Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра системного анализа

Практикум

«Стохастический анализ и моделирование»

Студент 415 группы
К. Ю. Егоров

Руководитель практикума
к.ф.-м.н., доцент С. Н. Смирнов

Москва, 2019

Содержание

1 Задание №1

1. Реализовать генератор схемы Бернулли с заданной вероятностью успеха p. На основе генератора схемы Бернулли построить датчик для биномиального распределения.

2. Реализовать генератор геометрического распределения. Проверить для данного распределения свойство отсутствия памяти.

3. Рассмотреть игру в орлянку — бесконечную последовательность испытаний с бросанием правильной монеты. Выигрыш S_n определяется как сумма по всем n испытаниям значений 1 и -1 в зависимости от выпавшей стороны. Проиллюстрировать (в виде ломаной) поведение нормированной суммы $Y(i) = \frac{S_i}{\sqrt{n}}$ как функции от номера испытания $i = 1, \ldots, n$ для одной отдельно взятой траектории. Дать теоретическую оценку для $Y(n)$ при $n \to \infty$.

1.1 Задача №1

Определение 1.1. *Схемой Бернулли* называется последовательность испытаний, в каждом из которых возможны два исхода — «успех» и «неудача», при этом «успех» в каждом испытании происхоит с одной и той же вероятностью $p \in (0,\, 1)$, а «неудача» — с вероятностью $q \equiv 1 - p$. На испытания в схеме Бернулли налагаются следующие требования: отсутствие взаимного влияния, воспроизводимость, а также сходные — но не идентичные — условия проведения.

Определение 1.2. Будем говорить, что случайная величина X имеет *распределение Бернулли*, если она принимает всего два значения: 1 и 0 с вероятностями p и $q \equiv 1 - p$ соответственно. Таким образом,

$$\mathbb{P}(X = 1) = p \qquad \text{и} \qquad \mathbb{P}(X = 0) = q,$$

то есть событие $\{\, X = 1 \,\}$ соответствует «успеху», а $\{\, X = 0 \,\}$ —«неудаче». Будем обозначать такую случайную величину

$$X \sim \mathrm{Bern}(p).$$

Реализуем генератор схемы Бернулли с заданной вероятностью успеха p следующим образом: пусть нам дана случайная величина ξ, равномерно распределённая на отрезке $[0,\, 1]$. Тогда случайная величина $X \sim \mathrm{Bern}(p)$ задаётся следующим образом:

$$X = \mathbb{I}(\xi < p) = \begin{cases} 1, & \text{при } 0 \leqslant \xi < p, \\ 0, & \text{при } p \leqslant \xi \leqslant 1. \end{cases}$$

Определение 1.3. Будем говорить, что случайная величина X имеет *биномиальное распределение* с параметрами n и p, если

$$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad \text{где } k \in \mathbb{N}_0.$$

В таком случае X интерпретируют как число «успехов» в серии из n испытаний схемы Бернулли с вероятностью успеха p. Будем обозначать такую случайную величину

$$X \sim \mathrm{Bi}(n,\, p).$$

Пусть теперь $X \sim \mathrm{Bi}(n,\, p)$, а $Y_i \sim \mathrm{Bern}(p)$, $i = \overline{1,\, n}$. Тогда, как видно из интерпретации биномиального распределения, датчик биномиальной случайной величины будет иметь вид:

$$X = \sum_{i=1}^{n} Y_i.$$

1.2 Задача №2

Определение 1.4. Будем говорить, что случайная величина X имеет *геометрическое распределение*, если

$$\mathbb{P}(X = k) = (1-p)^k p = q^k p, \quad \text{где } k \in \mathbb{N}_0.$$

Геометрически распределенная случайная величина интерпретируется как количество «неудач» до первого «успеха» в схеме испытаний Бернулли с вероятностью p. Будем обозначать такие случайные величины

$$X \sim \mathrm{Geom}(p).$$

Зная интерпретацию, мы легко строим датчик и для геометрического распределения.

Утверждение 1.1 (Свойство отсутствия памяти)**.** *Пусть $X \sim Geom(p)$, тогда для любых $n,\, m \in \mathbb{N}_0$ справедливо*

$$\mathbb{P}(X \geqslant m + n \mid X \geqslant m) = \mathbb{P}(X \geqslant n),$$

то есть количество прошлых «неудач» не влияет на количество будущих «неудач».

Д о к а з а т е л ь с т в о. Рассмотрим левую часть равенства из условия утверждения:

$$\mathbb{P}(X \geqslant m+n \mid X \geqslant m) = \frac{\mathbb{P}(X \geqslant m+n,\, X \geqslant m)}{\mathbb{P}(X \geqslant m)} = \\ = \frac{\mathbb{P}(X \geqslant m+n)}{\mathbb{P}(X \geqslant m)} = \frac{\sum_{i=m+n}^{\infty} q^i p}{\sum_{i=m}^{\infty} q^i p} = \frac{q^{m+n}}{q^m} = q^n.$$

Теперь рассмотрим правую часть равенства:

$$\mathbb{P}(X \geqslant n) = \sum_{i=n}^{\infty} q^i p = p \frac{q^n}{1-q} = q^n.$$

Таким образом, утверждение доказано. ■

1.3 Задача №3

Рассмотрим игру Орлянка, правила которой описаны в формулировке задания и построим траекторию заданного процесса $Y(i)$.

В данной нормированной сумме Y фигурируют независимые одинаково распределенные случйные величины X_i. Посчитаем их математическое ожидание и дисперсию.

$$\mathbb{E}\, X_i = 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0,$$
$$\mathbb{V}\mathrm{ar}\, X_i = \frac{1}{2} \cdot (1-0)^2 + \frac{1}{2} \cdot (-1-0)^2 = \frac{1}{4}.$$

Теперь можем воспользоваться следующей теоремой.

Теорема 1.1 (Центральная предельная теорема)**.** *Пусть* $X_1, \ldots, X_n, \ldots$ *есть бесконечная последовательность независимых одинаково распределенных случайных величин, имеющих конечное математическое ожидание* μ *и дисперсию* σ^2*. Пусть также* $S_n = \sum_{i=1}^{n} X_i$*. Тогда*

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \longrightarrow N(0,\, 1)$$

по распределению при $n \to \infty$*.*

Получается, что

$$2 \lim_{n\to\infty} Y(n) \xrightarrow{dist.} N(0,\, 1).$$

Для оценки этого значения воспользуемся «правилом трёх сигм».

