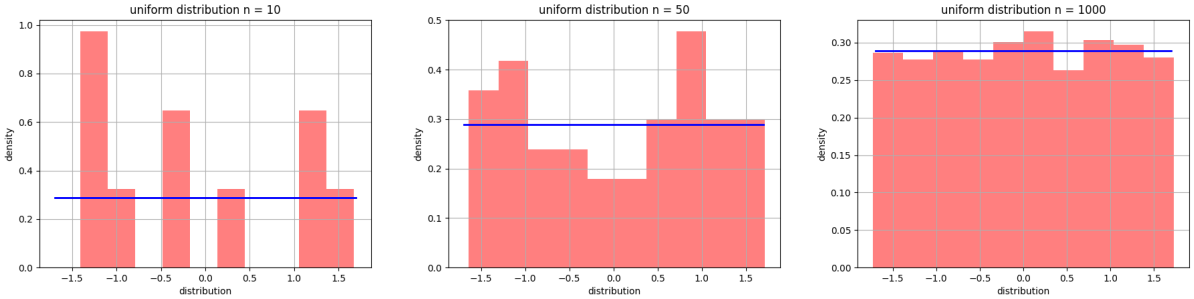
4.1 Гистограмма и график плотности распределения

Для распределения Коши при выборке n = 1000 взят логарифмический масштаб по оси ординат.



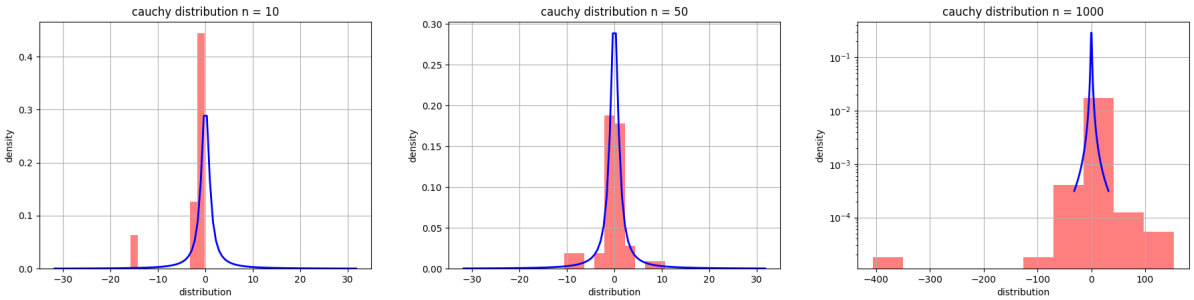
n = 10 n = 50 n = 1000

Рис. 1: Нормальное распределение



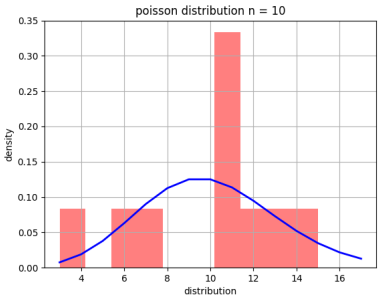
n = 10 n = 50 n = 1000

Рис. 2: Равномерное распределение

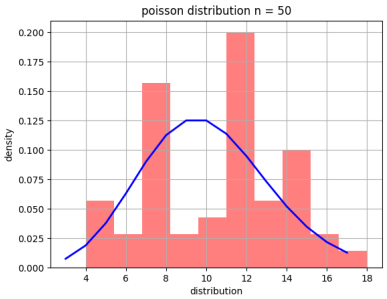


n = 10 n = 50 n = 1000

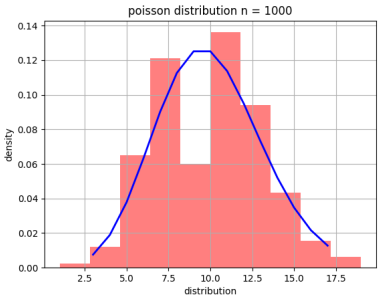
Рис. 3: Распределение Коши



n = 10



n = 50



n = 1000

Рис. 4: Распределение Пуассона

4.2 Характеристики положения и рассеяния

По поводу округления.
В оценке x=$< x >$ $\pm\sqrt{D} * k_\alpha$ вариации подлежит первая цифра после точки. Параметр k_α - зависит от доверительной вероятности и вида распределения. Произведем округление для k = 1. Величины для оценки вынесены в отдельную строку и приведены по модулю.

normal n = 10					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	0.006165	0.013676	-0.008412	0.006465	0.009960
D(z)	0.100815	0.140920	0.185873	0.112905	0.115106
$\pm\sqrt{D(z)}k_\alpha$	0.317514	0.375393	0.431130	0.336013	0.339273
normal n = 100					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	-0.001264	-0.004724	0.001399	-0.000205	-0.002978
D(z)	0.009302	0.015124	0.082152	0.011437	0.011049
$\pm\sqrt{D(z)}k_\alpha$	0.096448	0.122979	0.286621	0.106943	0.105114
normal n = 1000					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	0.001687	0.001536	0.009063	0.002175	0.001268
D(z)	0.001024	0.001658	0.059270	0.001295	0.001275
$\pm\sqrt{D(z)}k_\alpha$	0.031993	0.040721	0.243454	0.035987	0.035703

Таблица 2: Нормальное распределение

uniform n = 10					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	0.000838	0.005689	-0.001547	0.002410	0.003929
D(z)	0.095697	0.210199	0.044127	0.132373	0.151110
$\pm\sqrt{D(z)}k_\alpha$	0.309349	0.458475	0.210063	0.363831	0.388729
uniform n = 100					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	0.002191	0.003381	0.000456	0.001571	0.004332
D(z)	0.009682	0.029437	0.000577	0.014432	0.019047
$\pm\sqrt{D(z)}k_\alpha$	0.098396	0.171572	0.024024	0.120133	0.138010
uniform n = 1000					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	0.001194	0.004295	0.000017	0.000528	0.002185
D(z)	0.000909	0.002671	0.000006	0.001376	0.001763
$\pm\sqrt{D(z)}k_\alpha$	0.030142	0.051681	0.002474	0.037091	0.041990

Таблица 3: Равномерное распределение

cauchy n = 10					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	1.586289	-0.015230	8.032621	-0.014300	-0.014613
D(z)	3040.000210	0.354292	75626.572434	1.013246	0.560731
$\pm\sqrt{D(z)}k_\alpha$	55.136197	0.595224	275.002859	1.006601	0.748820
cauchy n = 100					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	-4.928030	-0.003462	-246.579602	-0.009702	-0.006692
D(z)	25836.313344	0.027038	64559740.914168	0.052778	0.027729
$\pm\sqrt{D(z)}k_\alpha$	160.736783	0.164432	8034.907648	0.229735	0.166519
cauchy n = 1000					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	-0.219684	-0.000402	-120.474938	-0.001302	-0.000229
D(z)	259.626832	0.002371	64658275.695761	0.004666	0.002415
$\pm\sqrt{D(z)}k_\alpha$	16.112940	0.048694	8041.036979	0.068307	0.049147

Таблица 4: Распределение Коши

poisson n = 10					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	10.009300	9.889500	10.278500	9.916875	9.902333
D(z)	0.920264	1.332540	1.779688	1.030762	1.035406
$\pm\sqrt{D(z)}k_\alpha$	0.959304	1.154357	1.334049	1.015265	1.017549
poisson n = 100					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	9.991500	9.825500	10.907000	9.901625	9.846340
D(z)	0.105896	0.215800	1.027351	0.151619	0.124123
$\pm\sqrt{D(z)}k_\alpha$	0.325416	0.464543	1.013583	0.389383	0.352310
poisson n = 1000					
	$\overline{x}$	medx	z_R	z_Q	z_{tr}
E(z)	10.009813	9.994000	11.654000	9.996625	9.868728
D(z)	0.009838	0.005964	0.635784	0.002067	0.010971
$\pm\sqrt{D(z)}k_\alpha$	0.099187	0.077227	0.797361	0.045461	0.104743

