

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

Отчет по компьютерному практикуму к курсу

«Стохастический анализ и моделирование»

Студент 415 группы
Е. В. Гуров

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

Москва, 2021

Содержание

Задание 1	4
Формулировка задания	4
Генератор схемы Бернулли и биномиального распределения	4
Геометрическое распределение	5
Свойство отсутствия памяти	6
Игра в орлянку	6
Задание 2	8
Формулировка задания	8
Датчик канторова распределения	8
Проверка корректности датчика и критерий Колмогорова	9
Свойство симметрии и самоподобия	10
Проверка однородности и критерий Смирнова	11
Математическое ожидание и дисперсия	12
Задание 3	14
Формулировка задания	14
Датчик экспоненциального распределения	14
Свойство отсутствия памяти	15
Случайная величина $Y = \min(X_1, X_2, \dots, X_n)$	16
Датчик пуассоновского распределения	17
Датчик пуассоновского распределения как предел биномиального распределения	18
Проверка корректности датчика и критерий хи-квадрат Пирсона	19
Датчик стандартного нормального распределения методом моделирования случайных величин парами с переходом в полярные координаты	21
Критерий Фишера и t-критерий Стьюдента	22
Задание 4	24
Формулировка задания	24
Датчик распределения Коши	24
Метод фон Неймана	24
Сравнение времени работы	27
Задание 5	27
Формулировка задания	27
Закон больших чисел и центральная предельная теорема для нормального распределения	28
Доверительные интервалы для математического ожидания и дисперсии	29
Закон больших чисел и распределение Коши	30

Задание 6	32
Формулировка задания	32
Метод Монте-Карло	33
Метод квадратур	35
Список литературы	37

Задание 1

Формулировка задания

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 независимых испытаний с бросанием правильной монеты. Выигрыш S_n определяется как сумма по всем n испытаниям 1 и -1 в зависимости от выпавшей стороны. Проиллюстрировать (в виде ломанной) поведение нормированной суммы $Y(i) = S_i/\sqrt{n}$, как функцию от номера испытания $i = 1, \dots, n$ для одной отдельно взятой траектории. Дать теоритическую оценку для $Y(n)$ при $n \rightarrow \infty$.

Генератор схемы Бернулли и биномиального распределения

Определение 1. *Схемой Бернулли называется эксперимент, в котором проводится, вообще говоря, неограниченное количество испытаний. При этом каждому испытанию присваивается бинарный признак (успех — 1 или неудача — 0), и выполняются следующие требования:*

1. *отсутствие взаимного влияния;*
2. *воспроизводимость;*
3. *испытания проводятся в сходных условиях.*

Определение 2. *Случайная величина X , принимающая значение 1 с вероятностью p и значение 0 с вероятностью $q = 1 - p$, называется случайной величиной с распределением Бернулли (или бернуллиевской случайной величиной).*

Для генератора схемы Бернулли реализуем генератор бернуллиевской случайной величины X . Для этого воспользуемся встроенным в библиотеку NumPy языка Python генератором равномерного распределения. Пусть тогда имеем случайную величину $Y \sim \mathbb{U}([0, 1])$. В таком случае X можно представить в виде: $X = \mathbb{I}(Y < p)$, где $\mathbb{I}()$ — индикаторная функция:

$$X = \mathbb{I}(Y < p) = \begin{cases} 1, & Y < p, \\ 0, & Y \geq p. \end{cases}$$

Генерация схемы Бернулли в таком случае будет происходить с помощью некоторого количества генераций бернуллиевской случайной величины.

Определение 3. *Случайная величина X имеет биномиальное распределение с параметрами n и p ($X \sim \text{Bin}(n, p)$), если*

$$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad k \in \mathbb{N} \cup \{0\}.$$

Случайную величину X обычно интерпретируют как число успехов в схеме из n испытаний Бернулли с вероятностью успеха p в каждом. Поэтому

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

где $Y_i \sim \text{Bern}(p)$, $i = 1, \dots, n$.

Промоделируем биномиальное распределение с параметрами $n = 50$, $p = 0.3$ с помощью генерации схемы Бернулли с n испытаниями:

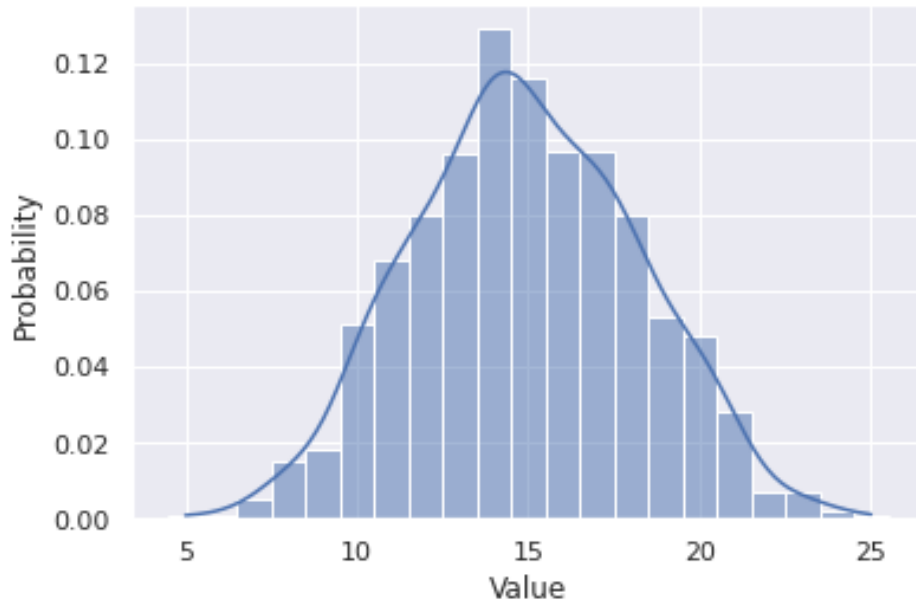


Рис. 1: Гистограмма биномиального распределения с $p = 0.3$, $n = 50$.

Геометрическое распределение

Определение 4. *Случайная величина X имеет геометрическое распределение с параметром p ($X \sim \text{Geom}(p)$), если*

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p = q^{k-1} p, \quad k \in \mathbb{N} \cup \{0\}.$$

Так же, как и в случае биномиального распределения, проводится некоторое количество испытаний Бернулли с одинаковой вероятностью успеха, до первого успеха. В качестве случайной величины с геометрическим распределением берется, как правило, количество неудач до первого успеха.

