Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчёт
по лабораторной работе №1
по дисциплине
«Математическая статистика»

Выполнил студент
В. А. Рыженко

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

Санкт-Петербург, 2020 г.

Содержание

Список иллюстраций

1. Постановка задачи

Для 5 распределений:

- Нормальное распределение N(x, 0, 1)
- Распределение Коши C(x, 0, 1)
- Распределение Лапласа L(x, 0, $\frac{1}{\sqrt{2}}$)
- Постановка задач исследованияРаспределение 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: $\overline{x}$ (8), $med\, x$ (9), z_R (10), z_Q (12), $z_t r$ (13). Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и 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 бокс-плот 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. Распределения

- Нормальное распределение

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

- Распределение Коши

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

- Распределение Лапласа

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

- Распределение Пуассона

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

- Равномерное распределение

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

2.2. Характеристики положения

- Выборочное среднее

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

- Выборочная медиана

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

- Полусумма экстремальных выборочных элементов

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

- Полусумма квартилей
  Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном} \\ x_{(np)} & \text{при } np \text{ целом} \end{cases} \tag{11}$$

  Полусумма квартилей

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

- Усечённое среднее

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

2.3. Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n} \sum_{i=1}^{n} x_i - \overline{x} \tag{14}$$

2.4. Боксплот Тьюки

2.4.1. Определение

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

2.4.2. 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.

2.4.3. 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

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

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

По формуле (15)можно вычислить теоретические нижнюю и верхнюю границы уса (X_1^T и X_2^T соответственно). Выбросами считаются величины x, такие что:

$$\begin{bmatrix} x < X_1^T \\ x \leq X_2^T \end{bmatrix} \tag{16}$$

Теоретическая вероятность выбросов для непрерывных распределений

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

где $F(X) = P(x > X)$ - функция распределения.

Теоретическая вероятность выбросов для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(X > X_2^T) = (F(X_1^T) - P(x = X_1^T)) + (1 - F(X_2^T)), \tag{18}$$

где $F(X) = P(x > X)$ - функция распределения.

2.6. Эмпирическая функция распределения

2.6.1. Определение

Эмпирической (выборочной) функцией распределения (э. ф. р.) называется относительная частота события $X < x$, полученная по данной выборке:

$$F_n^*(x) = P^*(X < x). \tag{19}$$

2.6.2. Описание

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

2.7. Оценки плотности вероятности

2.7.1. Определение

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

2.7.2. Ядерные оценки

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

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - x_i}{h_n}). \tag{20}$$

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 свойствами

$$h_n \underset{n\to\infty}{\longrightarrow} 0; \frac{h_n}{n^{-1}} \underset{n\to\infty}{\longrightarrow} \infty; \tag{21}$$

Гауссово (нормальное) ядро

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{22}$$

Правило Сильвермана

$$h_n = 1.06\hat{\sigma}n^{-1/5}, \tag{23}$$

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

3. Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки Jupyter Notebook и Visual Code. Исходный код лабораторной работы приведён в приложении.

4. Результаты

4.1. Гистограмма и график плотности распределения

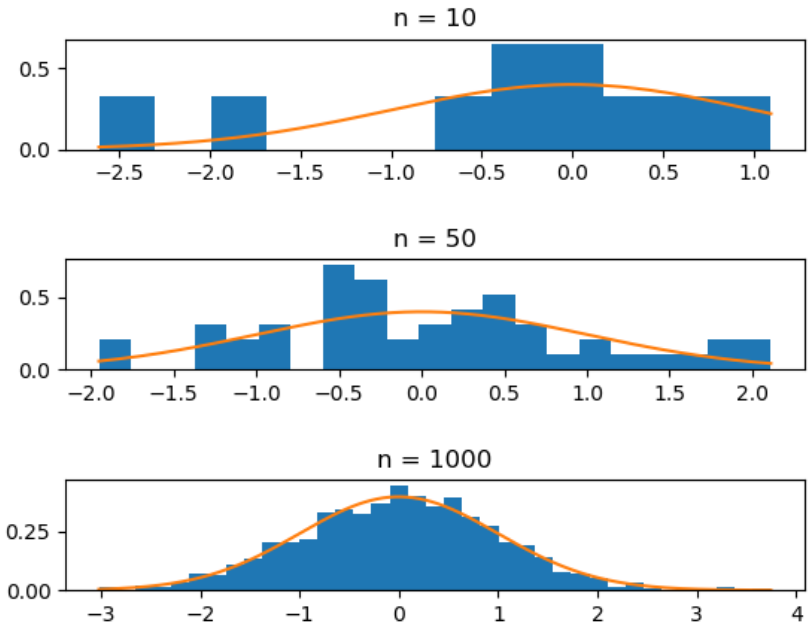

Рис. 1. Нормальное распределение (3)

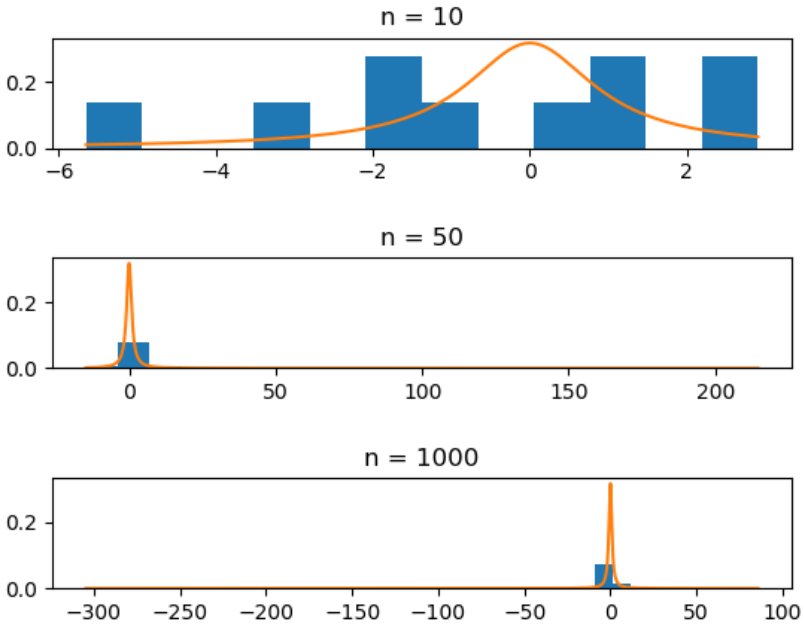

Рис. 2. Распределение Коши (4)

