Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

СВОДНЫЙ ОТЧЕТ

Тема: ***Одномерные распределения***

Направление: 01.03.02 Прикладная математика и информатика

Выполнил студент гр. 33631/4 Камалетдинова Ю.

Преподаватель Баженов А.

Санкт-Петербург
2019

Содержание

Постановка задачи

Данная группа лабораторных работ посвящена рассмотрению одномерных распределений. Выделим основные задачи:

- Исследовать такой способ представления набора статистических данных как гистограмма. Сгенерировать выборки разного размера для заданных распределений, построить гистограммы и сделать выводы о взаимосвязи функции плотности распределения и функции гистограммы.

- Вычислить некоторые из характеристик положения $N = 1000$ раз для выборок объемами $n = 20, 50, 100$ и проанализировать полученные результаты. Также необходимо установить, в каком соотношении находятся вычисленные характеристики для каждого из распределений, приведенных ниже

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

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

$$L(x, 0, \frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \text{ – Лаплас} \tag{3}$$

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

$$P(\lambda) = \frac{e^{-\lambda}}{k!} \lambda^k \text{ – Пуассон} \tag{5}$$

Приведем формулы для вычисления характеристик положения

$$\overline{x} = \overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ – выборочное среднее} \tag{6}$$

$$med\, x = \begin{cases} x_{(k+1)}, & n = 2k+1 \\ \frac{x_{(k)}+x_{(k+1)}}{2}, & n = 2k \end{cases} \text{ – медиана} \tag{7}$$

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \text{ – полусумма экстремальных значений} \tag{8}$$

$$z_Q = \frac{Q_1 + Q_3}{2} \text{ – полусумма квартилей} \tag{9}$$

$$z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_{(i)} \text{ – усеченное среднее} \tag{10}$$

Замечание: r – число наблюдений, оставшихся после усечения в характеристике (10), $r = \alpha n$, где α, как правило, равняется 0.1. В таком случае мы не вовлекаем 10% наибольших и 10% наименьших значений в вычисление усеченного среднего.

- Сгенерировать выборки размерами $n = 20, 100$ и построить для них боксплот Тьюки. Для каждого распределения определить процент выбросов экспериментально, сгенерировав выборку из распределения $N = 1000$ раз и вычислив средний процент выбросов, а затем сравнить с результатами, полученными теоретически. Рассматриваемые законы распределения – (1), (2), (3), (4), (5)

- Построить эмпирические функции распределения и ядерные оценки плотностей для распределений (1), (2), (3), (4), (5) с параметром $\lambda = 2$ на выборках размером $N = 20,\ 60,\ 100$ на отрезке $[-4,\ 4]$, а также сделать выводы о данных оценках законов распределений.

Описание методов

Гистограммы

Для построения гистограммы разобьем множество возможных значений ξ выбоки на непересекающиеся интервалы $\Delta_j = [t_j, t_{j+1}),\ j = \overline{0, m};\ t_0 = -\infty,\ t_{m+1} = +\infty$. Количество элементов X выборки из объема n, попавших в интервал Δ_j, обозначим через v_j, то есть $v_j = \sum_{i=1}^{n} I(X_i \in \Delta_j)$. По теореме Бернулли при $n \to \infty$

$$v_j/n \to P(\xi \in \Delta_j) = \int_{\Delta_j} f(u)du = |\Delta_j| f(c),\ c \in \Delta_j \tag{11}$$

Кусочно-постоянная функция, приведенная ниже, есть нормализованная гистограмма

$$f_n(t) = \frac{v_j}{n|\Delta_j|},\ t \in \Delta_j,\ j = \overline{1, m} \tag{12}$$

Итак, гистограмма является статистическим аналогом функции плотности распределения и функции распределения. Формулы (11), (12) указаны в пособии [1].

Выявление выбросов

Выброс – это некое наблюдение, которое нехарактерно далеко удалено от других значений в общей случайной выборке. В некотором смысле, это определение оставляет на усмотрение наблюдателя решение вопроса о том, что будет считаться выбросом.

Боксплот является удобным графическим способом описания поведения данных как в середине, так на концах распределения. Боксплот визуализирует положение медианы, нижнего и верхнего квартилей. Первый квартиль Q_1 определяется как медиана части выборки до медианного элемента всей выборки, третий квартиль Q_3 – медиана части выборки после медианного элемента всей выборки.

На графике присутствуют усы, границы которых определяются формулами

$$x_L = max(x_{(1)},\ Q_1 - 1.5 \cdot IQR) \text{ – нижняя граница уса} \tag{13}$$

$$x_U = min(x_{(n)},\ Q_3 + 1.5 \cdot IQR) \text{ – верхняя граница уса} \tag{14}$$

$$IQR = Q_3 - Q_1 \text{ – интерквартильная широта} \tag{15}$$

Будем считать элемент x_i выбросом, если $x_i \notin [x_L, x_U]$

Для сравнения теоретических и практических результатов вычислим квартили для непрерывных распределений. Квартили однозначно определяются уравнением

$$F_X(x_\alpha) = \alpha,\ \alpha = \left\{\frac{1}{4},\ \frac{3}{4}\right\} \tag{16}$$

Искомые квартили можно выразить, найдя обратную функцию к функции распределения.

Вычисление доли выбросов будем производить по формуле

$$p_{outliers} = \frac{\sum_i x_i}{\sum_{j=1}^{n} x_j},\ i : x_i \notin [x_L, x_U] \tag{17}$$

Ядерные оценки плотностей и функции плотностей распределения

Пусть имеется некоторая выборка объемом $n:\ x_1, \ldots,\ x_n;\ x_i \in \mathbb{R}$. Эмпирической функцией распределения называют

$$\hat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} u(x - x_i),\ \text{где} \tag{18}$$

$$u(z) = \begin{cases} 1,\ z \geq 0 \\ 0,\ z \leq 0 \end{cases} \text{ – фунция Хевисайда} \tag{19}$$

Ядерная оценка плотности определяется формулой:

$$\hat{f}_{n,h_n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right),\ \text{где } K(u) \text{ – ядро, } h = h_n \text{ – параметр сглаживания} \tag{20}$$

Ядро $K(u)$ – это вещественнозначная функция со следующими свойствами:

1. $K(u) \geq 0$
2. $K(-u) = K(u)$
3. $\int_{-\infty}^{+\infty} K(u)du = 1$

Ядерная оценка плотности сглаживает каждый элемент выборки до плавного участка, форма которого определяется функцией ядра $K(u)$. Затем функция суммирует все участки, чтобы получить оценку плотности. В данной работе будем использовать ядро Гаусса, заданное формулой

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{21}$$

Реализация

Для выполнения поставленных задач будем пользоваться библиотеками для языка Python: *numpy, scipy* – расчеты, законы распределения вероятностей; *matplotlib, seaborn* – визуализация результатов. Опишем ход работы для каждой задачи:

- Гистограммы
  - Производим построение графика теоретической функции распределения
  - При помощи метода `.rvs` берется выборка размера n из распределения с заданными параметрами
  - Определяем границы области интересующих нас значений и разделяем на равные интервалы (*bins*)
  - Строим гистограммы, представляющие собой столбцы высотой, пропорциональной числу попавших в их ширину значений

- Нормируем гистограммы и отрисовываем теоретическую функцию распределения в одних осях