Свойство отсутствия памяти

Случайная величина с геометрическим распределением обладает так называемым свойством отсутствия памяти. Неформально оно означает, что в момент проведения очередного испытания Бернулли количество прошлых неудач не влияет на количество будущих. Формально же это свойство можно сформулировать как

Утверждение 1. Пусть $Y \sim \text{Geom}(p)$, тогда $\forall m, n \in \mathbb{N} \cup \{0\}$ справедливо:

$$\mathbb{P}(Y > m + n \mid Y \geq m) = \mathbb{P}(Y > n),$$

Доказательство. Рассмотрим левую часть равенства:

$$\begin{aligned} \mathbb{P}(Y > m + n \mid Y \geq m) &= \frac{\mathbb{P}(Y > m + n, Y \geq m)}{\mathbb{P}(Y \geq m)} = \\ &= \frac{\mathbb{P}(Y > m + n)}{\mathbb{P}(Y \geq m)} = \frac{\sum_{i=m+n+1}^{\infty} q^i p}{\sum_{i=m}^{\infty} q^i p} = \frac{q^{m+n+1}}{q^m} = q^{n+1}. \end{aligned}$$

С другой стороны, правая часть равна:

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

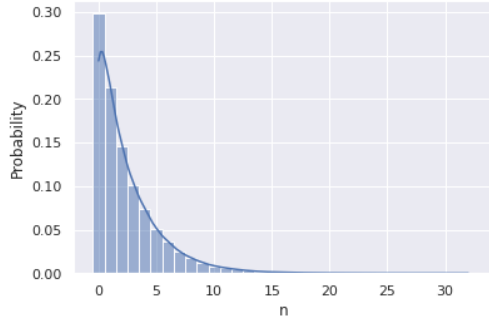
□

Для демонстрации этого свойства в Python сгенерируем массив некоторого достаточного количества геометрических случайных величин. С помощью него построим гистограмму геометрического распределения (Рис. (2а)). Зафиксируем некоторое m , и построим гистограмму распределения вектора геометрических случайных величин из первоначального набора, значения которых больше либо равны m . В результате увидим, что при достаточно большом количестве чисел в первоначальном наборе гистограммы двух распределений приблизительно совпадают (Рис. (2b)).

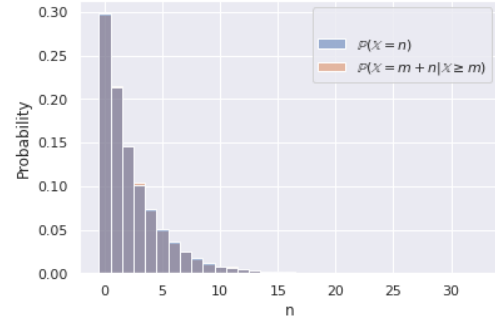
Игра в орлянку

Рассмотрим игру в орлянку. Для этого смоделируем последовательность случайных величин X_1, X_2, \dots , где

$$X_i = \begin{cases} 1, & p = \frac{1}{2}, \\ -1, & p = \frac{1}{2}, \end{cases} \quad i = 1, \dots, n.$$



(a) Гистограмма геометрического распределения при $p = 0.3$



(b) Демонстрация свойства отсутствия памяти

Рис. 2

Тогда необходимая сумма представляется в виде:

$$Y(i) = \frac{X_1 + \dots + X_i}{\sqrt{n}}, \quad i = 1, \dots, n,$$

где n — общее число генераций.

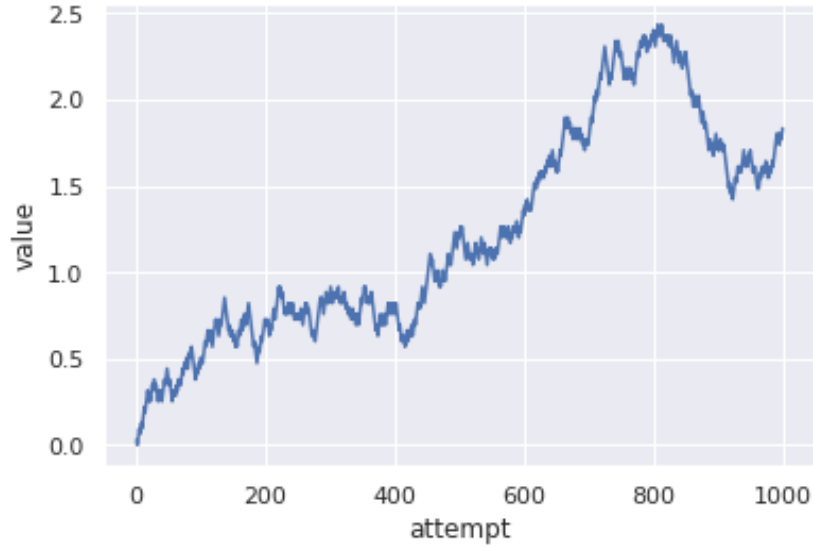


Рис. 3: Траектория суммы Y с $n = 1000$.

Оценим $Y(n)$ при $n \rightarrow \infty$. Для этого сформулируем необходимую теорему.

Теорема 1 (Центральная предельная теорема). Пусть X_1, \dots, X_n, \dots 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

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

по распределению при $n \rightarrow \infty$, где $N(0, 1)$ — нормальное распределение с нулевым математическим ожиданием и стандартным отклонением, равным единице.

В случае игры Орлянки:

$$\mu = \mathbb{E}[X_i] = 0, \quad \sigma^2 = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = 1, \quad i = 1, \dots, n$$

Тогда получим, что последовательность случайных величин

$$Y_n = Y(n) = \frac{S_n}{\sqrt{n}}$$

Удовлетворяет условиям теоремы 1. Таким образом получаем, что $Y(n) \rightarrow N(0, 1)$.

Задание 2

Формулировка задания

1. Построить датчик сингулярного распределения, имеющий в качестве функции распределения канторову лесницу. 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 $Y \in [0, 1/3]$ совпадает с распределением $\frac{Y}{3}$) с помощью критерия Смирнова.
3. Вычислить значение математическое ожидание и дисперсии для данного распределения. Сравнить теоритические значения с эмпирическими для разного объема выборки. Проиллюстрировать сходимость.

Датчик канторова распределения

Распределение, имеющее в качестве функции распределения канторову лесницу — это распределение сосредоточенное на канторовом множестве или канторово распределение. Рассмотрим алгоритм построения канторова множества:

Из единичного отрезка $C_0 = [0, 1]$ удалим интервал $(1/3, 2/3)$. Оставшееся множество обозначим через C_1 . Множество $C_1 = [0, 1/3] \cup [2/3, 1]$ состоит из двух отрезков; удалим теперь из каждого отрезка его среднюю треть, и оставшееся множество обозначим через C_2 . Повторив эту процедуру опять, удаляя средние трети у всех

четырёх отрезков, получаем C_3 . Действуя аналогично далее получаем последовательность вложенных множеств $C_0 \supset C_1 \supset C_2 \supset C_3 \supset \dots$.

Определение 5. *Пересечение*

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

называется канторовым множеством.

Из построения ясно, что канторово множество C можно определить как множество иррациональных чисел от нуля до единицы, представимое в троичной системе счисления лишь с помощью нулей и двоек. Это дает способ построения датчика канторова распределения.

$$X = \sum_{i=1}^{\infty} \frac{2}{3^i} \cdot Y_i, \quad i = 1, 2, \dots, \quad (1)$$

где $Y_i \sim \text{Bern}(0.5)$.

Для программной реализации датчика в таком случае можно использовать конечные суммы достаточно большого числа слагаемых. Сгенерируем n канторовых случайных величин и построим функцию распределения получившейся выборки (Рис. (4)). Отметим, что в силу возможности реализации лишь конечных сумм в (1), среди параметров генератора присутствует ϵ , имеющий смысл минимальной ширины ступеньки в канторовой лестнице.

Проверка корректности датчика и критерий Колмогорова

Для проверки корректности построенного датчика воспользуемся критерием Колмогорова. Статистикой критерия является величина

$$D_n = \sup_{-\infty < x < \infty} |\hat{F}_n(x) - F(x)|, \quad (2)$$

где $\hat{F}_n(x)$ — это выборочная функция распределения, а $F(x)$ — функция распределения элементов выборки. Теорема Гливенко-Кантели утверждает, что для произвольной функции распределения $F(x)$ имеет место сходимость $D_n \xrightarrow{\text{п.н.}} 0$. Поэтому в случае, когда гипотеза соответствия верна, значение D_n для выборки достаточно большого размера слабо отклоняется от нуля.