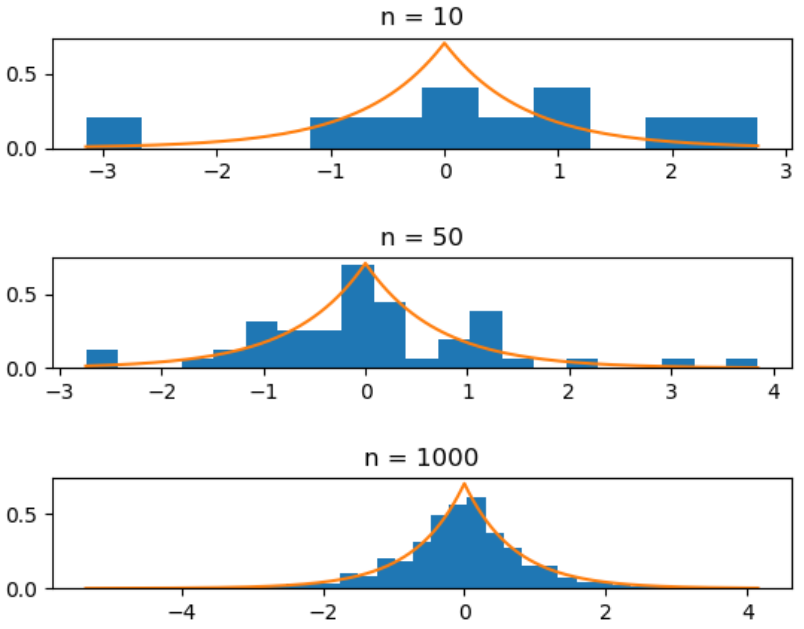

Рис. 3. Распределение Лапласа (5)

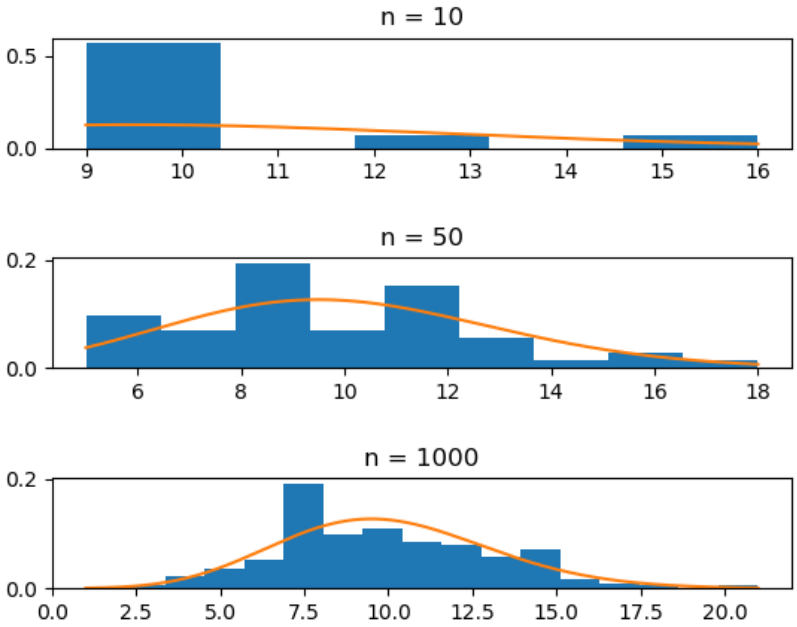

Рис. 4. Распределение Пуассона (6)

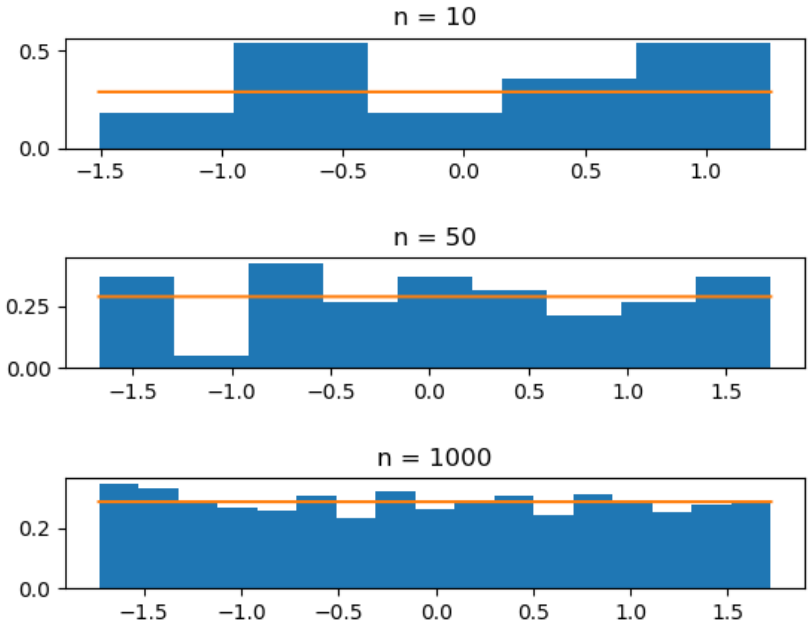

Рис. 5. Равномерное распределение (7)

4.2. Таблицы значений

Normal n = 10					
	$\overline{x}$ (8)	$medx$ (8)	z_R (10)	z_Q (12)	z_{tr} (13)
$E(z)$ (1)	-0.01	0.0	0.0	0.0	-0.4
$D(z)$ (2)	0.09935	0.13874	0.18181	0.11633	0.19166
Normal n = 100					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.01	-0.01	0.0	0.00	-0.53
$D(z)$	0.05471	0.07758	0.13405	0.06471	0.11674
Normal n = 1000					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.00	-0.01	0.0	0.01	-0.57
$D(z)$	0.03683	0.05231	0.11162	0.04357	0.08085

Таблица 1. Нормальное распределение

Cauchy n = 10					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	4	0	23	0	-4
$D(z)$	15908.09147	0.43148	397531.77972	1.26249	523.04326
Cauchy n = 100					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	1	0.0	-18	0.0	-7
$D(z)$	8637.35792	0.22896	1880864.30928	0.65602	2509.25432
Cauchy n = 1000					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	1	0.0	190	0.0	-6
$D(z)$	6106.00336	0.15343	87773120.69259	0.43918	1712.09914

Таблица 2. Распределение Коши

Laplace n = 10					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.0	0.00	0.0	0.0	-0.4
$D(z)$	0.10735	0.08220	0.41080	0.11063	0.18386
Laplace n = 100					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.00	0.00	0.0	0.00	-0.5
$D(z)$	0.05890	0.04395	0.40008	0.06062	0.11308
Laplace n = 1000					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.0	0.0	0.0	0.00	-0.53
$D(z)$	0.03961	0.02948	0.40734	0.04075	0.07833

Таблица 3. Распределение Лапласа

Poisson n = 10					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	10.0	10	10	10	12
$D(z)$	0.94092	1.30249	1.90001	1.19398	1.61149
Poisson n = 100					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	10.0	9.9	11	10.0	12
$D(z)$	0.51966	0.74990	1.51329	0.67369	1.14114
Poisson n = 1000					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	10.0	9.9	11	10.0	12.6
$D(z)$	0.34978	0.50642	1.43907	0.45051	0.81966

Таблица 4. Распределение Пуассона

Uniform n = 10					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.0	0.0	0.01	0.0	-0.4
$D(z)$	0.10099	0.23114	0.04599	0.13631	0.22124
Uniform n = 100					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.00	0.0	0.00	0.01	-0.5
$D(z)$	0.05538	0.13012	0.02329	0.07517	0.13984
Uniform n = 1000					
	$\overline{x}$	$medx$	z_R	z_Q	z_{tr}
$E(z)$	0.00	0.00	0.00	0.01	-0.57
$D(z)$	0.03726	0.08778	0.01553	0.05062	0.09711