Таблица 5: Распределение Пуассона

4.3 Боксплот Тьюки

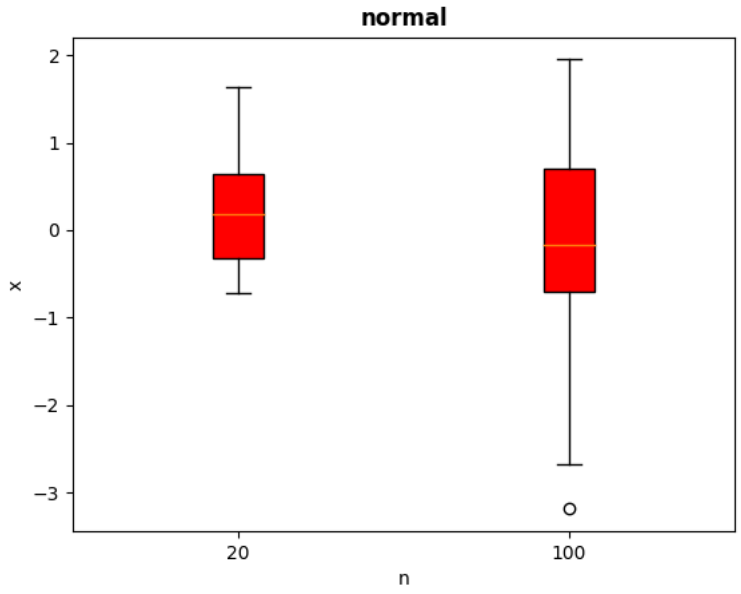


Рис. 5: Боксплот для нормального распределения

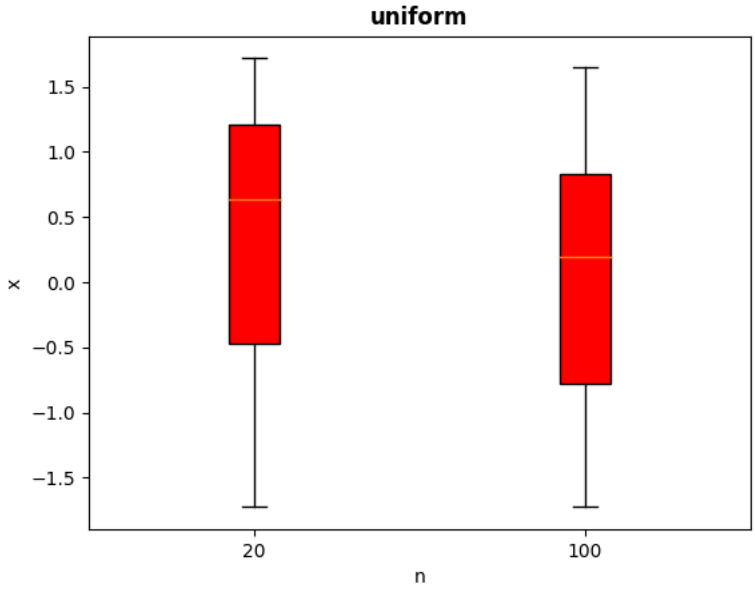


Рис. 6: Боксплот для равномерного распределения

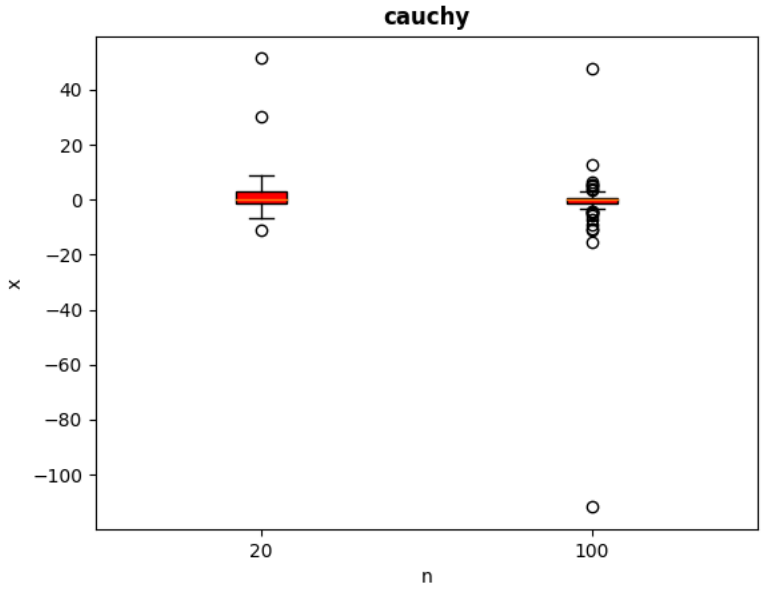


Рис. 7: Боксплот для распределения Коши

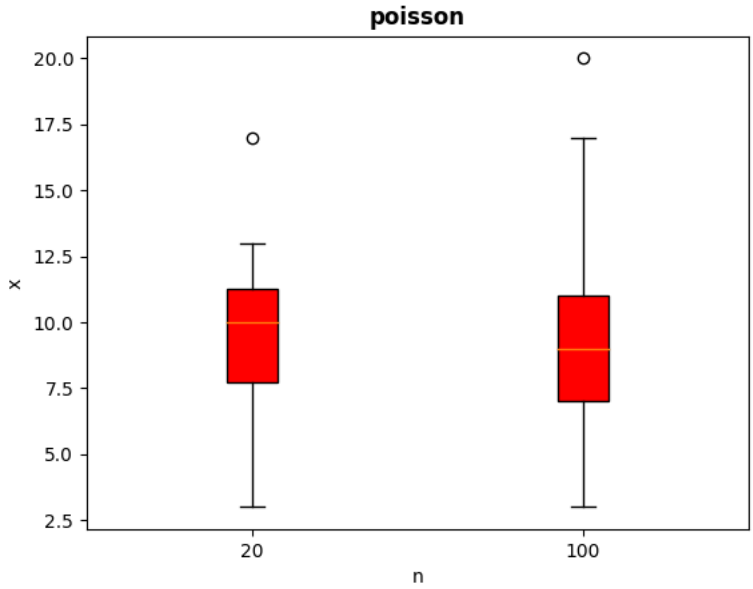


Рис. 8: Боксплот для распределения Пуассона

4.4 Доля выбросов

На счет округления. Выборка случайна, поэтому в качестве оценки рассеяния можно взять дисперсию пуассоновского потока: $D_n \approx \sqrt{n}$
Доля $p_n = D_n/n = 1/\sqrt{n}$
Для n = 20: $p_n = 1/\sqrt{20}$ - примерно 0.2 или 20%
Для n = 100: $p_n = 0.1$ или 10%
Исходя из этого можно решить, сколько знаков оставлять в доле выбросов.

Выборка	Доля Выбросов
normal n = 20	0.02
normal n = 100	0.01
uniform n = 20	0
uniform n = 100	0
cauchy n = 20	0.22
cauchy n = 100	0.21
poisson n = 20	0.11
poisson n = 100	0.09

Таблица 6: 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
Равномерное распределение	-0.866	0.866	-3.464	3.464	0
Распределение Коши	-1	1	-4	4	0.156
Распределение Пуассона	8	12	2	18	0.008

