Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторным работам № 1-4
по дисциплине: «Математическая статика».

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

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

Санкт-Петербург
2020 г.

Оглавление

Список иллюстраций

Список таблиц

Постановка задачи

Для каждого из 5 распределений:

1. Нормального $\mathcal{N}(x, 0, 1)$
2. Коши $\mathcal{C}(x, 0, 1)$
3. Лапласа $\mathcal{L}(x, 0, \frac{1}{\sqrt{2}})$
4. Пуассона $\mathcal{P}\ (k,\ 10)$
5. Равномерного $\mathcal{U}\left(x, -\sqrt{3}, \sqrt{3}\right)$

1. Сгенерировать выборки размеров: 10, 50, 1000; и построить графики плотности распределения вероятности и гистограмму на одном рисунке.

2. Сгенерировать выборки размеров: 10, 100, 1000. Вычислить следующие статистические характеристики для каждой выборки:
$$\bar{x}, med, x, z_R, z_Q, z_{tr}.$$
Для каждой выборки провести подобные вычисления по 1000 раз и найти среднее значение их характеристик положения и их квадратов:
$$E(z) = \bar{z}; \quad (1)$$
Вычислить оценку дисперсии:
$$D(z) = \overline{z^2} - \bar{z}^2; \quad (2)$$
Представить полученные данные в виде таблицы.

3. Сгенерировать выборки размеров: 10, 50, 1000; и построить графики плотности распределения вероятности и гистограмму на одном рисунке.

   Сгенерировать выборки размером 20, 100 элементов и построить для них боксплот Тьюки. Для каждого распределения экспериментально определить долю выбросов (сгенерировав выборку, соответствующую распределению, 1000 раз, и вычислить среднюю долю выбросов) и сравнить с результатами, полученными теоретически.

4. Сгенерировать выборки размеров: 20, 60, 100. 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.

Теория

1. Лабораторная

1.1. Распределения

Пусть задано вероятностное пространство $(\Omega, \mathcal{F}, P)$, на котором определена случайная величина $\xi : \Omega \longrightarrow \mathbb{R}$.

Функция $F_\xi(x) = P_\xi\ (-\infty,\ x]\ \ x \in \mathbb{R}$ называется функцией распределения случайной величины ξ.

В данной лабораторной работе рассматриваются следующие распределения:

1. Нормальное (абсолютно непрерывное)

   Распределение вероятностей, которое в одномерном случае задается функцией плотности вероятности, совпадающей с функцией Гаусса:

$$f(x,\ \mu,\quad \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

   μ — математическое ожидание

   σ — среднеквадратическое отклонение (σ^2 —дисперсия)

2. Коши (абсолютно непрерывное)

   Задается плотностью вероятности:

$$f(x, x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2} \tag{4}$$

   x_0 — параметр сдвига

   γ — параметр масштаба

3. Лапласа (абсолютно непрерывное)

   Задается плотностью вероятности:

$$f(x, \beta, \alpha) = \frac{\alpha}{2} e^{-\alpha\,|x-\beta|} \tag{5}$$

   α — параметр масштаба

   β — параметр сдвига

4. Пуассона (дискретное)

   Задается функцией вероятности:

$$P(k) \equiv \mathbb{P}(\,Y = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{6}$$

   λ — математическое ожидание случайной величины (среднее количество событий за фиксированный промежуток времени)

5. Равномерное (абсолютно непрерывное)

Задается плотностью вероятности:

$$f\ (x, a, b) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b] \\ 0\ \ , & x \notin [a, b] \end{cases} \tag{7}$$

1.2. Гистограмма

Гистограмма – функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него.

Построение гистограммы основывается на выделении интервалов и выстраивании пропорциональных прямоугольников. Множество значений, которые может принимать элемент выборки, разбивается на интервалы. Для каждого интервала на горизонтальной оси строится прямоугольник, его высота пропорциональна числу элементов выборки, попавших в этот интервал. (при разных интервалах: площадь прямоугольника пропорциональна числу элементов выборки в интервале). Также существует правило нормировки – общая площадь всех прямоугольников равна единице.

2. Лабораторная

2.1. Вариационный ряд

Вариационный ряд (или упорядоченная выборка) – последовательность

$X_{(1)} \leq\ X_{(2)} \leq\ \cdots \leq X_{(n-1)} \leq X_{(n)}$,

состоящая из одинаково распределенных случайных величин: $X_1, X_2, X_3, \cdots, X_n$, расположенных в неубывающем порядке. (Может быть получена из исходной выборки, в результате расположения элементов в неубывающем порядке)

Значение k-го элемента вариационного ряда $x_{(k)}$называется k-ой порядковой статистикой.

2.2. Выборочные числовые характеристики

Математическая статистика рассматривает приближенные методы отыскания законов распределения и числовых характеристик по результатам экспериментов.

Выборочные характеристики – случайные величины как борелевские функции от случайных величин.

Рассматриваются следующе характеристики:

- Выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{8}$$

- Выборочная медиана:

$$med\ x = \begin{cases} x_{(l+1)} & \text{при} \quad n = 2l+1 \\ \dfrac{x_{(l)} + x_{(l+1)}}{2} & \text{при} \quad n = 2l \end{cases} \tag{9}$$

- Полусумма экстремальных выборочных элементов:

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

- Выборочный квантиль порядка p:

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

- Полусумма квантилей:

$$z_Q = \frac{z_{0.25} + z_{0.75}}{2} \tag{12}$$

- Усечённое среднее:

$$z_{tr} = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}\text{, где } r = \left\lceil \frac{n}{4} \right\rceil \tag{13}$$

3. Лабораторная

3.1. Диаграмма размахов (« ящик с усами »)

Определение

— это график, позволяющий дать статистическую характеристику анализируемой совокупности.

Графики этого типа очень популярны, поскольку позволяют дать очень полную статистическую характеристику анализируемой совокупности.

Построение

Чтобы нарисовать ящик для одной группы про исходные данные необходимо знать всего три характеристики:

- Первый квартиль: $Q_{25} = X_{[1/4]}$
- Медиану: $Q_{50} = X_{[1/2]}$
- Третий квартиль $Q_{75} = X_{[3/4]}$

Диаграммы размахов, или "ящики с усами", получили свое название за характерный вид: 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 и полутора межквартильных расстояний. Формула имеет вид:

$$X_1 = Q_{25} - \frac{3}{2}(Q_{75} - Q_{25}), \tag{14}$$

$$X_2 = Q_{75} + \frac{3}{2}(Q_{75} - Q_{25}), \tag{15}$$

где X_1 – нижняя граница уса, X_2 – верхняя граница уса.

Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков.

3.2. Выбросы

Определение

Выброс – результат измерения, выделяющийся из выборки. Если элемент выборки не лежит в диапазоне $[X_1, X_2]$, то это и есть выброс.

Теоретическая вероятность выбросов

Для непрерывных распределений:

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

Для дискретных с учетом возможного скачка:

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

Где $F(X) = P(x \le X)$ – функция распределения.

4. Лабораторная

4.1. Эмпирическая функция распределения

Эмпирической функцией распределения, построенной на выборке $(x_1, \dots, x_n)$ объема n, называется случайная величина

$$F_n^*(y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i < y) \tag{18}$$

где I – индикатор события $x_i < y$.

4.2. Ядерная оценка плотности распределения

Если имеется выборка, полученная по распределению с некоторой плотностью f, то ядерной оценкой плотности этой функции называется [2]:

$$\hat{f}_h = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{19}$$

где K – ядро (неотрицательная функция), $h > 0$ – сглаживающий параметр (ширина полосы).

Чаще всего используется нормальное (гауссово) ядро, в силу его удобных математических свойств:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{20}$$

И если используется гауссово ядро, и оцениваемая плотность является гауссовой, оптимальный выбор для h определяется правилом Сильвермана [2]:

$$h_n = \left(\frac{4s_n^5}{3n}\right)^{\frac{1}{5}} \approx 1.06 s_n n^{-\frac{1}{5}} \tag{21}$$

где s_n – выборочное среднеквадратичное отклонение (корень из выборочной дисперсии).

Реализация

Код программы, реализующий данные задачи, был написан на языке Python в интегрированной среде разработке PyCharm.

Были использованы библиотеки:

- **Numpy** – библиотека для работы с данными.
- **Matplotlib** — комплексная библиотека для создания статических, анимированных и интерактивных визуализаций в Python.
- **Seaborn** — библиотека визуализации данных Python, основанная на matplotlib. Она обеспечивает высокоуровневый интерфейс для рисования привлекательной и информативной статистической графики.