Таблица 5. Равномерное распределение

4.3. Упорядоченные характеристики положения

1) Нормальное распределение:

$$z_{tr} < med\, x < \overline{x} = z_R < z_Q \tag{24}$$

2) Распределение Коши:

$$z_{tr} < med\, x < \overline{x} = z_Q < z_R \tag{25}$$

3) Распределение Лапласа:

$$z_{tr} < \overline{x} = med\, x = z_R = z_Q \tag{26}$$

4) Распределение Пуассона:

$$med\, x < \overline{x} = z_Q < z_R < z_{tr} \tag{27}$$

5) Равномерное распределение:

$$z_{tr} < \overline{x} = med\, x = z_R < z_Q \tag{28}$$

4.4. Боксплот Тьюки

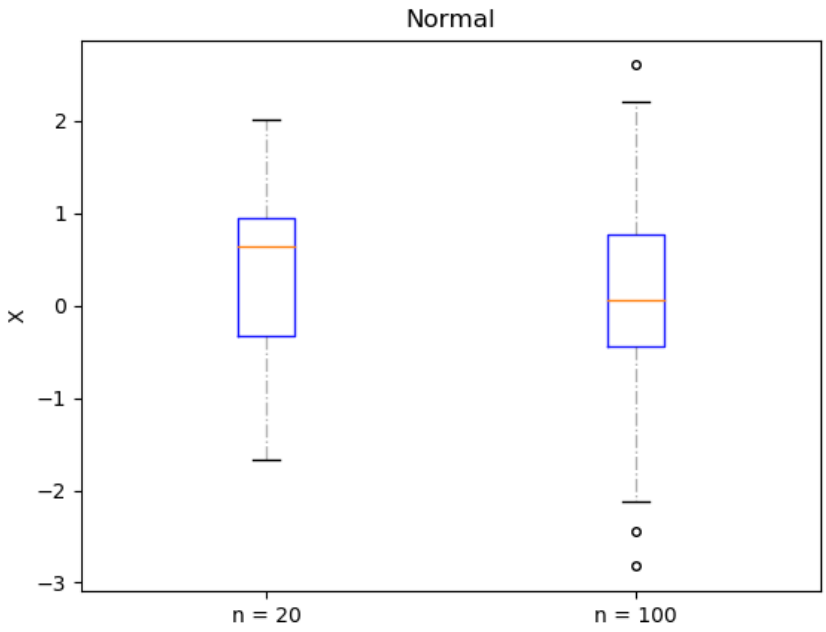

Рис. 6. Нормальное распределение (3)

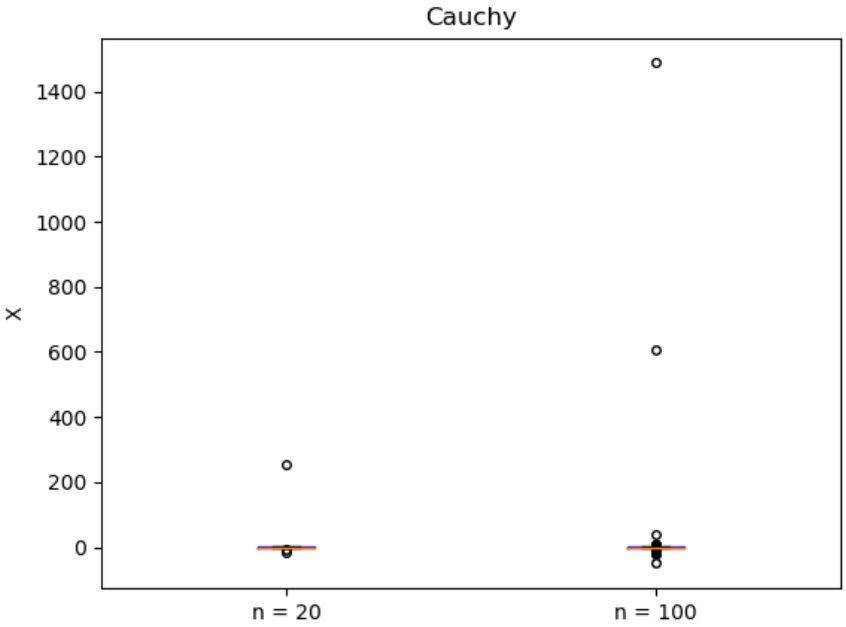

Рис. 7. Распределение Коши (4)

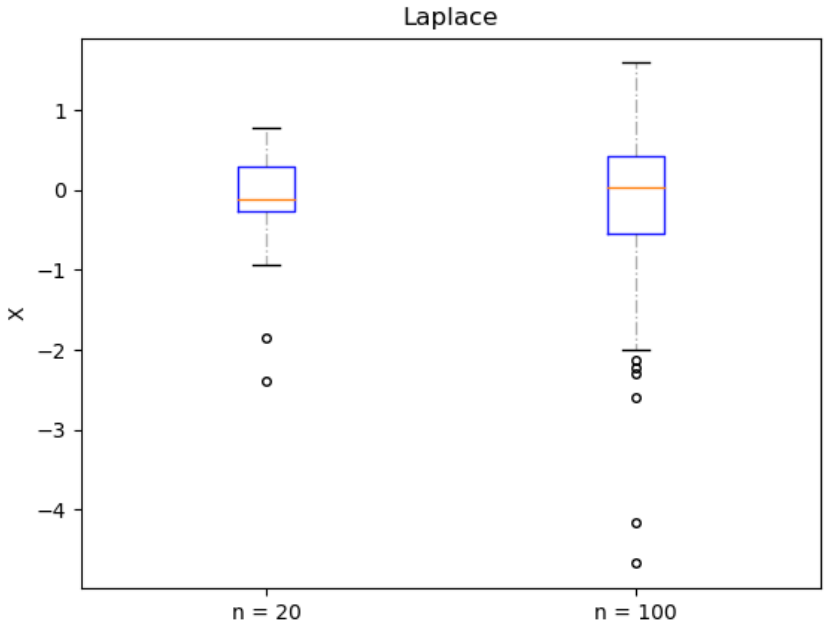

Рис. 8. Распределение Лапласа (5)