Таблица 7: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

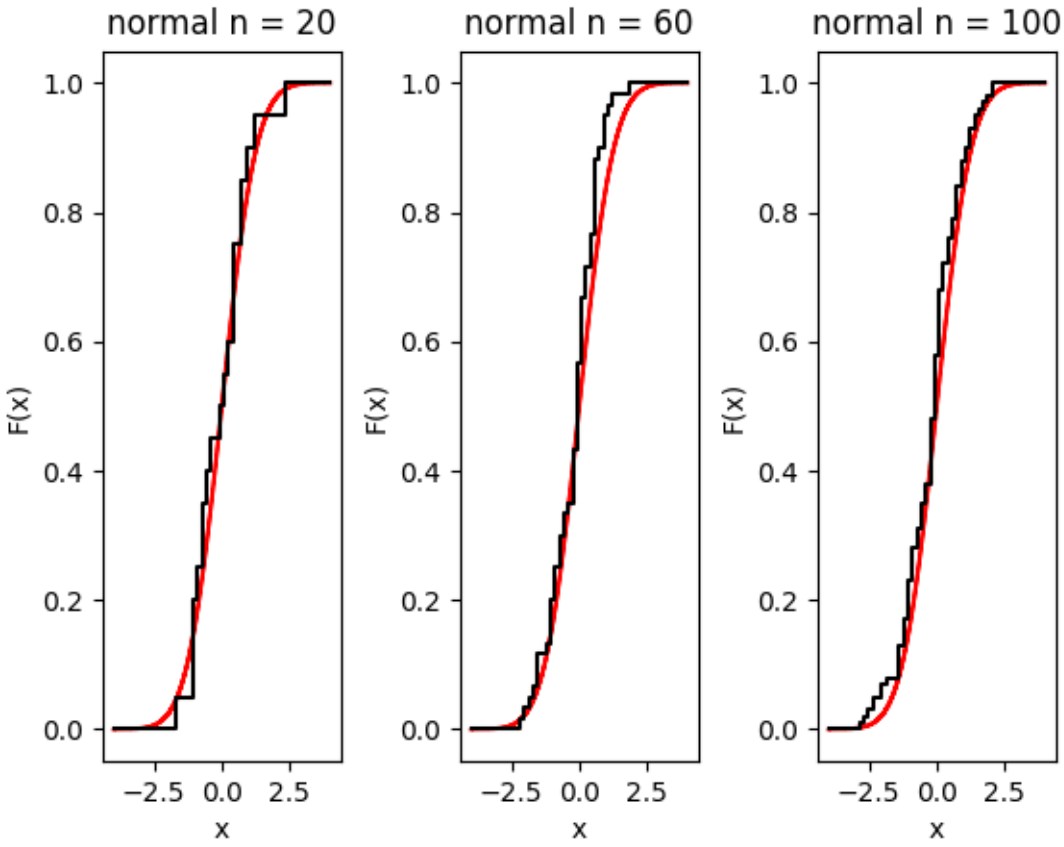


Рис. 9: Нормальное распределение

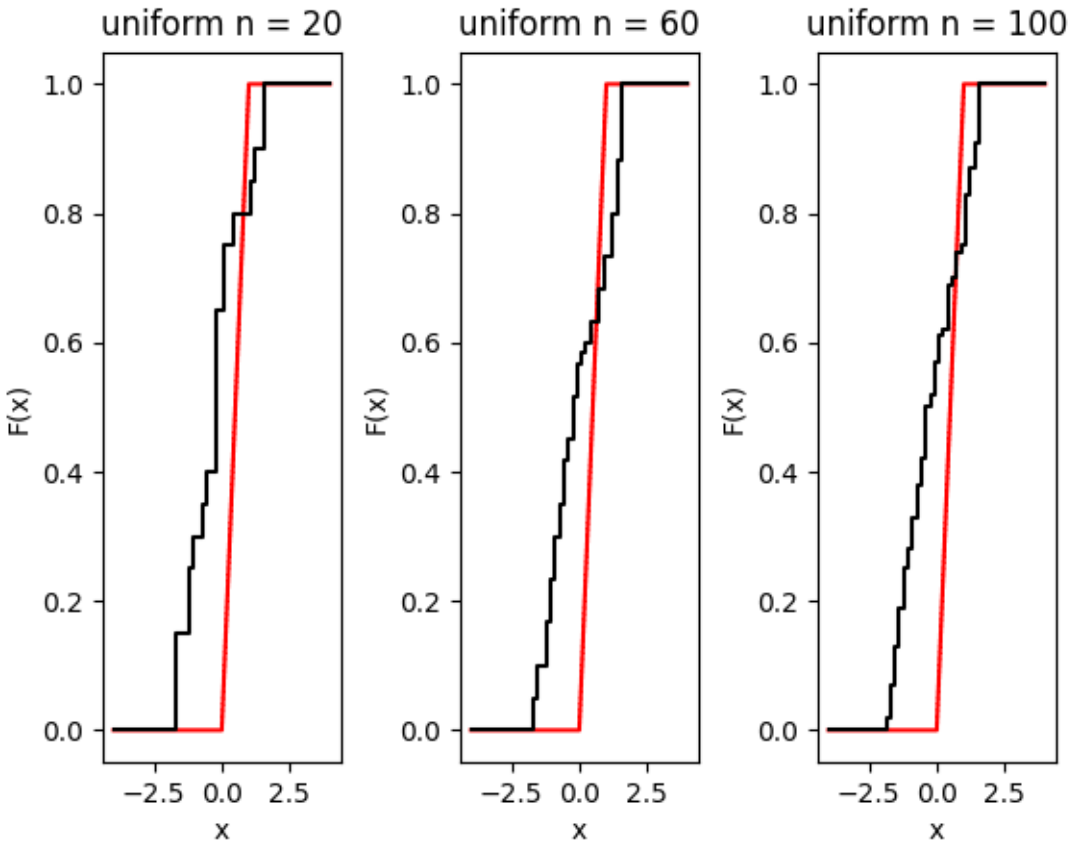


Рис. 10: Равномерное распределение

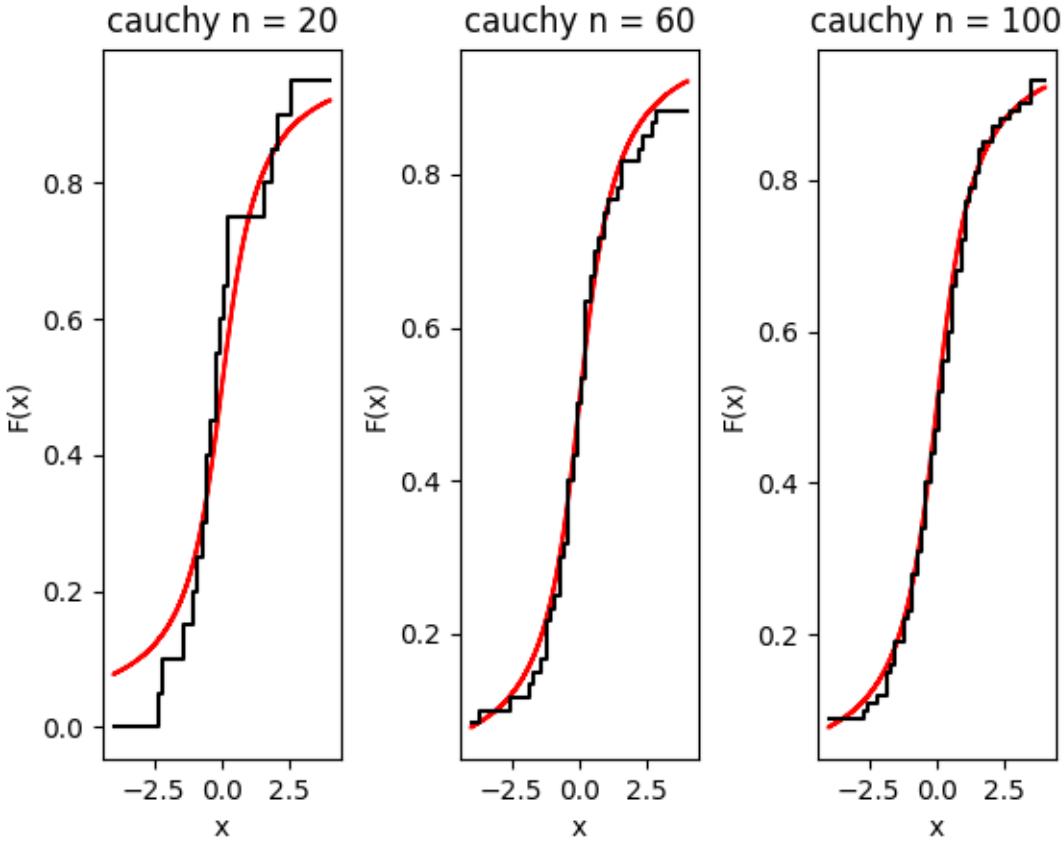


Рис. 11: Распределение Коши

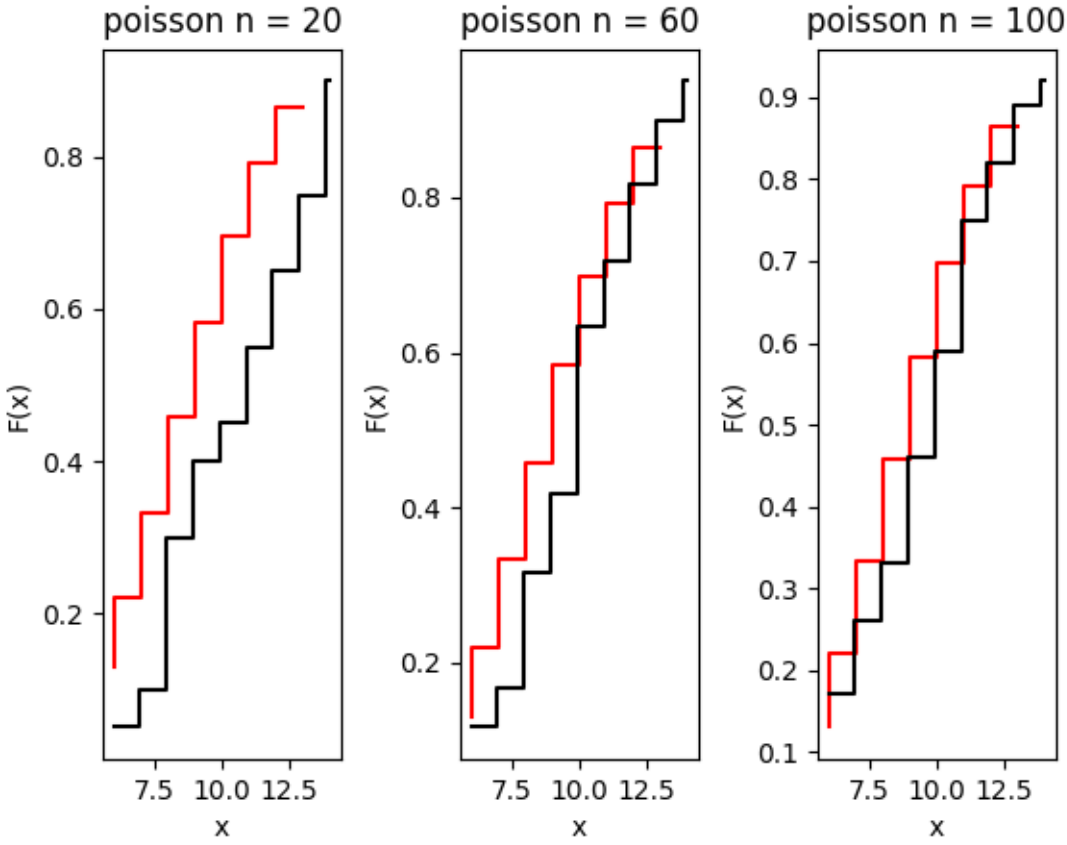


Рис. 12: Распределение Пуассона