Следующая теорема дает оценку для функции распределения величины $\sqrt{n}D_n$ и позволяет таким образом оценивать вероятность наблюдаемого отклонения эмпирической функции распределения от теоретической.

Теорема 2 (Теорема Колмогорова). *Если функция распределения элементов выборки $F(x)$ непрерывна, то для $x > 0$*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n}D_n \leq x) = K(x) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2}.$$

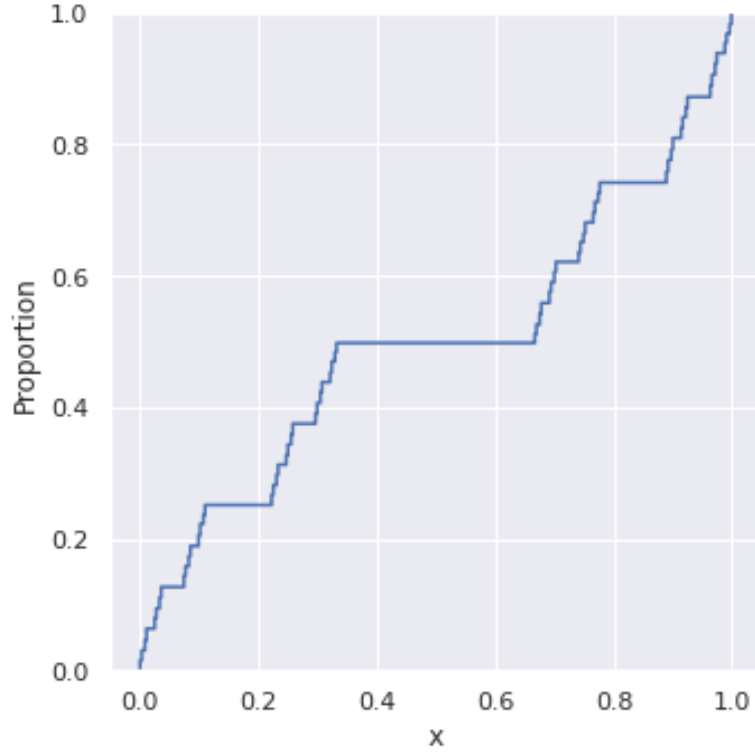


Рис. 4: Эмпирическая функция распределения сгенерированной выборки при $n = 100$

Таким образом проверка соответствия распределения может быть сведена к проверке $K(\sqrt{n}D_n)$, где D_n формируется для конкретной выборки. При заданном уровне значимости α гипотеза соответствия принимается при условии $1 - K(\sqrt{n}D_n) > \alpha$.

Так как функция распределения $F(x)$ непрерывна и неубывает, а $\hat{F}_n(x)$ — кусочно-постоянна, то \sup в (2) достигается в одной из точек разрыва функции \hat{F}_n . Отсюда получаем формулу для вычисления $D_n(x_1, \dots, x_n)$ заданной выборки (x_1, \dots, x_n) :

$$D_n(x_1, \dots, x_n) = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F(x_{(i)}), F(x_{(i)}) - \frac{i-1}{n} \right\}.$$

Здесь $x_{(i)}$ — i -ый элемент выборки, сортированной по возрастанию.

Свойство симметрии и самоподобия

Покажем свойство симметрии канторова распределения. Пусть имеется канторова случайная величина $X = \sum_{i=1}^{\infty} \frac{2}{3^i} Y_i$, где $Y_i \sim \text{Bern}(0.5)$. Рассмотрим случайную

величину $1 - X$:

$$1 - X = 1 - \sum_{i=1}^{\infty} \frac{2}{3^i} Y_i = \sum_{i=1}^{\infty} \frac{2}{3^i} - \sum_{i=1}^{\infty} \frac{2}{3^i} Y_i = \sum_{i=1}^{\infty} \frac{2(1 - Y_i)}{3^i} = \sum_{i=1}^{\infty} \frac{2}{3^i} Z_i.$$

Здесь $Z_i \sim \text{Bern}(0.5)$, поэтому случайные величины $1 - X$ и X распределены одинаково.

Покажем свойство самоподобия относительно деления на 3. Рассмотрим условное распределение канторовой случайной величины X на отрезке $\left[0; \frac{1}{3}\right]$. Это будет соответствовать тому, что $Y_1 = 0$. В таком случае:

$$X = \sum_{i=2}^{\infty} \frac{2}{3^i} Y_i = \sum_{i=1}^{\infty} \frac{2}{3^{i+1}} Y_{i+1} = \{Y_i = 0\} = \frac{1}{3} \sum_{i=1}^{\infty} \frac{2}{3^i} Y_i = \frac{1}{3} X.$$

Проверка однородности и критерий Смирнова

Пусть даны два набора наблюдений x_1, \dots, x_2 и y_1, \dots, y_m , являющиеся реализациями некоторых наборов случайных величин X_1, \dots, X_n и Y_1, \dots, Y_m , относительно которых выполнены следующие утверждения:

1. Случайные величины X_1, \dots, X_n независимы и имеют общую функцию распределения $F(x)$.
2. Случайные величины Y_1, \dots, Y_m независимы и имеют общую функцию распределения $G(x)$.
3. Обе функции F и G неизвестны, но являются непрерывными.
4. Все компоненты случайного вектора $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ независимы.

Определение 6. Два набора наблюдений, будем называть однородными, если для них выполнено:

$$G(x) = F(x)$$

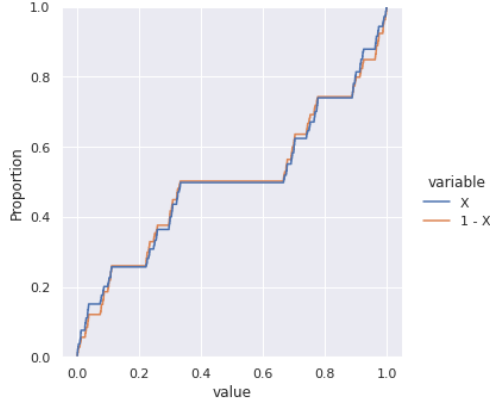
при всех x .

Для проверки гипотезы однородности против альтернативы неоднородности в случае выполнения утверждений (1)-(4) можно использовать критерий Смирнова, статистикой которого служит величина

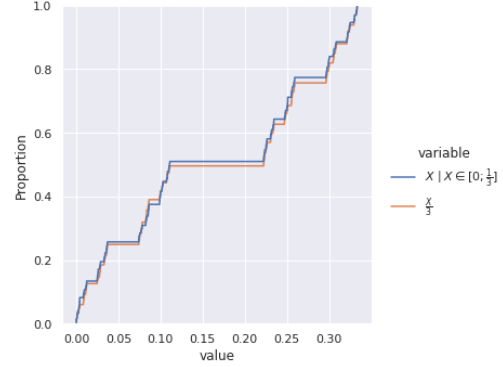
$$D_{n,m} = \sup_x \left| \hat{F}_n(x) - \hat{G}_m(x) \right|,$$

где $\hat{F}_n(x), \hat{G}_m(x)$ — выборочные функции распределения, то есть $D_{n,m}$ — расстояние в равномерной метрике между эмпирическими функциями выборок.

