**Санкт-Петербургский Политехнический университет Петра Великого**
**Институт прикладной математики и механики**
**Кафердра «Прикладная математика и информатика»**

Отчёт по проекту
Дисциплина: «Математическая статистика»
Лабораторные работы № 1-4

Выполнил студент гр. 3630102/70201 Н. А. Счастливцев
Преподаватель А. Н. Баженов

СПб 2020

# Содержание

# Список иллюстраций

# Список таблиц

# 1 Постановка задачи

Даны 5 распределений с параметрами:

1. Нормальное: $N(x, 0, 1)$
2. Коши: $C(x, 0, 1)$
3. Лапласа: $L(x, 0, \frac{1}{\sqrt{2}})$
4. Пуассона: $P(x, 10)$
5. Равномерное: $U(x, -\sqrt{3}, \sqrt{3})$

Требуется:

1. Сгенерировать массивы данных размером 10, 50, 1000 элементов. Для этих массивов построить графики плотности для каждого набора (плотность вероятности и гистограмму), всего 15 графиков.
2. Сгенерировать выборки размером 10, 100 и 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных: $\overline{x}, \operatorname{med} x, z_R, z_Q, z_{tr}$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

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

   Вычислить оценку дисперсии по формуле:

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

   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 (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 Плотности распределений

Требуемые плотности распределений [6] :

1. Нормальное (стандартное):

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

2. Коши:

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

3. Лапласа:

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

4. Пуассона:

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

5. Равномерное:

$$U(x,-\sqrt{3},\sqrt{3}) = \begin{cases} 0, & x \notin [-\sqrt{3},\sqrt{3}] \\ 1/(2\sqrt{3}) & x \in [-\sqrt{3},\sqrt{3}] \end{cases} \tag{7}$$

## 2.2 Гистограммы

*Определение.* Гистограммой выборки называется фигура, образованная прямоугольниками с основаниями $\Delta_i$ и высотами $n_i/(nh)$ , где $(i = 1,\ldots,k)$ [5].

*Замечание.* Тут $\Delta_i$ — длина $i$-го промежутка, всего промежутков $k$, эту величину можно выбрать полуэмпирически по формуле:

$$k = 1.72n^{1/3} \tag{8}$$

Все длины промежутков определяются по формуле:

$$h = \frac{x_{max} - x_{min}}{k} \tag{9}$$

Величины $n_i/n$ называются *относительными*, а $n_i/(nh)$ – *приведёнными частотами* группированного статистического ряда.

## 2.3 Выборочные числовые характеристики

Числовые характеристики случайной величины $X^*$ называются выборочными числовыми характеристиками, где $X^*$ — дискретная случайная величина, соответствующая наблюдениям [5].

### 2.3.1 Вариационный ряд

*Определение.* Вариационным рядом называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются. Запись вариационного ряда: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ [5].

### 2.3.2 Квартиль

Будем пользоваться понятием квартилей, частным случаем понятия квантиль.

*Определение.* Элементы вариационного ряда, на четверть отстоящие от краёв, называются соответственно нижней и верхней квартилями и обозначаются $z_{1/4}$ и $z_{3/4}$ [5].

### 2.3.3 Характеристики положения

1. Выборочное среднее:
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

2. Выборочная медиана:
$$\operatorname{med} x = \begin{cases} x_{(k+1)}, & n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, & n = 2k \end{cases} \tag{11}$$

3. Полусумма экстремальных выборочных элементов:
$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{12}$$

4. Полусумма выборочных квартилей:
$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{13}$$

5. Усечённое среднее:
$$z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_{(i)}, r = \frac{n}{4} \tag{14}$$

## 2.4 Боксплот Тьюки

### 2.4.1 Определение

Боксплот (или «ящик с усами», англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.[3]

### 2.4.2 Построение боксплота Тьюки

Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину усов определяет формула:

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

где $X_1$ — нижняя граница уса, $X_2$ — верхняя граница уса, $Q_1$ — первый квартиль, $Q_3$ — третий квартиль, $k$ — коэффициент, чаще всего равен 1.5.

Данные, выходящие за границы усов (*выбросы*), отображаются на графике в виде точек, маленьких кружков или звёздочек.[3]

### 2.4.3 Теоретическая вероятность выбросов

По формуле (15) можно вычислить границы «усов» боксплота $X_1$ и $X_2$. Выбросами считаются такие $x$ из выборки, что:

$$\begin{bmatrix} x < X_1 \\ x > X_2 \end{bmatrix} \tag{16}$$

Теоретическая вероятность выбросов:

- Для непрерывных распределений:

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

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

- Для дискретных распределений:

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

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

## 2.5 Эмпирическая функция распределения и ядерные оценки плотности

### 2.5.1 Статистический ряд

*Определение.* Статистическим рядом называется последовательность различных элементов $z_i$ вариационного ряда с указанием частот $n_i$ повторения элементов.

В общем случае статистический ряд можно записать в виде таблицы 1 ($n_1 + n_2 + \cdots + n_k = n$). [5]

| $z_i$ | $z_1$ | $z_2$ | $\cdots$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | $\cdots$ | $\cdots$ | $n_k$ |

Таблица 1: Статистический ряд

### 2.5.2 Эмпирическая функция распределения

*Определение.* Выборочной (эмпирической) функцией распределения называется относительная частота события $X < x$, полученная по выборке [5]:

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

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

$$P^*(X < x) = \frac{1}{n} \sum_{z_i < x} n_i \tag{20}$$

Получаем

$$F_n^*(x) = \frac{1}{n} \sum_{z_i < x} n_i \tag{21}$$

Это функция распределения дискретной случайной величины $X^*$, заданной таблицей распределения 2

| $X^*$ | $z_1$ | $z_2$ | $\ldots$ | $\ldots$ | $z_k$ |
|---|---|---|---|---|---|
| $P$ | $n_1/n$ | $n_2/n$ | $\ldots$ | $\ldots$ | $n_k/n$ |

Таблица 2: Таблица эмпирической функции распределения

Её графиком является восходящая ступенчатая линия, называемая кумулятой (линия накопленных относительных частот).

Так как относительная частота события приближается к вероятности события при увеличении $n$, то выборочная функция распределения $F_n^*(x)$ приближённо представляет функцию распределения $F(x)$ генеральной совокупности, как говорят, является её оценкой:

$$F_n^*(x) \approx F(x) \tag{22}$$

### 2.5.3 Ядерные оценки плотности вероятности

*Определение.* Оценкой плотности вероятности $f(x)$ называется функция $\overline{f}(x)$, построенная на основе выборки, приближённо равная $f(x)$ [4]

$$\overline{f}(x) \approx f(x) \tag{23}$$

Ранее была рассмотрена такая оценка $\overline{f}(x)$, заданная ступенчатой линией, ограничивающей гистограмму сверху. Для этой функции можно записать аналитическое выражение

$$\overline{f}(x) = \frac{1}{nh} \sum_{i=1}^{k} n_i \mathbb{K}\left(\frac{x - z_i}{h}\right) \tag{24}$$

Здесь функция $\mathbb{K}(u)$, называемая ядерной (ядром), определяется формулой

$$\mathbb{K}(u) = \begin{cases} 1; & -1/2 < u \leq 1/2, \\ 0; & u \leq -1/2 \text{ или } u > 1/2, \end{cases} \tag{25}$$

$n$ — объём выборки; $n_i$ — число элементов выборки, попавших в промежуток $\Delta_i$; $h$ — длина $\Delta_i$; $i = 1, ..., k$; $k$ — число промежутков.[4]