- Характеристики положения
  - Задаем распределение с заданными параметрами
  - Генерирем случайные выборки из распределений размерами $n = 20, 50, 100$
  - Для каждого из распределений вычисляем характеристики положения $N = 1000$ раз
  - Вычисляем математическое ожидание и дисперсию для каждой вычисленной характеристики по формулам:

$$E(z) = \overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{22}$$

$$D(z) = \overline{z^2} - (\overline{z})^2 \tag{23}$$

- Выявление выбросов
  - Задаем распределение с заданными параметрами
  - Генерируем случайные выборки из распределений размерами $n = 20, 100$
  - Для каждого из распределений вычисляем доли выбросов $N = 1000$ раз
  - Вычисляем теоретические квартили при помощи метода `.ppf(`α`)` (percent point function) - функция, обратная функции распределения, вычисляем доли выбросов с использованием данных квартилей
  - Усредняем полученные суммы долей выбросов, разделив на $N = 1000$

- Ядерные оценки плотностей и эмпирические распределения
  - Задаем распределение с заданными параметрами
  - Генерируем случайные выборки из распределений объемами $n = 20,\ 60,\ 100$

– Для отсортированных выборок из распределений задаем вектор значений $y = [\frac{1}{n}, \frac{2}{n}, \ldots, 1]$ и строим ступенчатый график - эмпирическую функцию распределения

– По формулам (20), (21) вычисляем ядерные оценки плотностей для параметров сглаживания h для всех выборок и строим графики

Результат

Нормальное распределение с параметрами 0, 1

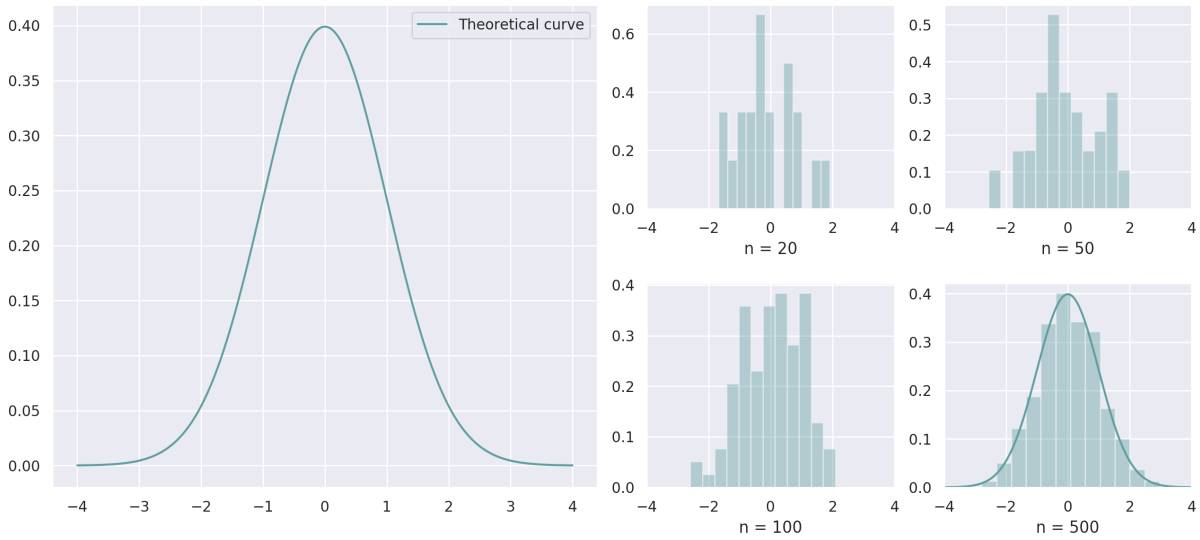


Рис. 1: Гистограмма нормального распределения

$n = 20$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0041	0.0053	0.0005	0.0046	0.0054
$D(z)$	0.0511	0.0753	0.1405	0.0586	0.0542

$n = 50$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0013	-0.0004	0.0170	0.0003	-0.0010
$D(z)$	0.0211	0.0313	0.1146	0.0251	0.0222

$n = 100$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0028	0.0026	0.0037	0.0034	0.0023
$D(z)$	0.0100	0.0152	0.0887	0.0122	0.0106

Таблицы моментов 1-го и 2-го порядков

Соотношение дисперсий при $n = 100$: $\overline{x} < z_{tr} < z_Q < med\ x < z_R$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.022	0.007
$n = 100$	0.009	0.007

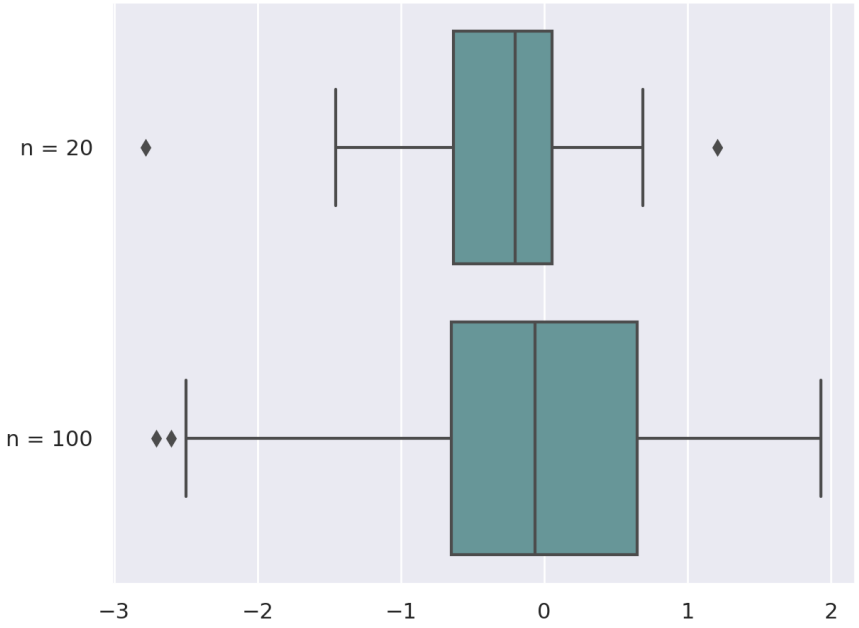
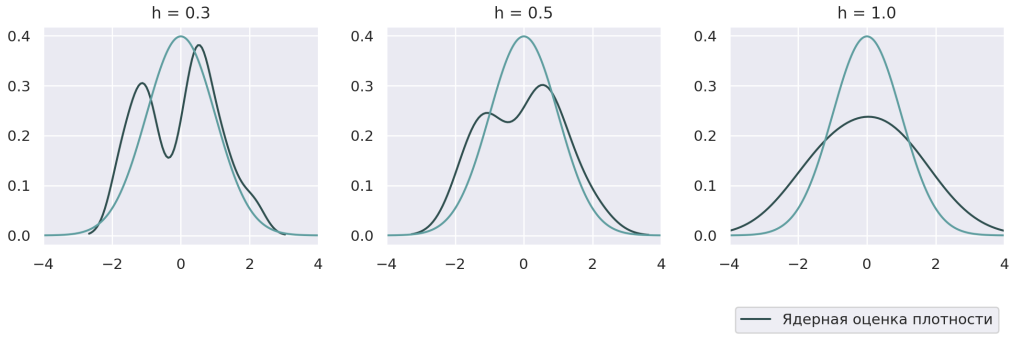