Следующая теорема аналогично теореме Колмогорова дает оценку для функции распределения статистики $\sqrt{\frac{nm}{n+m}} D_{n,m}$ и позволяет оценивать вероятность конкретного отклонения функций двух выборок.



(а) Эмпирические функции распределения выбокок из X и $1 - X$



(б) Эмпирические функции распределения выбокок из $X | X \in [0; \frac{1}{3}]$ и $\frac{X}{3}$

Рис. 5

Теорема 3 (теорема Смирнова). *Если гипотеза однородности верна, то при выполнении условий (1)-(4), для $x > 0$ имеет место:*

$$\lim_{n,m \rightarrow \infty} \mathbb{P} \left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leq x \right) = K(x),$$

где $K(x)$ — функция распределения Колмогорова из Теоремы 2.

Значения статистики на реализациях x_1, \dots, x_n и y_1, \dots, y_m можно находить следующим способом:

$$D_{n,m} = \max \{ D_{n,m}^+, D_{n,m}^- \},$$

где

$$D_{n,m}^+ = \sup_x (\hat{F}_n(x) - \hat{G}_m(x)) = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - \hat{G}_m(x_{(i)}) \right\},$$

$$D_{n,m}^- = \sup_x (\hat{G}_m(x) - \hat{F}_n(x)) = \max_{1 \leq j \leq m} \left\{ \frac{j}{m} - \hat{F}_n(y_{(j)}) \right\}.$$

Применим критерий Смирнова для проверки свойства симметрии. Для этого сформируем две выборки из распределений X и $1 - X$ и применим для них критерий. Получим, что при $n = 1000$, $\text{eps} = 0.00001$ и уровне значимости $\alpha = 0.05$ гипотеза однородности принимается. Выборочные функции распределения X и $1 - X$ представлены на Рис. (5а). Аналогично поступим для проверки свойства самоподобия. Выборочные функции соответствующих величин представлены на Рис. (5б).

Математическое ожидание и дисперсия

Вычислим математическое ожидание и дисперсию рассматриваемой случайной величины. Как упоминалось ранее, F обладает свойством самоподобия, то есть при

$0 < x < \frac{1}{3}$ выполнено соотношение $F(x) = \frac{F(3x)}{2}$, а при $\frac{2}{3} < x < 1$ имеет место равенство $F(x) = \frac{1}{2} + \frac{F(3x-2)}{2}$. Поэтому

$$\begin{aligned}\mathbb{E}[\xi] &= \int_{-\infty}^{+\infty} x \, dF(x) = \int_0^{\frac{1}{3}} x \, dF(x) + \int_{\frac{2}{3}}^1 x \, dF(x) = \\ &= \frac{1}{2} \int_0^{\frac{1}{3}} x \, dF(3x) + \frac{1}{2} \int_{\frac{2}{3}}^1 x \, d\left(\frac{1}{2} + F(3x-2)\right).\end{aligned}$$

Далее введем замену $y = 3x$ в первом интеграле и $y = 3x - 2$ во втором интеграле:

$$\begin{aligned}\mathbb{E}[\xi] &= \frac{1}{2} \int_0^1 \frac{y}{3} \, dF(y) + \frac{1}{2} \int_0^1 \frac{y+2}{3} \, dF(y) = \\ &= \frac{1}{6} \int_0^1 y \, dF(y) + \frac{1}{6} \int_0^1 y \, dF(y) + \frac{1}{3} \int_0^1 dF(y) = \frac{1}{3} \mathbb{E}[\xi] + \frac{1}{3}.\end{aligned}$$

Таким образом, получаем $\mathbb{E}[\xi] = \frac{1}{2}$.

Аналогичным способом с использованием свойства самоподобия вычислим дисперсию величины ξ , используя вычисленные значения математического ожидания.

$$\begin{aligned}\mathbb{E}[\xi^2] &= \int_0^{\frac{1}{3}} x^2 \, dF(x) + \int_{\frac{2}{3}}^1 x^2 \, dF(x) = \frac{1}{2} \int_0^1 \left(\frac{y}{3}\right)^2 \, dF(y) + \frac{1}{2} \int_0^1 \left(\frac{y+2}{3}\right)^2 \, dF(y) = \\ &= \frac{1}{9} \mathbb{E}[\xi^2] + \frac{2}{9} \mathbb{E}[\xi] + \frac{2}{9} = \frac{1}{9} \mathbb{E}[\xi^2] + \frac{1}{9} + \frac{2}{9}.\end{aligned}$$

То есть имеем $\mathbb{E}[\xi^2] = \frac{3}{8}$. Таким образом, получаем значение дисперсии $\mathbb{D}[\xi] = \frac{3}{8} - \left(\frac{1}{2}\right)^2 = \frac{1}{8}$.

На Рис. (6) демонстрируется сходимость выборочного математического ожидания и выборочной дисперсии к их теоретическим значениям, вычисленным выше, при увеличении размера выборки.

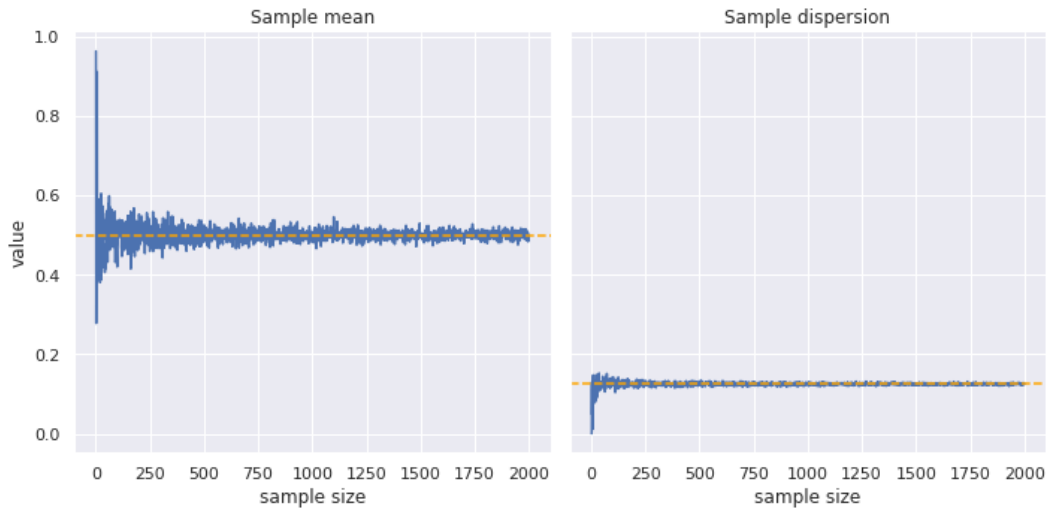


Рис. 6: Сходимость выборочных значений матожидания и дисперсии к теоретическим.

Задание 3

Формулировка задания

1. Построить датчик экспоненциального распределения. Проверить для данного распределения свойство отсутствия памяти. Пусть X_1, X_2, \dots, X_n — независимо экспоненциально распределенные с. в. с параметрами $\lambda_1, \lambda_2, \dots, \lambda_n$ соответственно. Найти распределение случайной величины $Y = \min(X_1, X_2, \dots, X_n)$.
2. На основе датчика экспоненциального распределения построить датчик пуассоновского распределения.
3. Построить датчик пуассоновского распределения как предел биномиального распределения. С помощью критерия хи-квадрат Пирсона убедиться, что получен датчик распределения Пуассона.
4. Построить датчик стандартного нормального 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.