Выбирая другие ядра из числа непрерывных функций, являющихся плотностями вероятности, можно получить непрерывные ядерные оценки подобного же вида с числом слагаемых в сумме, равным объёму выборки:

$$\overline{f_n}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{K}\left(\frac{x - x_i}{h_n}\right) \tag{26}$$

Здесь $x_1, ..., x_n$ — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, обладающих свойствами

$$h_n \xrightarrow[n\to\infty]{} 0 : \frac{h_n}{n^{-1}} \xrightarrow[n\to\infty]{} \infty \tag{27}$$

Ядро $\mathbb{K}(u)$ для ядерной оценки в общем случае — кусочно-непрерывная плотность, имеющая свойства [4]

$$\sup_{-\infty<u<\infty} \mathbb{K}(u) < \infty; \lim_{u\to\infty} \mathbb{K}(u) = 0 \tag{28}$$

# 3 Реализация

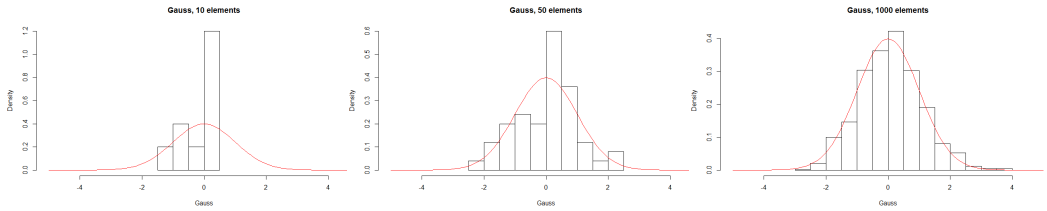
Результаты лабораторной получены с использованием средств языка программирования R и среды R-studio. Ссылка на репозиторий с кодом. Использованные библиотеки: stringr, LaplacesDemon.

# 4 Результаты

## 4.1 Гистограммы распределений

График плотности распределения на рисунках отмечен красной линией.

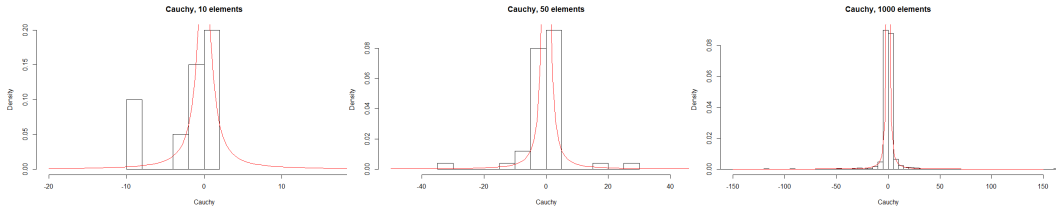
### 4.1.1 Нормальное распределение




(a) 10 элементов (b) 50 элементов (c) 1000 элементов

Рис. 1: Гистограммы плотности нормального распределения (3)

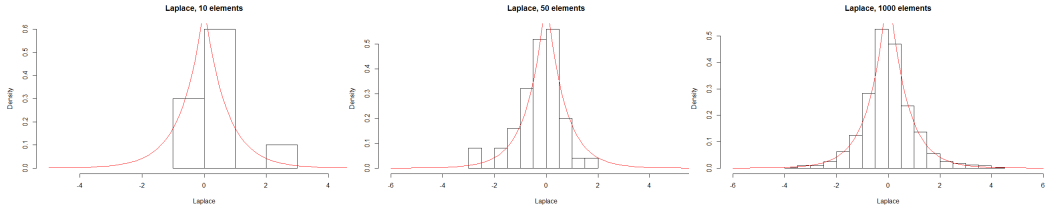
### 4.1.2 Распределение Коши




(a) 10 элементов (b) 50 элементов (c) 1000 элементов

Рис. 2: Гистограммы плотности распределения Коши (4)

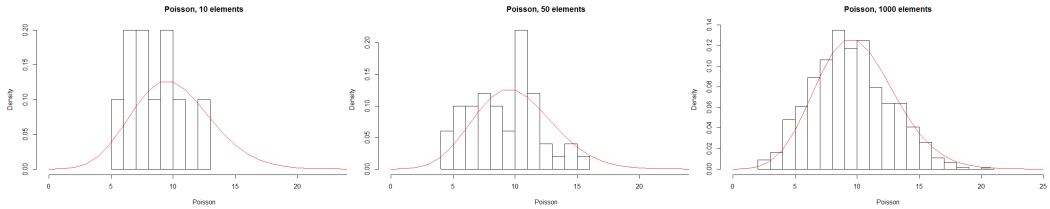
### 4.1.3 Распределение Лапласа




(a) 10 элементов (b) 50 элементов (c) 1000 элементов

Рис. 3: Гистограммы плотности распределения Лапласа (5)

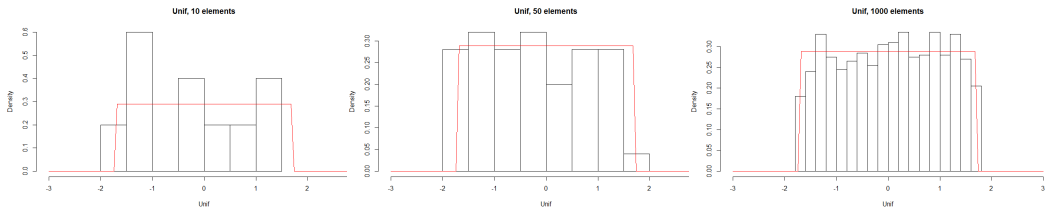
### 4.1.4 Распределение Пуассона




(a) 10 элементов (b) 50 элементов (c) 1000 элементов

Рис. 4: Гистограммы плотности распределения Пуассона (6)

### 4.1.5 Равномерное распределение




(a) 10 элементов (b) 50 элементов (c) 1000 элементов

Рис. 5: Гистограммы плотности равномерного распределения (7)

## 4.2 Характеристики положения и рассеяния

В таблицах приведены результаты эксперимента:

- Нормальное распределение:
  1. Таблица 3
  2. Таблица 4
  3. Таблица 5
- Распределение Коши:
  1. Таблица 6
  2. Таблица 7
  3. Таблица 8
- Распределение Лапласа:
  1. Таблица 9
  2. Таблица 10
  3. Таблица 11
- Распределение Пуассона:
  1. Таблица 12
  2. Таблица 13
  3. Таблица 14
- Равномерное распределение:

1. Таблица 15
2. Таблица 16
3. Таблица 17

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.02 | -0.01 | -0.02 | -0.02 | -0.01 |
| $\mathbb{D}(z)$ | 0.10 | 0.144 | 0.180 | 0.11 | 0.11 |

Таблица 3: Нормальное распределение (3) n = 10

| n = 100 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 0.0002 | -0.002 | -0.0003 | 0.001 | -0.0001 |
| $\mathbb{D}(z)$ | 0.0103 | 0.014 | 0.1 | 0.013 | 0.012 |

Таблица 4: Нормальное распределение (3) n = 100

| n = 1000 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.001 | -0.002 | -0.003 | -0.002 | -0.002 |
| $\mathbb{D}(z)$ | 0.001 | 0.0015 | 0.063 | 0.001 | 0.001 |

Таблица 5: Нормальное распределение (3) n = 1000

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -2 | -0.04 | -9 | 1 | -0.1 |
| $\mathbb{D}(z)$ | 1400 | 0.32 | 35000 | 1.31 | 0.50 |

