Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторным работам № 3, 4

По дисциплине «Теория вероятностей и Математическая статистика»

Выполнил
Студент гр. 3630201/80101 _____ Печеный Н. А.

Руководитель
к.ф.-м.н., доцент _____ Баженов А. Н.

«___»_____ 2020г.

Санкт-Петербург
2020

Содержание

1 Постановка задачи **5**

2 Теория **6**
2.1 Боксплот Тьюки 6
2.1.1 Построение 6
2.2 Теоретическая вероятность выбросов 6
2.3 Эмпирическая функция распределения 6
2.3.1 Статистический ряд 6
2.3.2 Эмпирическая функция распределения 6
2.3.3 Нахождение э. ф. р. 7
2.4 Оценки плотности вероятности 7
2.4.1 Определение 7
2.4.2 Ядерные оценки 7

3 Реализация **8**

4 Результаты **9**
4.1 Боксплот Тьюки 9
4.2 Сравнение теоретической вероятности и экспериментальной доли выбросов 11
4.3 Эмпирическая функция респределения 12
4.4 Ядерные оценки 17

5 Заключение **22**
5.1 Экспериментальная доля и теоретическая вероятность выбросов 22
5.2 Эмпирическая функция и ядерные оценки плотности распределения 22

Список Литературы **23**

Приложение А. Репозиторий с исходным кодом **24**

Список иллюстраций

1 Боксплоты выборок нормального распределения 9
2 Боксплоты выборок распределения Коши 9
3 Боксплоты выборок распределения Лапласа 10
4 Боксплоты выборок распределения Пуассона 10
5 Боксплоты выборок равномерного распределения 11
6 Э. ф. р. нормального распределения $N(x, 0, 1)$ 12
7 Э. ф. р. распределения Коши $C(x, 0, 1)$ 13
8 Э. ф. р. распределения Лапласа $L(x, 0, 1/\sqrt{2})$ 14
9 Э. ф. р. распределения Пуассона $P(k, 10)$ 15
10 Э. ф. р. равномерного распределения $U(x, -\sqrt{3}, \sqrt{3})$ 16
11 Ядерная оценка плотности нормального распределения $N(x, 0, 1)$ 17
12 Ядерная оценка плотности распределения Коши $C(x, 0, 1)$ 18
13 Ядерная оценка плотности распределения Лапласа $L(x, 0, 1/\sqrt{2})$ 19
14 Ядерная оценка плотности распределения Пуассона $P(k, 10)$ 20
15 Ядерная оценка плотности равномерного распределения $U(x, -\sqrt{3}, \sqrt{3})$ . . . 21

Список таблиц

1 Таблица распределения 7
2 Сравнение экспериментальной доли и теоретической вероятности выбросов 11

1 Постановка задачи

Для 5 распределений:

- Нормальное распределение $N(x, 0, 1)$
- Распределение Коши $C(x, 0, 1)$
- Распределение Лапласа $L(x, 0, 1/\sqrt{2})$
- Распределение Пуассона $P(k, 10)$
- Равномерное распределение $U(x, -\sqrt{3}, \sqrt{3})$

Требуется:

1. 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. 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 Боксплот Тьюки

2.1.1 Построение

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

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

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 [1].

2.2 Теоретическая вероятность выбросов

Можно вычислить теоретические первый и третий квартили распределений — Q_1^T и Q_3^T. По ф-ле (1) — теоретические нижнюю и верхнюю границы уса — X_1^T и X_2^T. Выбросы — величины x:

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

Теоретическая вероятность выбросов:

- для непрерывных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = F(X_1^T) + (1 - F(X_2^T)) \tag{2}$$

- для дискретных распределений

$$P_B^T = P(x < X_1^T) + P(x > X_2^T) = (F(X_1^T) - P(X = X_1^T)) + (1 - F(X_2^T)) \tag{3}$$

где $F(X) = P(x \leqslant X)$ — функция распределения.

2.3 Эмпирическая функция распределения

2.3.1 Статистический ряд

Статистический ряд — последовательность различных элементов выборки $z_1, z_2, \ldots, z_k$, расположенных в возрастающем порядке с указанием частот $n_1, n_2, \ldots, n_k$, с которыми эти элементы содержатся в выборке. Обычно записывается в виде таблицы [2].

2.3.2 Эмпирическая функция распределения

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

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

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

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

$F^*(x)$ — функция распределения дискретной случайной величины X^*, заданной таблицей распределения, приведённой ниже в Таблице 1.

Таблица 1: Таблица распределения

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

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

$$F_n^*(x) \approx F_X(x). \tag{4}$$

2.4 Оценки плотности вероятности

2.4.1 Определение

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

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

2.4.2 Ядерные оценки

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

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

Здесь функция $K(u)$, называемая ядерной (ядром), непрерывна и является плотностью вероятности, $x_1, \dots, x_n$ — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, облажающая свойствами

$$h_n \xrightarrow[n\to\infty]{} 0; \qquad \frac{h_n}{n^{-1}} \xrightarrow[n\to\infty]{} \infty.$$

Такие оценки называются непрерывными ядерными.
Гауссово (нормальное) ядро:

$$K(u) = \frac{1}{\sqrt{2\,\pi}} e^{-\frac{u^2}{2}}. \tag{5}$$

Правило Сильвермана:

$$h_n = 1.06\hat{\sigma} n^{-1/5}, \tag{6}$$

где $\hat{\sigma}$ — выборочное стандартное отклонение[2].

3 Реализация

Расчёты проводились в среде аналитических вычислений Maxima. Для генерации выборок и создания и отрисовки графиков были использованы библиотечные функции среды разработки. Код скрипта представлен в репозитории на GitHub, ссылка на репозиторий находится в **Приложении А**.

4 Результаты

4.1 Боксплот Тьюки

На каждом из представленных рисунков 1-5 по горизонатльной оси отложены значения элементов выборки, по вертикальной - размеры выборок.