Датчик экспоненциального распределения

Определение 7. Случайная величина X имеет экспоненциальное распределение с параметром $\lambda > 0$, если ее функция распределения имеет вид:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (3)$$

Теорема 4 (Метод обратной функции). Пусть функция распределения F имеет обратную F^{-1} . Тогда функцией распределения случайной величины

$$X = F^{-1}(Y),$$

где $Y \sim \mathbb{U}[0, 1]$, является F .

Доказательство. Найдем функцию распределения X :

$$F_X(x) = \mathbb{P}(X < x) = \mathbb{P}(F^{-1}(Y) < x) = \mathbb{P}(Y < F(x)) = F(x).$$

□

В случае экспоненциального распределения функция распределения (3) удовлетворяет условиям теоремы и обратная к ней легко выражается:

$$F_X^{-1}(y) = -\frac{1}{\lambda} \ln(1 - y).$$

Суперпозиция $F(Y)$, где $Y \sim \mathbb{U}[0, 1]$ является случайной величиной, имеющей экспоненциальное распределение с параметром λ :

$$X = -\frac{1}{\lambda} \ln(1 - Y) \sim \text{Exp}(\lambda).$$

На Рис.(7) приведено сравнение полученной эмпирически, с помощью построенного датчика, плотности экспоненциального распределения и его теоретической плотности, представимой в виде:

$$p(x) = \lambda e^{-\lambda x}$$

при $\lambda = 0.5$.

Свойство отсутствия памяти

Экспоненциальное распределение, как и его дискретный аналог — геометрическое, обладает свойством отсутствия мапяти, которое в данном случае можно сформулировать как

Утверждение 2. Случайная величина $X \sim \text{Exp}(\lambda)$ обладает свойством отсутствия памяти, то есть $\forall s, t \geq 0$ следует, что

$$\mathbb{P}(X \geq s + t \mid X \geq t) = \mathbb{P}(X \geq s). \quad (4)$$

Доказательство.

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

Таким образом, получаем:

$$\mathbb{P}(X \geq s + t) = \mathbb{P}(X \geq t) \mathbb{P}(X \geq s). \quad (5)$$

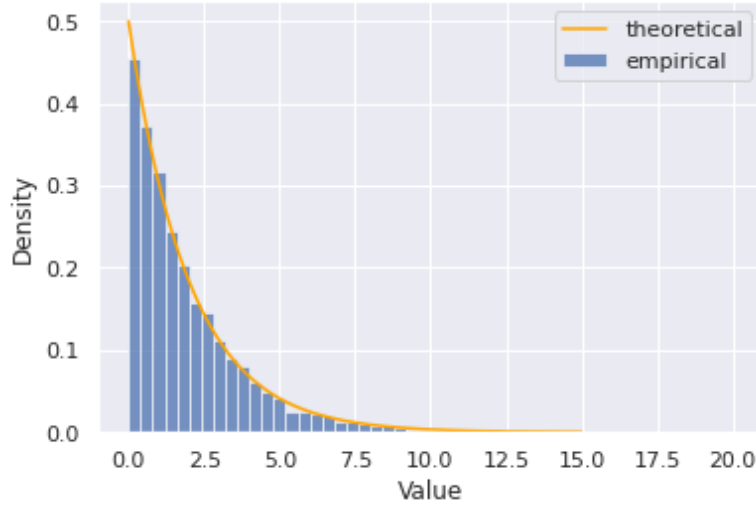


Рис. 7: Эмпирическая и теоретическая плотности экспоненциального распределения при $\lambda = 0.5$.

Для экспоненциально распределенной случайной величины верно, что:

$$\mathbb{P}(X \geq t) = 1 - F_X(t) = e^{-\lambda t}, \quad \mathbb{P}(X \geq s + t) = e^{-\lambda(s+t)}.$$

Следовательно, для (5) выполняется:

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

Следовательно, экспоненциальное распределение обладает свойством отсутствия памяти. \square

На Рис.(8), аналогично геометрическому распределению, данное свойство проиллюстрировано эмпирически.

Случайная величина $Y = \min(X_1, X_2, \dots, X_n)$

Утверждение 3. Пусть X_1, X_2, \dots, X_n — независимые экспоненциально распределённые случайные величины с параметрами $\lambda_1, \lambda_2, \dots, \lambda_n$ соответственно. Тогда случайная величина $Y = \min(X_1, X_2, \dots, X_n) \sim \text{Exp} \left(\sum_{i=1}^n \lambda_i \right)$.

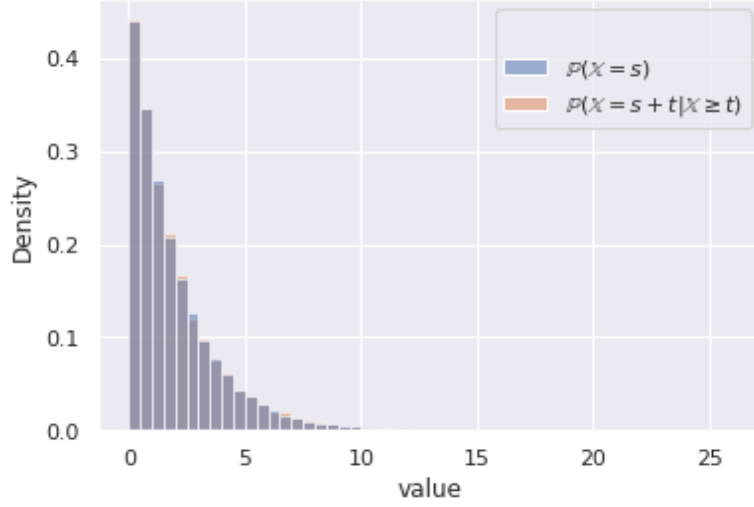


Рис. 8: Эмпирическая иллюстрация свойства отсутствия памяти при $t = 2$.

Доказательство.

$$\begin{aligned}
 F_Y(x) &= \mathbb{P}(Y \leq x) = 1 - \mathbb{P}(Y > x) = 1 - \mathbb{P}(\min(X_1, X_2, \dots, X_n) > x) = \\
 &= 1 - \mathbb{P}(X_1 > x, X_2 > x, \dots, X_n > x) = \{X_1, X_2, \dots, X_n \text{ независимы}\} = \\
 &= 1 - \mathbb{P}(X_1 > x) \cdot \mathbb{P}(X_2 > x) \cdot \dots \cdot \mathbb{P}(X_n > x) = \\
 &= 1 - (1 - F_{X_1}(x)) \cdot (1 - F_{X_2}(x)) \cdot \dots \cdot (1 - F_{X_n}(x)) = \\
 &= 1 - e^{-\lambda_1 x} \cdot e^{-\lambda_2 x} \cdot \dots \cdot e^{-\lambda_n x} = 1 - e^{-(\sum_{i=1}^n \lambda_i)x}.
 \end{aligned}$$

□

Эмпирическая демонстрация этого факта для $n = 4$, и случайно сгенерированных в интервале от 0 до 0.1 параметров λ_i приведена на Рис.(9).