Таблица 6: Распределение Коши (4) n = 10

| n = 100 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | $\operatorname{med} x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.2 | -0.01 | -9 | 1 | -0.004 |
| $\mathbb{D}(z)$ | 2100 | 0.024 | $5.1 \cdot 10^6$ | 0.1 | 0.027 |

Таблица 7: Распределение Коши (4) n = 100

| n = 1000 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | $\operatorname{med} x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 14 | 0.001 | $7 \cdot 10^3$ | 1 | 0.001 |
| $\mathbb{D}(z)$ | $1.2 \cdot 10^4$ | 0.0025 | $3 \cdot 10^{10}$ | 0.008 | 0.0026 |

Таблица 8: Распределение Коши (4) n = 1000

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | $\operatorname{med} x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.02 | -0.01 | -0.04 | 0.5 | -0.01 |
| $\mathbb{D}(z)$ | 0.1 | 0.07 | 0.45 | 0.12 | 0.07 |

Таблица 9: Распределение Лапласа (5) n = 10

| n = 100 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | $\operatorname{med} x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 0.002 | -0.0002 | -0.01 | 0.50 | 0.001 |
| $\mathbb{D}(z)$ | 0.01 | 0.006 | 0.41 | 0.016 | 0.006 |

Таблица 10: Распределение Лапласа (5) n = 100

| n = 1000 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | $\operatorname{med} x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.001 | $6 \cdot 10^{-5}$ | 0.01 | 0.50 | 0.0001 |
| $\mathbb{D}(z)$ | 0.001 | $5.3 \cdot 10^{-4}$ | 0.43 | 0.002 | 0.0006 |

Таблица 11: Распределение Лапласа (5) n = 1000

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 10 | 10 | 10 | 11 | 10 |
| $\mathbb{D}(z)$ | 1.03 | 1.45 | 1.9 | 1.9 | 1.2 |

Таблица 12: Распределение Пуассона (6) n = 10

| n = 100 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 10 | 10 | 11 | 12 | 9.85 |
| $\mathbb{D}(z)$ | 0.1 | 0.2 | 0.96 | 0.3 | 0.12 |

Таблица 13: Распределение Пуассона (6) n = 100

| n = 1000 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 10 | 10 | 11 | 12 | 9.86 |
| $\mathbb{D}(z)$ | 0.01 | 0.003 | 0.7 | 0 | 0.01 |

Таблица 14: Распределение Пуассона (6) n = 1000

| n = 10 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.02 | -0.03 | -0.01 | 0.7 | -0.02 |
| $\mathbb{D}(z)$ | 0.101 | 0.24 | 0.042 | 0.2 | 0.18 |

Таблица 15: Равномерное распределение (7) n = 10

| n = 100 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | -0.001 | $-9 \cdot 10^{-5}$ | 0.0002 | 0.90 | -0.002 |
| $\mathbb{D}(z)$ | 0.010 | $3 \cdot 10^{-2}$ | 0.0006 | 0.02 | 0.020 |

Таблица 16: Равномерное распределение (7) n = 100

| n = 1000 | | | | | |
|---|---|---|---|---|---|
| | $\overline{x}$ | med $x$ | $z_R$ | $z_Q$ | $z_{tr}$ |
| $\mathbb{E}(z)$ | 0.002 | 0.0002 | -9.0 $\cdot 10^{-5}$ | 0.86 | 0.0004 |
| $\mathbb{D}(z)$ | 0.001 | 0.003 | 6 $\cdot 10^{-6}$ | 0.002 | 0.002 |

Таблица 17: Равномерное распределение (7) n = 1000

## 4.3 Боксплоты и выбросы

### 4.3.1 Боксплоты выборок

1. Нормальное распределение (3), боксплот 6:

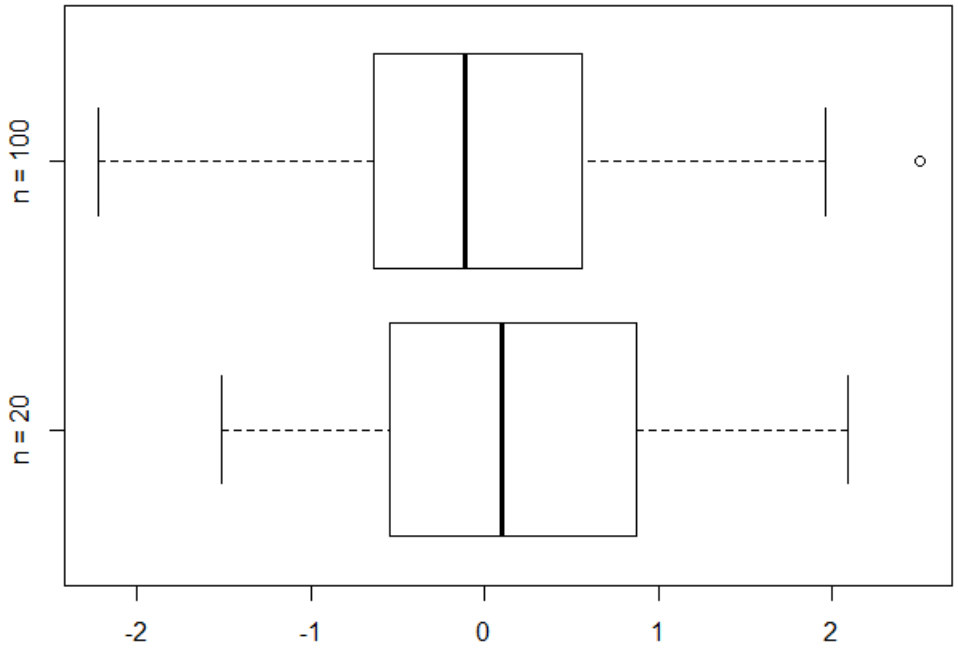
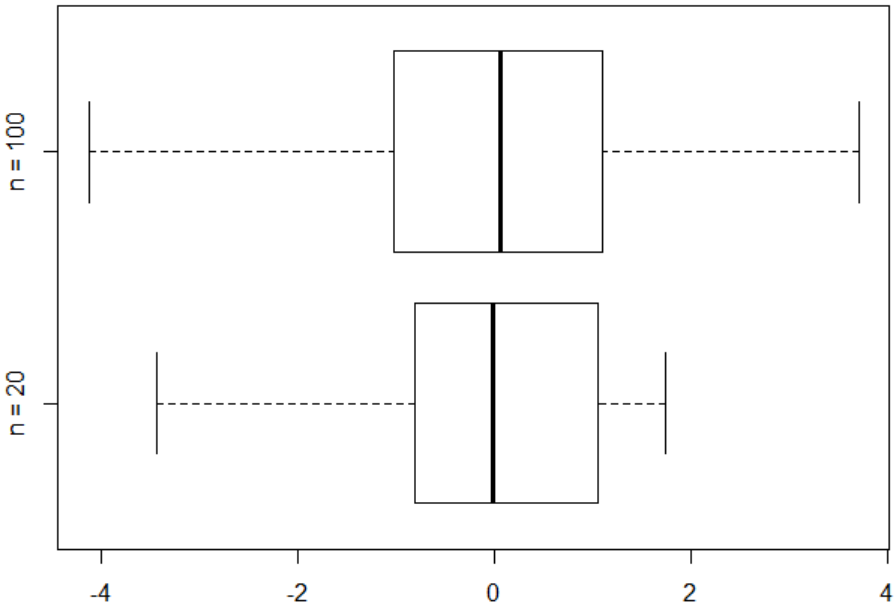
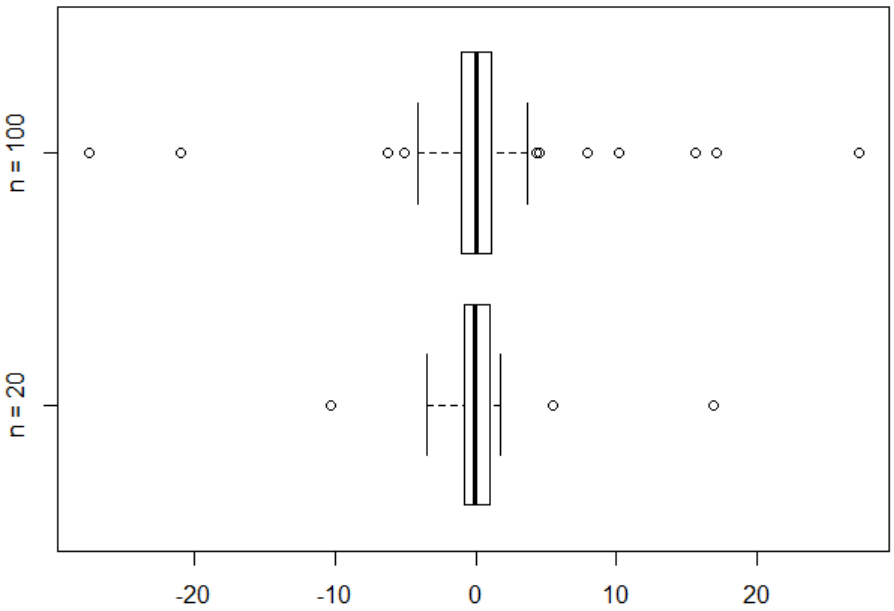