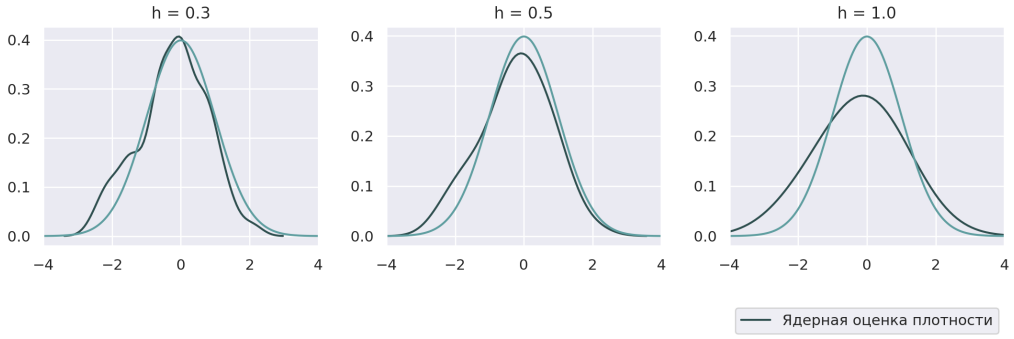
Рис. 2: Боксплот Тьюки для выборок из нормального распределения

(a) (b) (c)

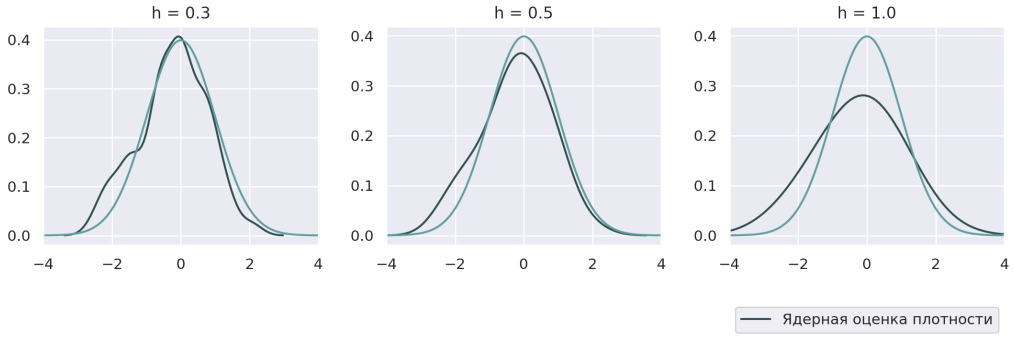
Рис. 3: Выборки из нормального распределения объемом: (a) 20; (b) 60; (c) 100



(a)



(b)



(c)

Рис. 4: Для выборок из нормального распределения объемом: (a) 20; (b) 60; (c) 100

Равномерное распределение на отрезке $[-\sqrt{3}, \sqrt{3}]$

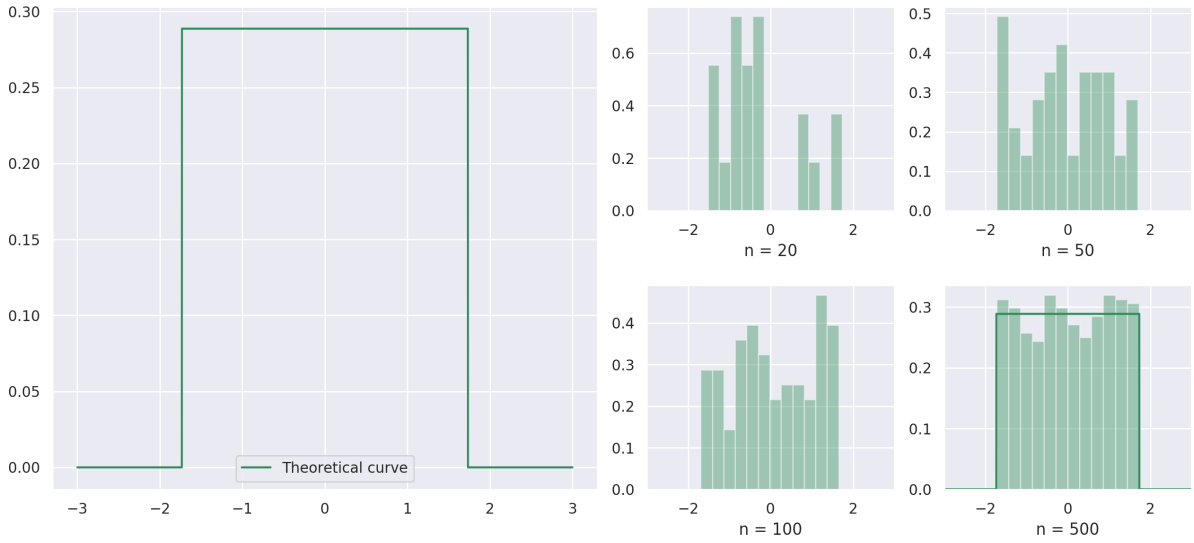


Рис. 5: Гистограмма равномерного распределения

$n = 20$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	-0.0068	-0.0119	0.0011	-0.0113	-0.0082
$D(z)$	0.0554	0.1404	0.0141	0.0728	0.0742

$n = 50$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0032	0.0049	-0.0004	0.0035	0.0040
$D(z)$	0.0209	0.0582	0.0022	0.0304	0.0288

$n = 100$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0015	0.0015	0.0011	0.0017	0.0019
$D(z)$	0.0098	0.0292	0.0007	0.0156	0.0142

Таблицы моментов 1-го и 2-го порядков

Соотношение дисперсий при $n = 100$: $z_R < \overline{x} < z_{tr} < z_Q < med\ x$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.0027	0
$n = 100$	0	0