Результаты

1. Гистограммы и плотности распределения

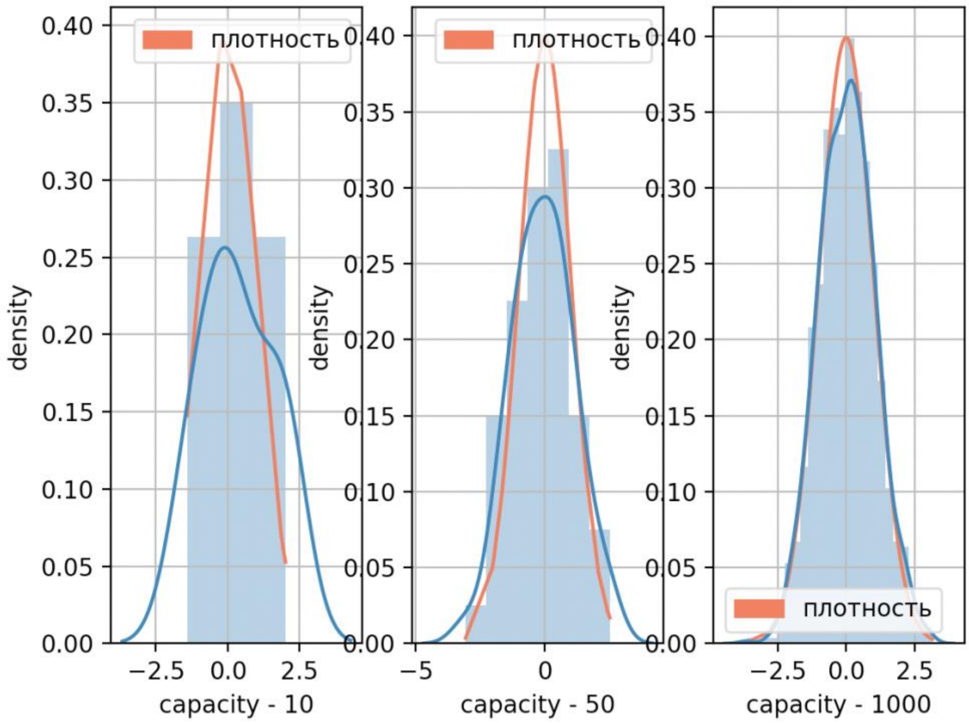

Рисунок 1: Нормальное распределение

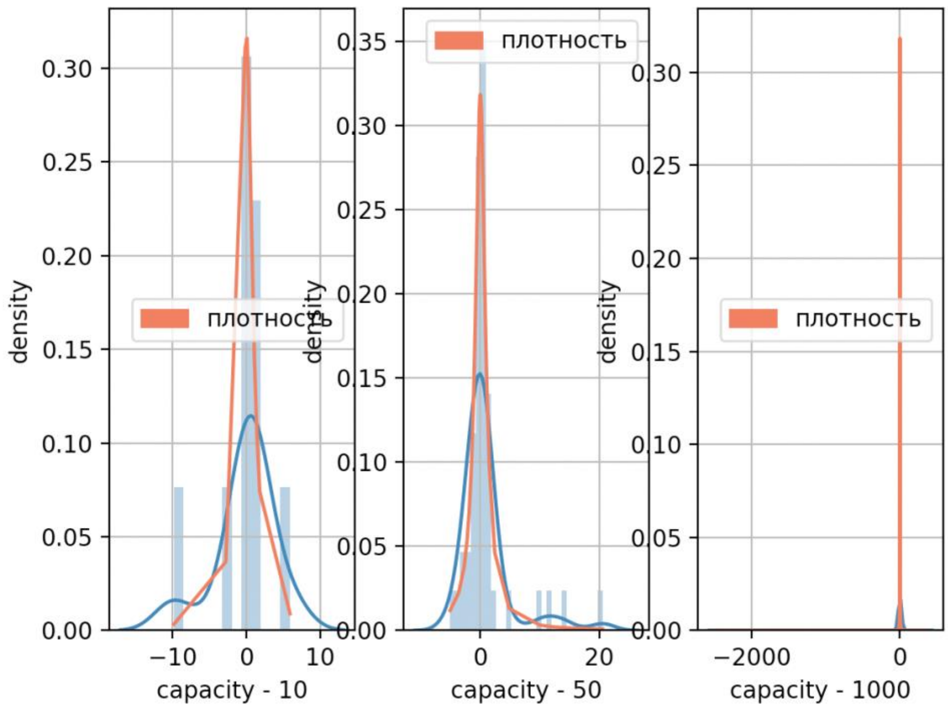

Рисунок 2: Распределение Коши

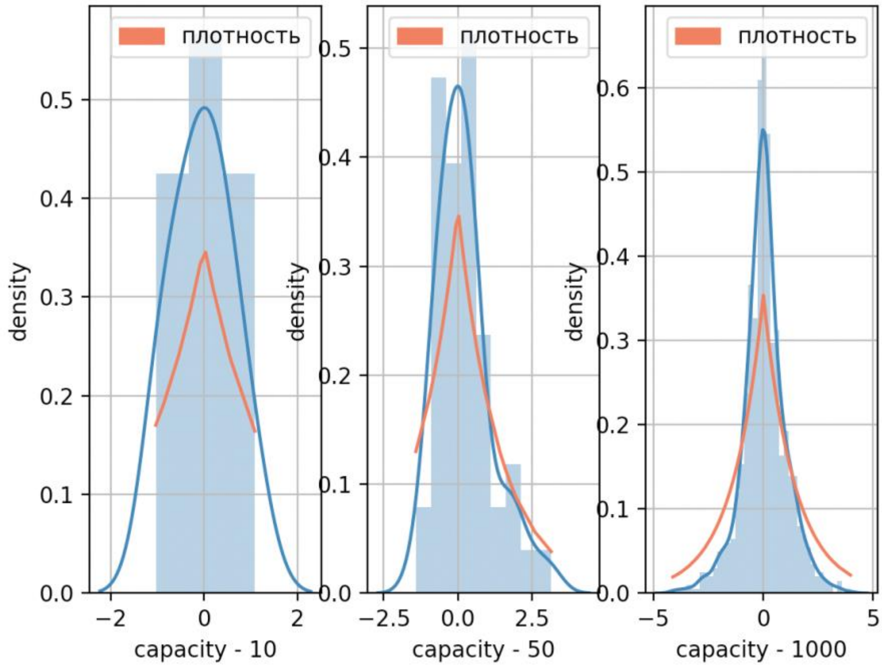

Рисунок 3: Распределение Лапласа

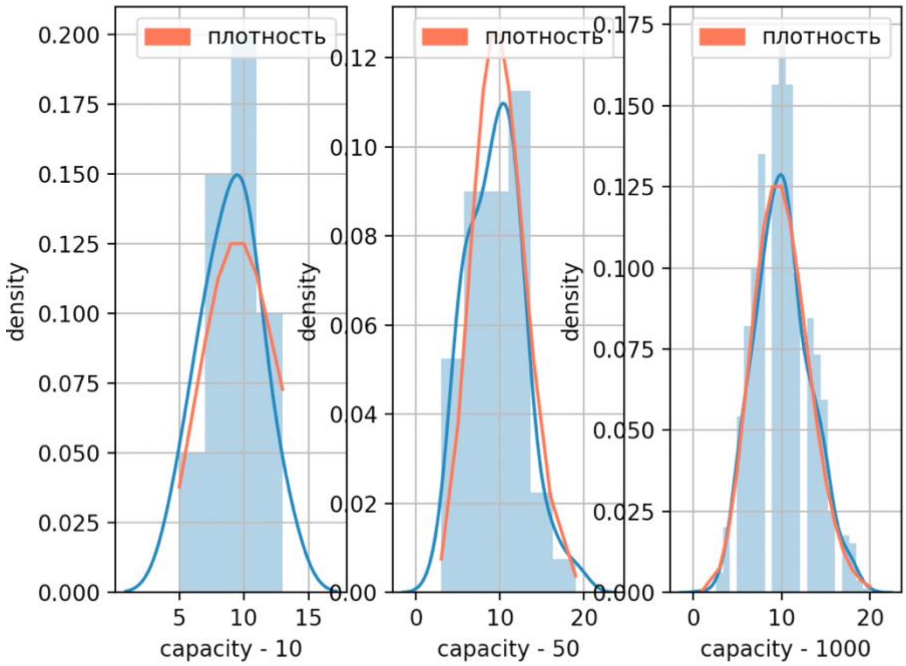

Рисунок 4: Распределение Пуассона

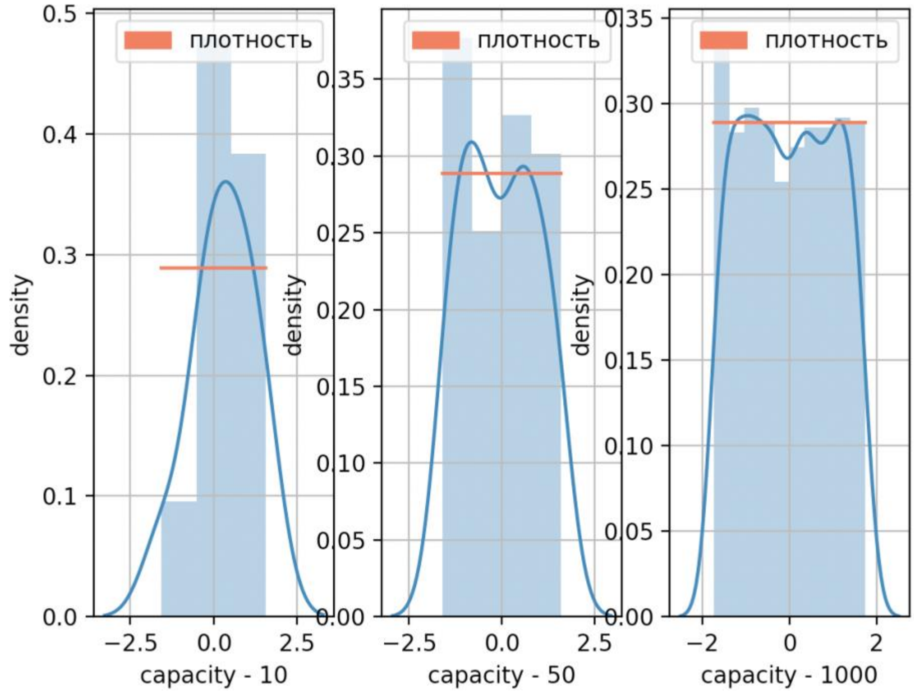

Рисунок 5: Равномерное распределение

2. Характеристики положения и рассеяния

	$\bar{x}$(3)	$med\ x$(4)	z_R(5)	z_Q(7)	z_{tr}(8)
$n = 10$					
$E(z)$	-0.01	0.0	0.0	0.3	0.0
$D(z)$	0.090461	0.145400	0.190432	0.122063	0.1860126
$n = 100$					
$E(z)$	0.001	0.00	-0.00	0.02	-0.00
$D(z)$	0.0092806	0.0162908	0.09482238	0.01335049	0.0214439
$n = 1000$					
$E(z)$	-1.147	0.000	-0.00	0.001	0.003
$D(z)$	0.00101	0.00156	0.061354	0.0012	0.0019

Таблица 1: Нормальное распределение

	$\bar{x}$(3)	$med\ x$(4)	z_R(5)	z_Q(7)	z_{tr}(8)
$n = 10$					
$E(z)$	-4.8	-0.0	-14.4	1.2	0.2
$D(z)$	22978.4851	0.35161	134103.5860	5.9101	304.4308
$n = 100$					
$E(z)$	-3.0	-0.00	51.4	0.04	0.3
$D(z)$	22963.8426	0.0224	3810135.8719	0.05342	1000.4721
$n = 1000$					
$E(z)$	1.8	0.001	819.1	0.005	5.5