Рис. 6: Боксплоты, нормальное распределение

2. Распределение Коши (4), боксплот 7:




(a) Без выбросов






(b) С выбросами

Рис. 7: Боксплоты выборок для распределения Коши

3. Распределение Лапласа (5), боксплот 8:

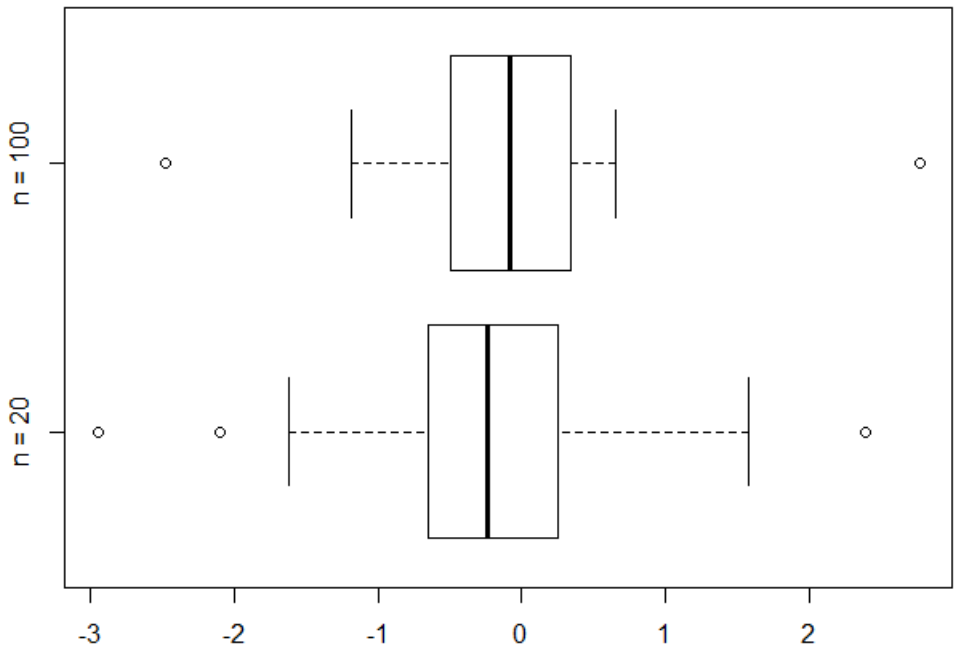



Рис. 8: Боксплоты, распределение Лапласа

4. Распределение Пуассона (6), боксплот 9:

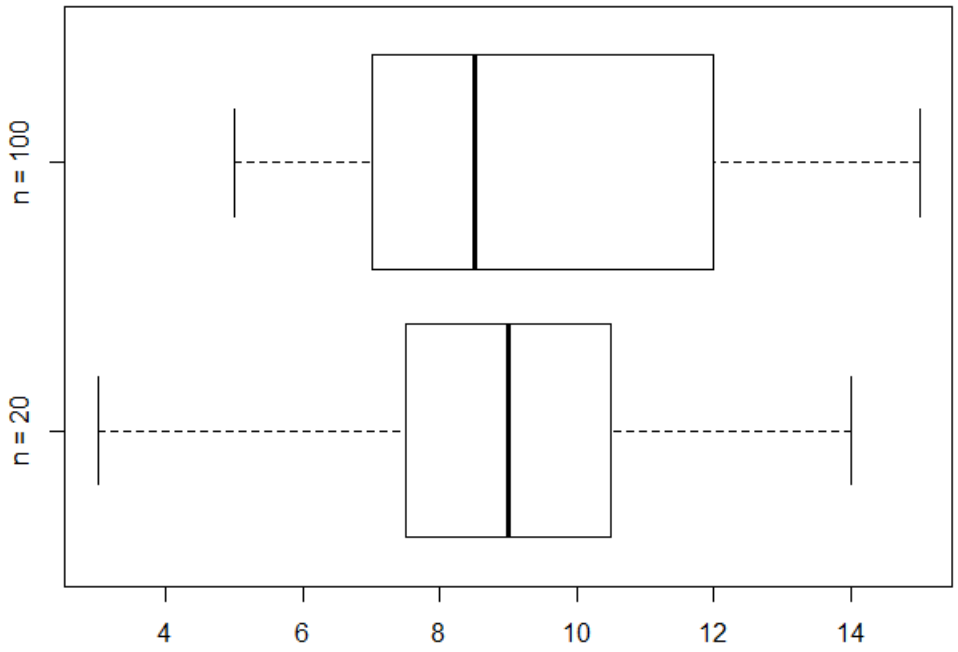



Рис. 9: Боксплоты, распределение Пуассона

5. Равномерное распределение (7), боксплот 10:

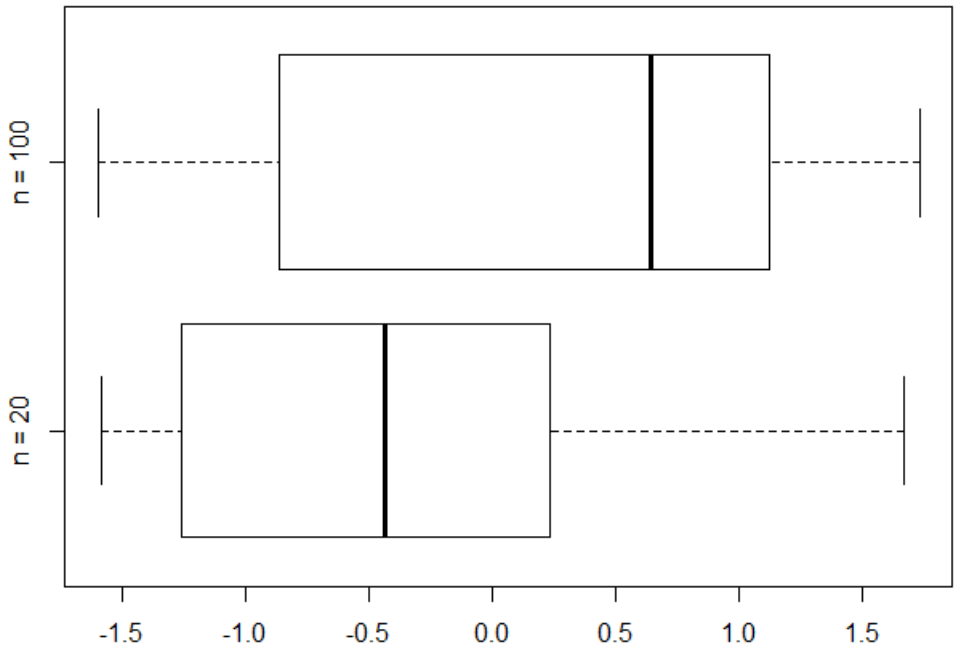



Рис. 10: Боксплоты, равномерное распределение

### 4.3.2 Выбросы

По формулам (17) и (18) с помощью средств языка R вычислим теоретические вероятности выбросов.

Теоретические и экспериментальные данные приведены в таблице 18.

| | Эксперимент | | Теория |
|---|---|---|---|
| Распределение | n = 20 | n = 100 | |
| Нормальное | 0.1 | 0.10 | 0.092 |
| Коши | 0.21 | 0.23 | 0.226 |
| Лапласа | 0.16 | 0.158 | 0.156 |
| Пуассона | 0.1 | 0.090 | 0.084 |
| Равномерное | 0.05 | 0.02 | 0 |

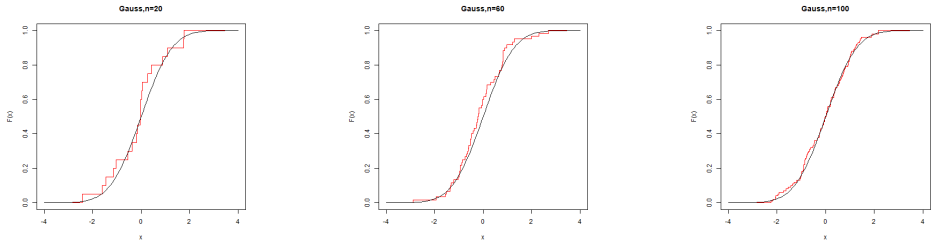
Таблица 18: Вероятность выбросов

## 4.4 Приближения распределений

### 4.4.1 Эмпирические функции распределения

График приближения функции распределения отмечен красной линией.

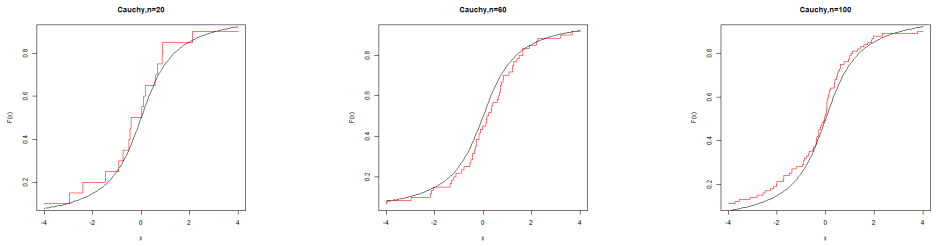
- Нормальное распределение (рис. 11)




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 11: Эмпирические функции нормального распределения

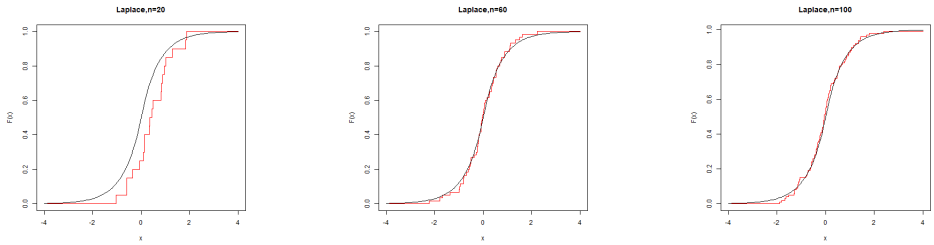
- Распределение Коши (рис. 12)




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 12: Эмпирические функции распределения Коши

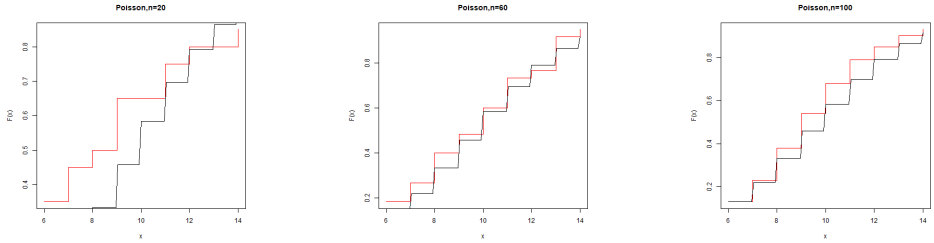
- Распределение Лапласа (рис. 13)




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 13: Эмпирические функции распределения Лапалса

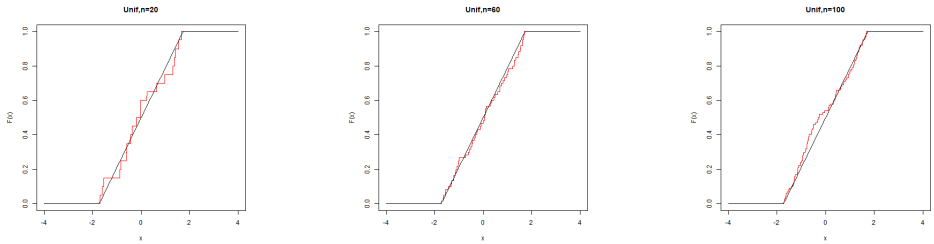
- Распределение Пуассона (рис. 14)




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 14: Эмпирические функции распределения Пуассона

- Равномерное распределение (рис. 15)




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 15: Эмпирические функции равномерного распределения

### 4.4.2 Ядерные оценки плотностей распределений

График приближения функции распределения отмечен красной линией. Для каждого распределения и выборки построены графики в трёх ва-

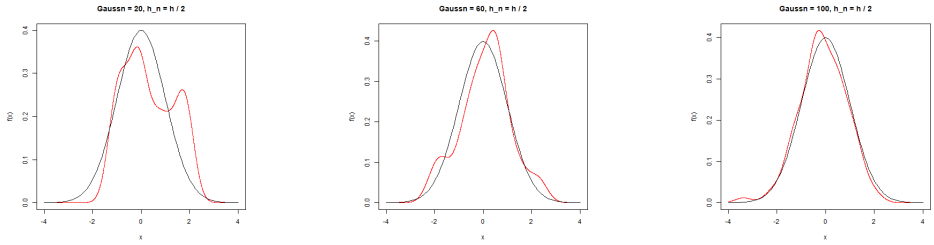
риациях параметра $h_n$ по отношению к параметру $h$, вычисленному по формулам (8) и (9).