Теорема 1.2. *Практически все значения нормально распределённой случайной величины* $\xi \sim N(\mu,\, \sigma^2)$ *лежат в интервале* $(\mu - 3\sigma,\, \mu + 3\sigma)$*. Более строго — приблизительно с вероятностью* $0,9973$ *значение нормально распределённой случайной величины лежит в указанном интервале.*

Таким образом, приблизительно с вероятностью $0,9973$

$$-\frac{3}{2} \leqslant \lim_{n\to\infty} Y(n) \leqslant \frac{3}{2}.$$

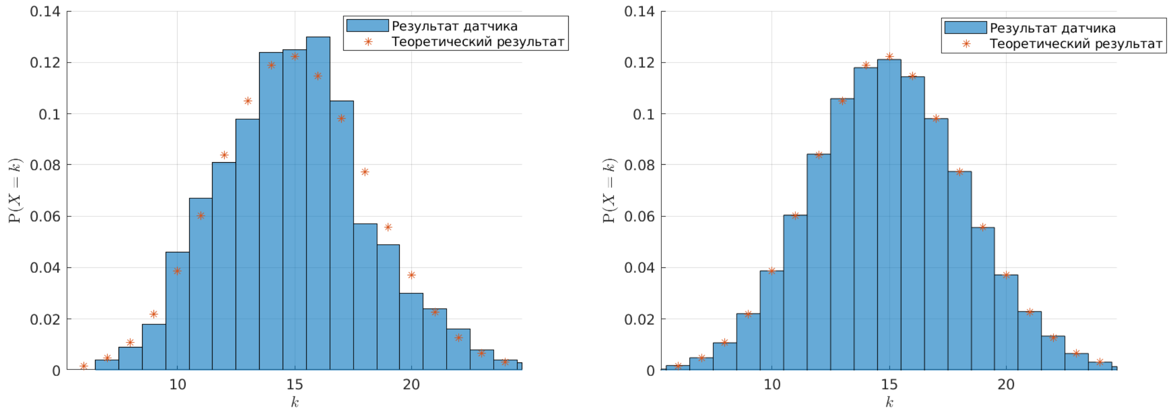

Рис. 1.1: Гистограмма биномиального распределения случайной величины с параметрами $n = 50$, $p = \frac{3}{10}$ при 10^3 (слева) и 10^5 (справа) испытаний.

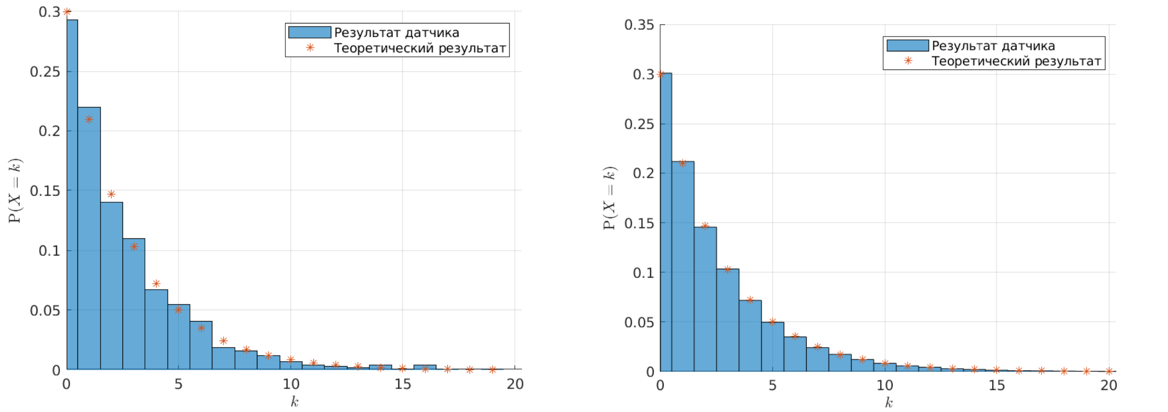

Рис. 1.2: Гистограмма геометрического распределения случайной величины с параметром $p = \frac{3}{10}$ при 10^3 (слева) и 10^5 (справа) испытаний.

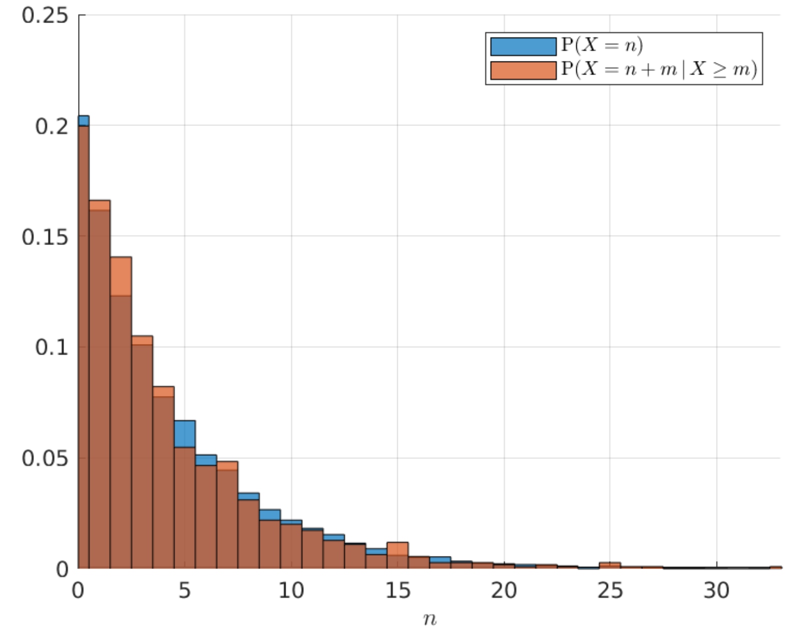

Рис. 1.3: Гистограмма геометрического распределения, демонстрирующая его свойство отсутствия памяти. Здесь задан параметр геометрического распределения $p = \frac{2}{10}$, а также «сдвиг» $m = 10$.

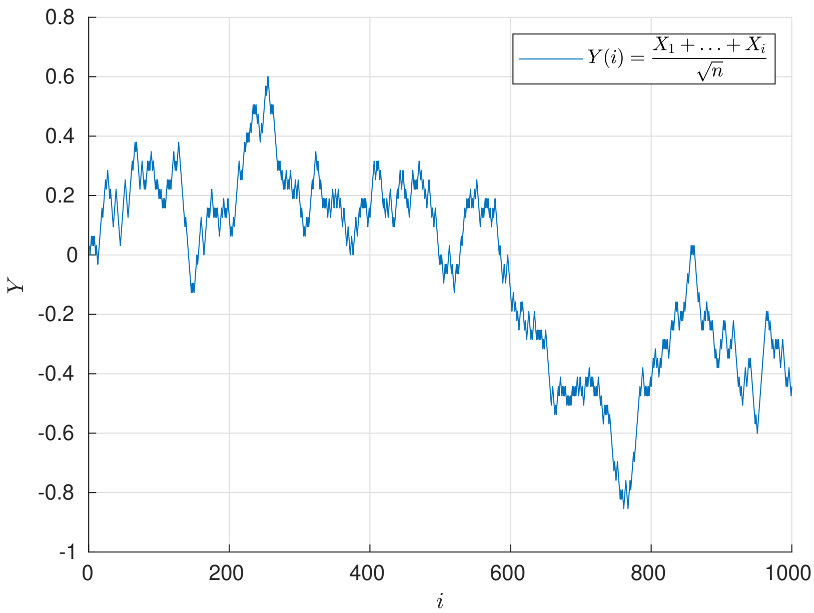

Рис. 1.4: Иллюстрация варианта поведения нормированной суммы $Y(i)$ игры в орлянку на отрезке $1 \leqslant i \leqslant 10^3$.

2 Задание №2

1. Построить датчик сингулярного распределения, имеющий в качестве функции распределения канторову лестницу. С помощью критерия Колмогорова убедиться в корректности работы датчика.

2. Для канторовых случайных величин проверить свойство симметричности относительно $\frac{1}{2}$ (X и $(1-X)$ распределены одинаково) и самоподобия относительно деления на 3 (условное распределение Y при условии $Y \in [0,\, \frac{1}{3}]$ совпадает с распределением $\frac{Y}{3}$) с помощью критерия Смирнова.

3. Вычислить значение математического ожидания и дисперсии с эмпирическими для разного объема выборок. Проиллюстрировать сходимость.