$D(z)$	10823.6848	0.0025	1589854120.7	0.0055	25780.6205

Таблица 2: Распределение Коши

	$\bar{x}$(3)	$med\ x$(4)	z_R(5)	z_Q(7)	z_{tr}(8)
$n = 10$					
$E(z)$	0.0	0.00	0.0	0.3	-0.0
$D(z)$	0.10458	0.07528	0.39273	0.123057	0.192249
$n = 100$					
$E(z)$	0.002	-0.001	0.0	0.02	0.00
$D(z)$	0.009718	0.005793	0.398961	0.0110906	0.02108
$n = 1000$					
$E(z)$	-0.002	-0.00	0.02	0.001	0.00
$D(z)$	0.001094	0.000493	0.439785	0.000954	0.002094

Таблица 3: Распределение Лапласа

	$\bar{x}$**(3)**	$med\ x$**(4)**	z_R**(5)**	z_Q**(7)**	z_{tr}**(8)**
$n = 10$					
$E(z)$	10	9.8	10.3	11	10
$D(z)$	0.96096	1.5011	1.9474	1.4409	2.0389
$n = 100$					
$E(z)$	10.0	9.9	10	10	9.9
$D(z)$	0.10364	0.19059	0.96729	0.16532	0.20728
$n = 1000$					
$E(z)$	9.996	9.99	11.7	9.994	10.00
$D(z)$	0.0099	0.0091	0.61238	0.00320	0.0190

Таблица 4: Распределение Пуассона

	$\bar{x}$**(22)**	$med\ x$	z_R	z_Q	z_{tr}
$n = 10$					
$E(z)$	0.01	-0.0	-0.00	0.3	-0.0
$D(z)$	0.098	0.2237	0.05196	0.13023	0.20549
$n = 100$					
$E(z)$	-0.003	-0.00	0.0014	0.01	-0.00
$D(z)$	0.0093	0.0295	0.000543	0.015398	0.0205049
$n = 1000$					
$E(z)$	0.0008	0.001	-8e-06	0.001	0.000
$D(z)$	0.000983	0.00318	6e-06	0.001490	0.00193