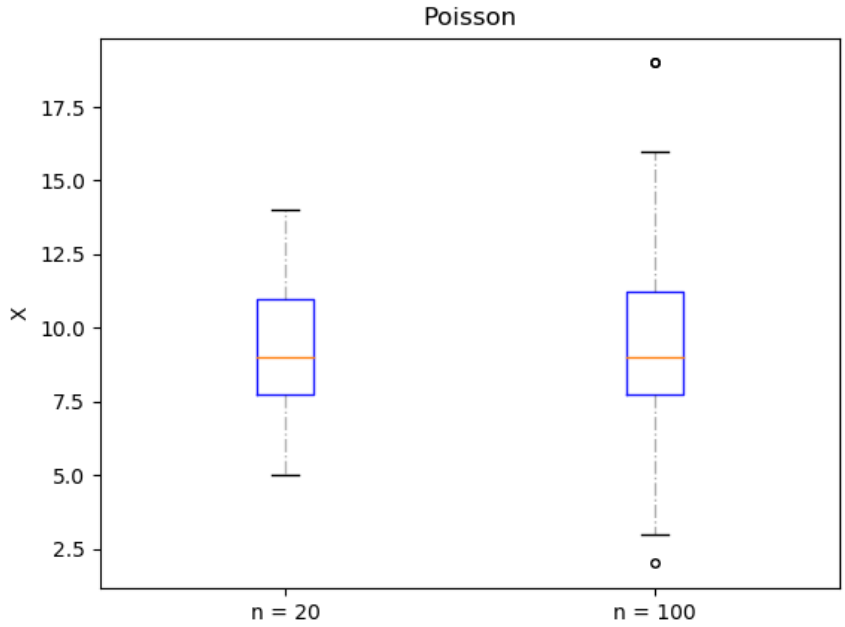

Рис. 9. Распределение Пуассона (6)

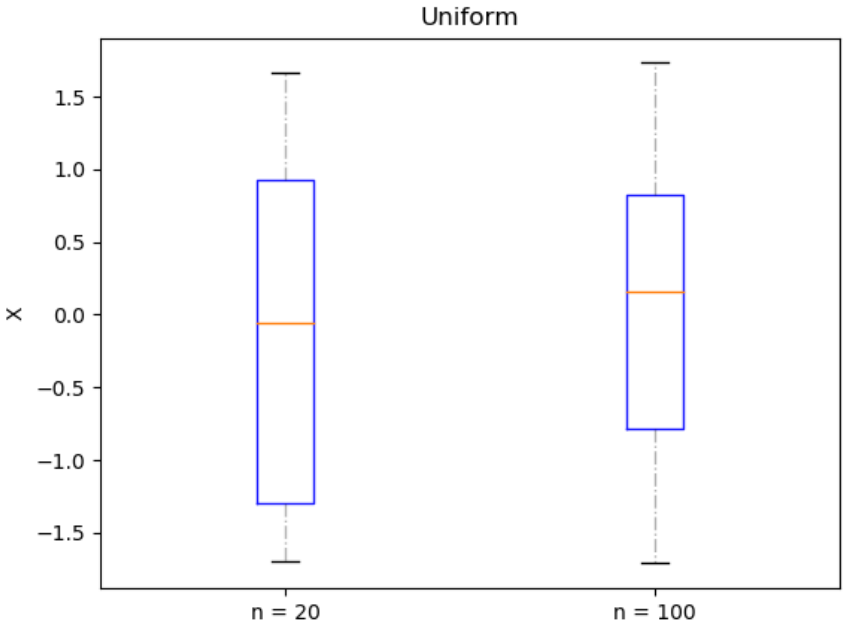

Рис. 10. Равномерное распределение (7)

4.5. Доля выбросов

Выборка	Среднее	Дисперсия
Normal, n = 20	0.017	0.001
Normal, n = 100	0.0095	0.0001
Cauchy, n = 20	0.139	0.005
Cauchy, n = 100	0.154	0.001
Laplace, n = 20	0.065	0.004
Laplace, n = 100	0.0642	0.0009
Poisson, n = 20	0.016	0.002
Poisson, n = 100	0.0101	0.0003
Uniform, n = 20	0	0
Uniform, n = 100	0	0

Таблица 6. Доля выбросов

4.6. Теоретическая вероятность выбросов

Распределение	Q_1^T	Q_3^T	X_1^T	X_2^T	P_B^T (17) (18)
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 7. Теоретическая вероятность выбросов

4.7. Эмпирическая функция распределения

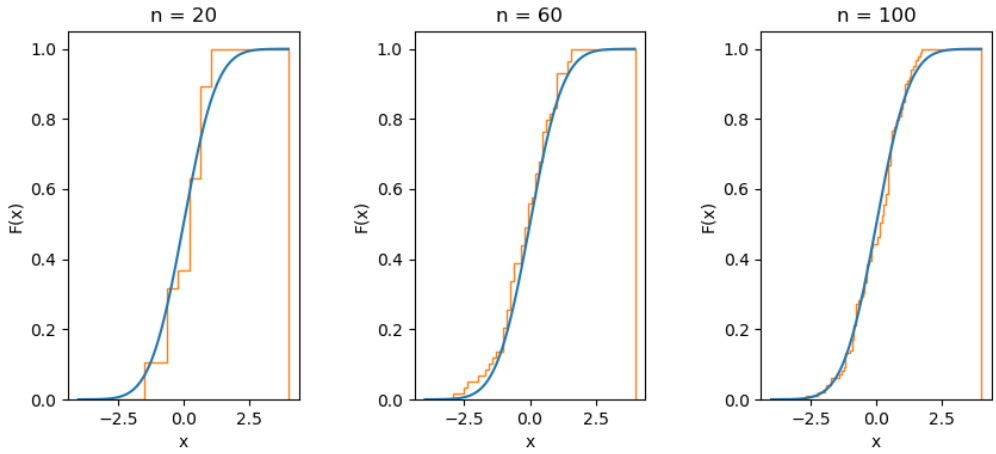

Рис. 11. Нормальное распределение (3)

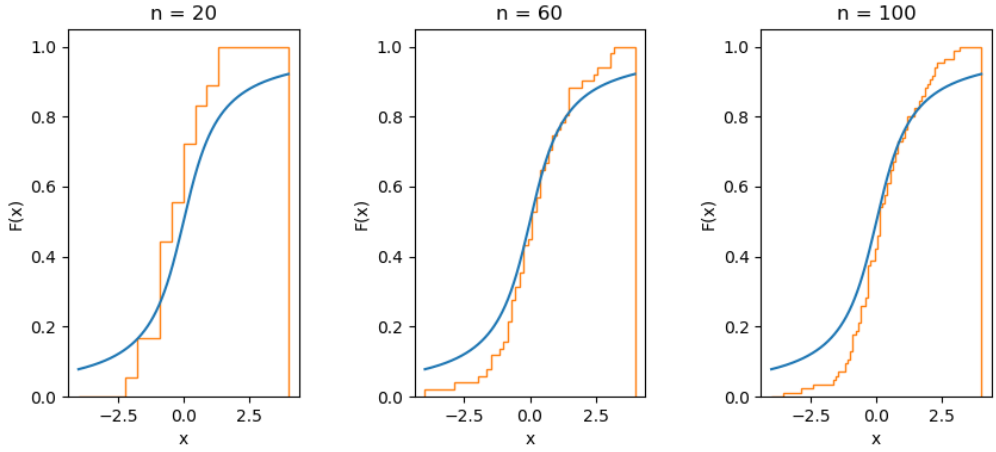

Рис. 12. Распределение Коши (4)