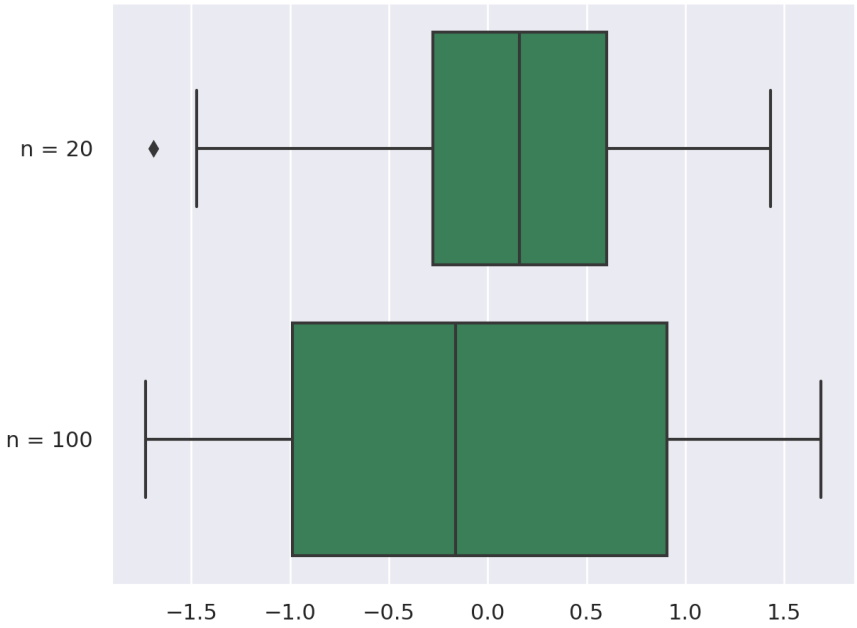
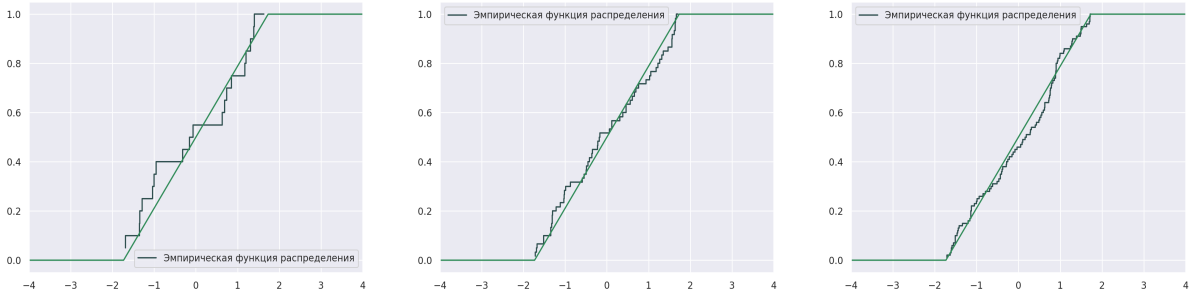
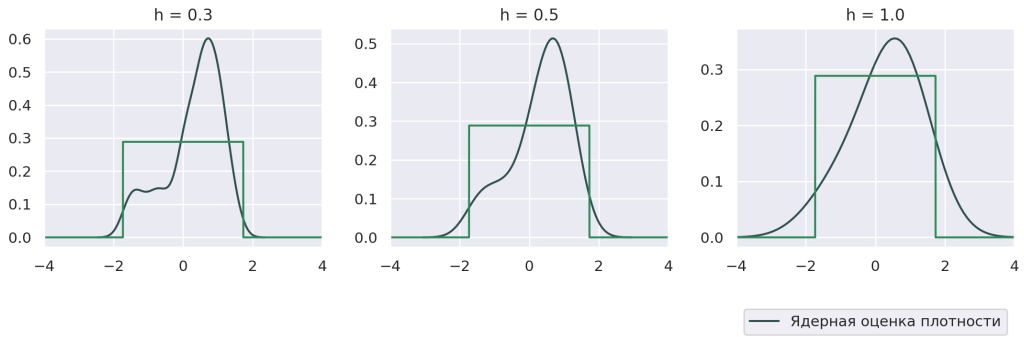


Рис. 6: Боксплот Тьюки для выборок из равномерного распределения

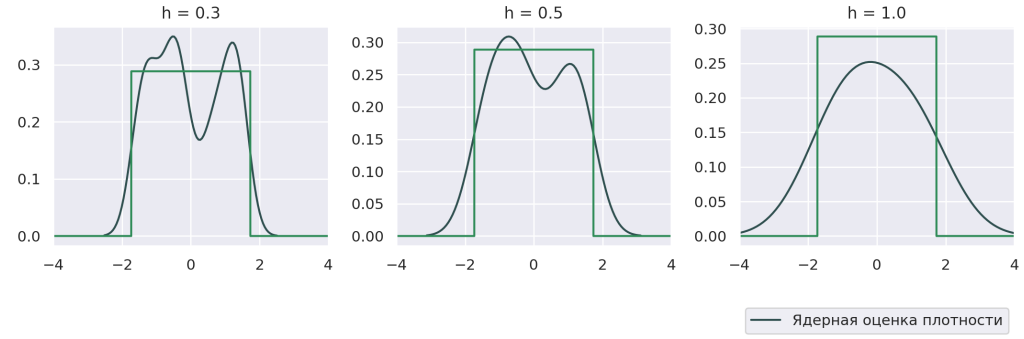


(a) (b) (c)

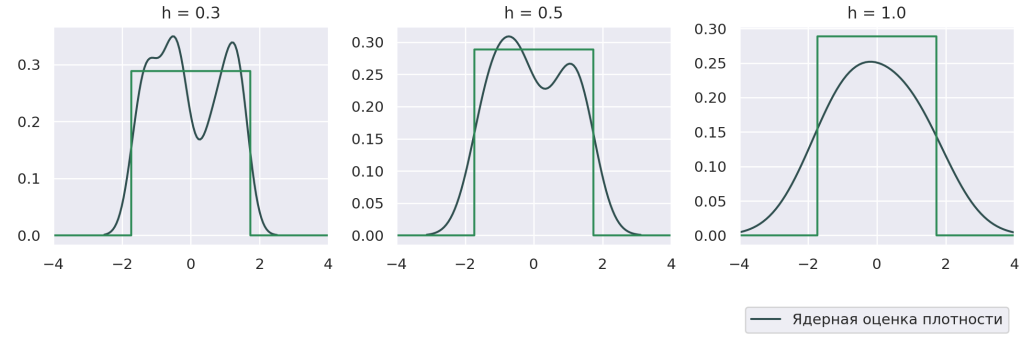
Рис. 7: Выборки из равномерного распределения объемом: (a) 20; (b) 60; (c) 100



(a)



(b)



(c)

Рис. 8: Для выборок из равномерного распределения объемом: (a) 20; (b) 60; (c) 100

Распределение Коши с параметрами 0, 1

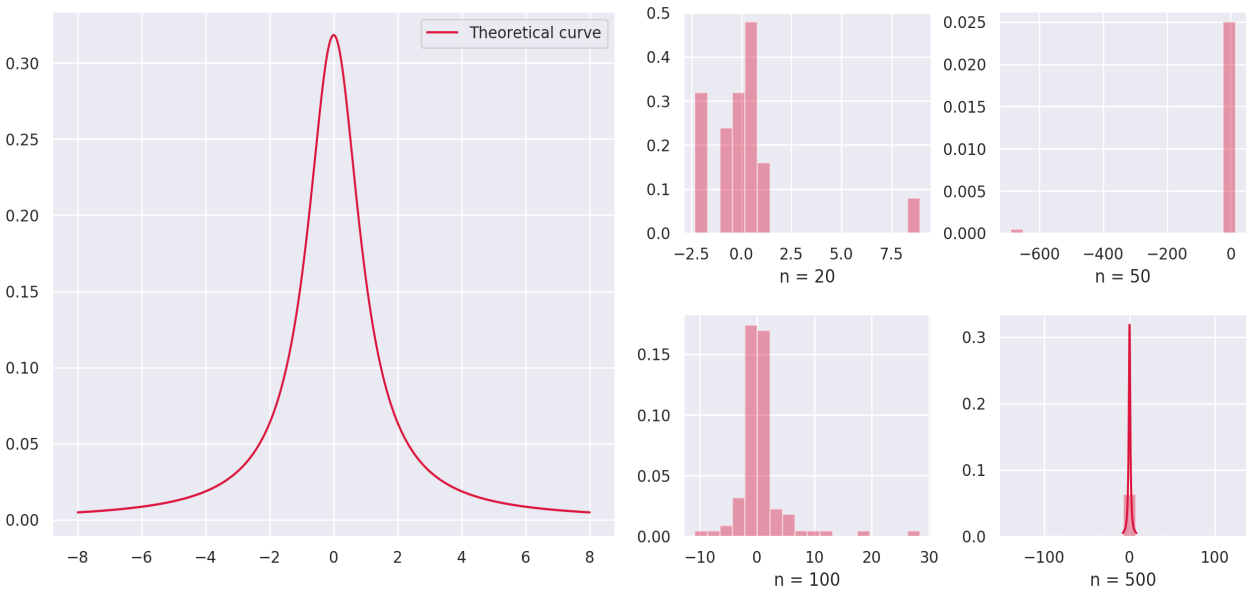


Рис. 9: Гистограмма распределения Коши

$n = 20$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	-4.1847	-0.0185	-41.9652	-0.0013	0.0043
$D(z)$	10^4	0.1450	10^6	0.3753	0.4334