Таблица 5: Равномерное распределение

3. Боксплот Тьюки

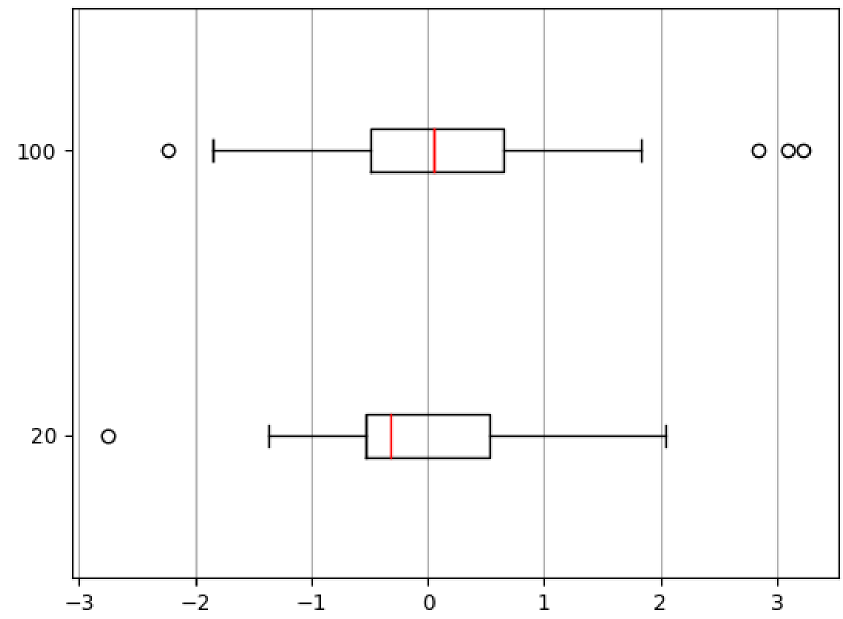

Рисунок 6: Нормальное распределение

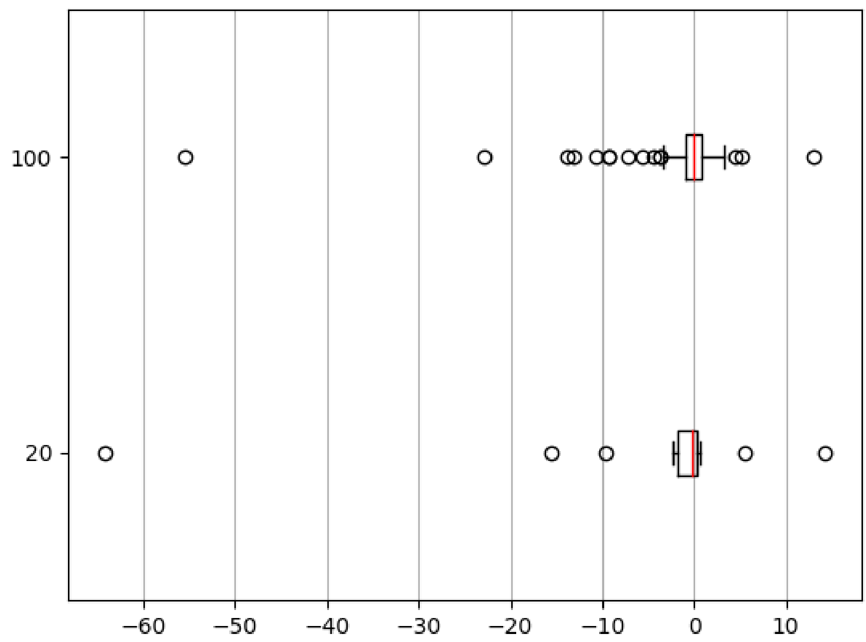

Рисунок 7: Распределение Коши

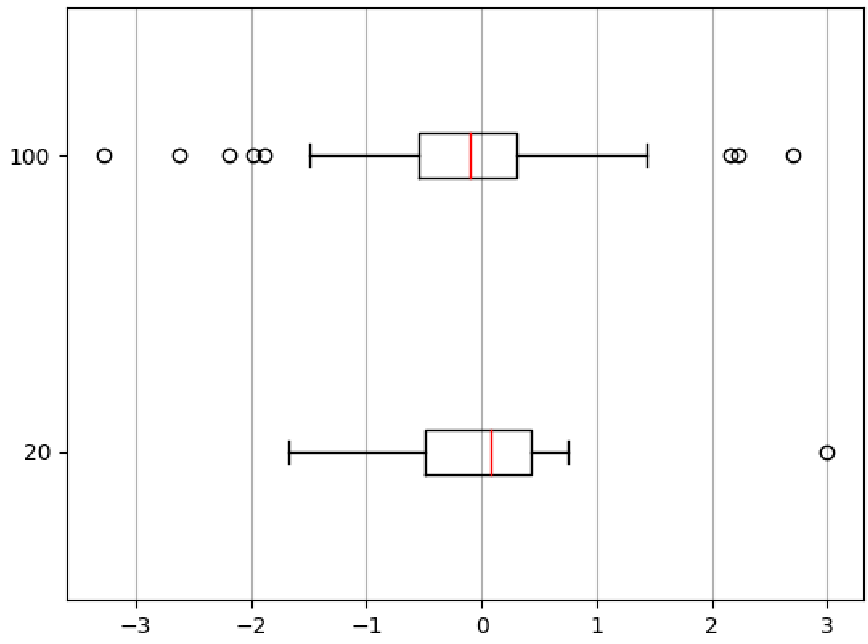

Рисунок 8: Распределение Лапласа

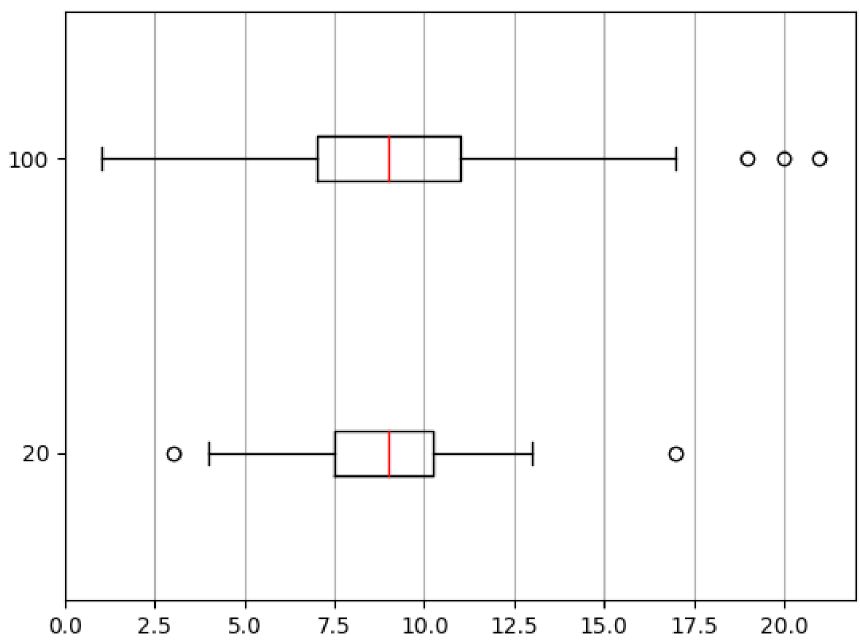

Рисунок 9: Распределение Пуассона

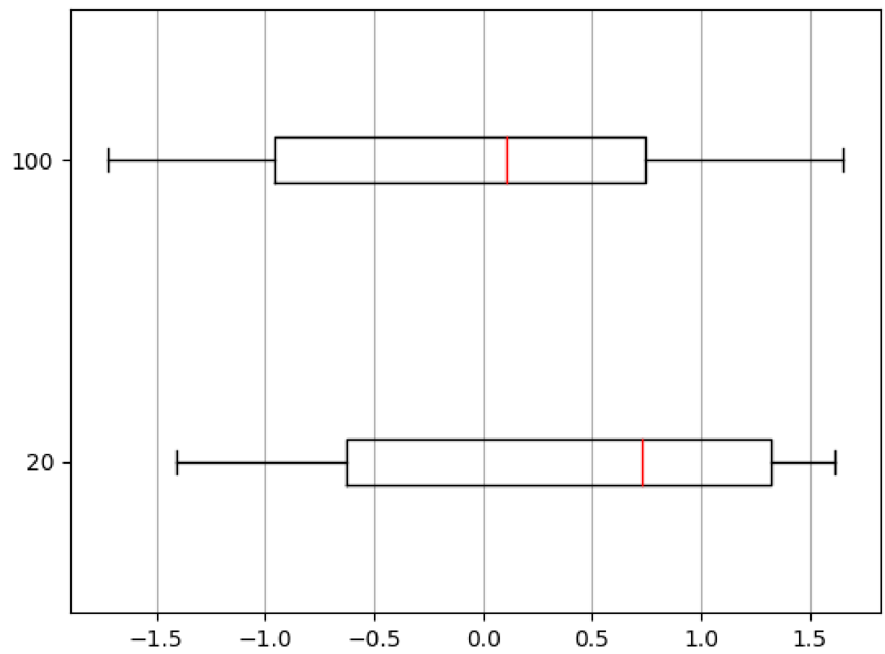

Рисунок 10: Равномерное распределение

3.1. Теоретическая вероятность выбросов

Распределение	Q_{25}^{T}	Q_{75}^{T}	X_1^T(1)	X_2^T(2)	P_B^T(3)(4)
Нормальное	-0.674	0.674	-2.698	2.698	0.007
Коши	-1	1	-4	4	0.156
Лапласа	-0.490	0.490	-1.961	1.961	0.063
Пуассона	8	12	2	18	0.008
Равномерное	-0.866	0.866	-3.464	3.464	0

Таблица 6: Теоретическая вероятность выбросов

3.2. Доля выбросов

Выборка	Доля выбросов	Дисперсия
Нормальное n = 20	0.02285	0.001715
Нормальное n = 100	0.01041	0.00017
Коши n = 20	0.15105	0.004681
Коши n = 100	0.15442	0.0011
Лаплас n = 20	0.07565	0.004625
Лаплас n = 100	0.06503	0.00096
Пуассон n = 20	0.0234	0.001802
Пуассон n = 100	0.01031	0.000223
Равномерное n = 20	0.00335	0.000401
Равномерное n = 100	0.0	0.0