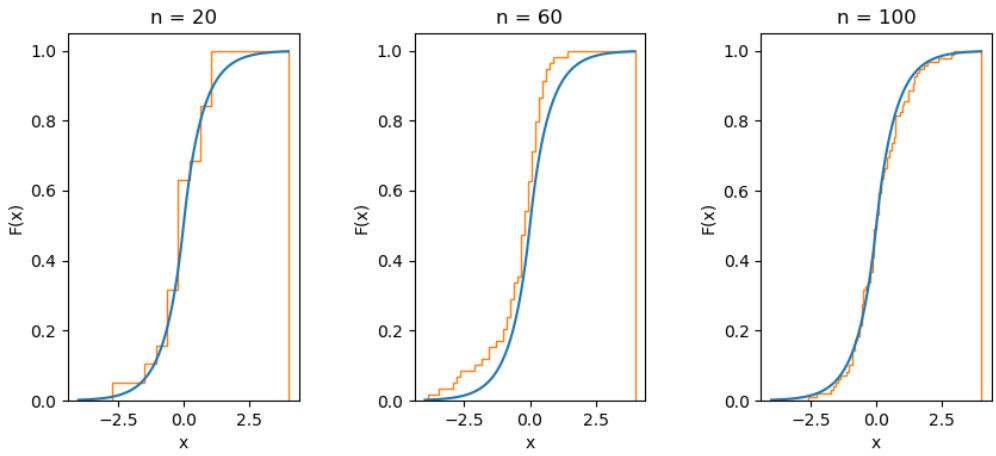

Рис. 13. Распределение Лапласа (5)

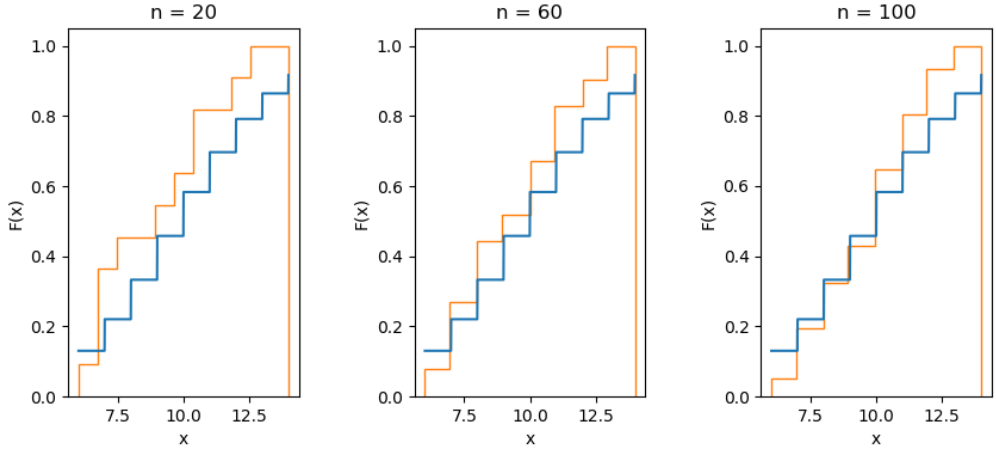

Рис. 14. Распределение Пуассона (6)

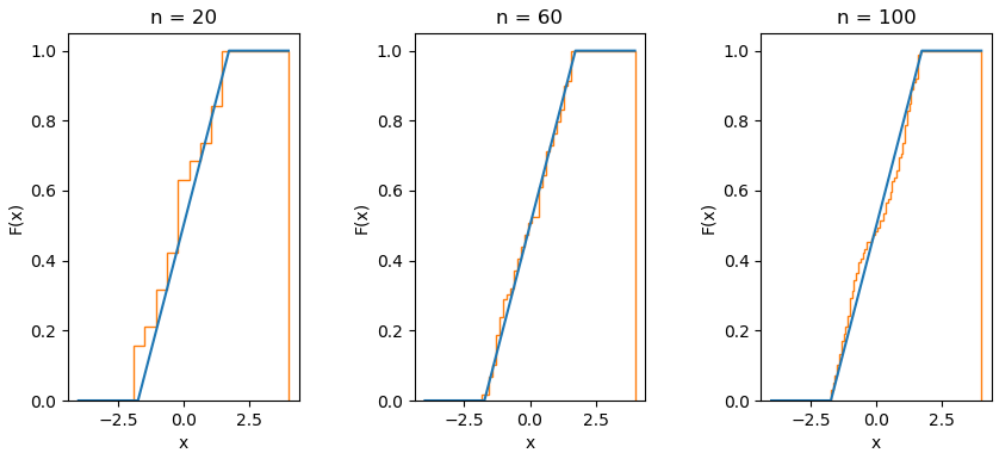

Рис. 15. Равномерное распределение (7)

4.8. Ядерные оценки плотности распределения

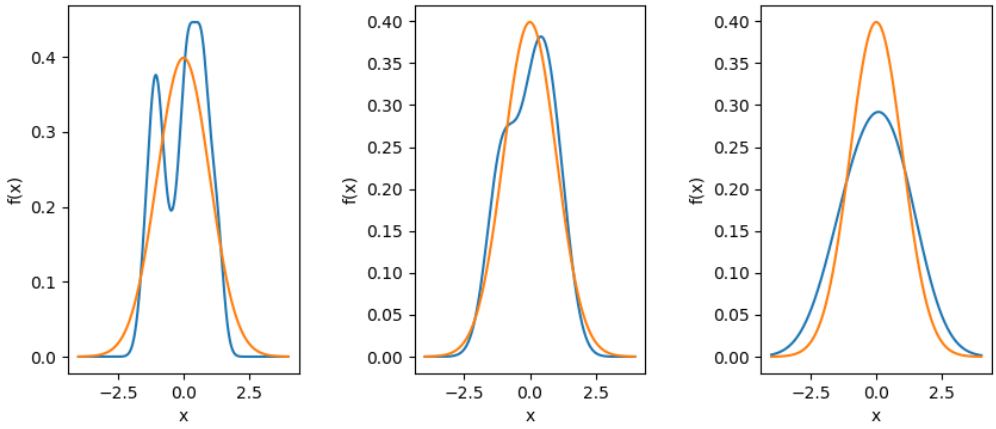

Рис. 16. Нормальное распределение, n = 20 (3)