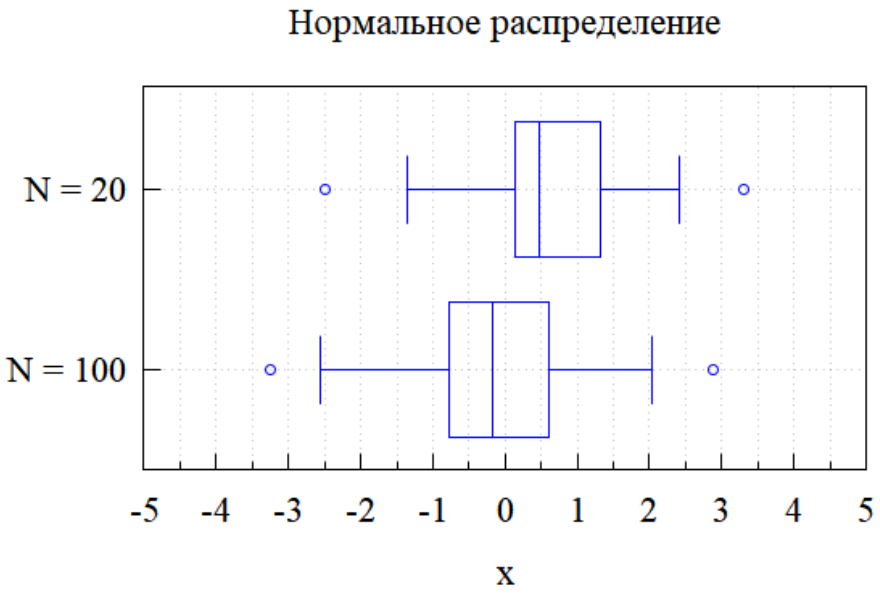


Рис. 1: Боксплоты выборок нормального распределения

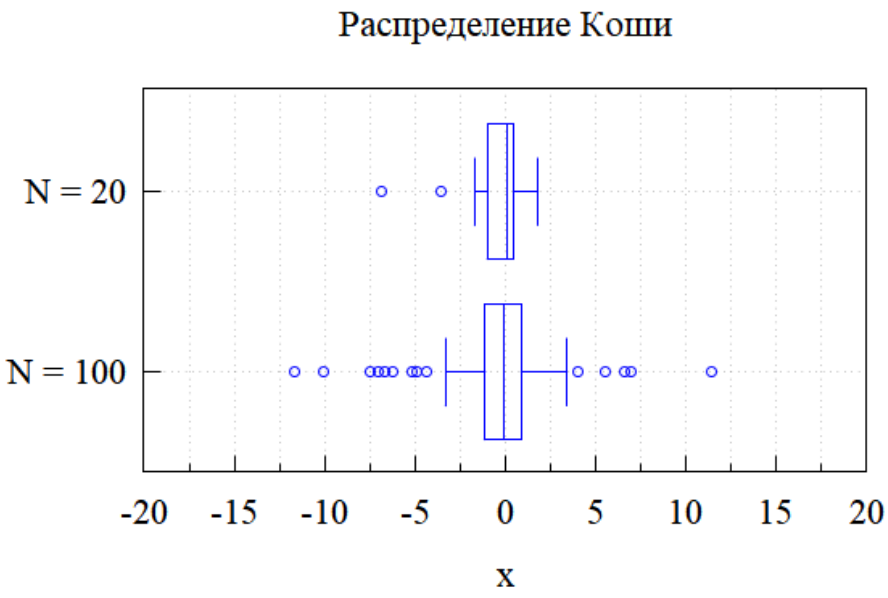


Рис. 2: Боксплоты выборок распределения Коши

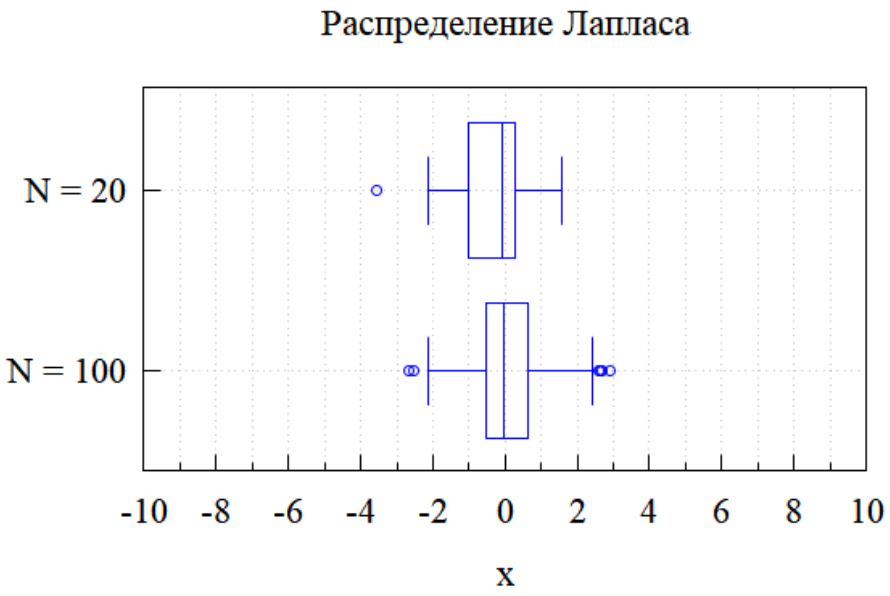


Рис. 3: Боксплоты выборок распределения Лапласа

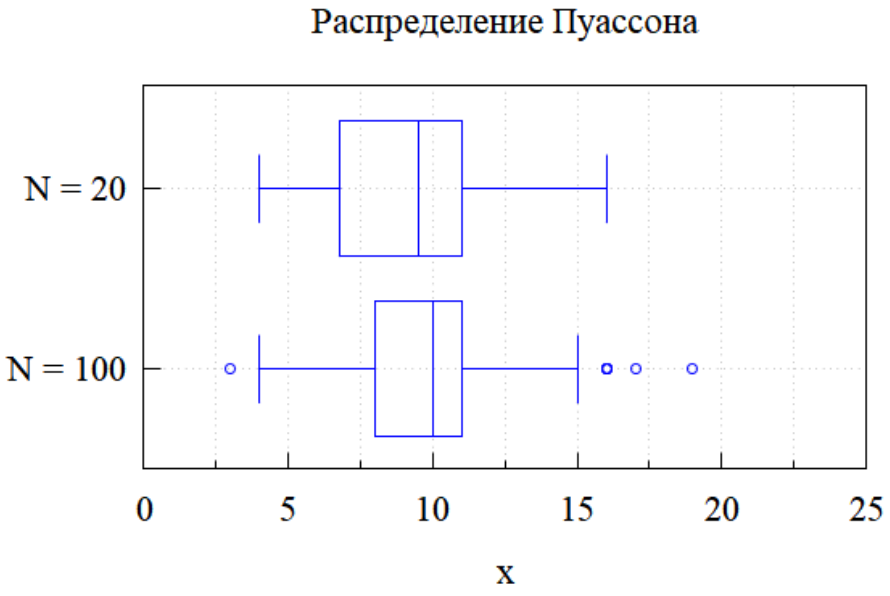


Рис. 4: Боксплоты выборок распределения Пуассона

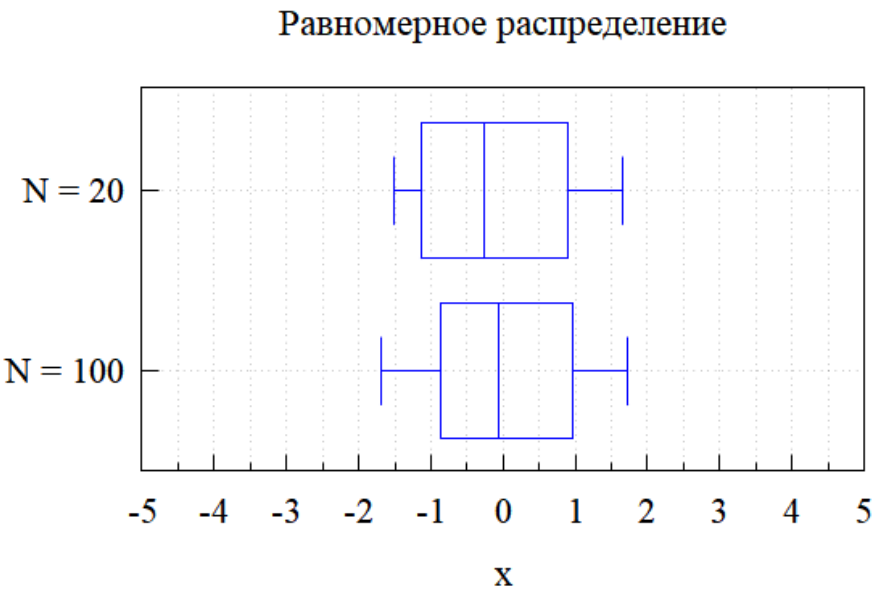


Рис. 5: Боксплоты выборок равномерного распределения

4.2 Сравнение теоретической вероятности и экспериментальной доли выбросов

Для каждого распределения были 1000 раз сгенерированы выборки, соответствующие распределению, была вычислена средня доля выбросов. Погрешность средней доли выбросов была рассчитана по формуле

$$\Delta_z = \sqrt{\left(\overline{z^2} - \overline{z}^2\right)} \tag{7}$$

Значения экспериментальной средней доли выбросов были округлены в соответствии с погрешностями. Результаты представлены ниже в Таблице 2.