Таблица 7: Доля выбросов

4. Эмпирическая функция выбросов

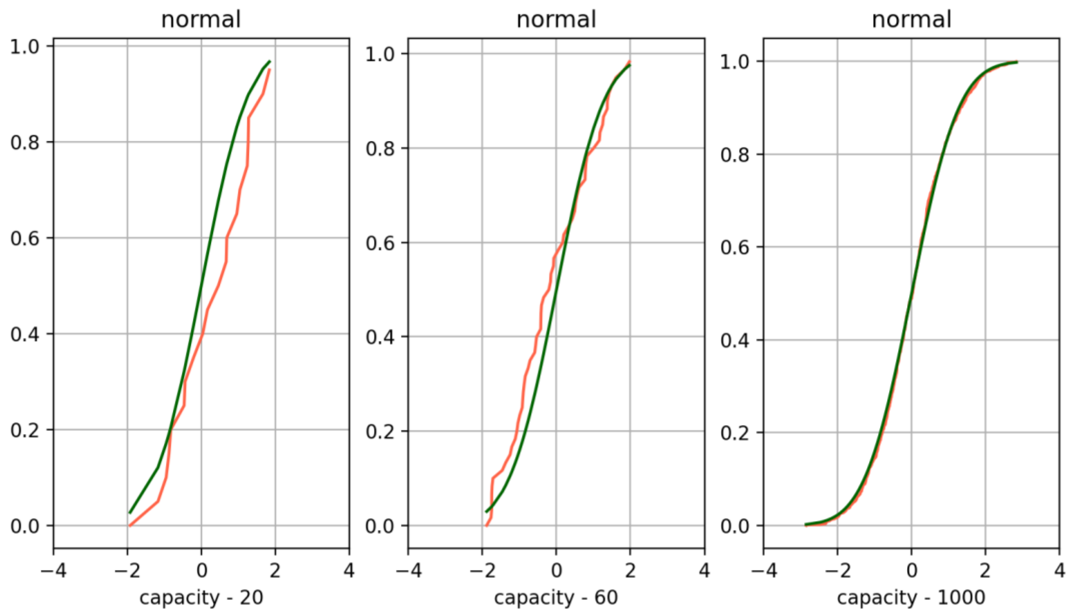

Рисунок 11: Нормальное распределение

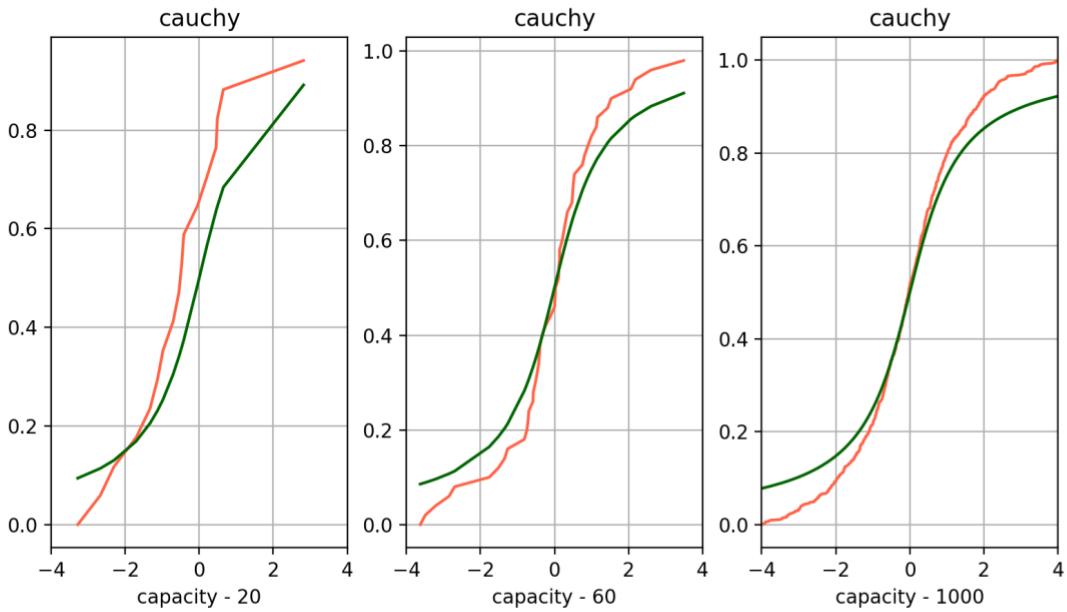

Рисунок 12: Распределение Коши

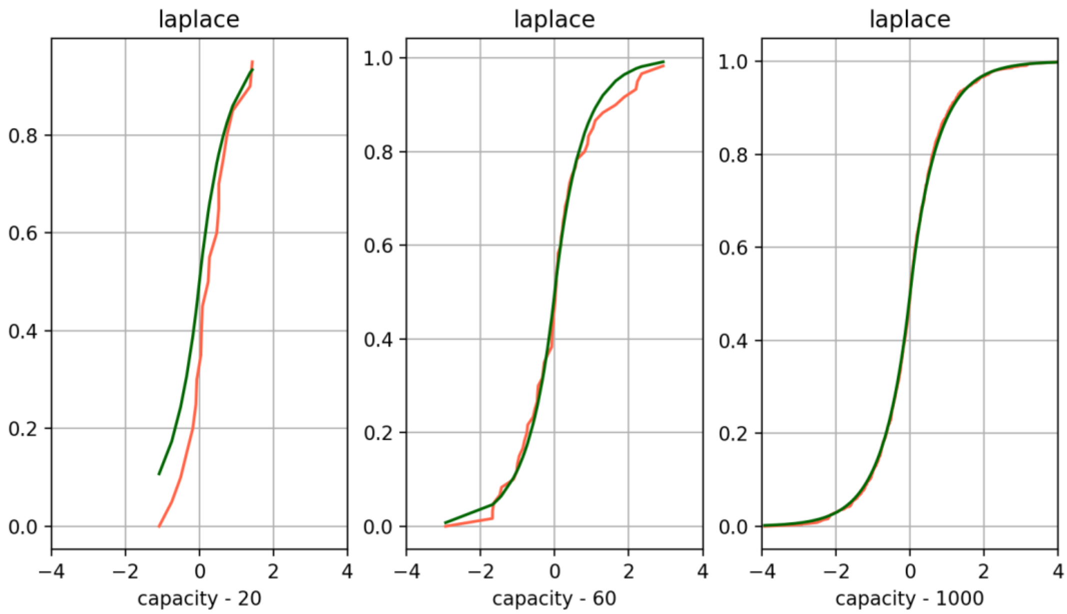

Рисунок 13: Распределение Лапласа

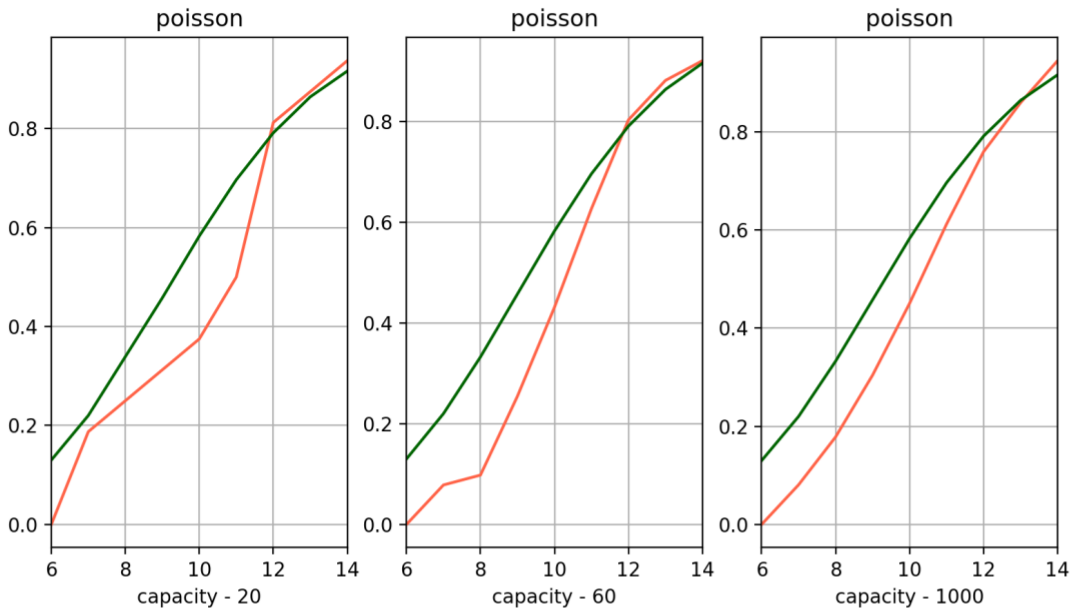

Рисунок 14: Распределение Пуассона

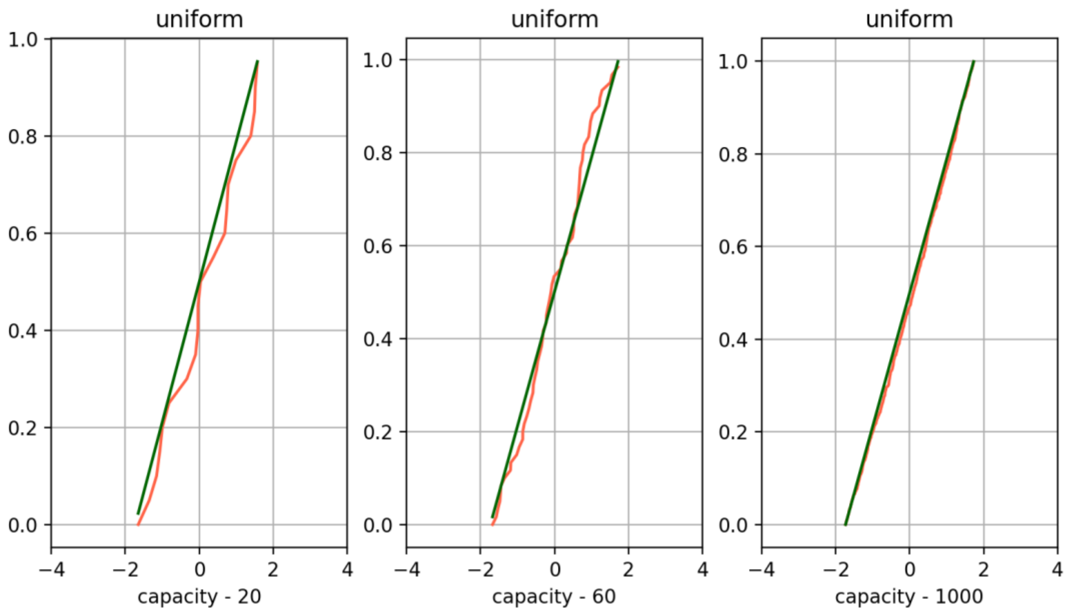

Рисунок 15: Равномерное распределение

6. Ядерные оценки плотности распределения

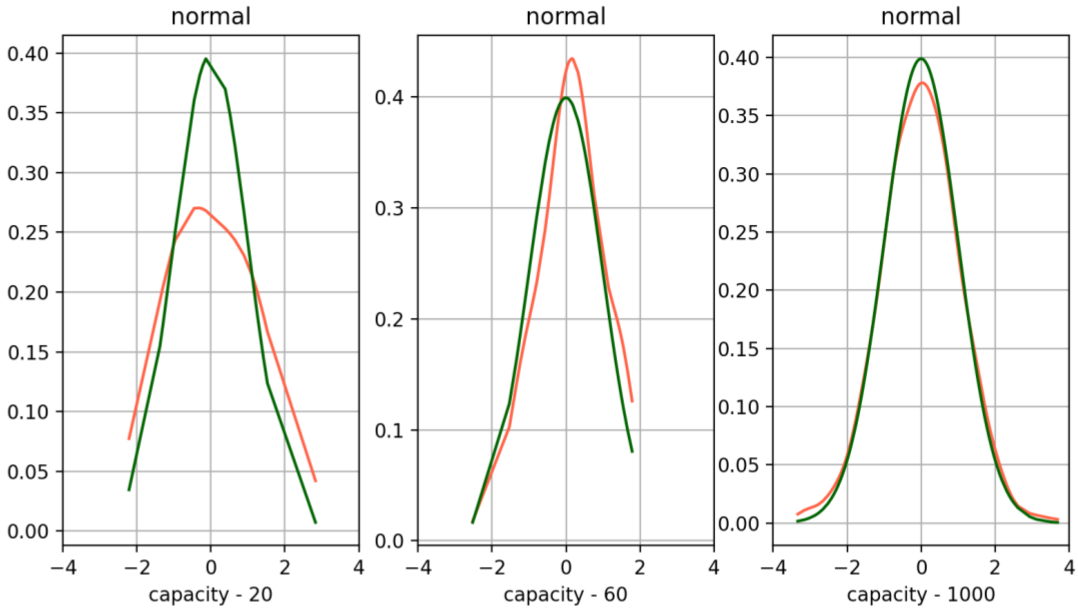

Рисунок 16: Нормальное распределение

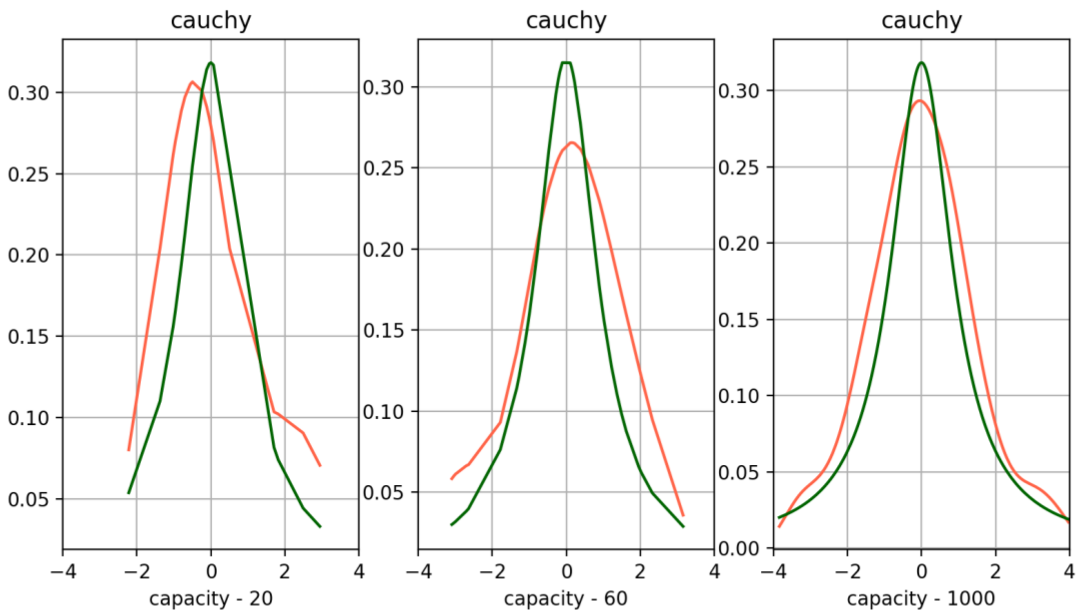

Рисунок 17: Распределение Коши

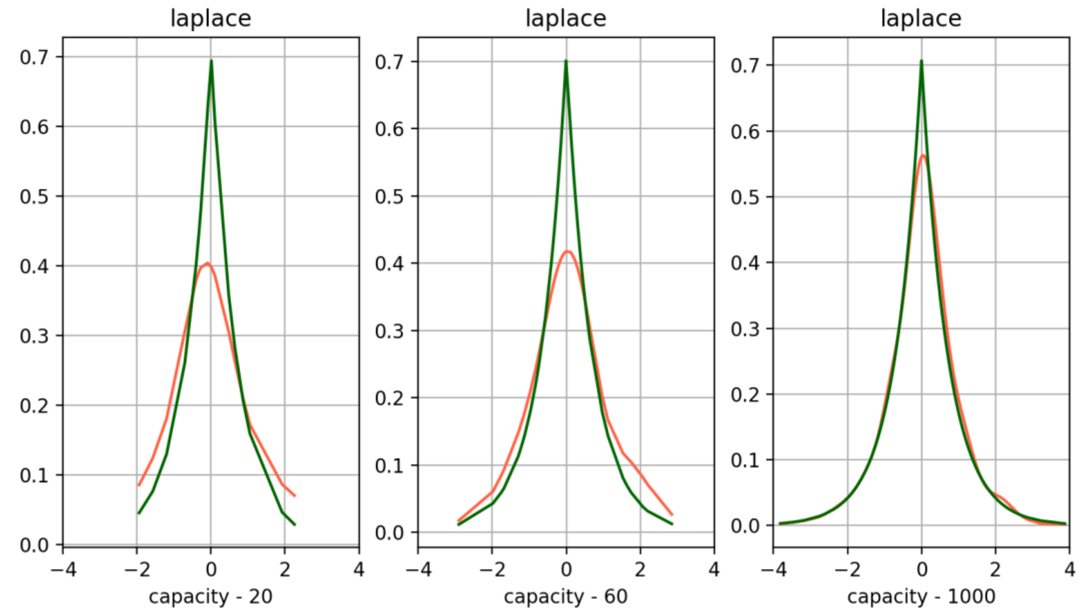

Рисунок 18: Распределение Лапласа

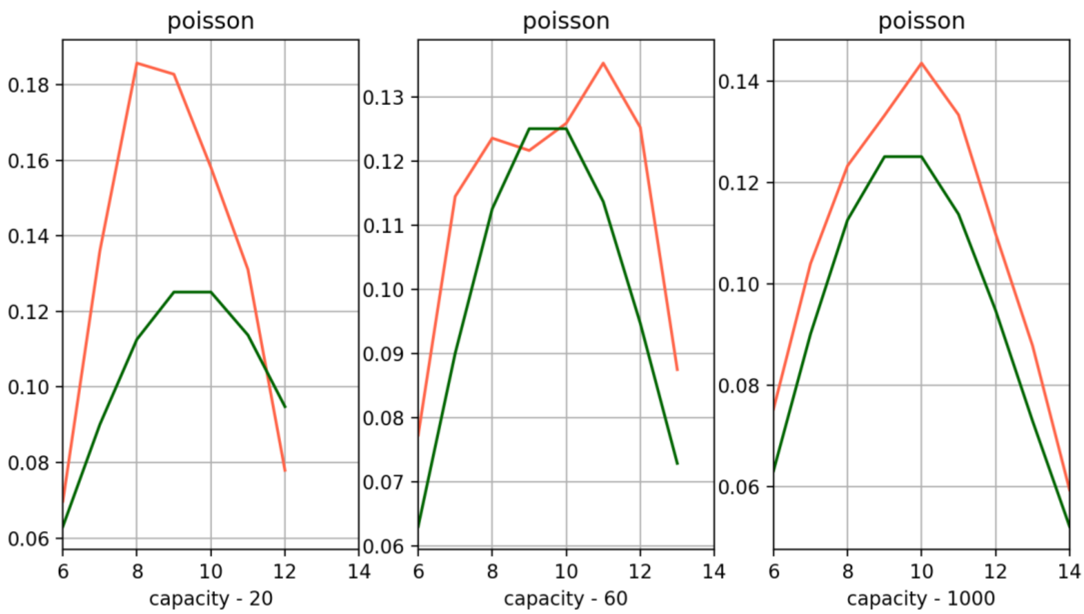

Рисунок 19: Распределение Пуассона

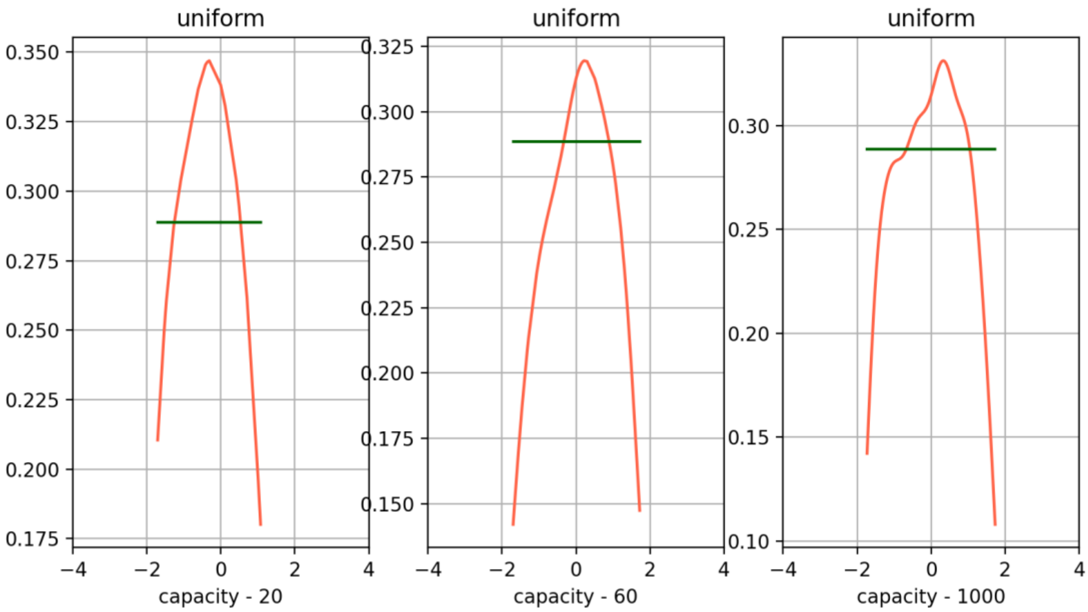