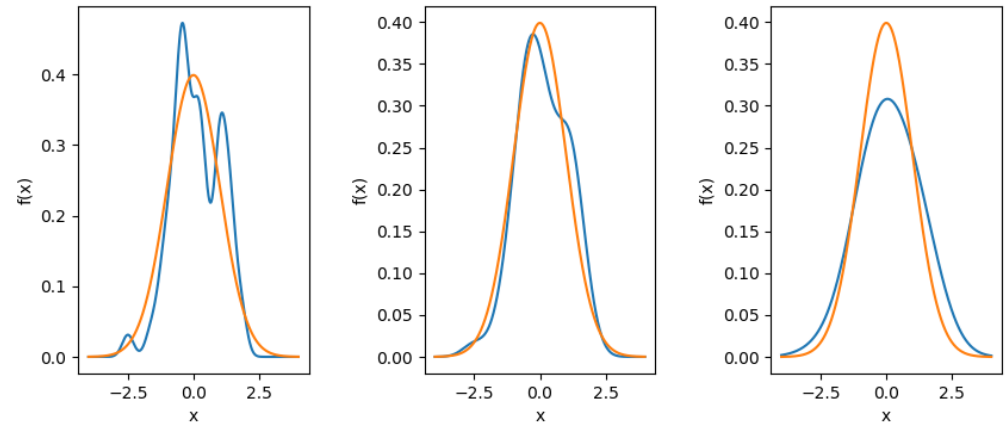

Рис. 17. Нормальное распределение, n = 60

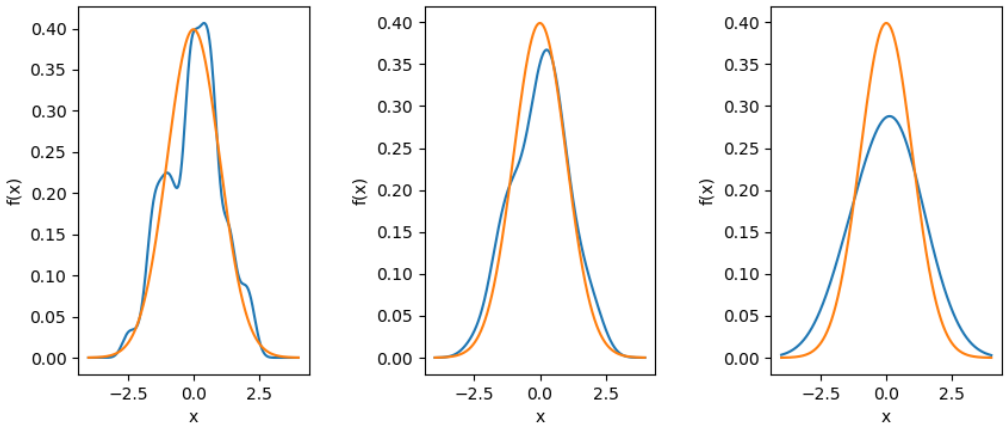

Рис. 18. Нормальное распределение, n = 100

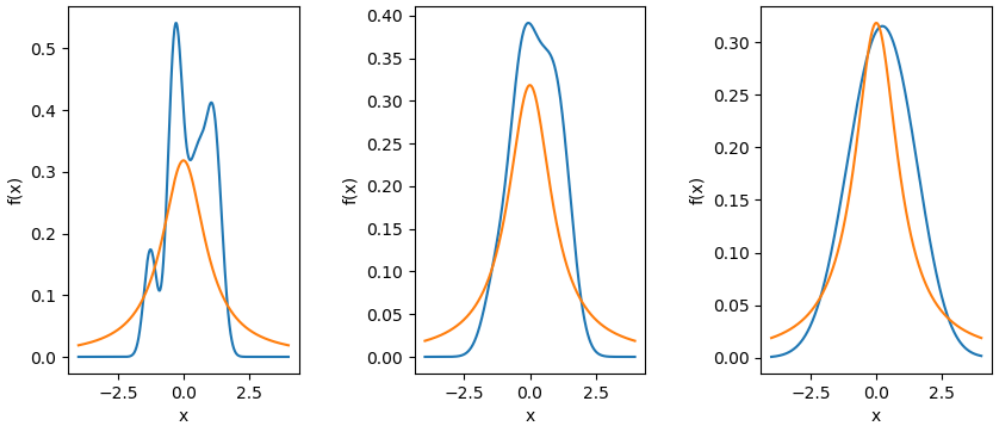

Рис. 19. Распределение Коши, n = 20 (4)

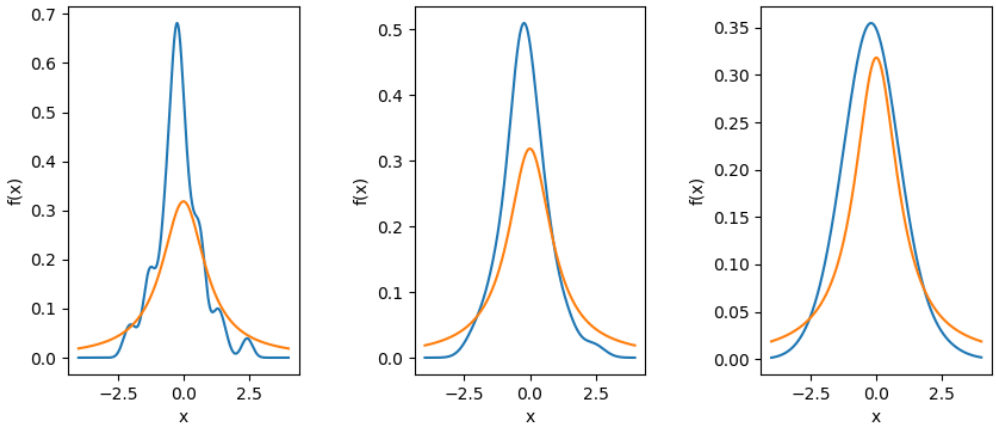

Рис. 20. Распределение Коши, n = 60

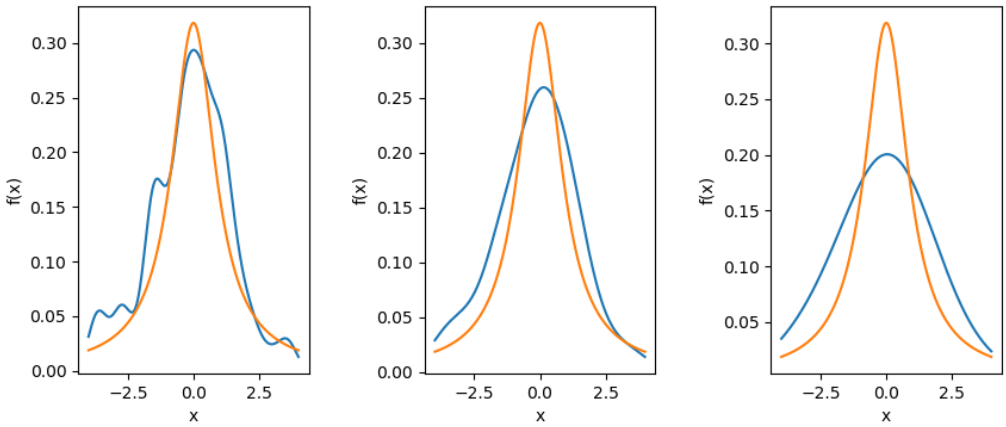

Рис. 21. Распределение Коши, n = 100

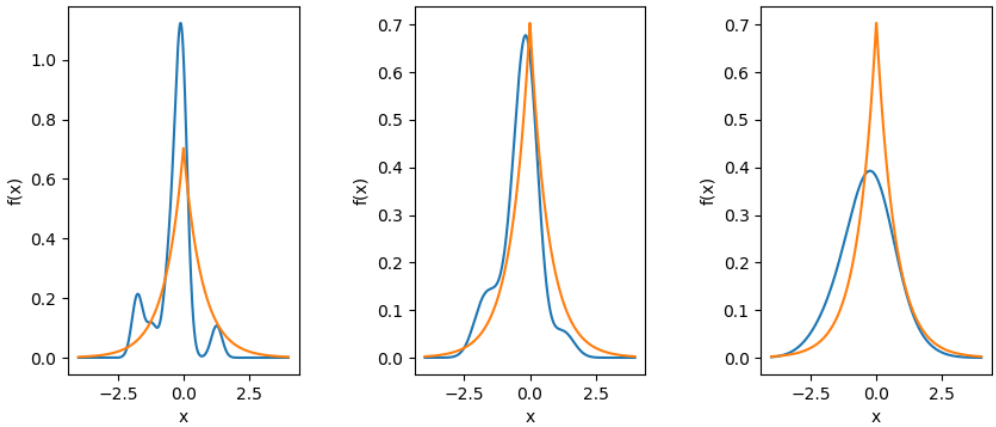

Рис. 22. Распределение Лапласа, n = 20 (5)