Зададим непрерывную функцию ядра:

$$\mathbb{K}(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{29}$$

1. Нормальное распределение:

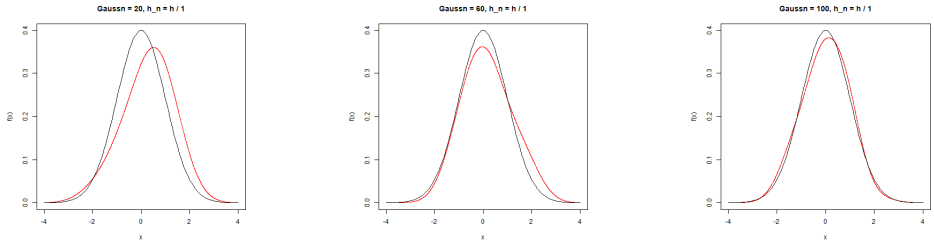
   - $h_n = \dfrac{h}{2}$ : рис. 16




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 16: Ядерная оценка плотности нормального распределения, $h_n = h/2$

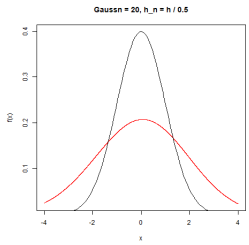
   - $h_n = h$ : рис. 17



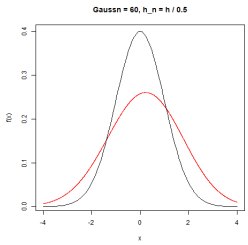

(a) n = 20 (b) n = 60 (c) n = 100

Рис. 17: Ядерная оценка плотности нормального распределения, $h_n = h$

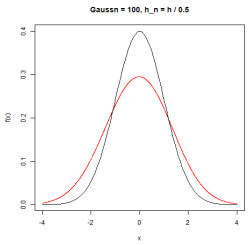
   - $h_n = 2h$ : рис. 18




(a) n = 20




(b) n = 60

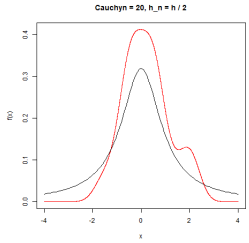



(c) n = 100

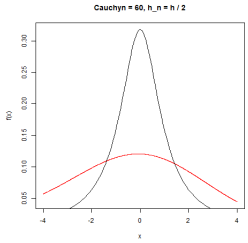
Рис. 18: Ядерная оценка плотности нормального распределения, $h_n = 2h$

2. Распределение Коши:

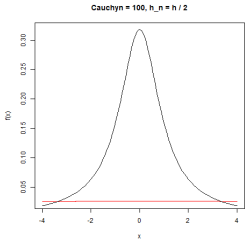
- $h_n = \dfrac{h}{2}$ : рис. 19




(a) n = 20



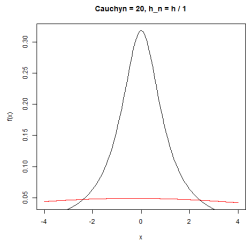

(b) n = 60



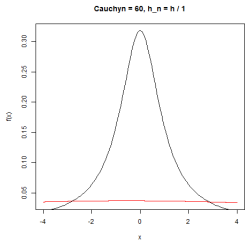

(c) n = 100

Рис. 19: Ядерная оценка плотности распределения Коши, $h_n = h/2$

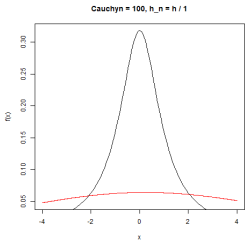
- $h_n = h$ : рис. 20




(a) n = 20



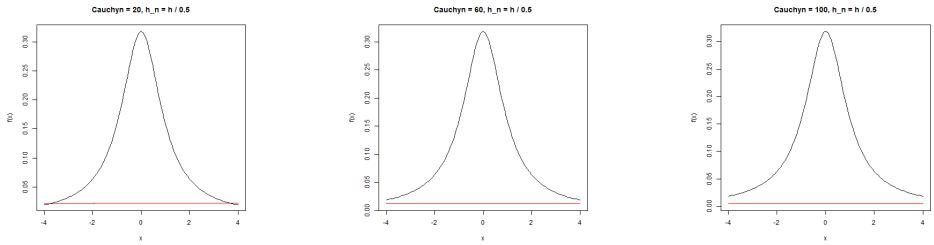

(b) n = 60




(c) n = 100

Рис. 20: Ядерная оценка плотности распределения Коши, $h_n = h$

- $h_n = 2h$ : рис. 21

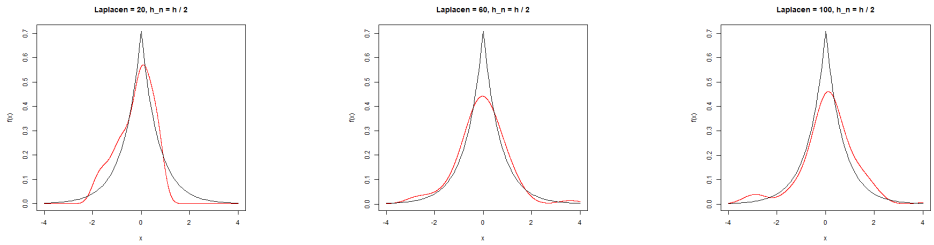



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 21: Ядерная оценка плотности распределения Коши, $h_n = 2h$

3. Распределение Лапласа:

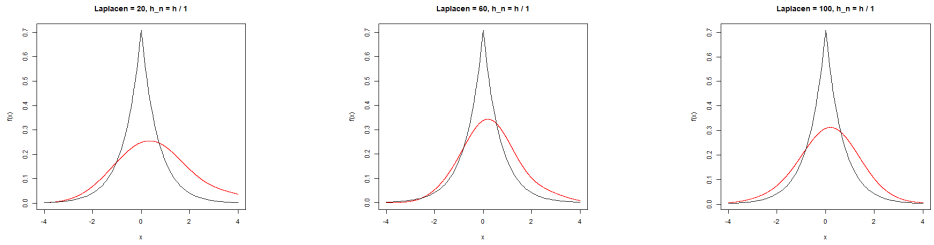
- $h_n = \dfrac{h}{2}$ : рис. 22




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 22: Ядерная оценка плотности распределения Лапласа, $h_n = h/2$

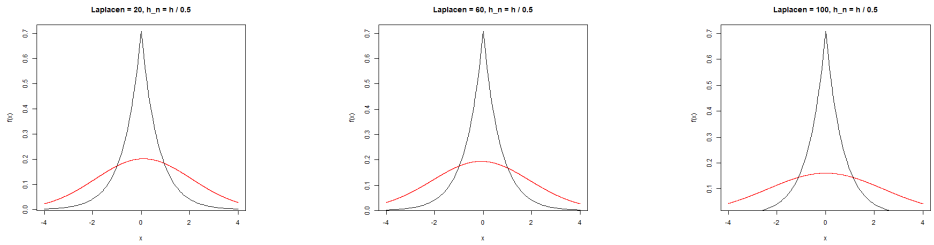
- $h_n = h$ : рис. 23




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 23: Ядерная оценка плотности распределения Лапласа, $h_n = h$

- $h_n = 2h$ : рис. 24

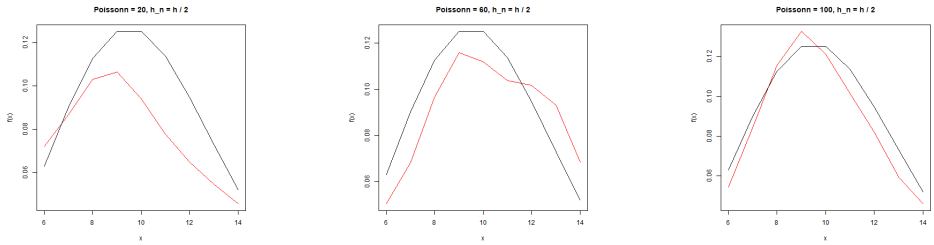



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 24: Ядерная оценка плотности распределения Лапласа, $h_n = 2h$