Рисунок 20: Равномерное распределение

<u>Обсуждение</u>

1. Лабораторная

Из полученных графиков можем подтвердить факт: если число интервалов гистограммы $k(n)$ устремить в бесконечности, таким образом, что $\lim_{n\to\infty}\frac{k(n)}{n}=0$, то имеет место сходимость по вероятности гистограммы к плотности.

2. Лабораторная

Исследование показало:

- действительно, математическое ожидание для распределения Коши не определено, а дисперсия – бесконечна;
- выборочное среднее при увеличении n стремится к математическому ожиданию;
- медиана у всех распределений определена;
- z_Q и z_R оценивают центр симметрии распределения

Упорядочение характеристик

- Нормального $\mathcal{N}(x,0,1)$: $\bar{x} < z_R < med\ x < z_Q < z_{tr}$
- Коши $\mathcal{C}(x,0,1)$: $med\ x < z_Q < \bar{x} < z_{tr} < z_R$
- Лапласа $\mathcal{L}(x,0,\ \frac{1}{\sqrt{2}})$: $\bar{x} < med\ x <\ z_{tr} < z_Q < z_R$
- Пуассона $\mathcal{P}\ (k,\ 10)$: $med\ x < z_Q < \bar{x} < z_{tr} < z_R$
- Равномерного $\mathcal{U}\left(x,-\sqrt{3},\sqrt{3}\right)$: $z_R < z_{tr} < \bar{x} < med\ x \le z_Q$

3. Лабораторная

4. Лабораторная

Ожидаемо, увеличение размера выборки улучшает приближение к теоретическим значениям. Так же видно, что равномерное распределение, в силу своей разрывности плохо приближается эмпирически.

Проведем визуальную оценку (см. ниже) для выбора h_n (сглаживающего параметра) для ядерной оценки. В случае распределения Лапласа мы имеем остроту на медиане, поэтому ${}^{h_n}/_2$ лучше приближает оценку. Для равномерного распределения оптимально само h_n. Возможно, это следует из того, что вне интервала никаких событий нет.

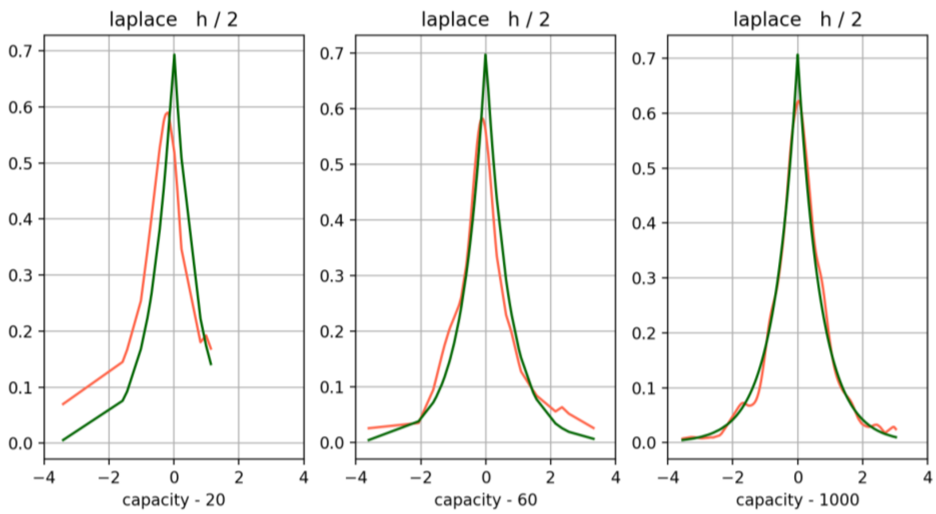

Рисунок 21: Распределение Лапласа (h/2)