4.7 Ядерные оценкци плотности распределения

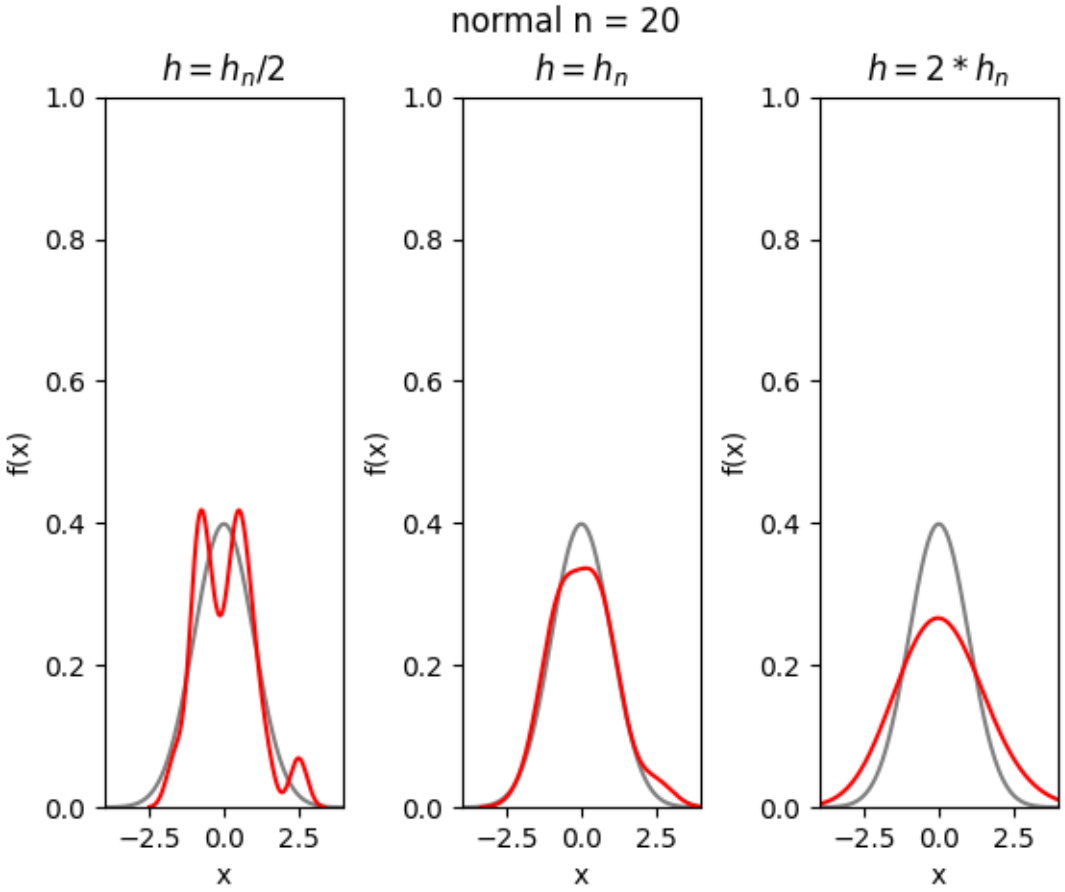


Рис. 13: Нормальное распределение, n = 20

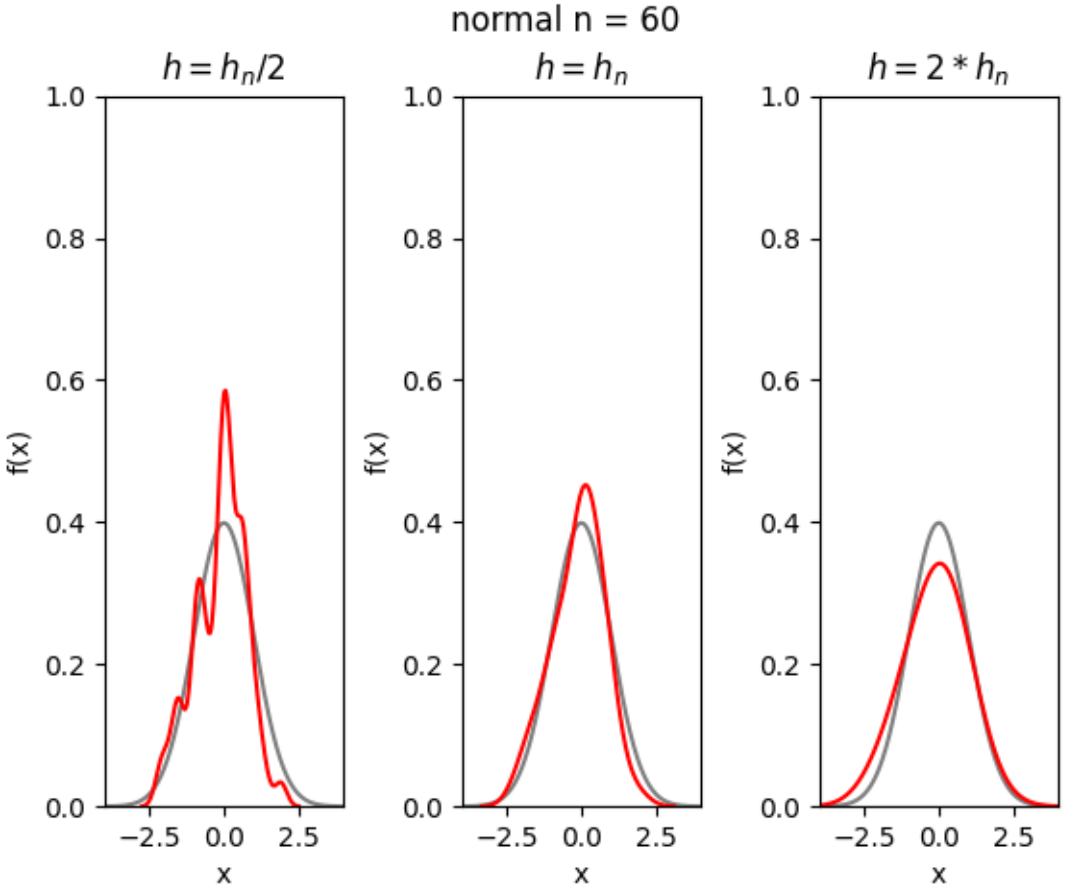


Рис. 14: Нормальное распределение, n = 60

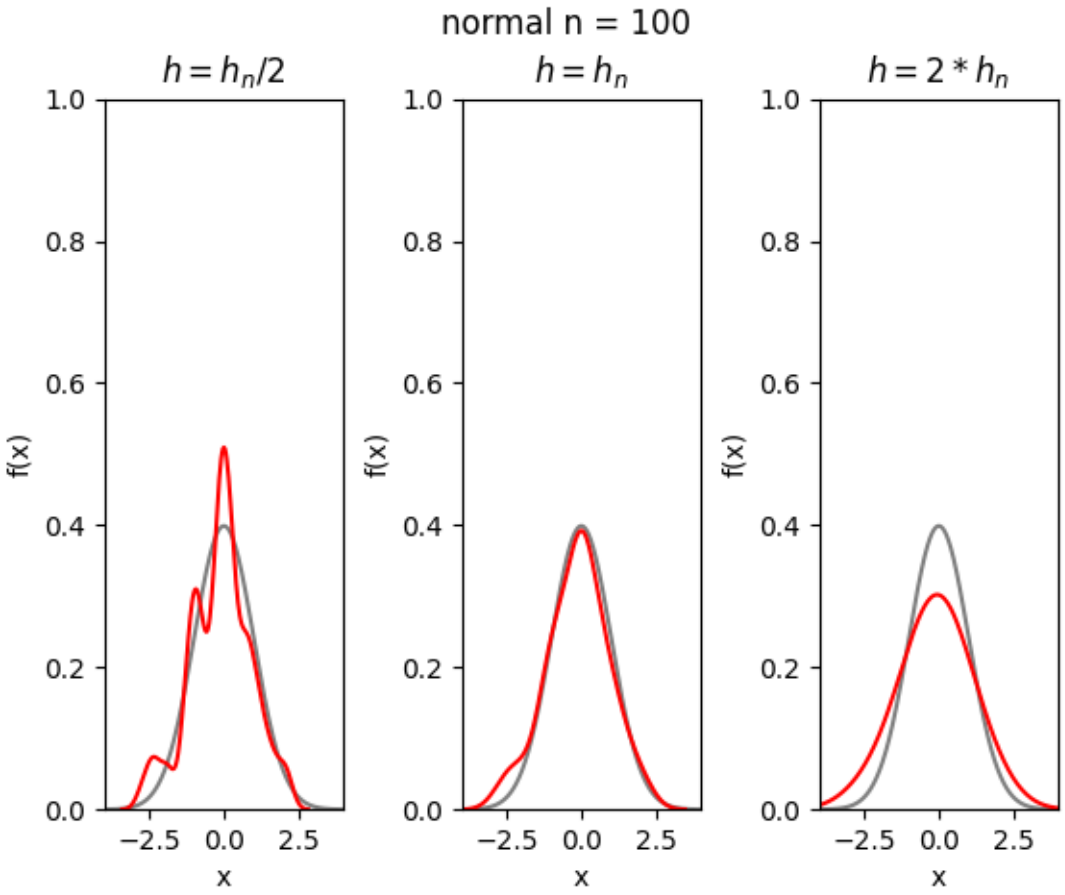


Рис. 15: Нормальное распределение, n = 100

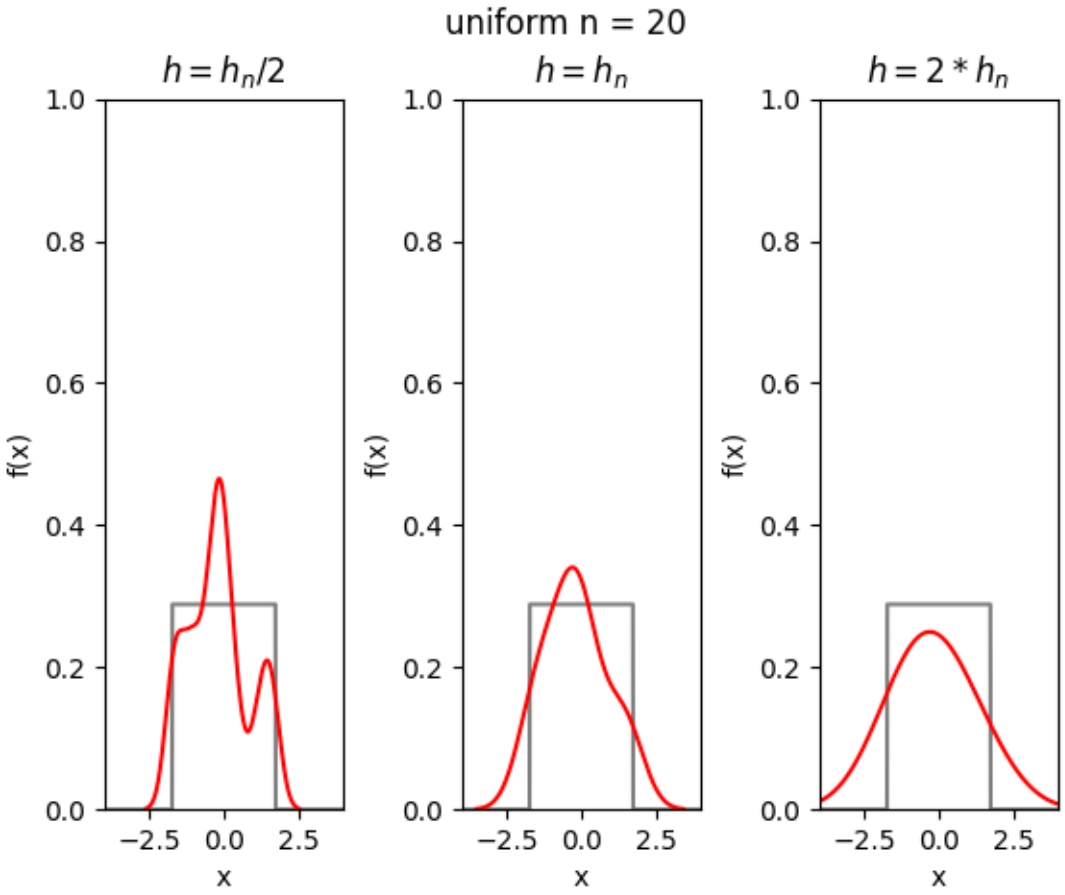