2.1 Построение датчика «канторовой» случайной величины

Определение 2.1. Пусть дано вероятностное пространство $(\mathbb{R},\, \mathcal{F},\, \mathbb{P})$, и на нем определена случайная величина ξ с распределением $\mathbb{P}_\xi$. Тогда *функцией распределения* случайной величины X называется функция $F_\xi : \mathbb{R} \to [0,\, 1]$, задаваемая формулой:

$$F_\xi(x) = \mathbb{P}(\xi \leqslant x) \equiv \mathbb{P}_\xi(\,(-\infty,\, x]\,).$$

Определение 2.2. Функция распределения некоторой случайной величины называется *сингулярной*, если она непрерывна и ее множество точек роста имеет нулевую меру Лебега.

Определение 2.3. Из единичного отрезка $C_0 = [0,\, 1]$ удалим интервал $\left(\frac{1}{3},\, \frac{2}{3}\right)$. Оставшееся множество обозначим за C_1. Множество $C_1 = \left[0,\, \frac{1}{3}\right] \cup \left[\frac{2}{3},\, 1\right]$ состоит из двух отрезков; удалим теперь из каждого отрезка его среднюю часть, и оставшееся множество обозначим за C_2. Повторив данную процедуру, то есть удаляя средние трети у всех четырех отрезков, получим C_3. Дальше таким же образом получаем последовательность замкнутых множеств $C_0 \supset C_1 \supset C_2 \supset \ldots \supset C_i \supset \ldots$. Пересечение

$$C = \bigcap_{i=0}^{\infty} C_i$$

называется *канторовым множеством.*

Замечание 2.1. Канторово множество так же можно определить как множество всех чисел от нуля до единицы, которые можно представить в троичной записи при помощи только нулей и двоек. То есть

$$C = \{\, 0,\alpha_1\,\alpha_2\, \ldots\, \alpha_i\, \ldots\,{}_3 \mid \alpha_i = 0,\, 2\,\}.$$

Утверждение 2.1. *Канторово множество имеет нулевую меру Лебега. [1]*

Определение 2.4. Рассмотрим функцию $K(x)$ такую, что в точках 0 и 1 значение функции принимается равным соответственно 0 и 1. Далее интервал $(0,\, 1)$ разбивается на три равные части $\left(0,\, \frac{1}{3}\right)$, $\left(\frac{1}{3},\, \frac{2}{3}\right)$ и $\left(\frac{2}{3},\, 1\right)$. На среднем сегменте полагаем $K(x) = \frac{1}{2}$. Оставшиеся два сегмента снова разбиваются на три равные части каждый, и на средних сегментах $K(x)$ полагается равной $\frac{1}{4}$ и $\frac{3}{4}$. Каждый из оставшихся сегментов снова делится на три части, и на внутренних сегментах $K(x)$ определяется как постоянная, равная среднему арифметическому между соседними, уже определенными значениями $K(x)$. На остальных точках единичного отрезка определяется по непрерывности. Полученная функция называется *канторовой лестницей.*

Замечание 2.2. Из определения канторовой лестницы $K(x)$ следует, что она действует на точки из канторова множества C по следующему правилу:

$$K(0,\alpha_1\,\alpha_2\,\dots\,\alpha_i\,\dots\,_3) = 0,\frac{\alpha_1}{2}\,\frac{\alpha_2}{2}\,\dots\,\frac{\alpha_i}{2}\,\dots\,_2.$$

Теперь рассмотрим случайную величину

$$Y = 0,\xi_1\,\xi_2\dots\xi_k\,\dots\,_2 = \sum_{k=1}^{\infty}\frac{\xi_k}{2^k},\quad \text{где } \xi_k \sim \mathrm{Bern}\left(\tfrac{1}{2}\right).$$

Такая случайная величина имеет равномерное распределение на откезке $[0,1]$, так как мы равновероятным образом выбираем знаки разложения числа ξ_k в двоичном представлении. Теперь рассмотрим искомую случайную величину X, имеющую в качестве функции распределения $F_X(x)$ канторову лестницу $K(x)$. Образ каждой случайной величины Y для такой функции будет равен

$$K^{-1}(Y) = \sum_{k=1}^{\infty}\frac{2\xi_k}{3^k}.$$

Эта точка лежит в канторовом множестве.

Теорема 2.1 (Метод обратной функции распределения)**.** *Пусть некоторая функция распределения F имеет обратную F^{-1}. Тогда функцией распределения случайной величины*

$$\eta = F^{-1}(\xi),$$

где ξ — равномерно распределенная на отрезке $[0,\,1]$ *случайная величина, является F.*

Доказательство. Найдем функцию распределения случайной величины η:

$$F_\eta(x) = \mathbb{P}(\eta \leqslant x) = \mathbb{P}(F^{-1}(\xi) \leqslant x) = \mathbb{P}(\xi \leqslant F(x)) = F(x).$$

Таким образом, теорема доказана. ■

Из теоремы вытекает, что при помощи построенного ранее (см. раздел 1) генератора схемы Бернулли мы можем смоделировать случайную величину X, принимающую с вероятностью 1 значения из канторова множества C и имеющую канторову лестницу $K(x)$ в качестве функции распределения $F_X(x)$, следующим образом:

$$X = \sum_{k=1}^{\infty}\frac{2\xi_k}{3^k},\quad \text{где } \xi_k \sim \mathrm{Bern}\left(\tfrac{1}{2}\right).$$

В программной реализации будем рассматривать частичные суммы. Для этого этого введем погрешность ε и найдем такое число n, при котором частичная сумма будет отличаться от бесконечной не более, чем на заданную погрешность.

$$\sum_{k=n}^{\infty}\frac{2\xi_k}{3^k} \leqslant 2\sum_{k=n}^{\infty}\frac{1}{3^k} = \frac{1}{3^{n-1}} \leqslant \varepsilon,$$

$$\Downarrow$$

$$n \geqslant 1 - \lceil \log_3 \varepsilon \rceil \ \forall \varepsilon < 1.$$

Замечание 2.3. Из выведенной формулы также видно, что для столь малой погрешности как $\varepsilon = 10^{-9}$ достаточно использовать всего $n = 20$ первых членов ряда.

2.2 Проверка корректности работы датчика

Определение 2.5. Пусть задана выборка $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$, где $\xi_i \in \Xi, i = \overline{1, n}$. *Эмпирической (выборочной) функцией распределения*, построенной на этой выборке, называется функция $F_n(x)$, равная доле таких значений ξ_i, что $\xi_i < x$. Или другими словами

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,\, x)}(\xi_i), \quad \text{где } \mathbb{I}_{(-\infty,\, x)}(\xi_i) = \begin{cases} 1, & \text{при } \xi_i < x, \\ 0, & \text{иначе.} \end{cases}$$

Определение 2.6. Пусть в некотором эксперименте доступна наблюдению случайная величина ξ, распределение которой $\mathbb{P}$ полностью или частично неизвестно. Тогда любое утверждение относительно $\mathbb{P}$ называется *статистической гипотезой* H.

Теорема 2.2 (Критерий согласия Колмогорова)**.** *Обозначим нулевую гипотезу H_0 как гипотезу о том, что выборка подчиняется распределению $F(\xi) \in C^1(\Xi)$. Введем статистику критерия*

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Тогда если гипотеза H_0 верна, то $\sqrt{n}D_n$ с ростом n сходится по распределению к случайной величине K с функцией распределения Колмгорова

$$F_K(x) = 1 + 2\sum_{i=1}^{+\infty}(-1)^k e^{-2k^2x^2}.$$

Замечание 2.4. Гипотеза H_0 отвергается, если при большом объеме выборки n статистика $\sqrt{n}D_n$ превышает квантиль распределения K_α, заданного *уровня значимости* α, и принимается в противном случае. Здесь $K_\alpha = F_K^{-1}(1 - \alpha)$.