Таблица 2: Сравнение экспериментальной доли и теоретической вероятности выбросов

Распределение	N	Доля выбросов	$P_B^T(2)$
Нормальное	20	0.00 ± 0.02	0.0069
	100	0.007 ± 0.008	
Коши	20	0.16 ± 0.08	0.156
	100	0.16 ± 0.04	
Лапласа	20	0.06 ± 0.06	0.0625
	100	0.06 ± 0.02	
Пуассона	20	0.01 ± 0.02	0.0099
	100	0.008 ± 0.009	
Нормальное	20	0.0 ± 0.0	0.0
	100	0.0 ± 0.0	

4.3 Эмпирическая функция респределения

На рисунках 6-10 представлены графики, по которым можно оценить отклонение эмпирических функций (4) распределений от теоретических.

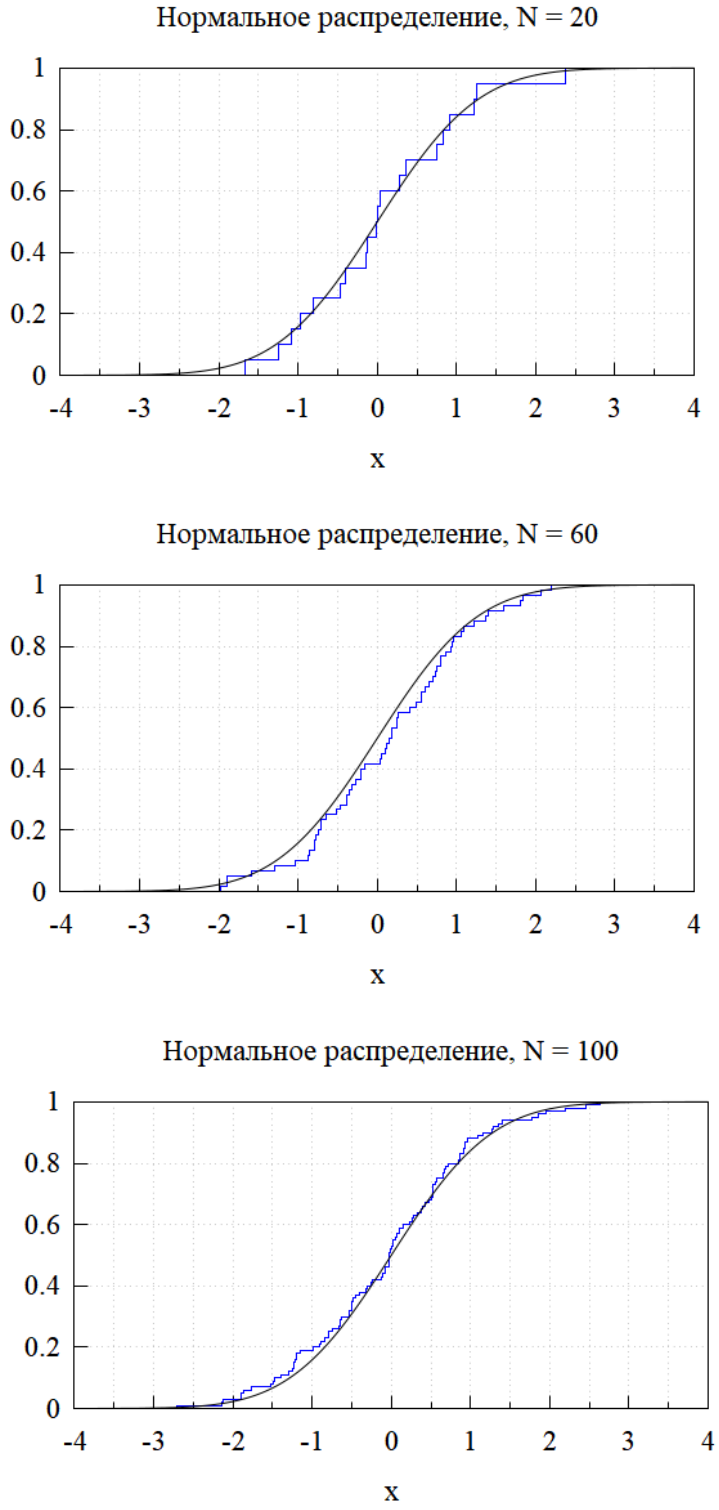


Рис. 6: Э. ф. р. нормального распределения $N(x, 0, 1)$

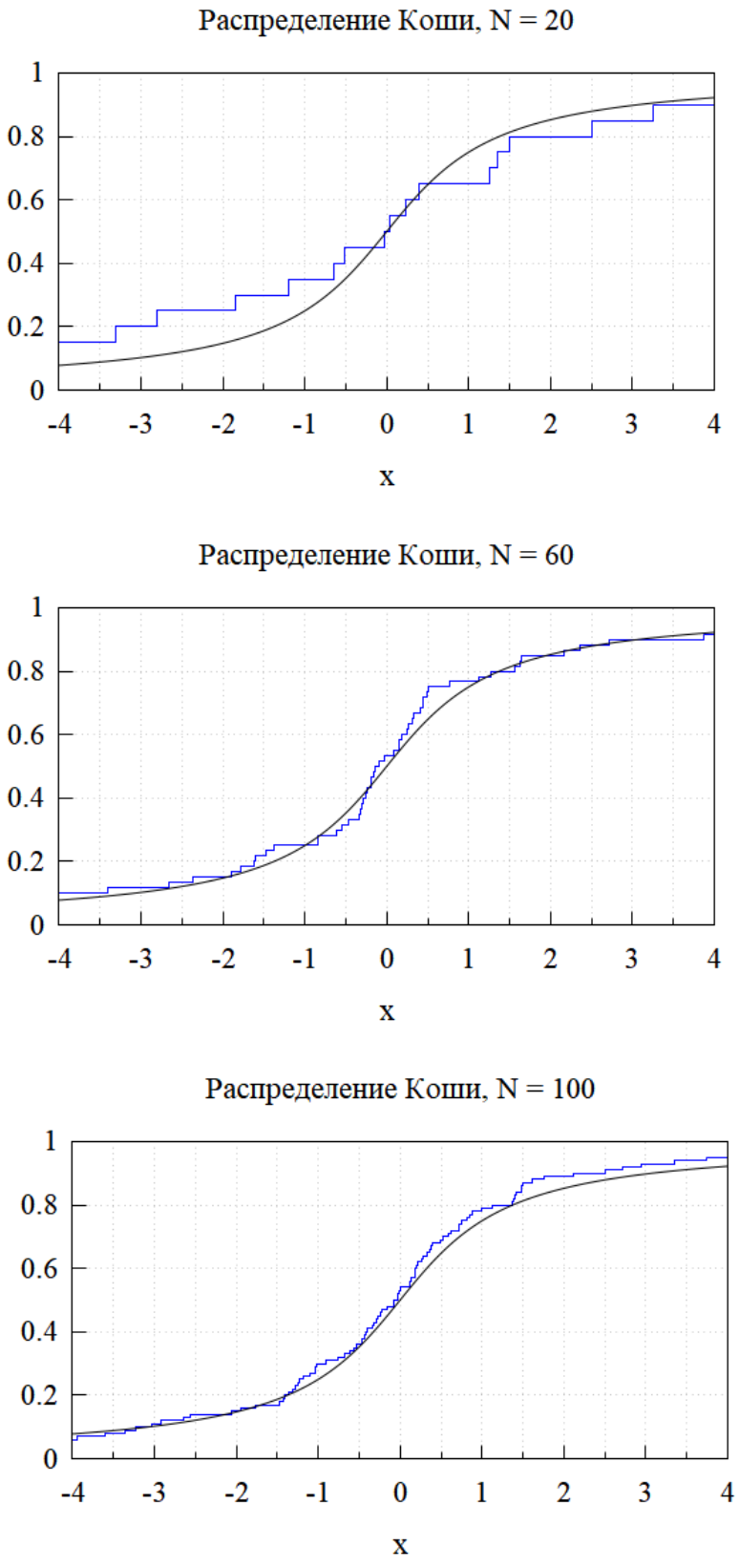


Рис. 7: Э. ф. р. распределения Коши $C(x, 0, 1)$

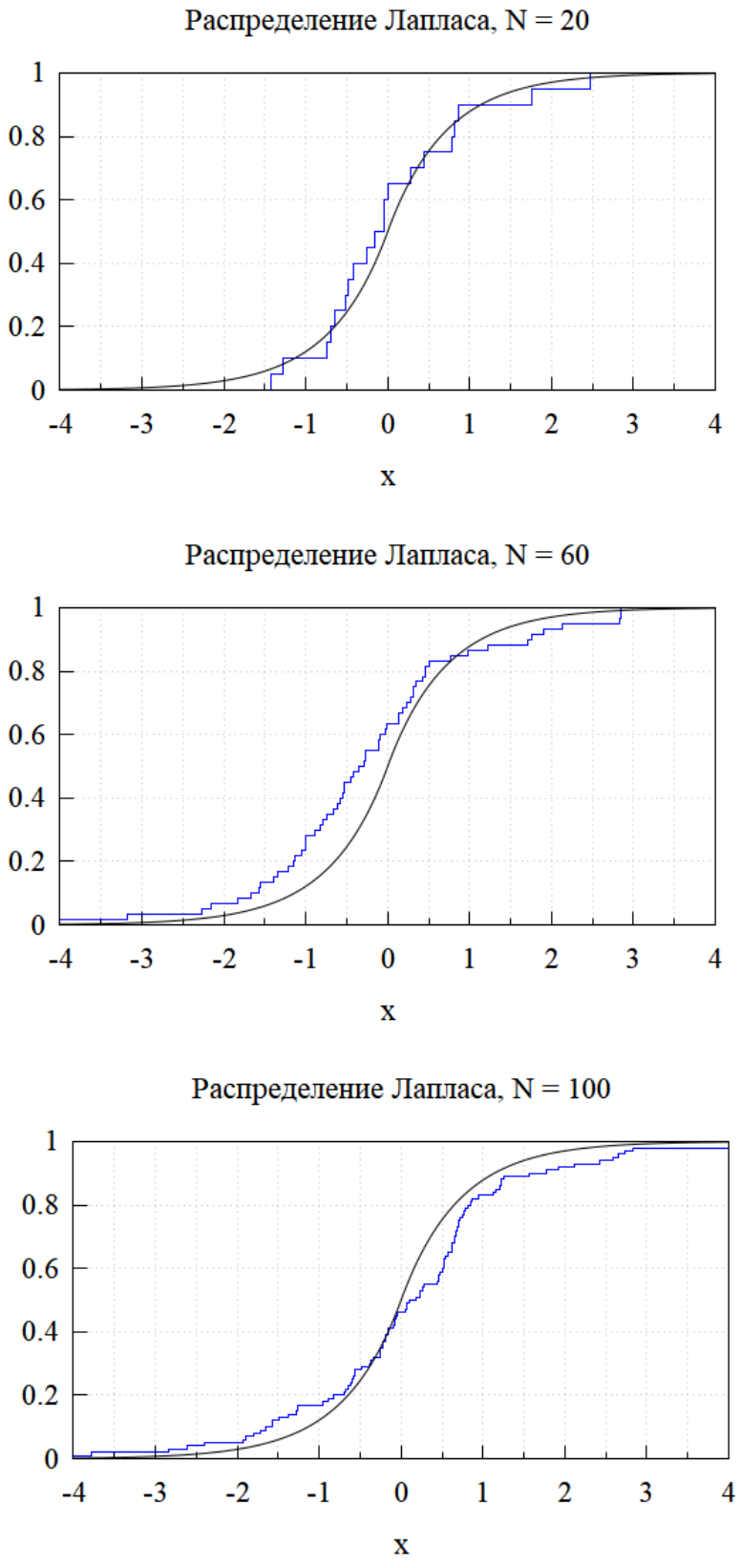
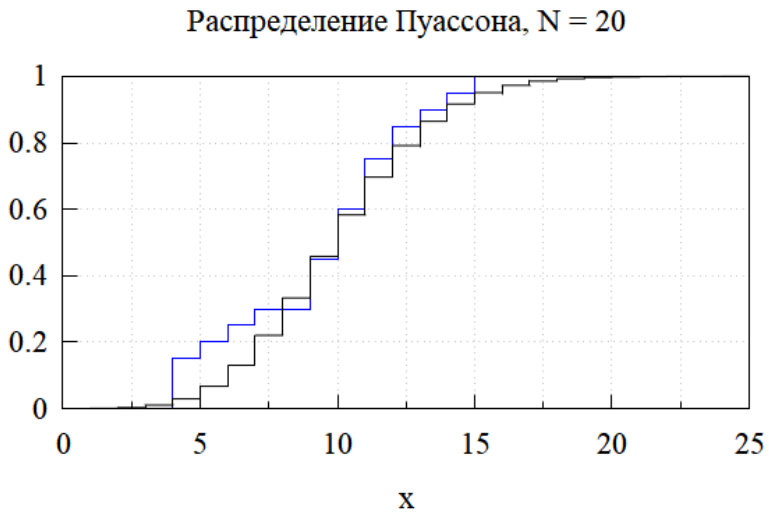
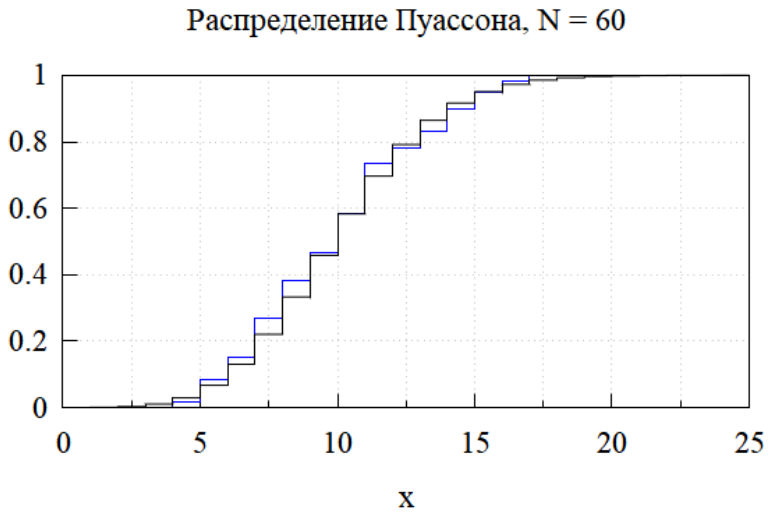