Датчик пуассоновского распределения

Определение 8. *Случайная величина X имеет распределение Пуассона с параметром $\lambda > 0$, если*

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{N} \cup \{0\}.$$

Удобный метод построения датчика пуассоновского распределения даёт следующая

Теорема 5. ¹ Пусть $X_1, X_2, \dots, X_n, \dots \sim \text{Exp}(\lambda)$ — независимые одинаково распределённые случайные величины. Тогда случайная величина, определённая следую-

¹Доказательство теоремы можно найти в [2] на стр. 34.

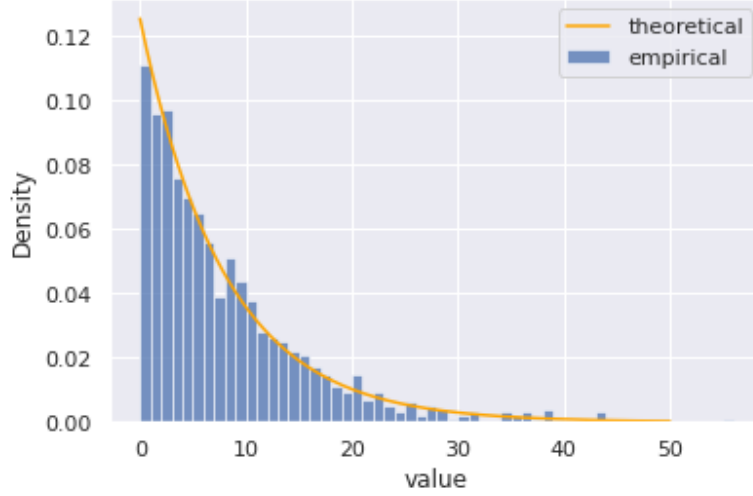


Рис. 9: Распределение $Y = \min(X_1, \dots, X_n)$.

щим образом:

$$Y = \max(n \mid S_n = X_1 + X_2 + \dots + X_n < 1)$$

имеет распределение Пуассона с параметром λ . При этом полагается $Y = 0$, если таких n не существует.

Таким образом для моделирования случайной величины Пуассона можно последовательно генерировать показательные случайные величины, пока их сумма не станет больше единицы. Количество сгенерированных экспоненциальных величин минус один и будет значением пуассоновской случайной величины. На Рис.(10) изображено сравнение распределения выборки полученной с помощью построенного вышеописанным способом датчика и теоретической функции вероятности.

Датчик пуассоновского распределения как предел биномиального распределения

Другой способ моделирования пуассоновской случайной величины основывается на следующей предельной теореме, связывающей распределение Пуассона с биномиальным распределением. Пусть

$$P_n(k) = \begin{cases} C_n^k p^k q^{n-k}, & k = 0, 1, \dots, n, \\ 0, & k = n+1, n+2, \dots, \end{cases}$$

и пусть p является функцией от n , $p = p(n)$.

Теорема 6 (Пуассона). ² Пусть $p(n) \rightarrow 0, n \rightarrow \infty$, причем так, что $np(n) \rightarrow \lambda$, где

²Доказательство этой теоремы можно найти в [3] на стр. 90.

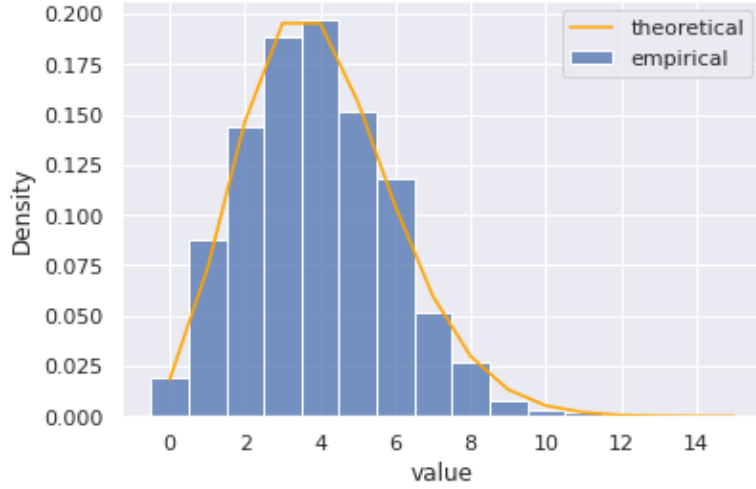


Рис. 10: Эмпирическая и теоретическая плотности распределения Пуассона при $\lambda = 4$.

$\lambda > 0$. Тогда для любого $k = 0, 1, \dots$

$$P_n(k) \rightarrow \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \dots$$

Таким образом строить датчик распределения Пуассона с параметром λ можно с помощью датчика биномиального распределения при $p = \frac{\lambda}{n}$ и больших значениях n . На Рис.(11) проиллюстрировано достаточно хорошее совпадение распределений $\text{Bin}\left(n, \frac{\lambda}{n}\right)$ и $\text{Pois}(\lambda)$ при $n = 10000, \lambda = 10$.

Проверка корректности датчика и критерий хи-квадрат Пирсона

Проверим корректность построенного с помощью биномиального распределения датчика. Для этого воспользуемся критерием хи-квадрат Пирсона, но для начала дадим необходимые определения.

Определение 9. Пусть случайные величины Z_1, \dots, Z_k распределены по стандартному нормальному закону $\mathcal{N}(0, 1)$ и независимы. Тогда распределение случайной величины $R_k^2 = Z_1^2 + \dots + Z_k^2$ называют распределением хи-квадрат с k степенями свободы (кратко: $R_k^2 \sim \chi_k^2$).

Пусть X_1, \dots, X_n — выборка из закона с функцией распределения $F(x)$. Разобьем множество значений X_1 на N промежутков (возможно бесконечных) $\delta_j = (a_j, b_j]$, $j = 1, \dots, N$. В случае дискретных распределений вместо промежутков значений можно рассматривать отдельные значения. Положим $p_j = \mathbb{P}(X_1 \in \delta_j)$, а

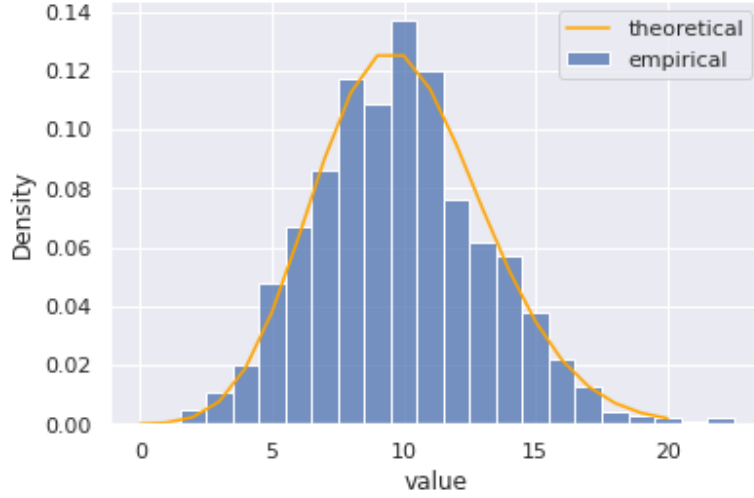


Рис. 11: Демонстрация предельного совпадения биномиального и пуассоновского распределений при $n = 10000$, $\lambda = 10$.