В рамках реализации подсчет K_α в явном виде весьма трудоемкая задача. Поэтому мы будем расчитывать *p-значение* для нашей статистики $p_{value} = 1 - F_K(\sqrt{n}D_n)$. Если p-значение оказалось ниже или равно установленному уровню значимости α, то наша гипотеза отвергается и применяется альтернативная.

2.3 Симметричность «канторовой» случайной величины относительно $^1/_2$

Утверждение 2.2 (Свойство симметричности относительно $^1/_2$)**.** *Пусть X — случайная величина, с канторовой лестницей в качестве функции распределения. Тогда верно*

$$F_X(x) = F_{1-X}(x).$$

Число испытаний	Размер выборки	Уровень значимости	Частота принятия гипотезы
10^3	10^3	$0,05$	$0,958$
10^4	10^3	$0,05$	$0,9536$
10^3	10^4	$0,05$	$0,959$
10^3	10^3	$0,1$	$0,907$

Таблица 1: Частота принятия гипотезы H_0 о том, что построенный датчик случайной величины имеет канторову лестницу в качестве функции распределения. Из таблицы видно, что вероятность отклонить гипотезу не превышает допустимый уровень значимости.

Д о к а з а т е л ь с т в о. Как мы помним из построения датчика «канторовой» случайной величины, она представима в виде:

$$X = 2\sum_{k=1}^{\infty}\frac{\xi_k}{3^k}, \quad \text{где } \xi_k \sim \mathrm{Bern}\left(1/2\right), k = \overline{1, n}.$$

Тогда

$$1 - X = 1 - 2\sum_{k=1}^{\infty}\frac{\xi_k}{3^k} = 2\sum_{k=1}^{\infty}\frac{1}{3^k} - 2\sum_{k=1}^{\infty}\frac{\xi_k}{3^k} = 2\sum_{k=1}^{\infty}\frac{1-\xi_k}{3^k} = 2\sum_{k=1}^{\infty}\frac{\eta_k}{3^k},$$

при этом случайная величины $\eta_k = 1 - \xi_k$ также имеют распределение Бернулли с параметром $p = 1/2$. Значит, случайные величины X и $1-X$ имеют одинаковое распределение. ■

Теперь эмперическим путем убедимся в выполнении этого свойства. Для этого нам потребуется *критерий Смирнова*. Он используется для проверки гипотезы о принадлежности двух независимых выборок одному и тому же распределению.

Теорема 2.3 (Критерий однородности Смирнова)**.** *Обозначим за нулевую гипотезу H_0 гипотезу о том, что две исследуемые выборки объемами n и m с эмпирическими функциями распределения $F_n(x)$ и $F_m(x)$ соответственно распределениы по одному закону. Введем статистику критерия*

$$D_{n,m} = \sup_x |F_n(x) - F_m(x)|.$$

Тогда если гипотеза H_0 верна, то при увеличении объемов выборок n и m случайная величина $\sqrt{\frac{nm}{n+m}}D_{n,m}$ будет сходиться по распределению к случайной величине K с функцией распределения Колмогорова

$$F_K(x) = 1 + 2\sum_{i=1}^{+\infty}(-1)^k e^{-2k^2x^2}.$$

2.4 Самоподобие «канторовой случайной величины относительно деления на 3

Утверждение 2.3 (Свойство самоподобия относительно деления на 3)**.** *Пусть X — случайная величина с канторовой лестницей в качестве функции распределения. Тогда верно*

$$F_{X/3}(x) = F_{X \mid X\in[0,\,1/3]}(x).$$

Д о к а з а т е л ь с т в о. Заметим, что из построения датчика «канторовой» случайной величины вытекает, что случайная величина $Y = X \mid X \in [0, 1/3]$ задается в виде

$$X = 2\sum_{k=1}^{\infty} \frac{\xi_k}{3^k}, \quad \text{где } \xi_1 = 0,\, \xi_k \sim \mathrm{Bern}(1/2), k = \overline{2, n}.$$

Тогда получается, что

$$Y = 2\sum_{k=2}^{\infty} \frac{\xi_k}{3^k} = 2\sum_{k=1}^{\infty} \frac{\xi_{k+1}}{3^{k+1}} = \frac{2}{3}\sum_{k=1}^{\infty} \frac{\xi_k}{3^k} = \frac{X}{3}.$$

■

Так же проверим получившийся результат при помощи критерия однородности Смирнова.

Число испытаний	10^3	10^3	10^3
Размер первой выборки	10^2	10^3	10^3
Размер второй выборки	10^3	10^3	10^3
Уровень значимости	$0,05$	$0,05$	$0,01$
Частота принятия гипотезы	$0,956$	$0,955$	$0,912$

Таблица 2: Частота принятия гипотезы H_0 о том, что случайные величины X и $(1 - X)$ имеют одинаковое распределение.

Число испытаний	10^3	10^3	10^3
Размер первой выборки	10^2	10^3	10^3
Размер второй выборки	10^3	10^3	10^3
Уровень значимости	$0,05$	$0,05$	$0,01$
Частота принятия гипотезы	$0,956$	$0,955$	$0,915$

Таблица 3: Частота принятия гипотезы H_0 о том, что случайные величины $X \mid X \in [0, 1/3]$ и $X/3$ имеют одинаковое распределение.

2.5 Математическое ожидание и дисперсия «канторовой» случайной величины

Вычислим значение математического ожидания для построенной случайной величины X:

$$\mathbb{E}\, X = \mathbb{E}\sum_{k=1}^{\infty} \frac{2\xi_k}{3^k} = \sum_{k=1}^{\infty} \frac{2}{3^k}\mathbb{E}\xi_k = \sum_{k=1}^{\infty} \frac{2}{3^k} \cdot \frac{1}{2} = \frac{\frac{1}{3}}{1 - \frac{1}{3}} = \frac{1}{2}.$$

Теперь, помня о независимости случайных величин ξ_k $k \in \mathbb{N}$, вычислим значение дисперсии

$$\mathbb{V}\mathrm{ar}\, X = \mathbb{V}\mathrm{ar}\sum_{k=1}^{\infty} \frac{2\xi_k}{3^k} = \sum_{k=1}^{\infty} \left(\frac{2}{3^k}\right)^2 \mathbb{V}\mathrm{ar}\, \xi_k = \sum_{k=1}^{\infty} \frac{4}{9^k} \cdot \frac{1}{4} = \frac{\frac{1}{9}}{1 - \frac{1}{9}} = \frac{1}{8}.$$

Замечание 2.5. При подсчете мы использовали известные значения для математического ожидания и дисперсии бернуллиевой случайной величины $\xi \sim \mathrm{Bern}(p)$:

$$\mathbb{E}\,\xi = p \quad \text{и} \quad \mathbb{V}\mathrm{ar}\,\xi = p(1-p).$$

Для сравнения практического и теоретического результатов построим также графики *выборочного среднего* $\overline{X}$ и *несмещенной выборочной дисперсии* S^2, задаваемых формулами:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{и} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

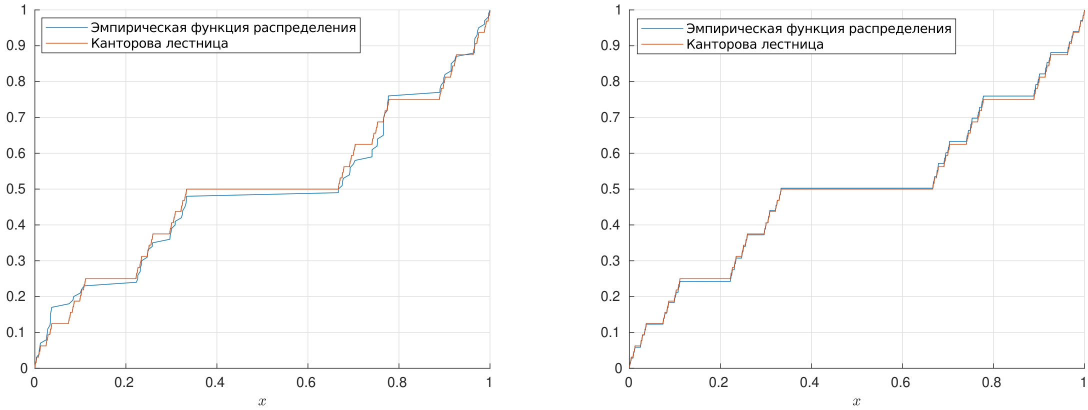

Рис. 2.1: Эмпирическая и теоретическая функции распределения «канторовой» случайной величины X при выборке из 100 испытаний (слева) и 10^4 испытаний (справа).