Рис. 8: Э. ф. р. распределения Лапласа $L(x, 0, 1/\sqrt{2})$





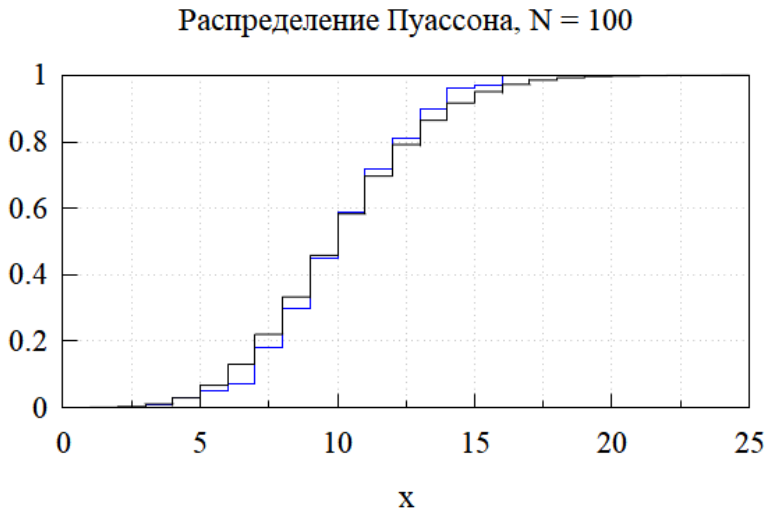


Рис. 9: Э. ф. р. распределения Пуассона $P(k, 10)$

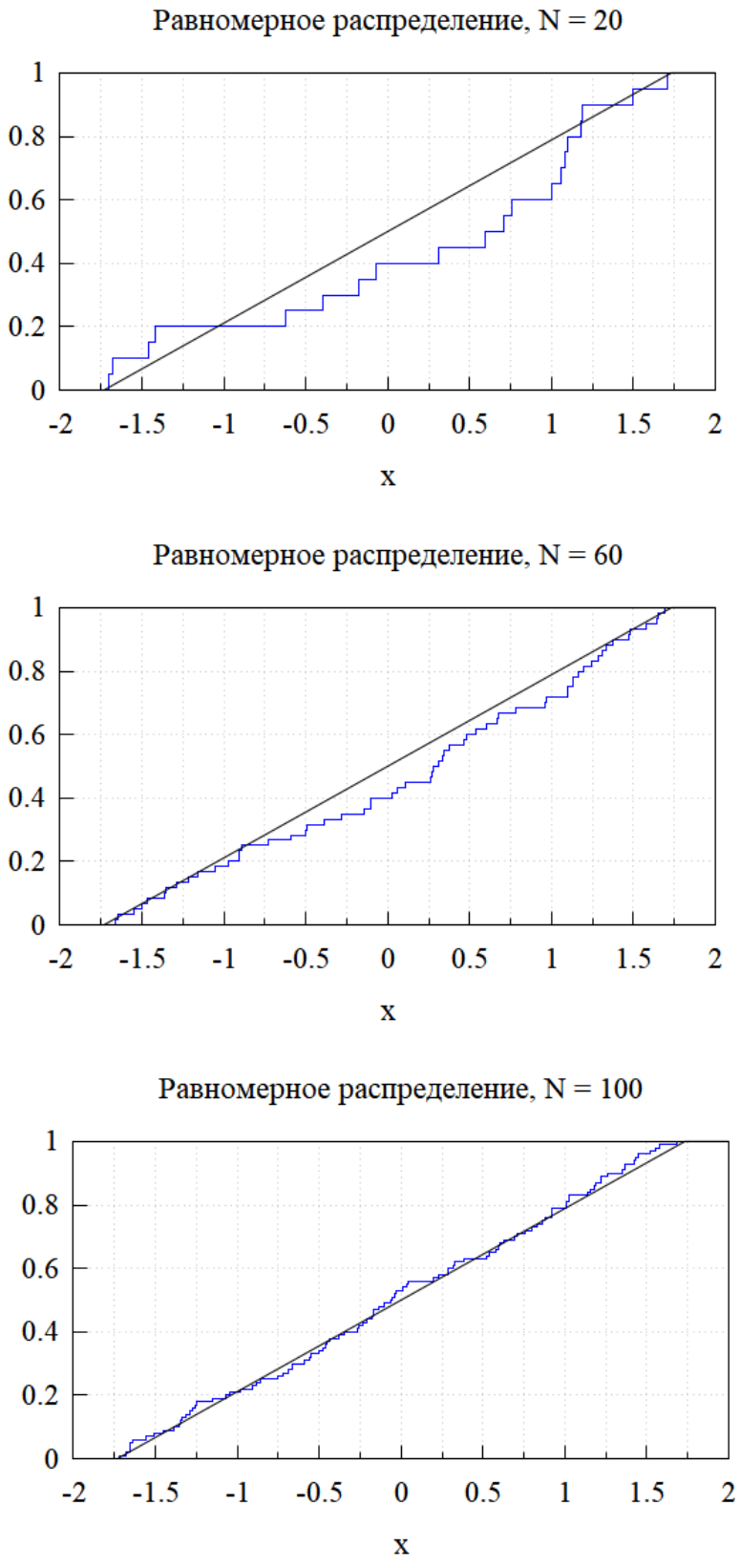


Рис. 10: Э. ф. р. равномерного распределения $U(x, -\sqrt{3}, \sqrt{3})$

4.4 Ядерные оценки

На рисунках 11-15 представлены графики ядерных оценок плотности для выборок, соответствующих заданным распределениям. Ядерная функция имеет вид (5). Чёрной линией обозначена теоретическая плотность вероятности, красной линией обозначена ядерная оценка с параметром сглаживания $h_n/2$, зелёной — h_n, синей — $2h_n$. Параметр сглаживания рассчитан по формуле (6).