случайные величины ν_j — равными количеству элементов выборки в δ_j ($\nu_1 + \dots + \nu_N = n$). Функция F неизвестна и проверяется гипотеза

$$H_0 : F(x) = F_0(x),$$

где F_0 — заданная функция распределения. Если гипотеза верна, то согласно закону больших чисел частоты попадания в промежутки $\hat{p}_j = \frac{\nu_j}{n}$ при достаточно больших n должны быть близки к соответствующим вероятностям $p_j^0 = F_0(b_j) - F_0(a_j)$. В качестве меры отклонения от гипотезы H_0 принимается статистика

$$X_n^2 = n \sum_{j=1}^N \frac{1}{p_j^0} (\hat{p}_j - p_j^0)^2 = \sum_{j=1}^N \frac{(\nu_j - np_j^0)^2}{np_j^0},$$

которая по сути является взвешенной суммой квадратов отклонений частот от гипотетических вероятностей. В силу центральной предельной теоремы каждое отклонение асимптотически нормально и имеет порядок малости $\frac{1}{\sqrt{n}}$, поэтому представляется правдоподобной следующая

Теорема 7. ³ Если $0 < p_j^0 < 1$, $j = 1, \dots, N$, то при $n \rightarrow \infty$

$$X_n^2 \xrightarrow{d} \zeta \sim \chi_{N-1}^2.$$

Здесь сходимость понимается в смысле сходимости по распределению. Аналогично теореме Колмогорова, данная теорема позволяет оценивать вероятность отклонения, задаваемого статистикой Пирсона, посчитанного для конкретной выборки и,

³Доказательство этой теоремы можно найти в [1] на стр. 274.

в зависимости от необходимого уровня значимости, принимать или отвергать гипотезу H_0 .

Датчик стандартного нормального распределения методом моделирования случайных величин парами с переходом в полярные координаты

Определение 10. Случайная величина X имеет нормальное распределение вероятностей с параметрами μ и σ^2 , $X \sim \mathcal{N}(\mu, \sigma^2)$ (μ — математическое ожидание X , σ^2 — дисперсия X), если ее плотность распределения задается формулой

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty.$$

Определение 11. Нормальное распределение с параметрами $\mu = 0$ и $\sigma^2 = 1$ называется стандартным нормальным распределением, и его плотность распределения имеет следующий вид:

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

Рассмотрим способ точного моделирования, базирующийся на нелинейном преобразовании пары независимых равномерно распределенных на $[0, 1]$ случайных величин η_1, η_2 в пару независимых $\mathcal{N}(0, 1)$ случайных величин X, Y :

$$X = \sqrt{-2 \ln \eta_1} \cos(2\pi \eta_2), \quad Y = \sqrt{-2 \ln \eta_1} \sin(2\pi \eta_2)$$

Доказательство. Для независимых $\mathcal{N}(0, 1)$ случайных величин X и Y плотность вектора (X, Y) служит

$$p_{(X,Y)}(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

Обозначим через R и Φ полярные координаты точки (X, Y) : $X = R \cos \Phi$, $Y = R \sin \Phi$. Воспользуемся далее формулой преобразования плотности:

$$p_\eta(y) = |J(y)| p_\xi(f^{-1}(y)),$$

где $J(y) = \det \begin{pmatrix} \frac{\partial f_1^{-1}}{\partial y_1} & \dots & \frac{\partial f_k^{-1}}{\partial y_1} \\ \dots & \dots & \dots \\ \frac{\partial f_1^{-1}}{\partial y_k} & \dots & \frac{\partial f_k^{-1}}{\partial y_k} \end{pmatrix}$ — якобиан f^{-1} .

Находим (в данном случае якобиан замены равен r)

$$p_{(R,\Phi)}(r, \varphi) = \frac{1}{2\pi} e^{-\frac{r^2}{2}} r, \quad r > 0, \quad 0 < \varphi < 2\pi.$$

Так как она распадается в произведение плотностей

$$p_R(r) = re^{-\frac{r^2}{2}} \mathbb{I}_{\{r>0\}} \text{ и } p_\Phi(\varphi) = \frac{1}{2\pi} \mathbb{I}_{\{0<\varphi<2\pi\}},$$

то R и Φ независимы. Интегрируя плотности, вычисляем функцию распределения

$$F_R(r) = 1 - e^{-\frac{r^2}{2}}, \quad \text{при } r \geq 0 \text{ и } F_\Phi(\varphi) = \frac{\varphi}{2\pi}, \quad \text{при } 0 \leq \varphi \leq 2\pi.$$

Методом обратной функции (Теорема 4) получаем формулы для моделирования случайных величин R и Φ : $R = \sqrt{-2 \ln \eta_1}$, $\Phi = 2\pi\eta_2$, которые остается подставить в формулы замены координат. \square

Будем генерировать стандартные нормально-распределенные случайные величины с помощью полученных явно их выражений. Сравнение плотности полученной выборки и теоретической плотности приведено на Рис.(12).

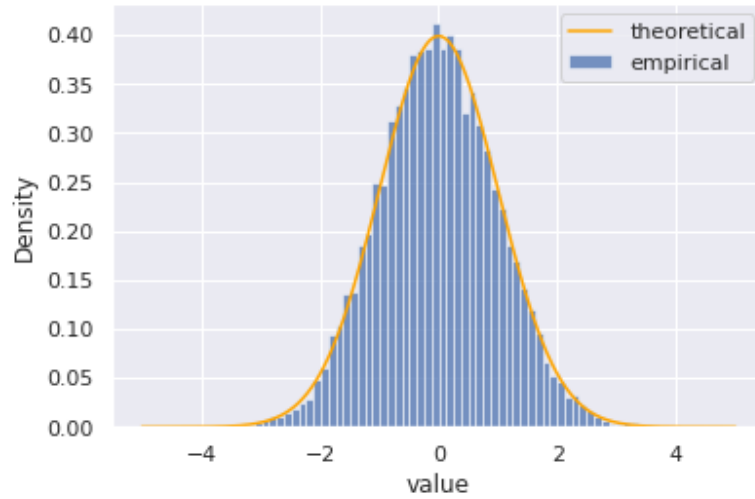


Рис. 12: Демонстрация совпадения сгенерированного стандартного нормального распределения с теоретическим при размере выборки $n = 10000$.

Критерий Фишера и t-критерий Стьюдента

Проверим равенство дисперсий и матожиданий пары случайных величин построенных с помощью такого датчика. Для этого воспользуемся критерием Фишера и t-критерием Стьюдента.

Определение 12. Случайная величина ζ имеет F -распределение(Фишера-Снедекора) с k_1 и k_2 степенями свободы(обозначается $\zeta \sim F_{k_1, k_2}$), если

$$\zeta = \left(\frac{1}{k_1} \xi \right) / \left(\frac{1}{k_2} \eta \right),$$

где $\xi \sim \chi_{k_1}^2$, $\eta \sim \chi_{k_2}^2$, ξ и η независимы.

Определение 13. Пусть случайные величины Z и R_k^2 независимы и распределены согласно законам $\mathcal{N}(0, 1)$ и χ_k^2 соответственно. Тогда распределение случайной величины $T_k = Z/\sqrt{R_k^2/k}$ называют распределением Стьюдента с k степенями свободы или t -распределением (кратко $T_k \sim t_k$).

Критерий Фишера по сути может быть сформулирован следующим образом. Если гипотеза $H' : \sigma_1 = \sigma_2$, μ_1 и μ_2 — любые верна, то статистика S_1^2/S_2^2 распределена по закону $F_{n-1, m-1}$. Здесь

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2$$

— несмещенные оценки для дисперсий σ_1^2 и σ_2^2 . Это утверждение опирается на определение распределения Фишера и следующую теорему.