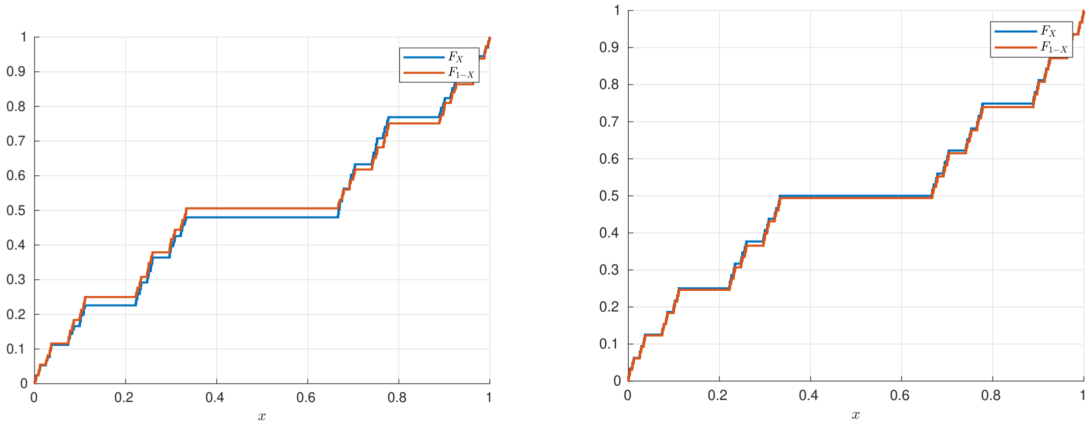

Рис. 2.2: График, иллюстрирующий свойство симметричности относительно $^1/_2$ «канторовой» случайной величины. Представлены эмперические функции распределения при объемах выборок 10^3 (слева) и 10^4 (справа).

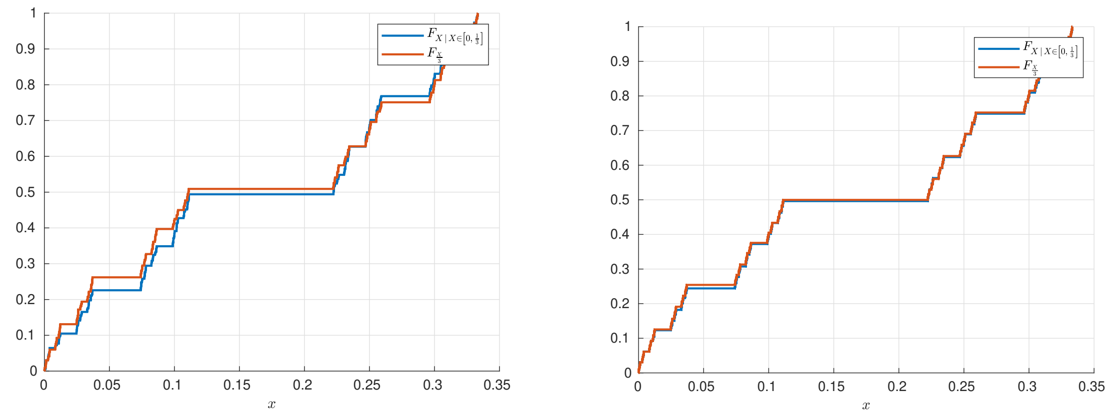

Рис. 2.3: График, иллюстрирующий свойство самоподобия относительно деления на 3 «канторовой» случайной величины. Представлены эмпирические функции распределения при объемах выборок 10^3 (слева) и 10^4 (справа).

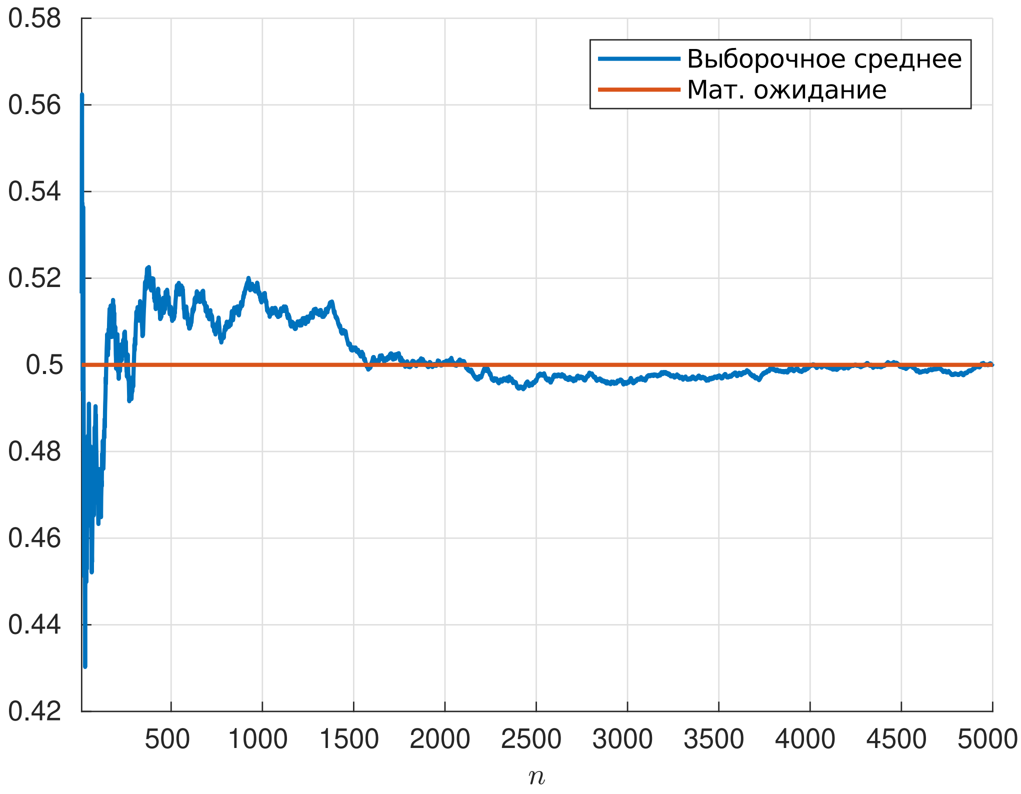

Рис. 2.4: Эмпирическое значение математического ожидания «канторовой» случайной величины X в зависимости от объема выборки $10 \leqslant n \leqslant 5000$.

Рис. 2.5: Несмещенная выборочная дисперсия и теоретическая дисперсия «канторовой» случайной величины X в зависимости от объема выборки $100 \leqslant n \leqslant 10000$.