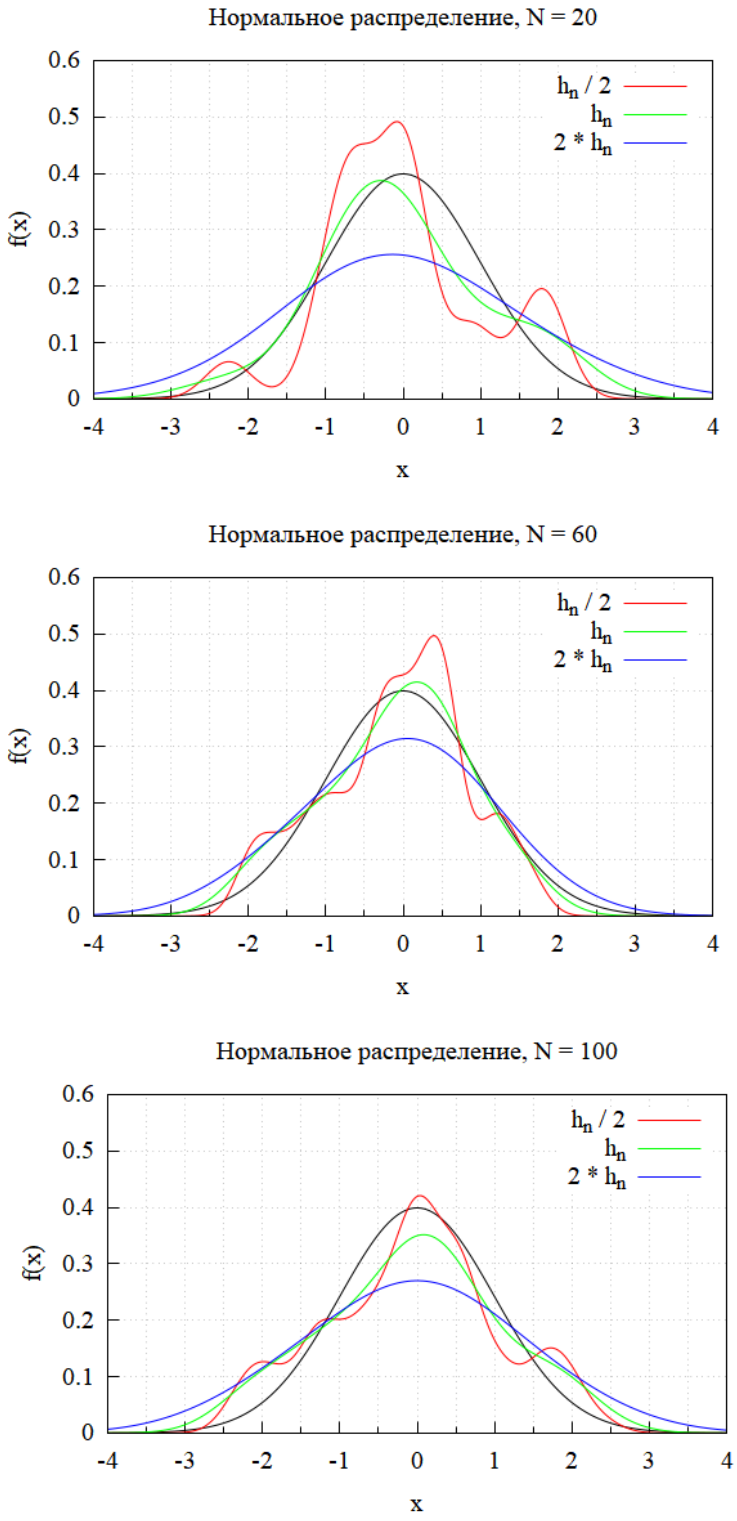


Рис. 11: Ядерная оценка плотности нормального распределения $N(x, 0, 1)$

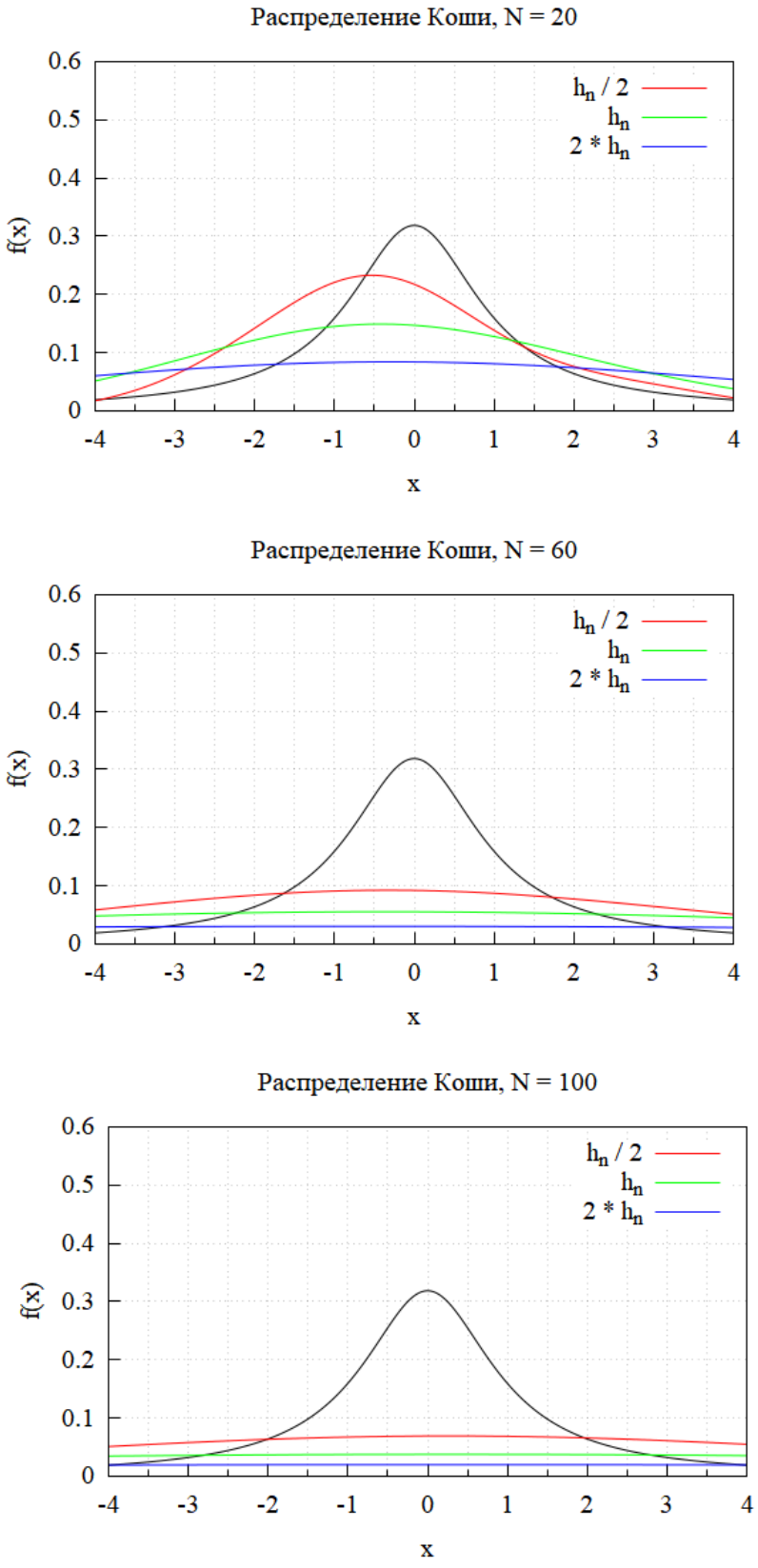


Рис. 12: Ядерная оценка плотности распределения Коши $C(x, 0, 1)$

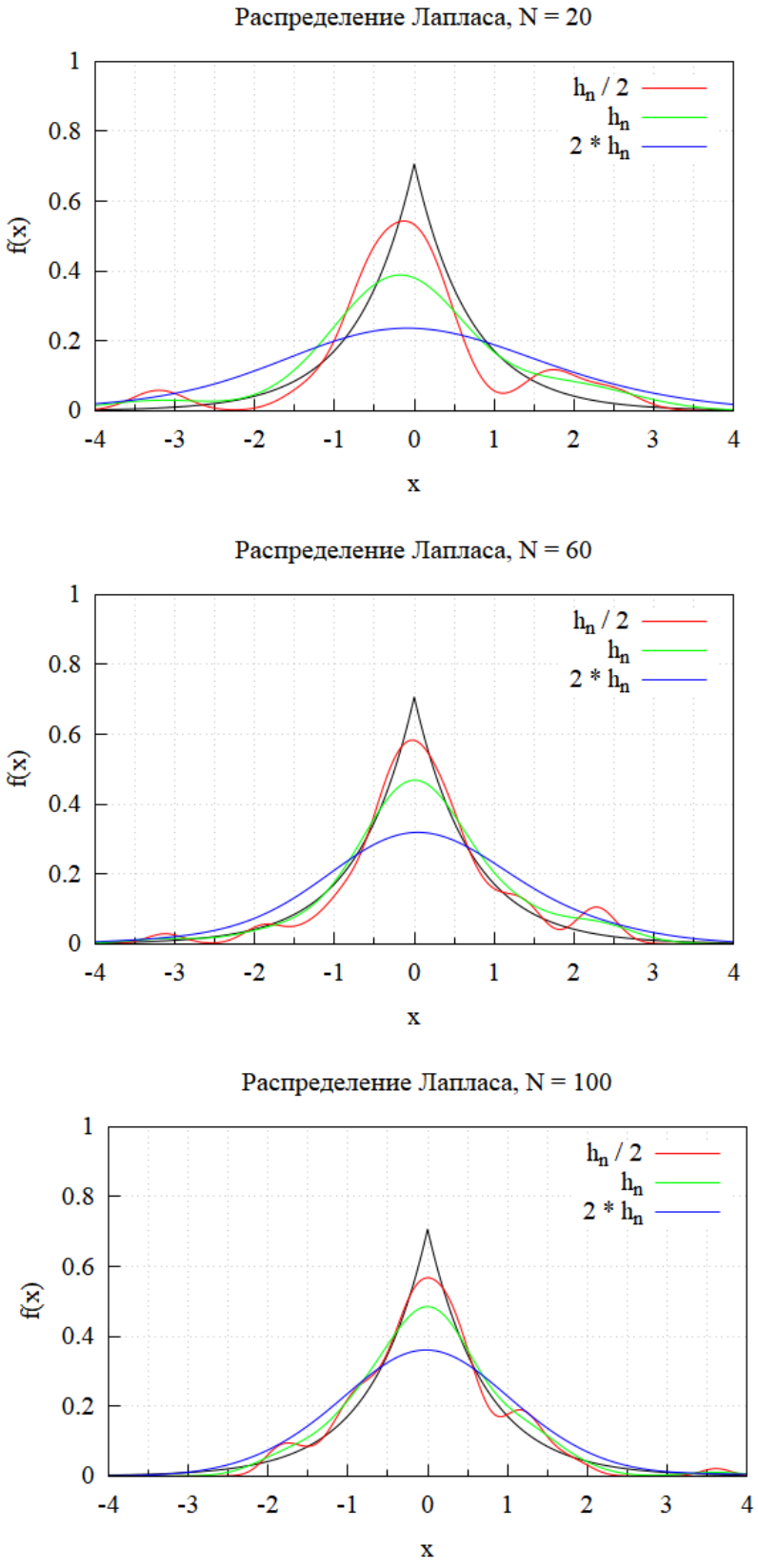


Рис. 13: Ядерная оценка плотности распределения Лапласа $L(x, 0, 1/\sqrt{2})$

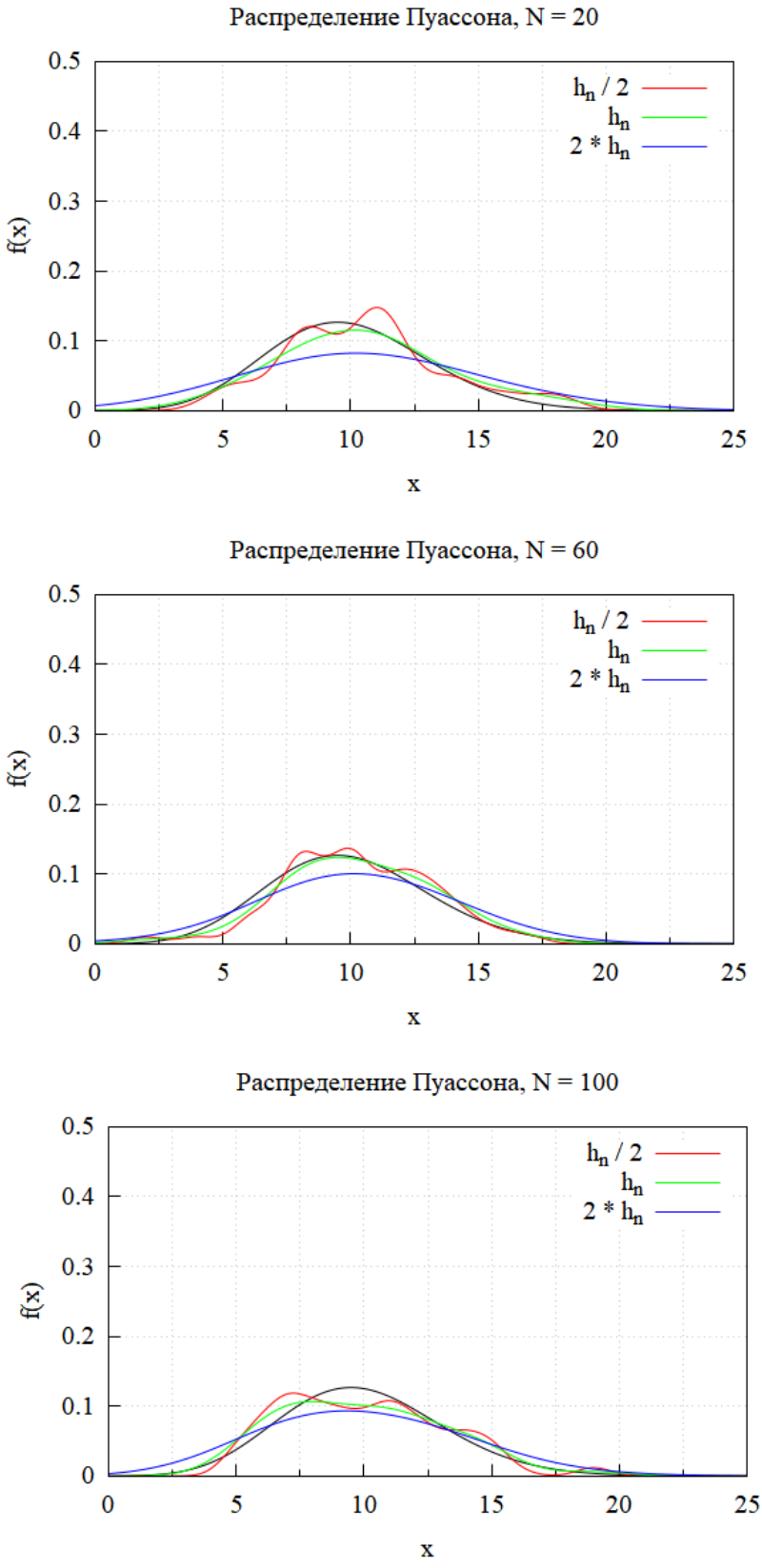


Рис. 14: Ядерная оценка плотности распределения Пуассона $P(k, 10)$

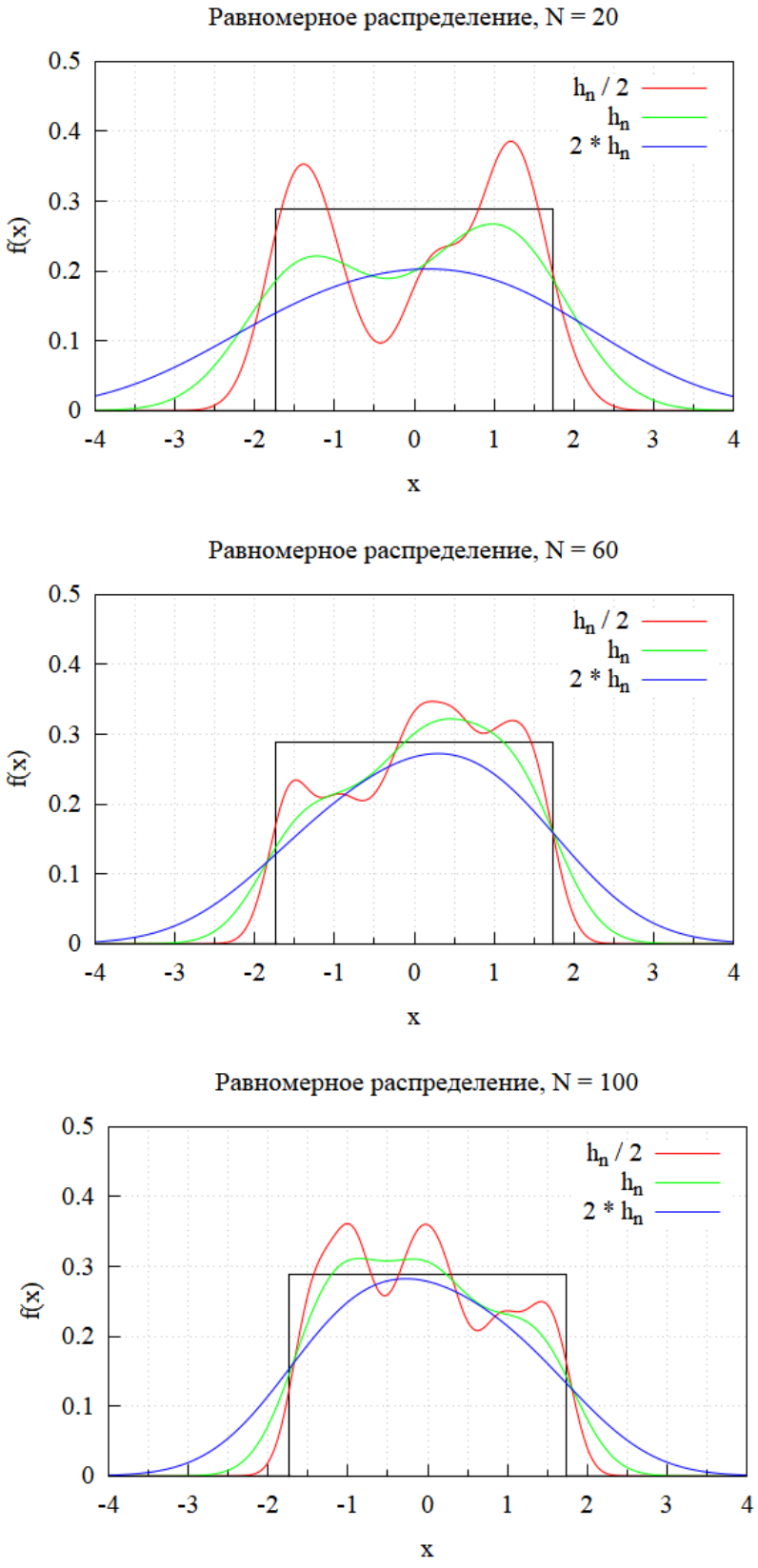


Рис. 15: Ядерная оценка плотности равномерного распределения $U(x, -\sqrt{3}, \sqrt{3})$

5 Заключение

5.1 Экспериментальная доля и теоретическая вероятность выбросов

В ходе выполнения лабораторной работы были построены боксплоты Тьюки для выборок разного размера, соответствующих заданным распределениям. По построенным боксплотам удалось визуально оценить мощность выбросов соответствующих распределений и сделать вывод о том, что наиболее подвержены выбросам выборки, построенные по распределению Коши.