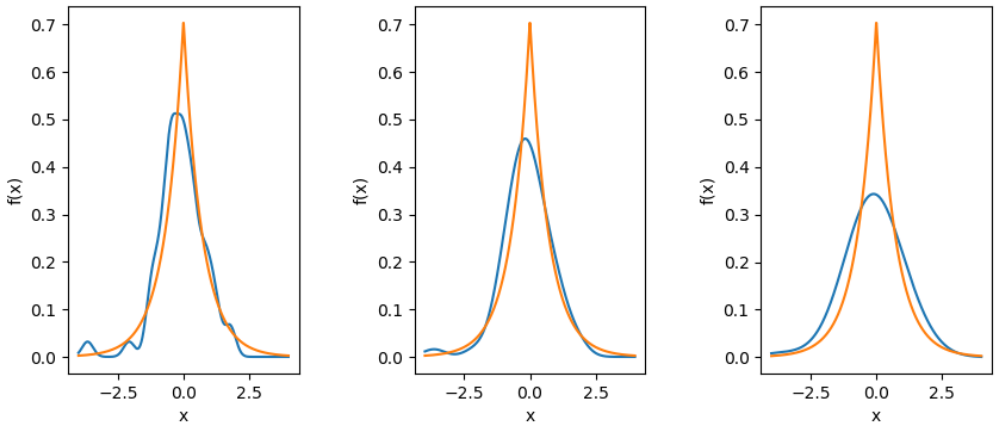

Рис. 23. Распределение Лапласа, n = 60

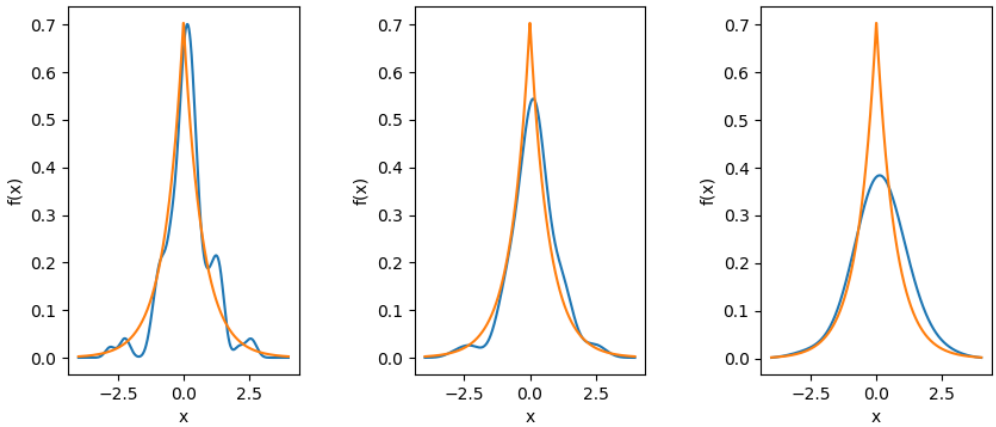

Рис. 24. Распределение Лапласа, n = 100

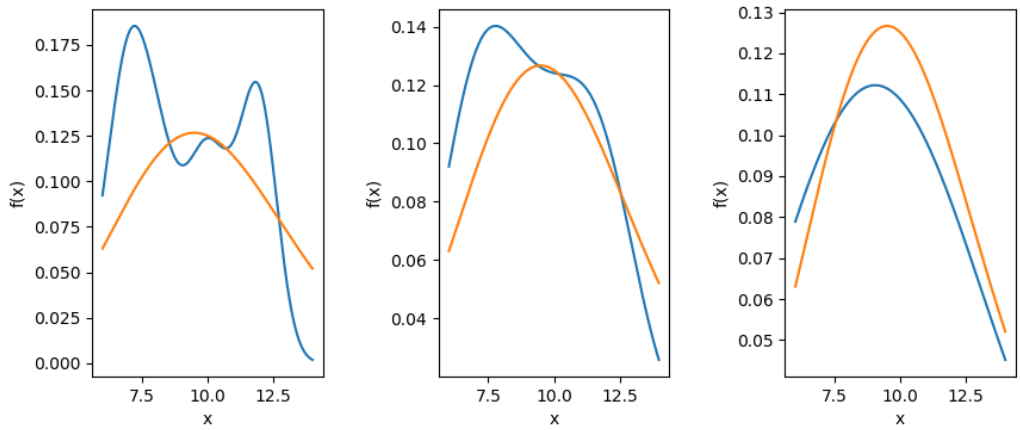

Рис. 25. Распределение Пуассона, n = 20 (6)