$n = 50$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	3.7413	0.0051	95.1768	-0.0011	-0.0022
$D(z)$	10^4	0.0483	10^7	0.1102	0.1094

$n = 100$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	2.6928	-0.0082	133.6930	-0.0080	-0.0116
$D(z)$	10^3	0.0247	10^8	0.0522	0.0513

Таблицы моментов 1-го и 2-го порядков

Соотношение дисперсий при $n = 100$: $med\ x < z_{tr} < z_Q < \overline{x} < z_R$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.15	0.15
$n = 100$	0.15	0.15

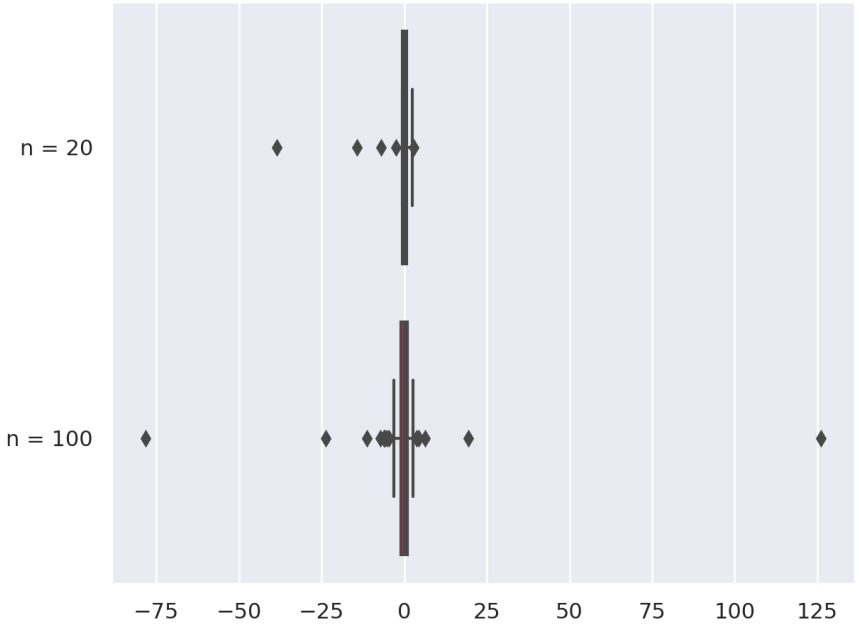
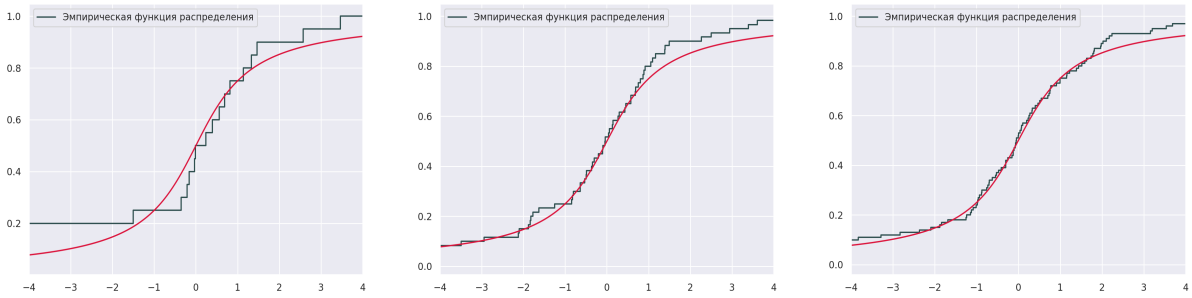
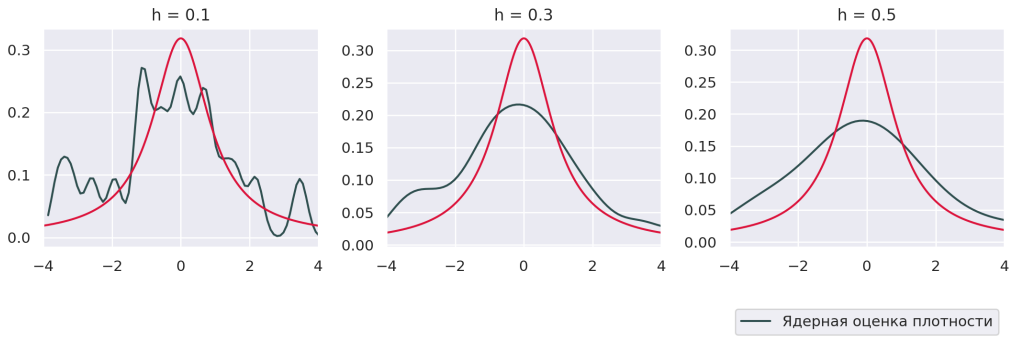


Рис. 10: Боксплот Тьюки для выборок из распределения Коши



(a) (b) (c)

Рис. 11: Выборки из распределения Коши объемом: (a) 20; (b) 60; (c) 100



(a)

(b)

(c)

Рис. 12: Для выборок из распределения Коши объемом: (a) 20; (b) 60; (c) 100

Распределение Лапласа с параметрами 0, $\frac{1}{\sqrt{2}}$

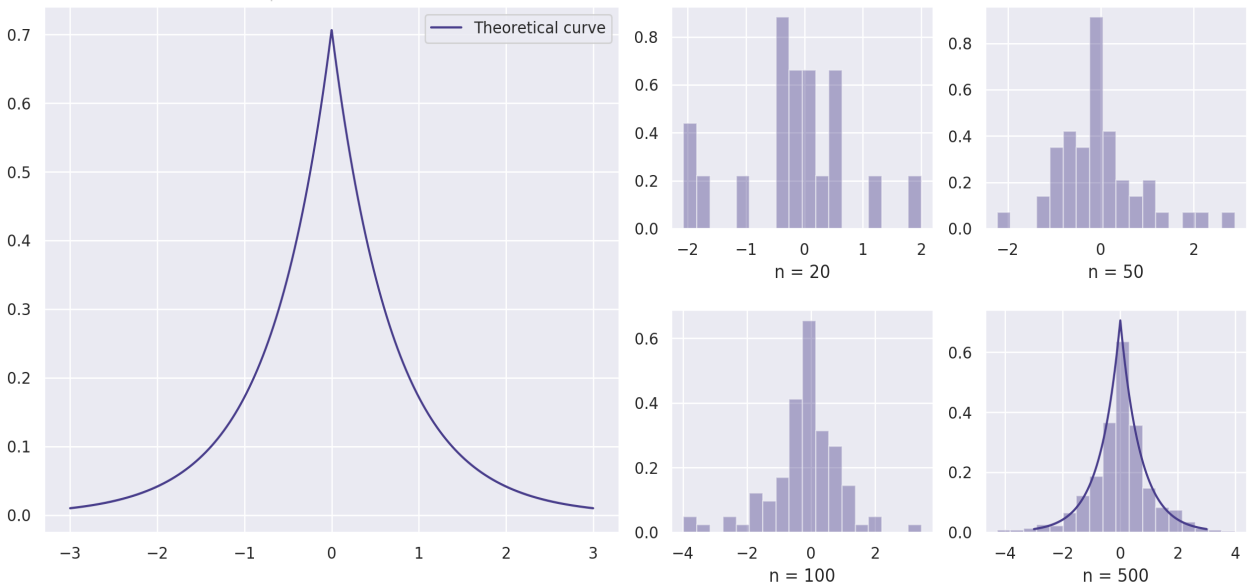


Рис. 13: Гистограмма рампределения Лапласа

$n = 20$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	-0.0113	-0.0035	-0.0344	-0.0087	-0.0070
$D(z)$	0.0550	0.0348	0.4406	0.0519	0.0427