Рис. 16: Равномерное распределение, n = 20

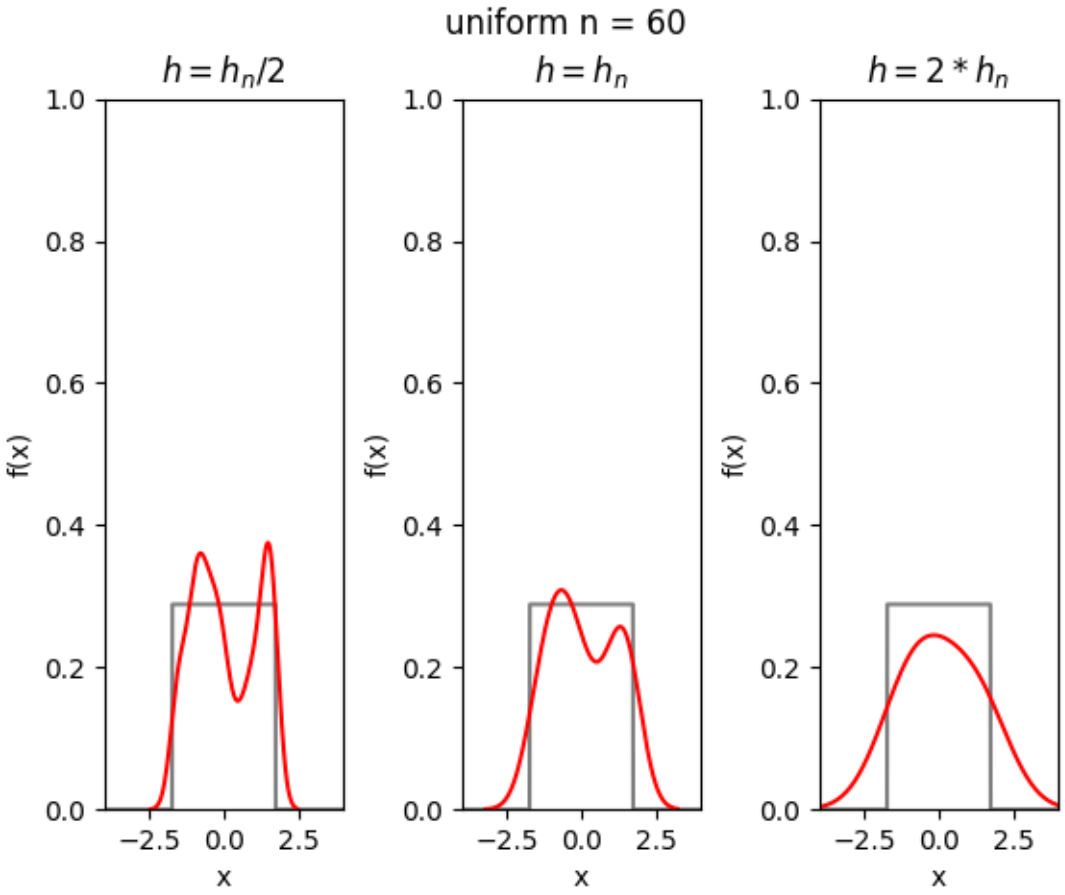


Рис. 17: Равномерное распределение, n = 60

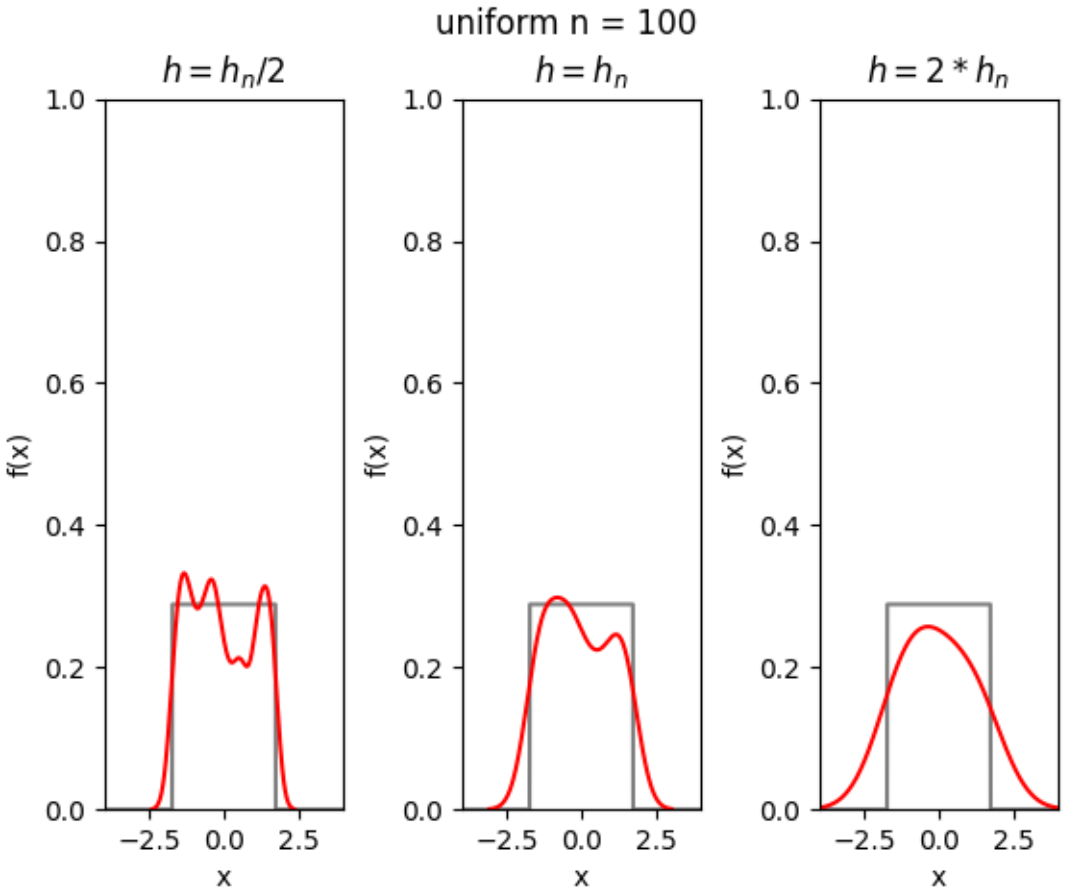


Рис. 18: Равномерное распределение, n = 100

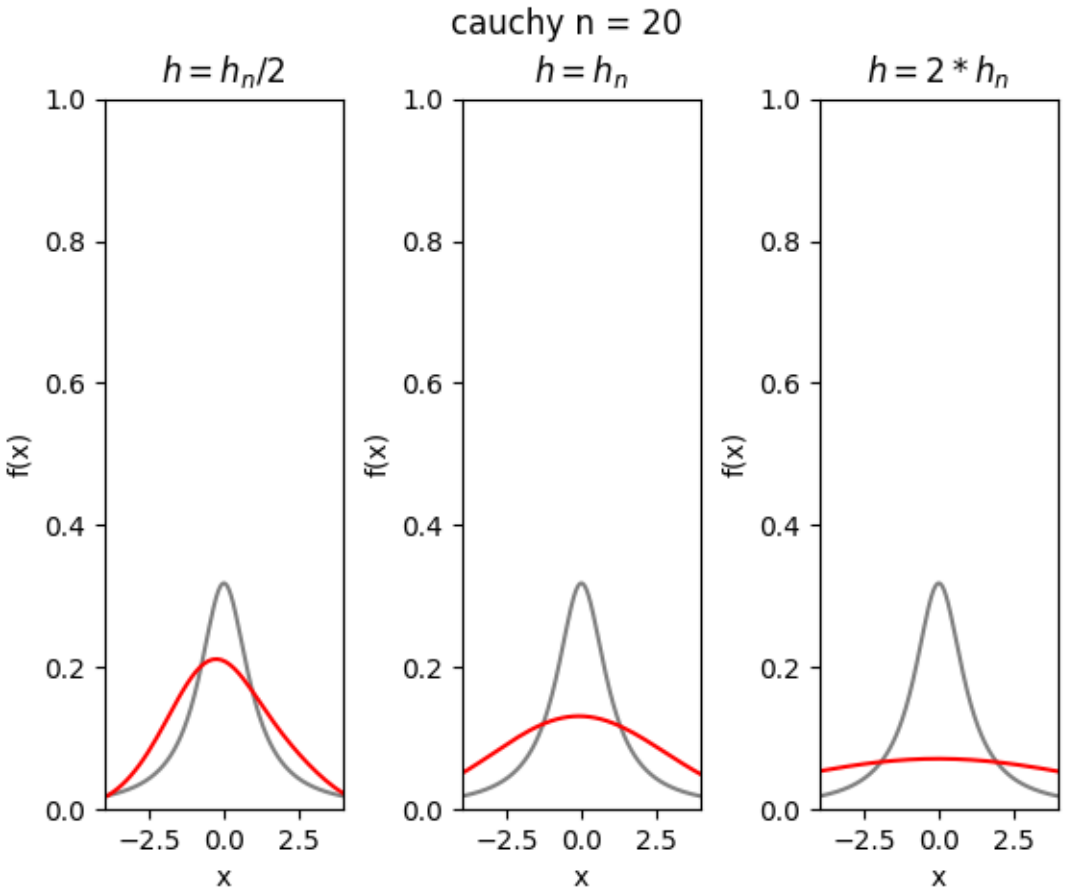


Рис. 19: Распределение Коши, n = 20

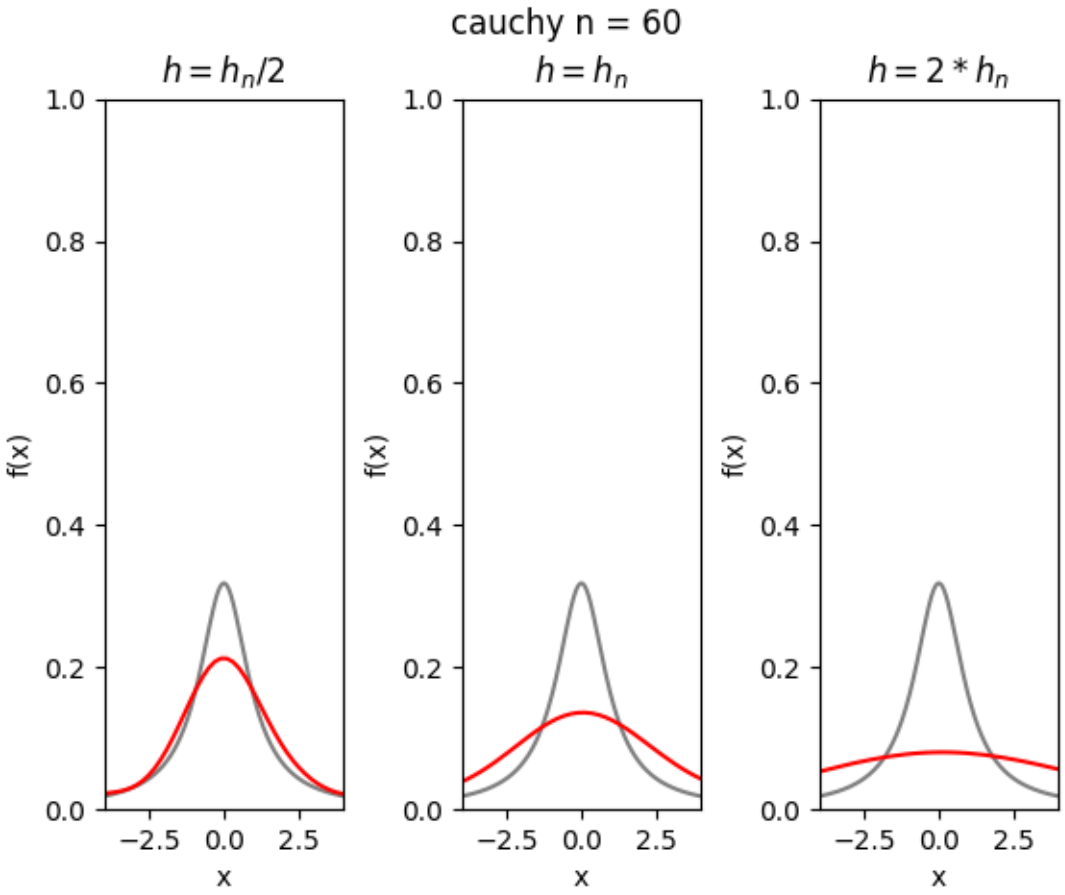