3 Задание №3

1. Построить датчик экспоненциального распределения. Проверить для данного распределения свойство отсутствия памяти. Пусть $X_1, X_2, \ldots, X_n$ независимо распределенные случайные величины с параметрами $\lambda_1, \lambda_2, \ldots, \lambda_n$ соответственно. Найти распределение случайной величины $Y = \min\{ X_1, X_2, \ldots, X_n \}$.

2. На основе датчика экспоненциального распределения построить датчик пуассоновского распределения.

3. Построить датчик пуассоновского распределения как предел биномеального распределения. С помощью критерия хи-квадрат Пирсона убедиться, что получен датчик распределения Пуассона.

4. Построить датчик стандартного распределения методом моделирования случайных величин парами с переходом в полярные координаты. Проверить при помощи t-критерия Стьюдента равенство математических ожиданий, а при помощи критерия Фишера — равенство дисперсий.

3.1 Построение датчика экспоненциальной случайной величины

Определение 3.1. Будем говорить, что случайная величина X *имеет экспоненциальное распределение с параметром* $\lambda > 0$, если ее функция распределения имеет вид

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{при } x \geqslant 0, \\ 0, & \text{при } x < 0. \end{cases}$$

Будем обозначать такие случайные величины

$$X \sim \mathrm{Exp}(\lambda).$$

Для того чтобы построить датчик экспоненциально распределенной с параметром λ случайной величины X, воспользуемся методом обратной функции распределения (смотри теорему 2.1). Получается, что такую случайную величину можно представить в виде:

$$X = F_x^{-1}(\xi) = -\frac{1}{\lambda} \ln(1 - \xi),$$

где ξ — равномерно распределенная на отрезке $[0, 1]$ случайная величина.

3.2 Отсутствие памяти у экспоненциального распределения

Утверждение 3.1 (Свойство отсутствия памяти)**.** *Пусть $X \sim Exp(\lambda)$, тогда для любых $t \neq 0$ и s справедливо:*

$$\mathbb{P}(X \geqslant s + t \,|\, X \geqslant t) = \mathbb{P}(X \geqslant s).$$

Д о к а з а т е л ь с т в о. Рассмотрим левую часть равенства:

$$\mathbb{P}(X \geqslant s + t \,|\, X \geqslant t) = \frac{\mathbb{P}(X \geqslant s + t,\, X \geqslant t)}{\mathbb{P}(X \geqslant t)} = \frac{\mathbb{P}(X \geqslant s + t)}{\mathbb{P}(X \geqslant t)}.$$

Таким образом получаем, утверждение эквивалентно тому, что

$$\mathbb{P}(X \geqslant s+t) = \mathbb{P}(X \geqslant t)\mathbb{P}(X \geqslant s).$$

Из определения функции распределения $F_X(t) = \mathbb{P}(X < t) = 1 - \mathbb{P}(X \geqslant t)$ получаем, что

$$e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t}.$$

Последнее равенство точно верно. Таким образом, утверждение доказано. ■

3.3 Распределение минимума экспоненциальных случайных величин

Рассмотрим теперь случайную величину $Y = \min\{\, X_1,\, X_2,\, \ldots,\, X_n \,\}$, где $X_i, i = \overline{1,\, n}$ есть независимо распределенные экспоненциальные случайные величины с параметрами $\lambda_1,\, \lambda_2,\, \ldots,\, \lambda_n$ соответственно, и найдем её функцию распределения:

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y < x) = \mathbb{P}(\min_{i=\overline{1,\, n}} X_i < x) = 1 - \mathbb{P}(\min_{i=\overline{1,\, n}} X_i \geqslant x) = \\ &= 1 - \mathbb{P}(X_1 \geqslant x,\, X_2 \geqslant x,\, \ldots,\, X_n \geqslant x) = 1 - \prod_{i=1}^{n} \mathbb{P}(X_i \geqslant x) = 1 - \prod_{i=1}^{n} (1 - F_{X_i}(x)). \end{aligned}$$

Таким образом функция распределения случайной величины Y представима в виде

$$F_Y(x) = 1 - \prod_{i=1}^{n} (1 - 1 + e^{\lambda_i x}) = 1 - \prod_{i=1}^{n} e^{\lambda_i x} = 1 - e^{(\lambda_1 + \lambda_2 + \ldots + \lambda_n)x}.$$

Получается, что заданная случайная величина Y имеет экспоненциальное распределение с параметром $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

3.4 Построение датчика распределения Пуассона

Определение 3.2. Будем говорить, что случайная величина X *имеет распределение Пуассона с параметром* $\lambda > 0$, если

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{где } k \in \mathbb{N}_0.$$

Будем обозначать такие случайные величины

$$X \sim \mathrm{Pois}(\lambda).$$

Для построения датчика Пуассоновской случайной величины докажем вспомогательное утверждение.

Утверждение 3.2 (О распределении суммы экспоненциальных случайных величин)**.** *Пусть* $S_n = \sum_{i=1}^{n} \xi_i$, *где* $\xi_i \sim Exp(\lambda)$, $i = \overline{1,\, n}$. *Тогда*

$$F_{S_n}(x) = 1 - e^{-\lambda x} \sum_{k=0}^{n-1} \frac{\lambda^k x^k}{k!}.$$

Д о к а з а т е л ь с т в о. Для доказательства достаточно показать, что случайная величина S_n имеет плотность распределения, равную

$$\rho_{S_n}(x) = e^{-\lambda x}\frac{\lambda^n x^{n-1}}{(n-1)!}, \quad x \geqslant 0.$$

Докажем это методом математической индукции. База индукции очевидна. Теперь пусть для шага n выполнена предыдущая формула. Воспользуемся формулой свертки плотностей распределений для нахождения $\rho_{S_{n+1}}$

$$\rho_{S_{n+1}}(x) = \int_0^x \rho_{S_1}(x-t)\cdot\rho_{S_n}(t)\,dt = \frac{\lambda^{n+1}}{(n-1)!}e^{-\lambda x}\int_0^x (x-t)^{n-1}\,dt = \frac{\lambda^{n+1}}{(n-1)!}e^{-\lambda x}\frac{x^n}{n}.$$

Теорема доказана.

■

Пусть $t > 0$. Рассмотрим независимые случайные величины $\{\xi_k\}_{k\geqslant 1}$, имеющие показательное распределение с параметром λ. Как и в предыдущем утверждении, положим $S_n = \sum_{k=1}^n \xi_k$. Наконец, обозначим $X = \max\{\, n \geqslant 0 \mid S_n < t\}$, полагая $S_0 = 0$. Докажем теперь, что $X \sim \mathrm{Pois}(\lambda t)$.

Для этого найдем вероятность того, что $X = n$. При $n = 0$

$$\mathbb{P}(X = n) = \mathbb{P}(\xi_1 \geqslant t) = e^{\lambda t}.$$

При $n > 0$, поскольку $\xi_k \geqslant 0$, то согласно утверждению 3.2

$$\mathbb{P}(X = n) = \mathbb{P}(S_n < t,\, S_{n+1} \geqslant t) = \mathbb{P}(S_n < t) - \mathbb{P}(S_n < t,\, S_{n+1} < t) =$$
$$= \mathbb{P}(S_n < t) - \mathbb{P}(S_{n+1} < t) = F_{S_n}(t) - F_{S_{n-1}}(t) = \frac{\lambda^n t^n}{n!}e^{-\lambda t}.$$

Таким образом мы получали способ построения пуассоновской случайной величины. Нужно просто брать показательные случайные величины с параметром $\lambda = 1$ и смотреть сумма скольки первых из них меньше параметра пуассоновского распределения.