4. Распределение Пуассона:

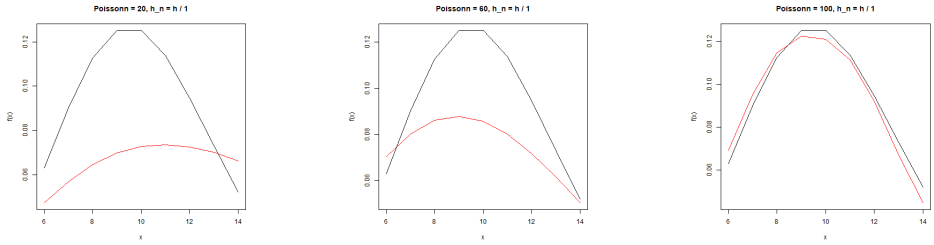
- $h_n = \dfrac{h}{2}$ : рис. 25




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 25: Ядерная оценка плотности распределения Пуассона, $h_n = h/2$

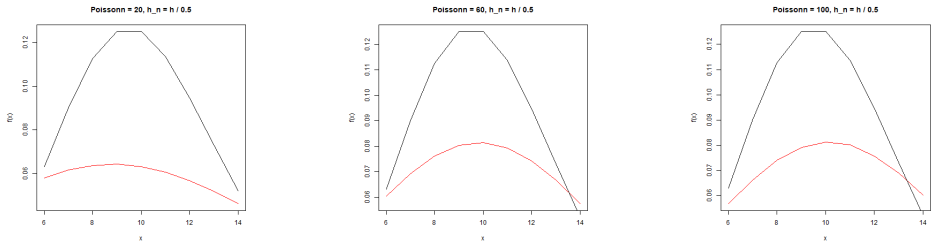
- $h_n = h$ : рис. 26




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 26: Ядерная оценка плотности распределения Пуассона, $h_n = h$

- $h_n = 2h$ : рис. 27

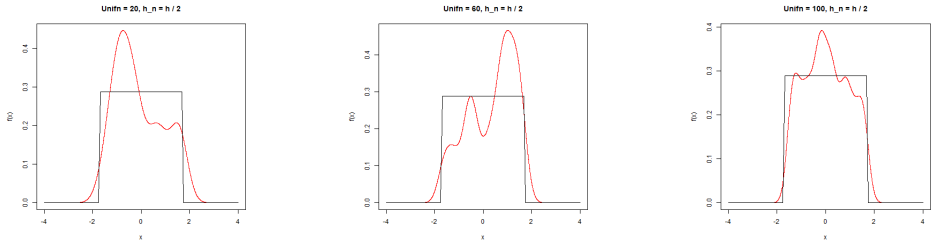



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 27: Ядерная оценка плотности распределения Пуассона, $h_n = 2h$

5. Равномерное распределение:

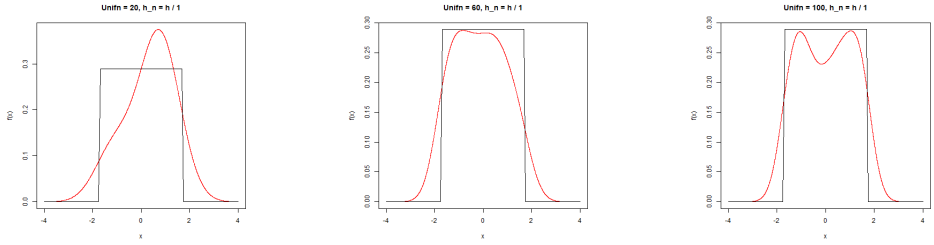
- $h_n = \dfrac{h}{2}$ : рис. 28




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 28: Ядерная оценка плотности равномерного распределния, $h_n = h/2$

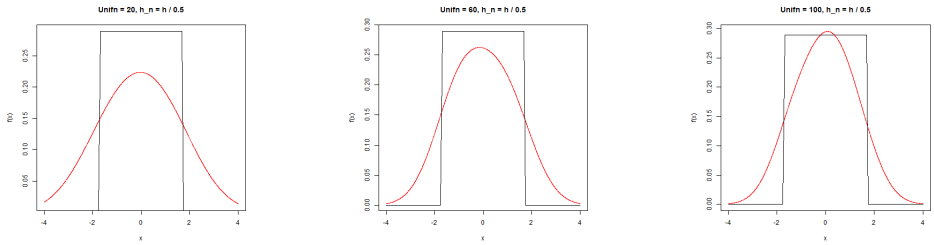
- $h_n = h$ : рис. 29




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 29: Ядерная оценка плотности равномерного распределния, $h_n = h$

- $h_n = 2h$ : рис. 30




(a) n = 20 (b) n = 60 (c) n = 100

Рис. 30: Ядерная оценка плотности равномерного распределния, $h_n = 2h$

# 5 Обсуждение

## 5.1 Анализ гистограмм

Из приведённых графиков и гистограмм видно, что для всех распределений для небольшой выборки (10 элементов) гистограмма очень плохо приближает плотность распределения, отклонение велико. Это свидетельствует о недостаточном объёме данных (мало элементов, статистику составлять практически бессмысленно). Для 50 элементов опыт уже почти соответствует плотности вероятности для каждого графика, отклонения есть, они заметны, но они носят «флуктуационный» характер, в «целом» зависимость неплохо приближается. Наконец, для 1000 элементов достигнута высокая точность, гистограммы составляют практически всю область под графиком аналитической плотности. Такие зависимости опытного приближения плотности вероятности от числа испытаний объясняются результатами теории вероятностей, а именно, законом больших чисел (хотя, для распределения Коши он не применим, так как у распределения Коши не существует мат. ожидания) — относительная частота стремится к вероятности при большом числе испытаний.

Стоит заметить, что для распределения Коши на гистограмме для 1000 элементов появляются выбросы (и для конкретной выборки в рассматриваемый промежуток $[-150, 150]$ ещё не попали числа 207 и -205). Такой результат можно обосновать характеристиками самого распределения Коши — оно не имеет математического ожидания. Можно взять интеграл в смысле главного значения, тогда в данном случае «математическое ожидание» будет нулём, но дисперсия бесконечна [2]. Рассмотрим неравенство Чебышёва в классической форме:

$$P\{|\xi - \mathbb{E}\xi| \geq \varepsilon\} \leq \frac{\mathbb{D}\xi}{\varepsilon^2} \tag{30}$$

Таким образом, для большого отклонения от нуля (для данного распределения) вероятность не столь мала, как для остальных распределений (т. к. дисперсия бесконечно велика).

## 5.2 Анализ характеристик положения и рассеяния

Для нормального распределения все характеристики имеют один порядок для фиксированного размера выборки. По мере увеличения размера выборки, все характеристики уточняются на 1 порядок.

Для распределения Коши характерны «выбросы», поэтому характеристики среднего положения разнятся, а характеристики рассеяния имеют разные порядки. В данном случае можно руководствоваться такими характеристиками, как выборочная медиана и усечённое среднее, так как они устойчивы к выбросам. При увеличении размеров выборки характеристики рассеяния неограниченно растут.

Для распределения Лапласа характеристики ведут себя так же, как и в нормальном распределении, за исключением выборочной медианы — она горздо ближе к нулю, чем остальные характеристики, и её рассеяние меньше. Это связано с «изломом» плотности вероятности в нуле из-за негладкости функции (5): медиана является очень устойчивой характеристикой (по аналогии: медиана распределения вероятности существует даже тогда, когда моментов нет).