Рис. 20: Распределение Коши, n = 60

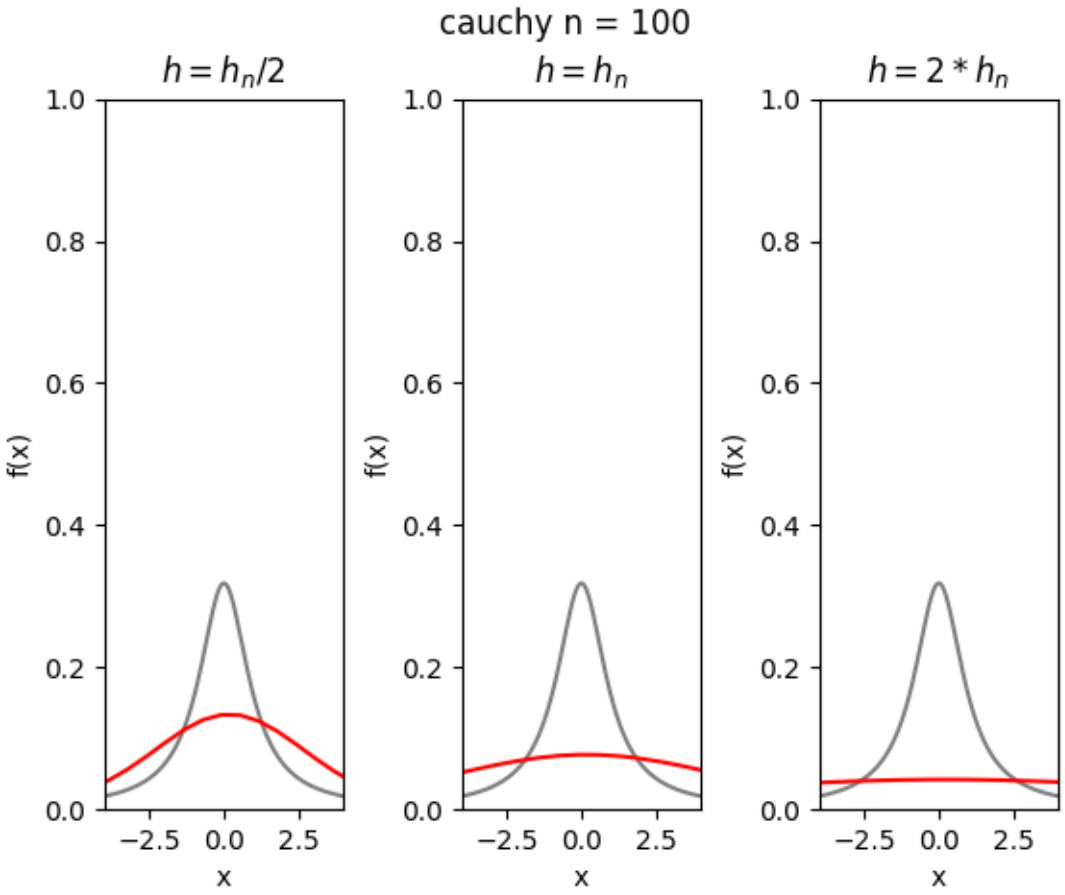


Рис. 21: Распределение Коши, n = 100

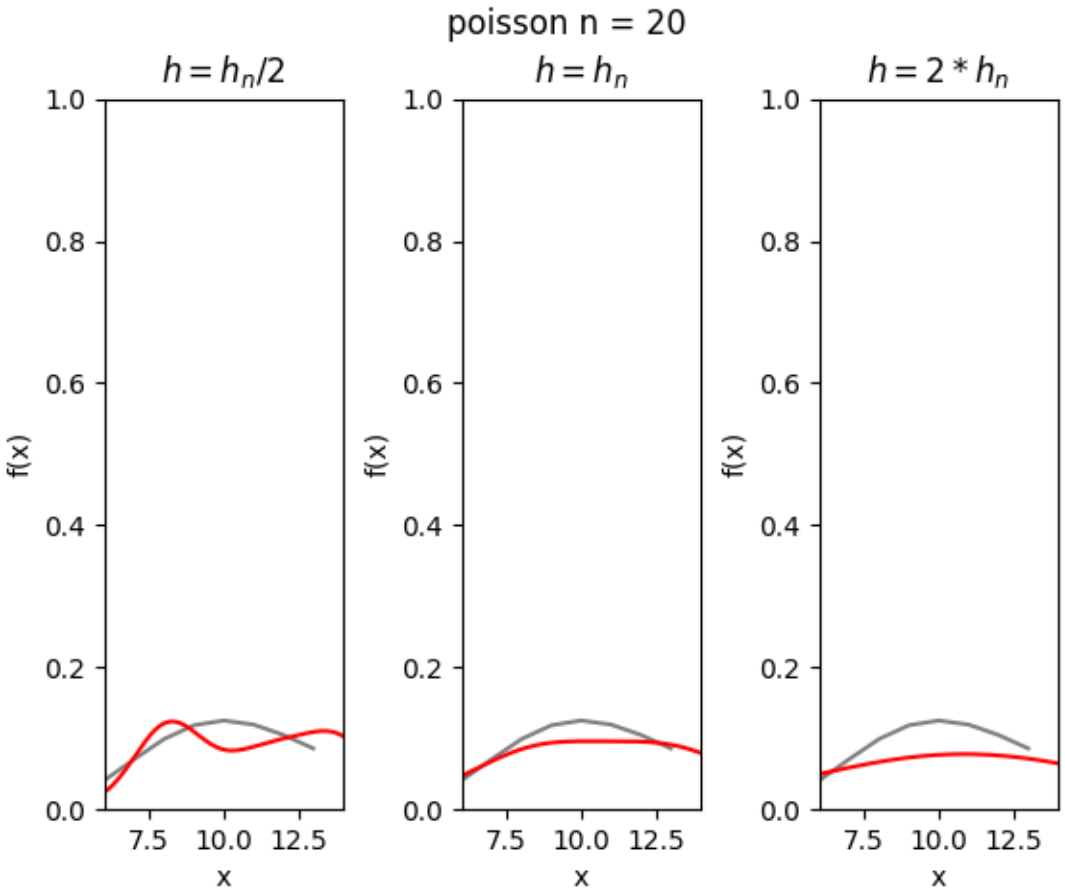


Рис. 22: Распределение Пуассона, n = 20

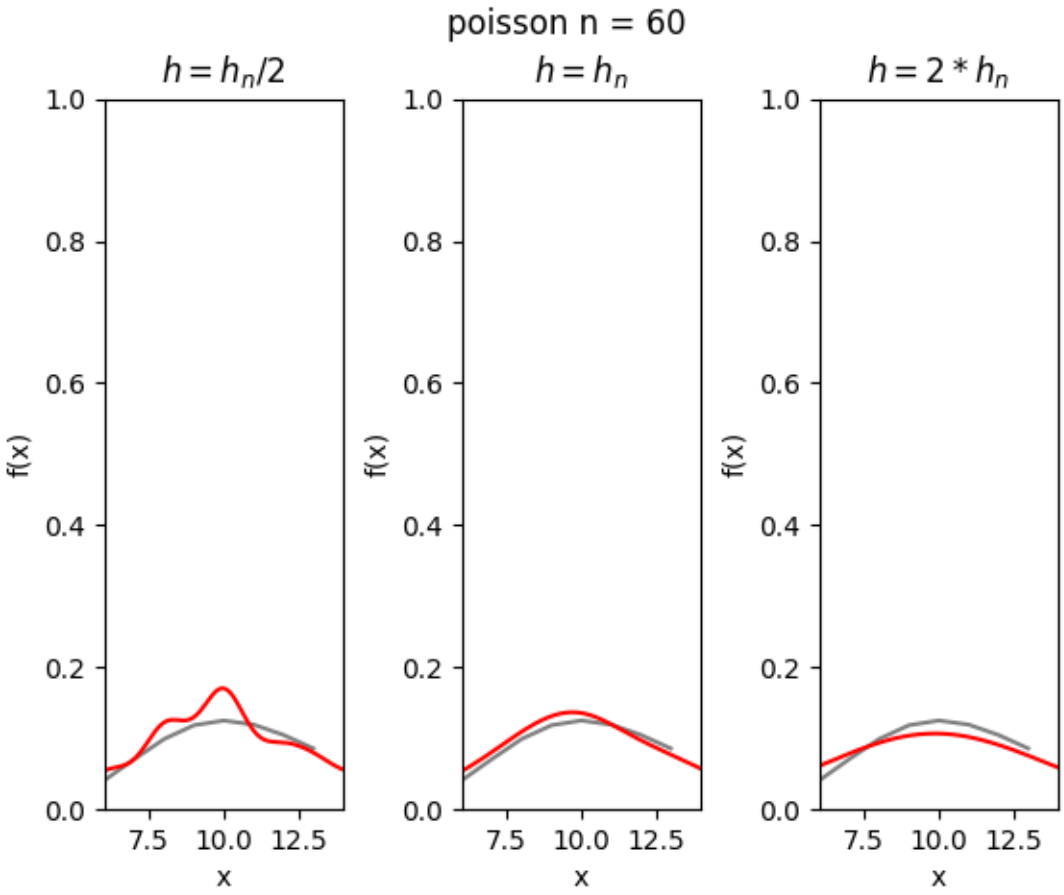


Рис. 23: Распределение Пуассона, n = 60

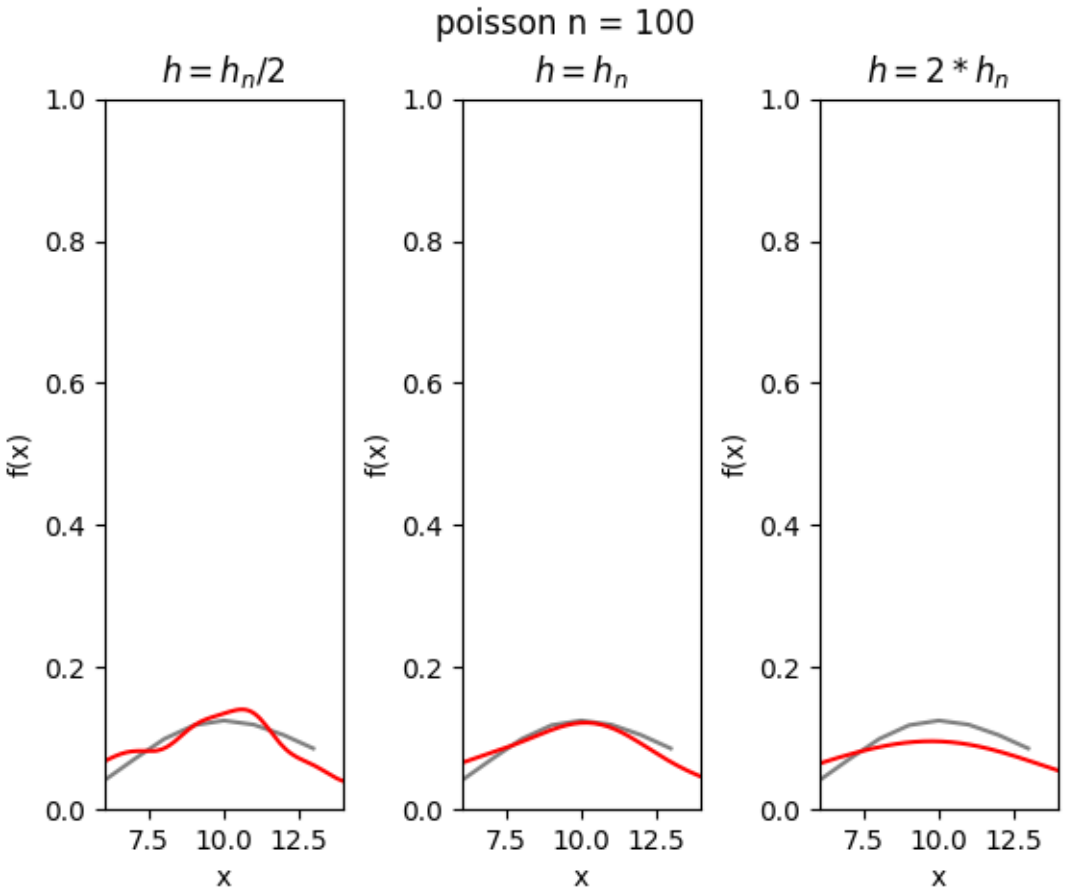


Рис. 24: Распределение Пуассона, n = 100

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 Характеристика 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 получили.

Боксплот Тьюки действительно позволяет более наглядно и с меньшими усилиями оценивать важные характеристики распределений. Так, исходя из полученных рисунков, наглядно видно то, что мы довольно трудоемко 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 также, 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 отклонение 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 распределения и распределения Коши подойдет параметр $h = h_n$. 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 Ссылки

https://github.com/AvitusCode/AvitusStatistics/Lab14