3.5 Построение датчика распределения Пуассона как предел биномиального распределения

Биномиальное распределение сходится к распределению Пуассона, так как число испытаний уходит в бесконечность, в то время как произведение np остается фиксированным или, по крайней мере, p стремится к нулю. Поэтому распределение Пуассона с параметром $\lambda = np$ можно использовать как приближение к Bin (n, p) биномиального распределения, если n достаточно велико, а p достаточно мало. Подробное доказательство этого факта, а также эмперические правила выбора параметров биномиального распределения и оценки точности приближения можно посмотреть в [2] и [3][Глава 4]. Мы же попытаемся лишь обосновать сам факт этого.

Пусть случайная величина ξ имеет биномиальное распределение с параметрами n и p. И введем обозначение:

$$\mathbb{P}_n(\xi = k) = C_n^k p^k (1-p)^{n-k}.$$

Зафиксируем значение $\lambda = np$, которое является математическим ожиданием биномиального распределения и будем устремлять параметр n к бесконечности:

$$\lim_{n\to\infty} \mathbb{P}_n(k) = \lim_{n\to\infty} \left[C_n^k \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \right] =$$
$$= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot \lim_{n\to\infty} \left[\frac{n!}{n^k (n-k)!} \left(1 - \frac{\lambda}{n}\right)^{-k} \right] =$$
$$= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot \lim_{n\to\infty} \left[\frac{n(n-1)\dots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{-k} \right] = \frac{\lambda^k}{k!} e^{-\lambda}.$$

3.6 Проверка корректности работы датчика

Для проверки правильности работы датчика пуассоновской случайной величины воспользуемся *критерием согласия Пирсона* (или критерием согласия χ-квадрат). Это один из наиболее популярных критериев для проверки гипотезы о принадлежности наблюдаемой выборки некорому теоретическому закону распределения.

Теорема 3.1 (Критерий согласия Пирсона)**.** *Обозначим нулевую гипотезу H_0 как гипотезу о том, что выборка $\xi_1, \xi_2 \dots \xi_n$ подчиняется закону распределения $\mathbb{P}$. Обозначим за n_k количество элементов в выборке, равных k. За r обозначим количество различных элементов выборки. А за p_k — вероятность выпадения значения в теоретическом распределении $p_k = \mathbb{P}(\xi = k)$. Введем статистику критерия*

$$X_n^2 = n \sum_{k=1}^{r} \frac{\left(\frac{n_k}{n} - p_k\right)^2}{p_k}.$$

Тогда если гипотеза H_0 верна, то при увеличении n статистика X_n^2 стремится к распределению χ^2 с $r-1$ степенью свободы.

Замечание 3.1. Функция распределения χ^2 с l степеней свододы выглядит следующим образом:

$$F_{\chi^2_l} = \frac{\gamma(l/2,\, x/2)}{\Gamma(l/2)}, \quad \text{где } \Gamma \text{ и } \gamma \text{ обозначают полную и неполную гамма-функции.}$$

Такое определение делает почти невозможным аналитический поиск квантилей этой функции распределения. По этой причине для данных расчетов мы будем пользоваться встроенной в систему `Matlab` функцией `chi2inv(l, alpha)`, которая возвращает квантиль χ^2-распределения с `l` степенями свободы порядка `alpha`.

3.7 Задача №4

Определение 3.3. `Нормалное распределение -- это ...`

Рассмотрим случайную величину $Z = \sqrt{2\xi}\sin\eta$, где $\xi \sim \mathrm{Exp}\,(1)$, $\eta \sim \mathrm{U}[0, 2\pi] \sim$

2πU[0, 1]. Тогда

$$\mathbb{P}(Z < x) = \mathbb{P}(\sqrt{2\xi}\sin\eta < x) = \iint\limits_{\{\,(\xi,\eta)\,|\,\sqrt{2\xi}\sin\eta<x\,\}} \frac{e^{-\xi}}{2\pi}\,d\xi d\eta = \left\{\, \xi = \frac{\psi^2}{2} \,\right\} =$$

$$= \iint\limits_{\{\,(\psi,\eta)\,|\,\psi\sin\eta<x\,\}} \frac{e^{-\frac{\psi^2}{2}}}{2\pi}\psi\,d\psi d\eta = \{\, X = \psi\cos\eta,\, Y = \psi\sin\eta \,\} =$$

$$= \iint\limits_{\{\,(X,Y)\,|\,Y<x\,\}} \frac{e^{-\frac{X^2}{2}}e^{-\frac{Y^2}{2}}}{2\pi}\,dXdY = \int\limits_{-\infty}^{+\infty} \frac{e^{-\frac{X^2}{2}}}{\sqrt{2\pi}}\,dX \int\limits_{-\infty}^{x} \frac{e^{-\frac{Y^2}{2}}}{\sqrt{2\pi}}\,dY = \int\limits_{-\infty}^{x} \frac{e^{-\frac{Y^2}{2}}}{\sqrt{2\pi}}\,dy.$$

Таким образом случайная величина Z имеет стандартное нормальное распределение.

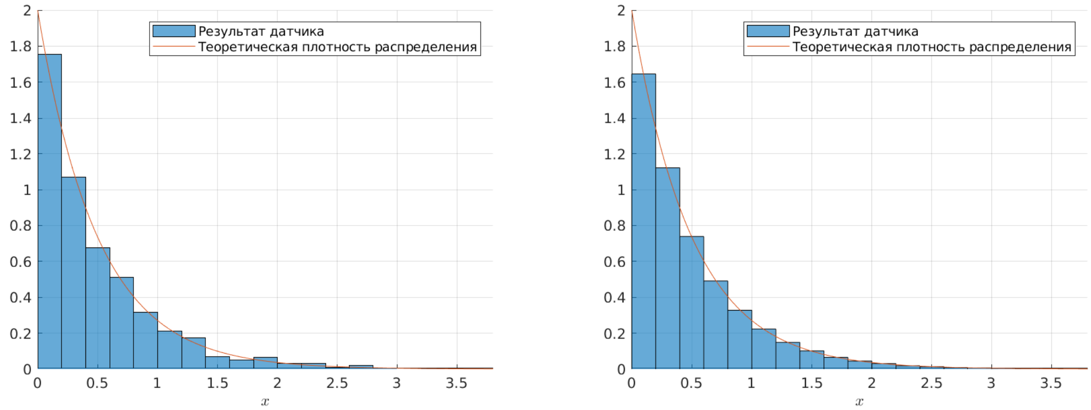

Рис. 3.1: Гистограмма экспоненциального распределения случайной величины с параметром $\lambda = 2$ при 10^3 (слева) и 10^5 (справа) испытаний.