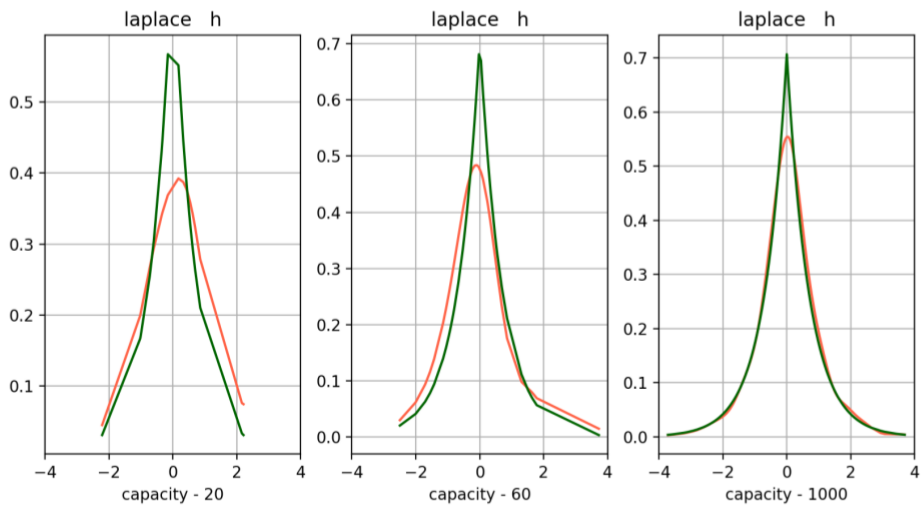

Рисунок 22: Распределение Лапласа (h)

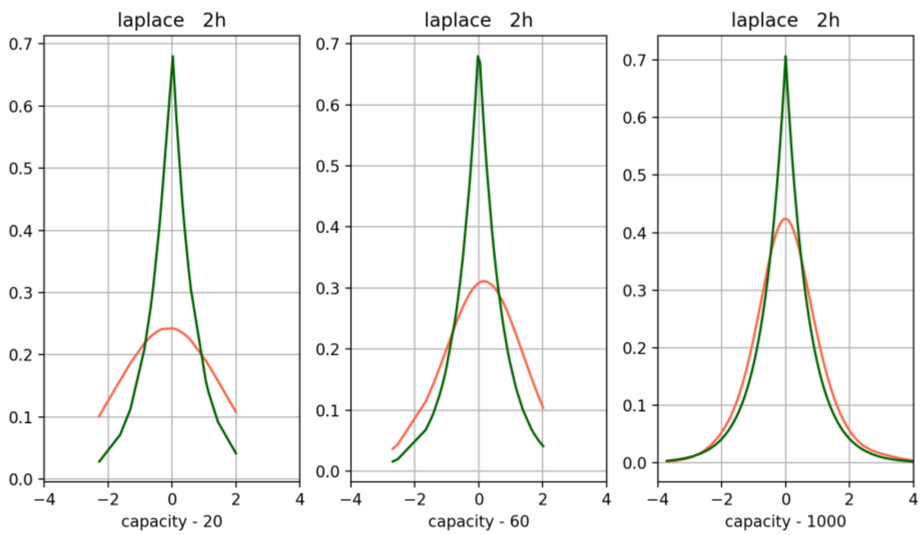

Рисунок 23: Распределение Лапласа (2h)

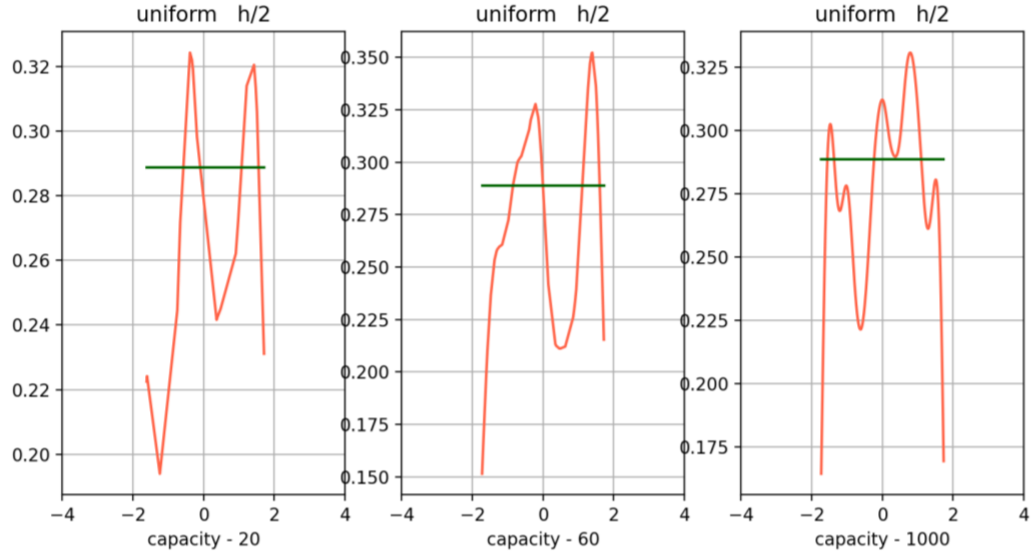

Рисунок 24: Равномерное распределение (h/2)

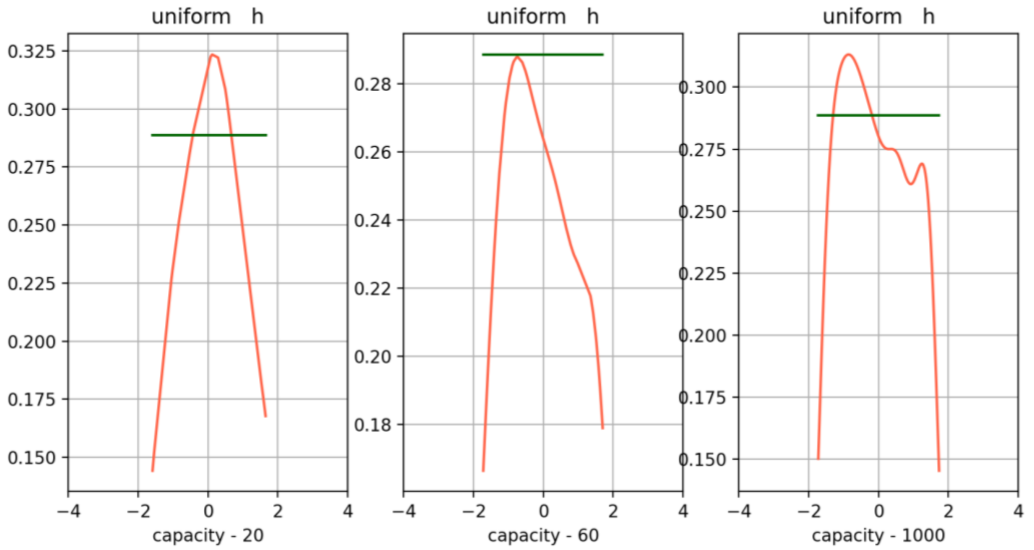

Рисунок 25: Равномерное распределение (h)

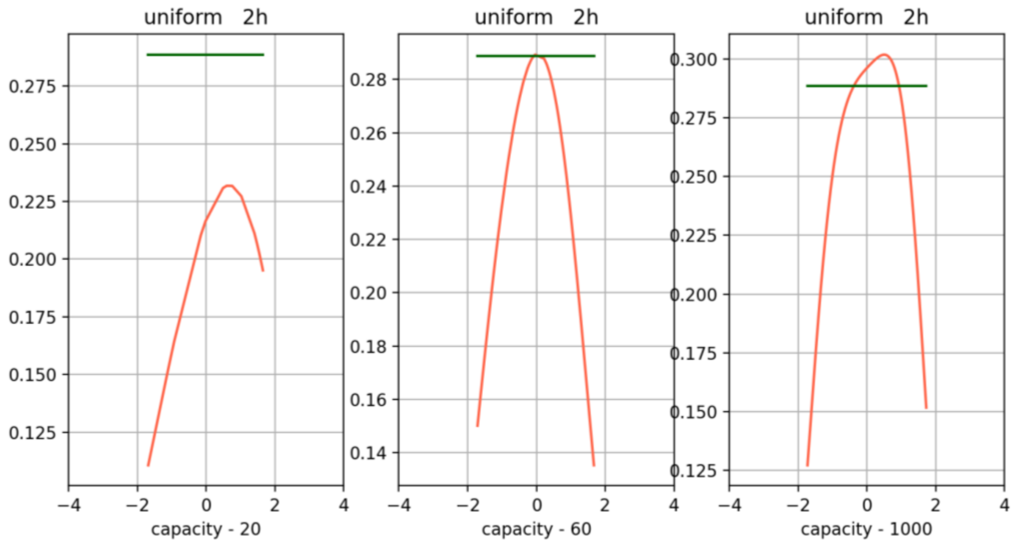

Рисунок 26: Равномерное распределение (2h)

Литература

1. Конспекты лекции
2. Википедия: https://ru.wikipedia.org/wiki

Ссылка на github: https://github.com/KateZabolotskih/MathStatLabs