$n = 50$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	0.0028	-0.0010	0.0085	0.0035	0.0019
$D(z)$	0.0193	0.0130	0.3839	0.0206	0.0156

$n = 100$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	-0.0059	-0.0029	-0.0104	-0.0047	-0.0051
$D(z)$	0.0097	0.0056	0.3931	0.0093	0.0071

Таблицы моментов 1-го и 2-го порядков

Соотношение дисперсий при $n = 100$: $med\ x < z_{tr} < z_Q < \overline{x} < z_R$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.07	0.06
$n = 100$	0.06	0.06

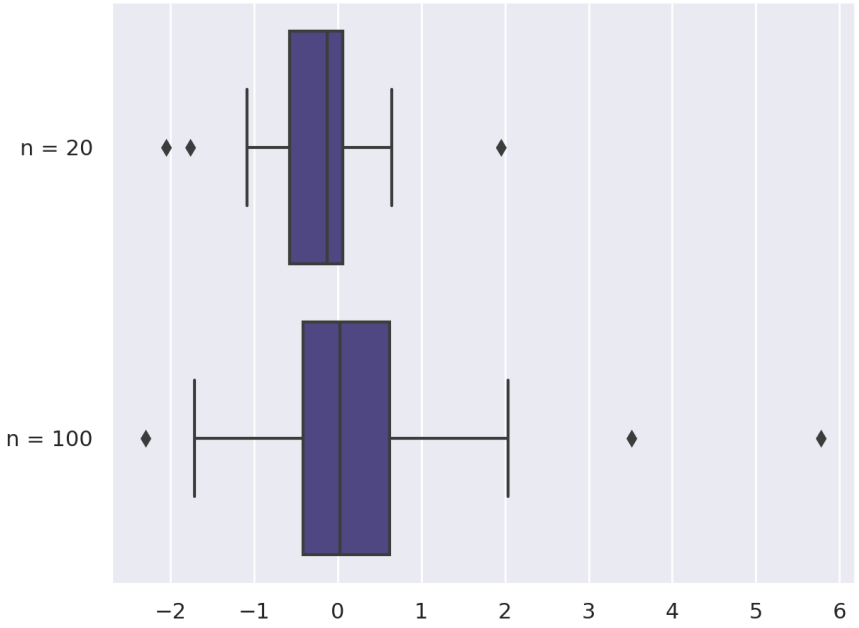
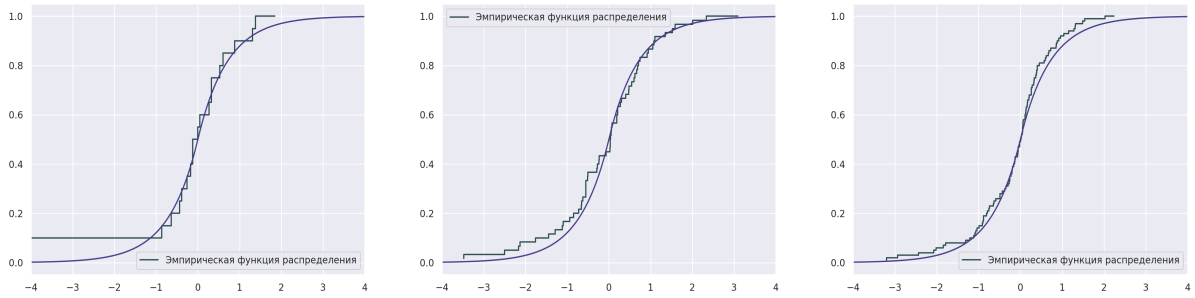
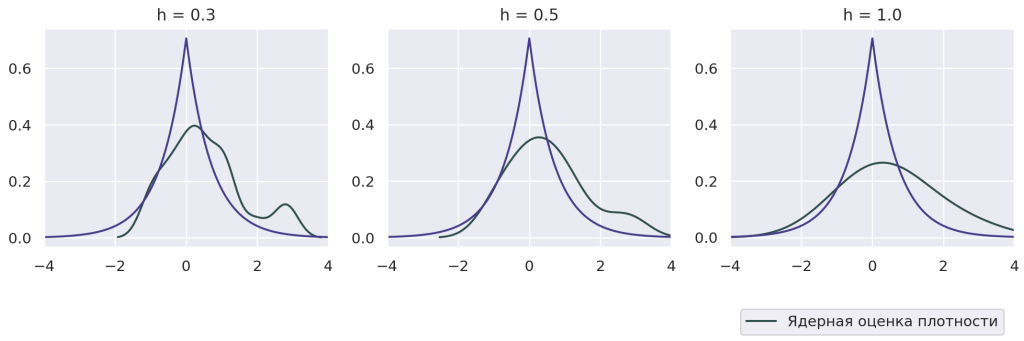


Рис. 14: Боксплот Тьюки для выборок из распределения Лапласа

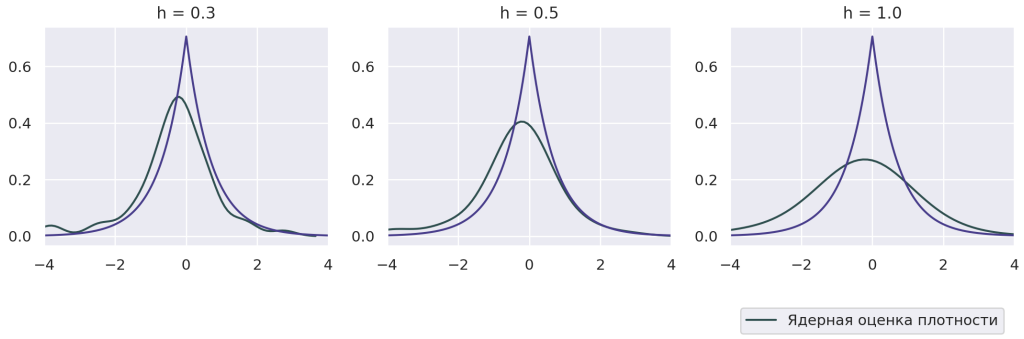


(a) (b) (c)

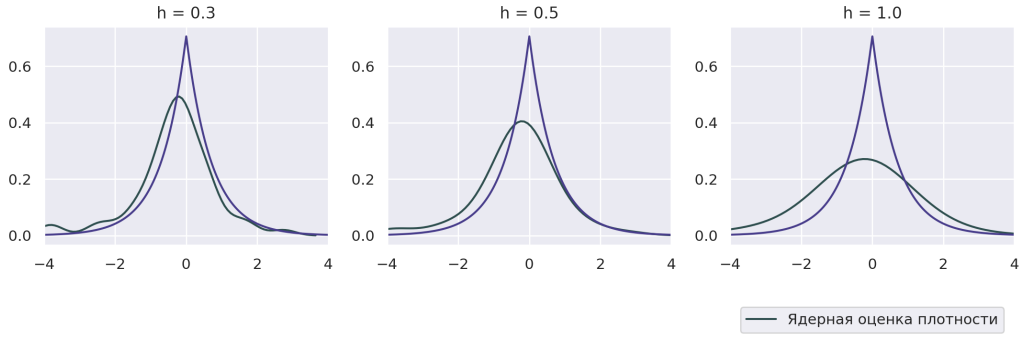
Рис. 15: Выборки из распределения Лапласа объемом: (a) 20; (b) 60; (c) 100