Также были рассчитаны теоретические вероятности выбросов для каждого распределения. После этого на основе генерации 1000 выборок размерами в 20 и 100 элементов для каждого распределения были вычислены экспериментальные доли выбросов. На основании полученных результатов можно сделать вывод о том, что экспериментальные доли выбросов достаточно близки к теоретическим вероятностям и тем сильнее к ним приближаются, чем больше размер выборки.

5.2 Эмпирическая функция и ядерные оценки плотности распределения

В ходе выполнения лабораторной работы были построены эмпирические функции распределения для выборок разного размера, соответствующих пяти заданыым распределениям. Из построенных графиков можно сделать вывод, что эмпирическая функция распределения тем ближе к теоретической, чем больше размер выборки.
Также были построены ядерные оценки плотности вероятностей распределений. Из графиков можно сделать вывод, что чем шире полоса пропускания (больше параметр сглаживания), тем более более график сглажен и менее чувствителен к выбросам. Для нормального распределения и распределения Пуассона больше всего подходит выбор параметра сглаживания h_n, для равномерного распределения — $2h_n$, для распределения Лапласа — $h_n/2$. Для распределения Коши не подходит ни один из параметров сглаживания, при любом его выборе ядерная оценка достаточно далека от теоретической плотности вероятности.

Список литературы

[1] Box plot `https://en.wikipedia.org/wiki/Box_plot` Дата обращения 7.12.2020

[2] Теоретическое приложение к лабораторным работам №1-4 по дисциплине «Математическая статистика». – СПб.: СПбПУ, 2020. – 12 с

Приложение А. Репозиторий с исходным кодом

Ссылка на репозиторий GitHub с исходным кодом: `https://github.com/pchn/TeorVer`