Для распределения Пуассона характерно малое изменение средних значений при изменении размера выборки, это обусловлено самим характером распределения: оно дискретно. При этом, рассеяние уменьшается от порядка $10^0$ до $10^{-2}$ тоже ввиду дискретности распределнения Пуассона. Интересный результат получается для выборок размером 1000 элементов, где рассеяние полусуммы выборочных квартилей равно нулю. Это тоже связано с дискретностью распределения – от эксперимента к эксперименту квартили получаются почти одними и теми же (так как плотность распределения принимает только целочисленные значения).

Для равномерного распределения все характеристики стремятся к нулю, что характеризует ситуацию неверно. Это связано с тем, что у равномерного распределения нет одной выделенной моды, как у остальных рассмотренных распределений, попадение в каждую точку промежутка равновероятно, поэтому выборочные характеристики могут «скакать» от эксперимента к эксперименту. Так как в данном эксперименте посчитано «среднее от средних», то эти скачки остались незамеченными, они в совокупности друг с другом дали близкие к нулю итоговые средние характеристики (так как распределение симметрично относительно нуля).

Из проведённого эксперимента следует, что использованный метод

оценки характеристик положения выборки и рассеяния годится, когда у распределения есть одна мода; наиболее предпочтительной характеристикой является выборочная медиана, так как она наименее чувствительна к выбросам.

## 5.3 Боксплоты и выбросы

### 5.3.1 Анализ боксплотов

Боксплоты являются полезным инструментом, так как очень наглядно представляют информацию о параметрах выборки. При этом, выборки можно сравнивать, сравнивая построенные по ним боксплоты, располагая их друг над другом.

По виду боксплотов можно судить о том, насколько распределение наблюдений из выборки «близко» к нормальному. Из теоретических соображений нормальное распределение симметрично, его квартили симметричны, выбросы встречаются редко (пользуемся правилом «трёх сигм») — боксплот выборки, которая порождается нормальным распределением выглядит почти симметрично, медиана отмечена почти по середине «ящика», выбросы редки (см. боксплот 6). К примеру, боксплоты для выборки, порождённой распределением Коши, таким свойством не обладают: у них «много» выбросов, «усы» часто несимметричны, однако медиана находится близко к центру «ящика» (см. боксплот 7). Для выборок из равномерного распределения характерны одни и те же границы «усов», так как выбросов минимум. Если строить боксплоты по разным выборкам, размеры «ящиков» будут скакать, отметка медианы будет сильно меняться (см. п. 5.2).

Способ построения боксплотов даёт ответ на вопрос о том, что считать выбросом.

### 5.3.2 Анализ выбросов

Из рассмотрения таблицы 18 можно произвести следующие выводы. Для нормального, равномерного, пуассоновского распределений и распределений Лапласа и Коши наблюдается уточнение экспериментальной относительной частоты выброса с возрастанием размера выборки, уточнение в сторону теоретически вычисленной вероятности выбросов. В случае равномерного распределения относительная частота даже в случае $n = 100$ далека от вероятности. Это связано с тем, что распределение равномерное, квантильные оценки положения и рассеяния плохо его характеризуют (см. п. 5.2), однако именно эти характеристики используются

при построении боксплотов Тьюки и определении выбросов. В распределении Пуассона в таблице виден серьёзный скачок от n = 20 к n = 100, это связано с дискретным характером распределения Пуассона.

В распределении Коши наибольшая теоретическа вероятность выбросов. На практике получили достаточно близкий результат к теоретической вероятности, что позволяет судить о применимости метода отбраковывания выбросов в случае распределения Коши.

## 5.4 Приближения распределений

### 5.4.1 Эмпирические функции распределния

Из опыта видно, при увеличении мощности выборки приближения всё точнее описывают реальную функцию распределения. Лучшие результаты получены на нормальном, равномерном распределениях и распределении Лапласа.

Распределение Коши подвержено выбросам, поэтому для него приближения функции распределений не достигают ни нуля, ни единицы на рассматриваемом промежутке $[-4; 4]$.

Распределение Пуассона дискретно, поэтому отклонения относительных частот от вероятностей слишком сильно сказываются друг на друге и, как итог, на приближении функции распределения — не удаётся повторить график с той же точностью, как в случае непрерывных распределений.

Из полученных результатов следует, что эмпирическая функция распределения наиболее точно приближает распределение в следующих случаях:

- размер выборки достаточно большой
- выборка порождена абсолютно непрерывным распределением или дискретным распределением (но число возможных результатов наблюдений для дискретного распределения достаточно велико)
- доля выбросов в выборке низкая

### 5.4.2 Ядерные оценки плотности

Из рассмотренных параметров сразу видно, что в большинстве случаев оценка наиболее точная в случае $n = 60, h_n = h/2$. Такой результат (лучшие оценки при не наибольшем $n$) объясняется взаимосвязью между $n$ и $h_{\text{opt}}$: чем больше $n$, тем меньше следует выбрать $h$, это следует из

условия (27). К примеру, можно руководствоваться правилом "Rule of thumb"[1]:

$$h(n) \approx 0.9 \min\{s, \frac{\text{IQR}}{1.34}\} n^{-\frac{1}{5}} \tag{31}$$

Тем не менее, выбор параметра $h_n$ в каждом случае индивидуален, например, для равномерного распределения в эксперименте оказалось точной ядерная оценка при $n = 100$, $h = h_n$ (рис. 29).

В случае оценок плотности распределений Коши получилось слишком мало экспериментов, где ядерная оценка «пытается» приблизить истинную плотность вероятности. Для выборок, порождённых распределением Пуассона результат немного лучше, но всё равно при выбранных параметрах ядерная оценка неточна в большинстве экспериментов.

Распределение Пуассона дискретное, поэтому оценку его плотности надо строить по-другому, однако на практике зачастую неизвестно, действительно ли распределение является дискретным (из эксперимента в этом случае видно, что можно попробовать сперва построить ядерную оценку плотности, а потом судить о дискретности распределения по «изломанности» графика).

Из приведённых рассуждений следует, что выбор параметра $h_n$ очень влияет на точность ядерной оценки, при этом он индивидуален для каждой конкретной выборки (важен размер выборки и порождающее её распределение). Для выбора $h_n$ в некоторых конкретных случаях можно решить некоторую задачу оптимизации или взять формулы наподобие (31), которые иногда дают оптимальное значение $h_n$.

# 6 Литература

# Список литературы


[1] Wikipedia contributors. *Kernel density estimation — Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Kernel_density_estimation&oldid=946548066`. [Online; accessed 4/04/z20]. 2020.

[2] Википедия. *Распределение Коши — Википедия, свободная энциклопедия*. [Online; accessed 27/02/20]. 2019. URL: `https://ru.wikipedia.org/?oldid=99601782`.

[3] Википедия. *Ящик с усами — Википедия, свободная энциклопедия*. [Online; accessed 19/03/20]. 2020. URL: `https://ru.wikipedia.org/?oldid=104502300`.

[4] Ю. Д. Максимов. “Вероятностные разделы математики”. в: *СПб.: Иван Федоров* (2001).

[5] Ю. Д. Максимов. “Математическая статистика”. в: *СПб.: СПбГПУ* (2004).

[6] В Феллер. “Введение в теорию вероятностей и её приложения, пер. с англ., 2 изд., т. 1”. в: 1967.