(a)



(b)



(c)

Рис. 16: Для выборок из распределения Лапласа объемом: (a) 20; (b) 60; (c) 100

Распределение Пуассона с параметром $\lambda = 7$

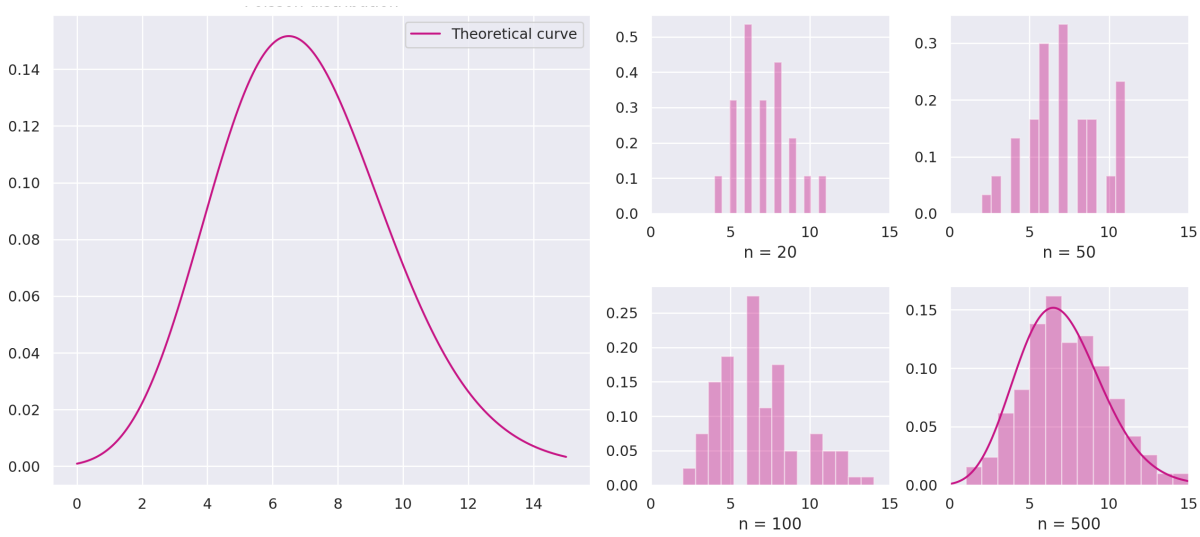


Рис. 17: Гистограмма рампределения Пуассона

$n = 20$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	6.9844	6.8310	7.4820	6.9067	6.8978
$D(z)$	0.3285	0.5259	1.0587	0.4049	0.3441

$n = 50$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	7.0165	6.8640	7.7515	6.9400	6.9287
$D(z)$	0.1561	0.2665	0.8315	0.2299	0.1639

$n = 100$	$\overline{x}$	$med\ x$	z_R	z_Q	z_{tr}
$E(z)$	7.0063	6.8755	7.9230	6.9088	6.9181
$D(z)$	0.0714	0.1482	0.7266	0.1074	0.0746

Таблицы моментов 1-го и 2-го порядков

Соотношение дисперсий при $n = 100$: $\overline{x} < z_{tr} < z_Q < med\ x < z_R$

	Практическая доля выбросов	Теоретическая доля выбросов
$n = 20$	0.027	0.003
$n = 100$	0.012	0.002

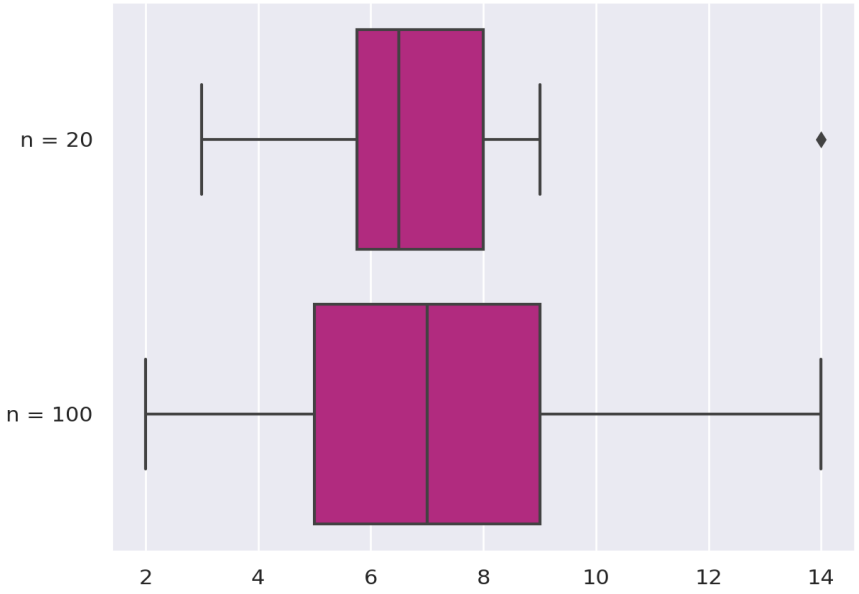
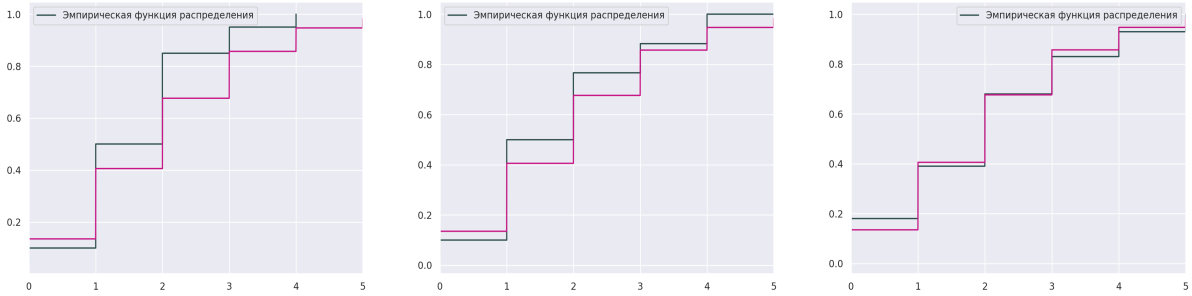
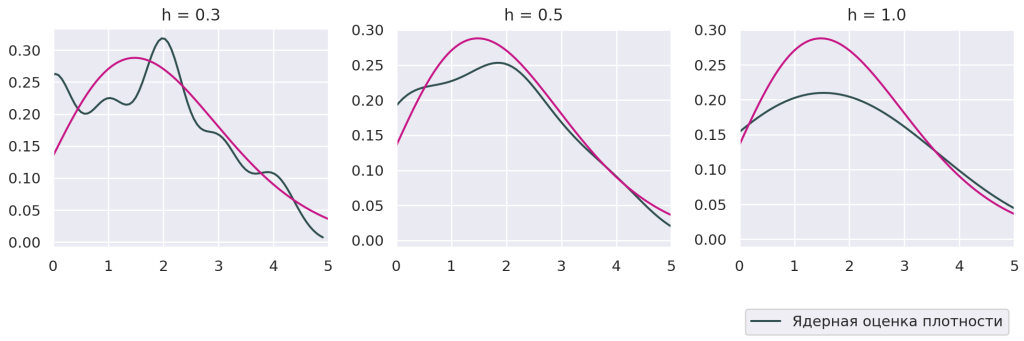