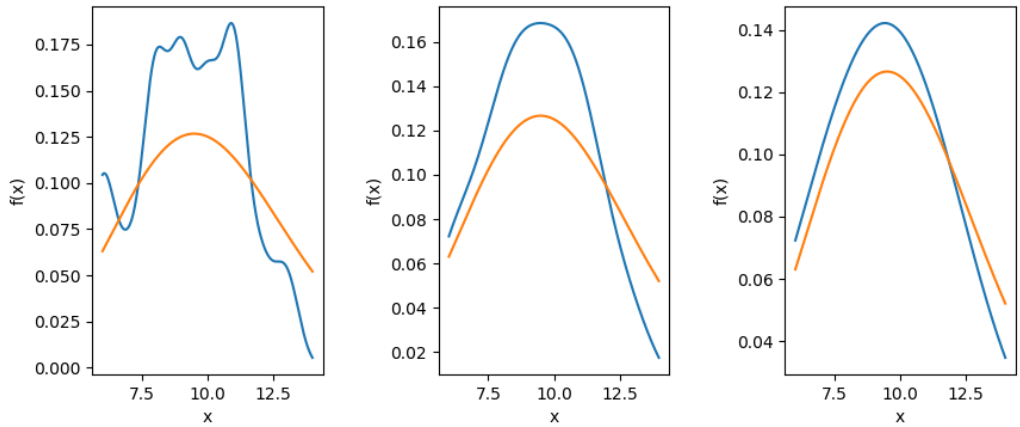

Рис. 26. Распределение Пуассона, n = 60

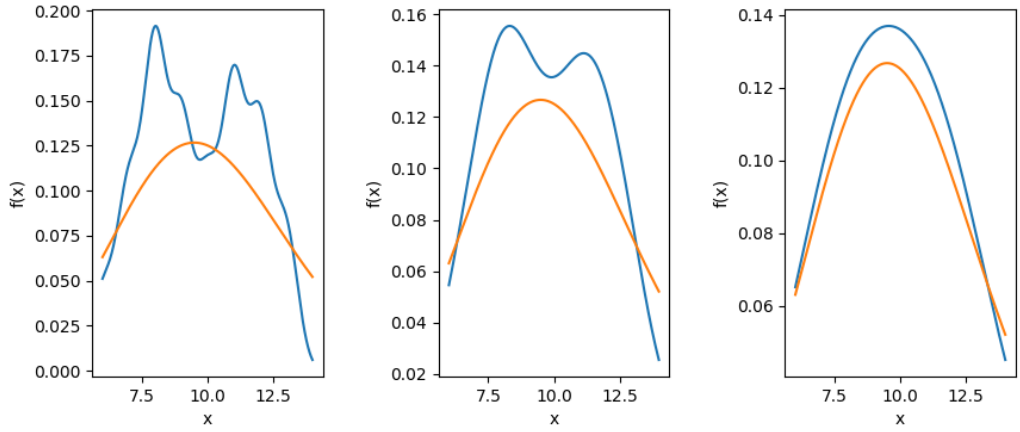

Рис. 27. Распределение Пуассона, n = 100

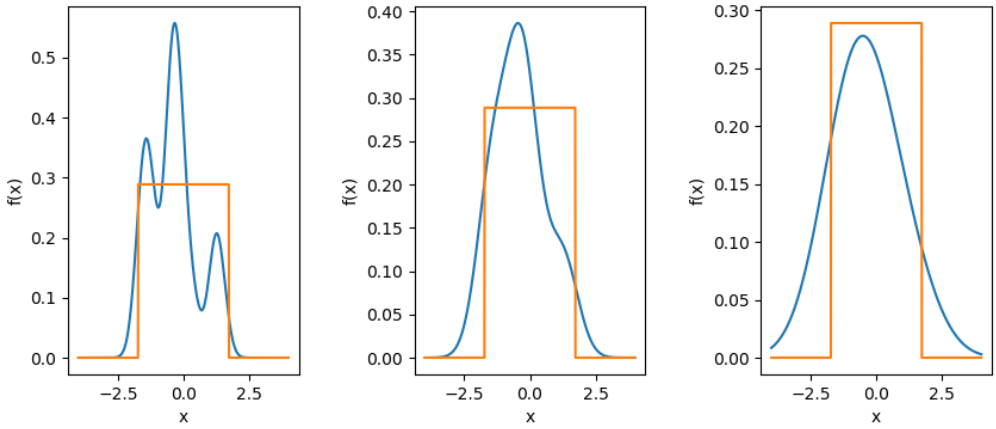

Рис. 28. Равномерное распределение, n = 20 (7)

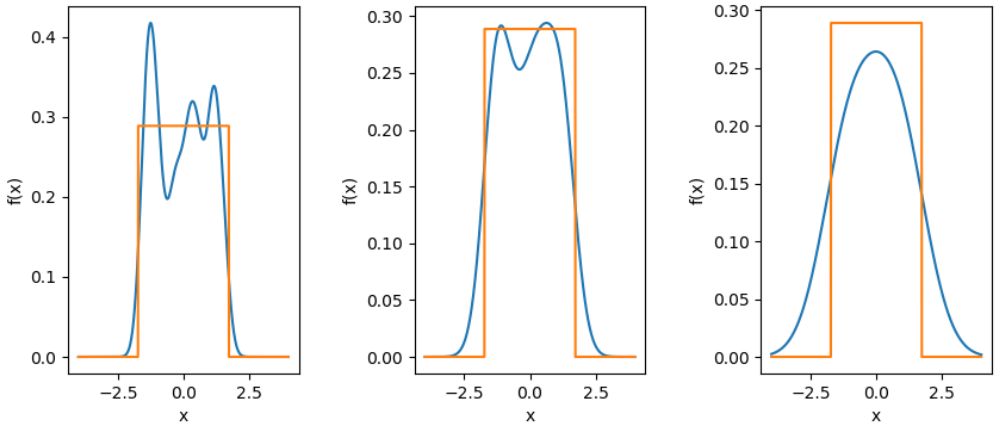

Рис. 29. Равномерное распределение, n = 60

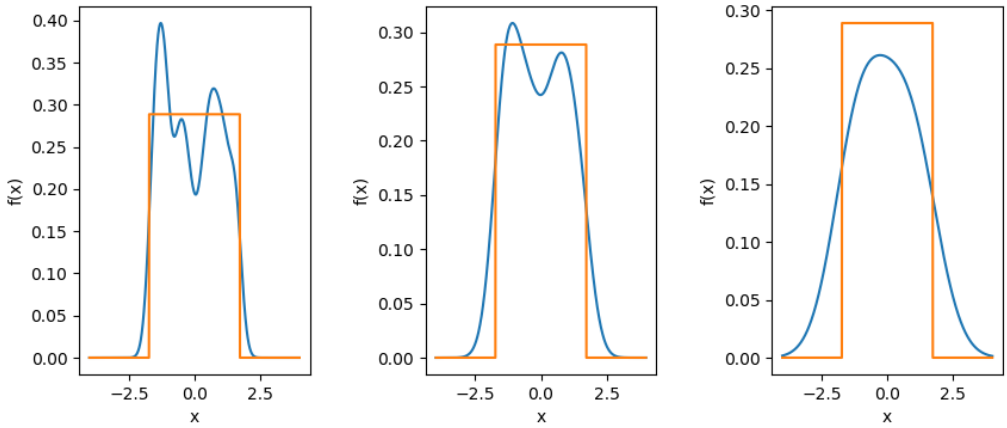

Рис. 30. Равномерное распределение, n = 100

5. Обсуждение

5.1. Гистограмма и график плотности распределения

Из графиков видна чёткая зависимость, увеличение выборки увеличивает точность аппроксимации исходного распределения для всех распределений кроме Коши (4).

5.2. Характеристики положения и рассеяния

Из полученных данных видно, что среднее (1) всех характеристик стремится к теоретическому, а оценка дисперсии (2) к нулю при увеличении размера выборки. В случае распределения Коши (4) это верно только для медианы (9).

5.3. Доля и теоретическая вероятность выбросов

Из полученных данных видно, что средняя доля выбросов (для 1000 эксперементов) стремится к теоретической оценке при увеличении размера выборки, для всех распределений, кроме распределения Пуассона (6). Дисперсия в свою очередь стремится к нулю уже для всех распределений. Также можно заметить, что для равномерного распределения отсутвуют выбросы, а вероятность их появления равна 0.

5.4. Эмпирическая функция и ядерные оценки плотности распределения

Из полученных данных видно, что точность аппроксимации эмпирической функцией (19) распределения увеличивается с увеличением выборки. Для распределения Пуассона (6) точность аппроксимации наименьшая. При увеличении размера выборки увеличевается точность аппроксимации плотности распределения для всех распределений кроме распределения Пуассона. Для нормального и равномерного распределения и рапределения Лапласа лучше подходит параметр $h = h_n$. Для рапределения коши и Пуассона лучше подходит параметр $h = \frac{h_n}{2}$

6. Приложения

Репозиторий на GitHub с релизацией: github.com.