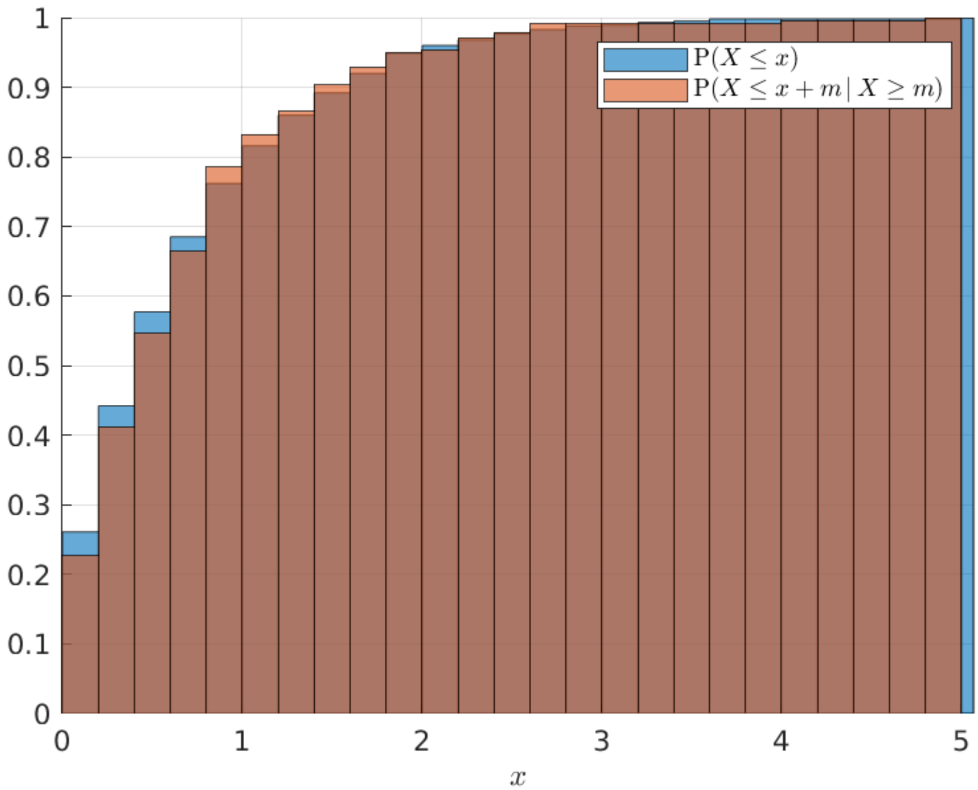

Рис. 3.2: Гистограмма экспоненциального распределения, демонстрирующая его свойство отсутствия памяти. Здесь задан параметр распределения $\lambda = \frac{3}{2}$, а также «сдвиг» $m = 1$. Проведено 10^3 испытаний.

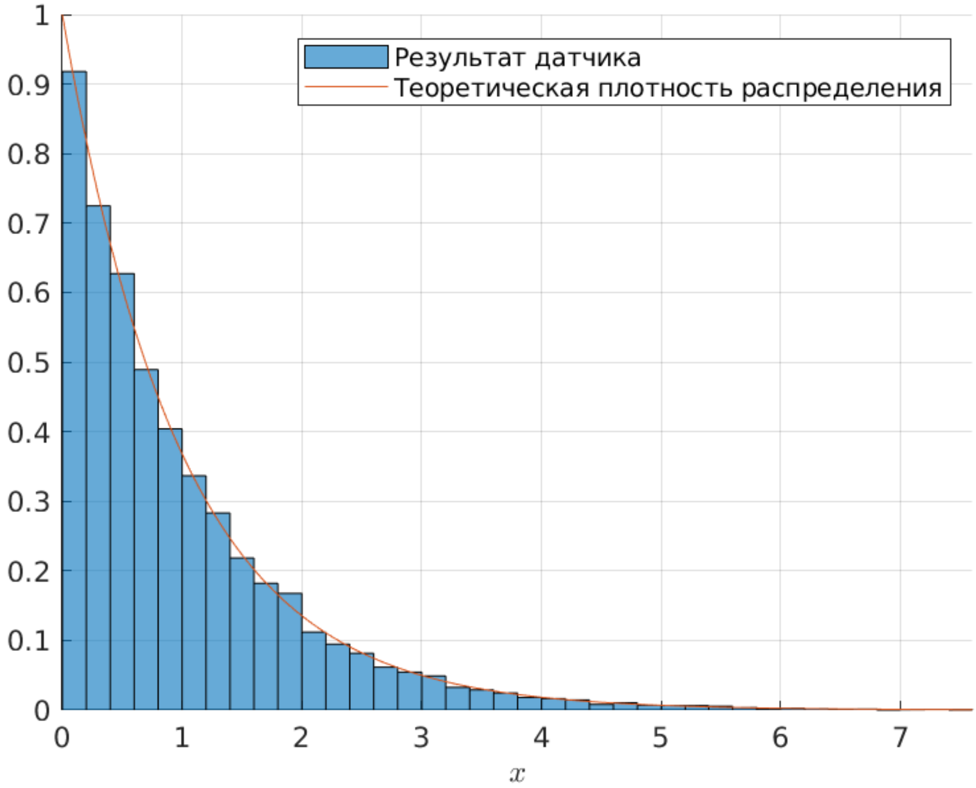

Рис. 3.3: Гистограмма распределения случайной величины $Y = \min_{i=\overline{1,n}} X_i$. Здесь $\lambda_1 = \lambda_2 = \ldots = \lambda_n = \frac{1}{10}$, $n = 10$. Число испытаний 10^4.

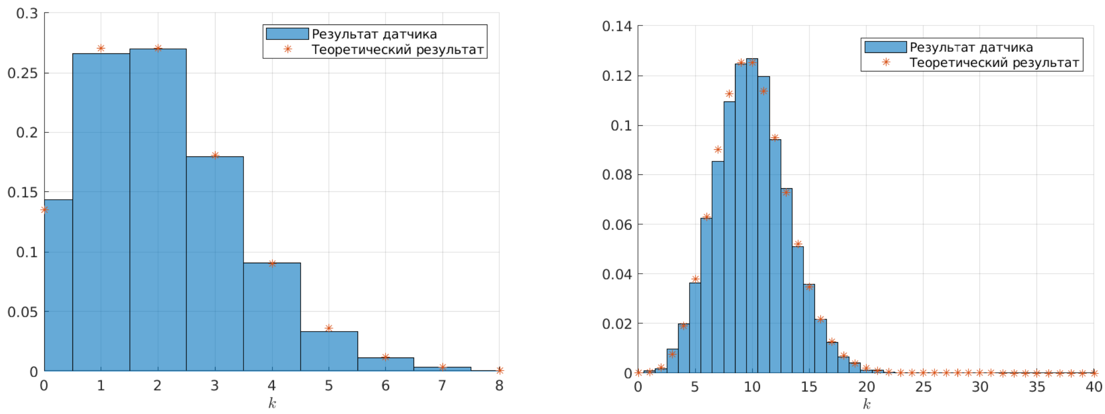

Рис. 3.4: Гистограмма распределения Пуассона случайной величины с параметром $\lambda = 2$ (слева) и $\lambda = 10$ (справа) при 10^4 испытаний.

4 Задание №5

1. Пусть $X \sim \mathrm{N}(\mu, \sigma^2)$. Убедиться в справедливости ЗБЧ и ЦПТ, то есть исследовать поведение суммы S_n и эмпирического распределения случайной величины
$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right).$$

2. Считая μ и σ неизвестными, для пункта 1 построить доверительные интервалы для среднего и дисперсии.

3. Пусть $X \sim K(a, b)$ имеет распределение Коши со сдвигом a и масштабом b. Проверить эмпирически, как ведут себя суммы $\frac{S_n}{n}$. Результат объяснить, а также найти закон распределения данных сумм.

Список литературы

[1] Ширяев А. Н. Вероятность, в 2-х кн. — 4-е изд., переработанное и дополненное — М.: МЦНМО, 2007.

[2] NIST/SEMATECH, “6.3.3.1. Counts Control Charts”, e-Handbook of Statistical Methods.

[3] Novak S.Y. Extreme value methods with applications to finance. London: CRC/ Chapman and Hall/Taylor and Francis, 2011.