Рис. 18: Боксплот Тьюки для выборок из распределения Пуассона

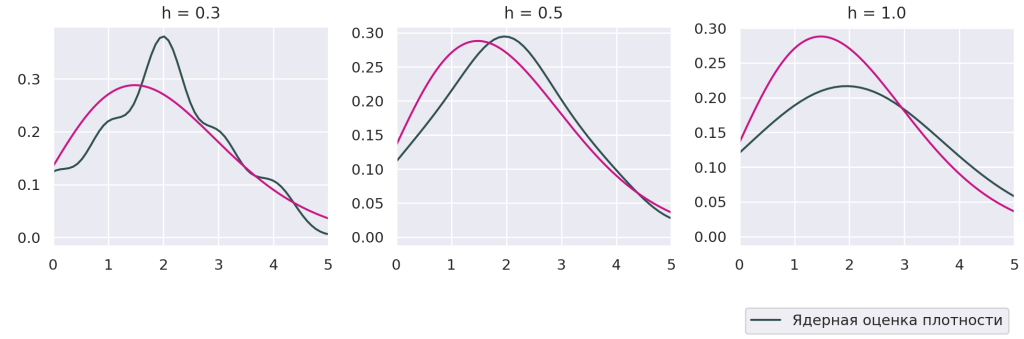


(a) (b) (c)

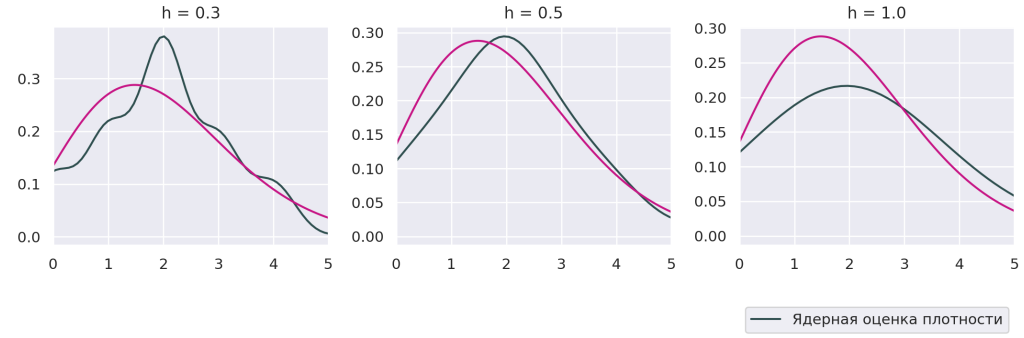
Рис. 19: Выборки из распределения Пуассона с параметром $\lambda = 2$ объемом: (a) 20; (b) 60; (c) 100



(a)



(b)



(c)

Рис. 20: Для выборок из распределения Пуассона $\lambda = 2$ объемом: (a) 20; (b) 60; (c) 100

Выбор параметра обоснован стремлением распределения к нормальному при увеличении λ. Для наглядной визуализации подходит число 7, к тому же являющееся числом Миллера, или предельной порцией информации, обрабатываемой человеком за раз (оригинальная статья - [2]). Этот факт является довольно символичным, учитывая область применения распределения Пуассона. Чтобы построить эмпирическую функцию распределения и ядерную оценку плотности на указанном отрезке, будем использовать $\lambda = 2$.

Выводы

При достаточной мощности выборки из распределений (n = 100) гистограмму можно рассматривать как аналог плотности распределения непрерывной случайной величины.

По полученным соотношениям дисперсий выборок из распределений можно сделать вывод о том, что полусумма экстремальных значений z_R имеет наибольший разброс относительно математического ожидания. Данное суждение не сходится с результатами для случая равномерного распределения на отрезке. Медиана равномерного распределения на $[a, b]$ есть $med = \frac{a+b}{2}$, что и является полусуммой значений в крайних точках отрезка.

Также в случае симметричного распределения, исходя из дисперсий, наиболее выгодно использовать выборочное среднее, чем медиану, хотя они и оценивают одну и ту же величину в данном случае. Но для распределения Лапласа медиана становится более эффективной.

Дисперсии таких характеристик, как усеченное среднее или полусумма квартилей, показывают среднее отклонение относительно остальных. Усеченное среднее представляет собой некий баланс между медианой и выборочным средним (является ими в частных случаях параметра α)

Взглянув на полученные боксплоты Тьюки, можно сделать вывод о том, что равномерное распределение не имеет выбросов. Ненулевое значение в малой выборке при расчете квантилей с использованием значений выборки есть показатель того, что нельзя судить о характере распределения по выборке размером $n = 20$ окончательно. Отсутствие

выбросов объясняется тем, что функция плотности такого респределения постоянна, и элемент выборки никогда будет вне отрезка, ограниченного (13), (14)

Наибольший процент выбросов установлен для выборов из распределения Коши. Также увеличение выборки не влияет на изменение результата, математического ожидания нет. Обратимся к результатам предыдущей лабораторной работы. Дисперсия размаха распределения принимает значения порядка 10^8. Это объясняет тот факт, что доля выбросов достаточно высока. Распределение Коши имеет тяжелые хвосты, и доля элементов, принимающих далекие от центра распределения значения, на порядок выше по сравнению с остальными распределениями, такими как нормальное, Лапласа, Пуассона.

Описывая полученные статистические функции, можно заключить, что чем больше выборка, тем точнее эмпирическая функция распределения оценивает теоретическую.

Точность ядерной оценки плотности сильно варьируется в зависимости от значения сглаживающего параметра h. Так, при $h \to 0$ оценка плотности точна на выборочных данных, но только на них, и такая функция не способна описать характер распределения. Выбрав $(n+1)$-ое значение из распределения, мы можем столкнуться с тем, что статистическая функция плохо оценит значение теоретической функции плотности вероятности для данного выборочного элемента. В таком случае можно говорить о плохой способности функции к обобщению.

Напротив, при увеличении параметра h ядерная оценка плотности может показывать себя плохо даже на выборочных данных и вообще не позволяет понять характера распределения.

Выбор параметра сглаживания следует производить исходя из того, насколько плотно распределение объектов выборки; большей плотности соответствует выбор меньшего параметра и наоборот.

Список литературы

[1] *Кадырова Н. О.* Теория вероятностей и математическая статистика. Статистический анализ данных: учеб. пособие / *Н. О. Кадырова, Л. В. Павлова, И. Е. Ануфриев.* - СПб.: Изд-во Политехн. ун-та, 2010. -54с.

[2] *Miller, G. A.* The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 1956; 63:81–97.

[3] *Chrisman, L.* How the strange Cauchy distribution proved useful. *Lumina Decision Systems* (2018). URL: http://www.lumina.com/blog/how-the-strange-cauchy-distribution-proved-useful

[4] *Conlen, M.* Kernel Density Estimation (2019). URL: https://mathisonian.github.io/kde/