Теорема 8. ⁴ Для нормальной выборки $X_i \sim \mathcal{N}(\theta_1, \theta_2^2)$ Выборочное среднее $\bar{X} = \frac{1}{n} \sum X_i$ и выборочная дисперсия $S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ независимы, причем $nS^2/\theta_2^2 \sim \chi_{n-1}^2$, а $\sqrt{n-1}(\bar{X} - \theta_1)/S \sim t_{n-1}$.

В силу этой теоремы $(n-1)S_1^2/\sigma_1^2 \sim \chi_{n-1}^2$, $(m-1)S_2^2/\sigma_2^2 \sim \chi_{m-1}^2$, и следовательно формулировка критерия Фишера верна. Отметим, что критерий Фишера имеет двустороннюю критическую область, поэтому сравнение статистики для отвержения или принятия гипотезы в этом случае нужно проводить и с $\frac{\alpha}{2}$ -квантилью и с $1 - \frac{\alpha}{2}$ -квантилью распределения Фишера-Снедекора.

Проверим теперь равенство математических ожиданий с помощью критерия Стьюдента. Обозначим неизвестную общую дисперсию через σ^2 . Так как распределение хи-квадрат является частным случаем гамма-распределения ($\chi_k^2 \sim \Gamma(k/2, 1/2)$), получаем

$$\sigma^{-2} [(n-1)S_1^2 + (m-1)S_2^2] \sim \chi_{n+m-2}^2.$$

Поскольку математическое ожидание закона χ_{n+m-2}^2 равно $n+m-2$, статистика $S_{tot}^2 = [(n-1)S_1^2 + (m-1)S_2^2] / (n+m-2)$ несмещенно оценивает σ^2 по объединенной выборке.

При справедливости гипотезы $H'' : \mu_1 = \mu_2$ ввиду независимости выборок имеем: $\bar{X} - \bar{Y} \sim \mathcal{N}(0, (1/n + 1/m)\sigma^2)$. Отсюда согласно определению закона Стьюдента:

$$T = (\bar{X} - \bar{Y}) / \left(S_{tot} \sqrt{\frac{1}{n} + \frac{1}{m}} \right) = \sqrt{\frac{nm}{n+m}} (\bar{X} - \bar{Y}) / S_{tot} \sim t_{n+m-2}.$$

Это приводит к критерию Стьюдента, позволяющему проверить гипотезу H'' . Отметим также, что данный критерий, как и критерий Фишера имеет двустороннюю критическую область.

⁴Доказательство этой теоремы можно найти в [1] на стр. 149.

Задание 4

Формулировка задания

1. Построить датчик распределения Коши.
2. На основе датчика распределения Коши с помощью метода фон Неймана построить датчик стандартного нормального распределения. При помощи функции `normal probabity plot` убедиться в корректности построенного датчика и обосновать наблюдаемую линейную зависимость.
3. Сравнить скорость моделирования стандартного нормального распределения в заданиях 3 и 4.

Датчик распределения Коши

Определение 14. *Случайная величина X имеет распределение Коши с параметрами a и b , если ее функция распределения имеет вид:*

$$F_X(x) = \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right) + \frac{1}{2}.$$

Плотность распределения Коши:

$$p_X(x) = \frac{1}{\pi} \frac{b}{(x-a)^2 + b^2}.$$

Функция распределения $F_X(x)$ обладает обратной, а значит в данном случае для моделирования распределения можно пользоваться методом обратной функции (Теорема (4)). Обратная функция для $F_X(x)$ равна $F_X^{-1}(y) = a + b \tan\left(\pi\left(y - \frac{1}{2}\right)\right)$. Следовательно, в качестве датчика распределения Коши можно построить датчик случайной величины $X = F_X^{-1}(Y)$, где $Y \sim U[0, 1]$. На Рис.(13) продемонстрировано совпадение эмпирической и теоретической функций распределения для распределения Коши, полученного построенным датчиком.

Метод фон Неймана

Метод фон Неймана заключается в моделировании нормального распределения путём мажорирования плотностью распределения Коши с параметрами a и b . Для достижения наилучшей оценки, будем подбирать параметры a и b . Плотность стандартного нормального распределения $p_1(x)$ и плотность распределения Коши $p_2(x)$ выглядят следующим образом:

$$p_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$$p_2(x) = \frac{1}{\pi} \frac{b}{(x-a)^2 + b^2}.$$

При моделировании будем следовать алгоритму:

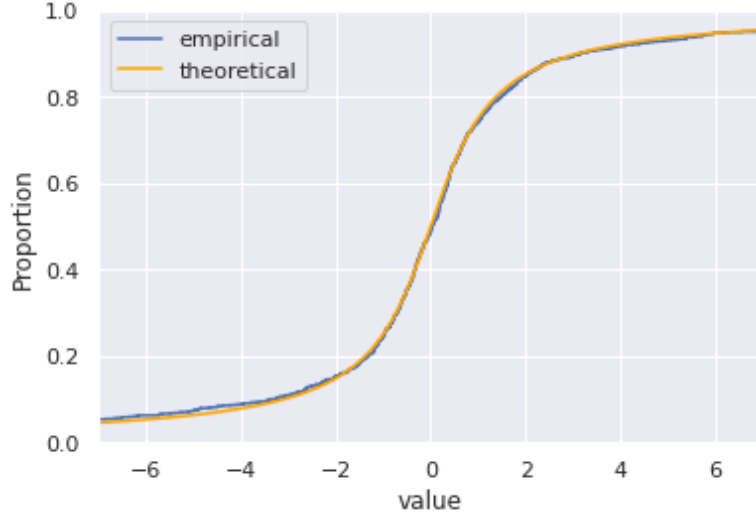


Рис. 13: Демонстрация совпадения эмпирической и теоретической функций распределения для распределения Коши. Размер выборки: $n = 1000$.

1. возьмем некоторое число $k > 0$, такое что $p_1(x) \leq kp_2(x), \forall x \in \mathbb{R}$,
2. рассмотрим значение случайной величины $x = X, X \sim \text{Cauchy}(a, b)$,
3. сгенерируем случайную величину $y = Y(x) \sim \text{Bern}\left(\frac{p_1(x)}{kp_2(x)}\right)$,
4. если $y = 1$, то x — значение из распределения с плотностью $p_1(x)$, иначе — продолжаем моделирование, начиная с пункта 2).

Данный алгоритм работает тем быстрее, чем ближе отношение $\frac{p_1(x)}{kp_2(x)}$ к единице, поэтому в качестве k возьмем $k^* = \min_{a,b} \max_x \frac{p_1(x)}{p_2(x)}$. Рассмотрим отношение

$$\frac{p_1(x)}{p_2(x)} = \frac{\sqrt{\pi}}{\sqrt{2}b} e^{-\frac{x^2}{2}} ((x-a)^2 + b^2).$$

Пусть $a = 0$. Рассмотрим вспомогательную функцию:

$$g(x) = e^{-\frac{x^2}{2}} (x^2 + b^2).$$

Найдем максимум этой функции:

$$g'(x) = e^{-\frac{x^2}{2}} x (2 - b^2 - x^2) = 0,$$

следовательно, точки экстремума:

$$\begin{cases} x = 0, |b| > \sqrt{2}, \\ x = \pm \sqrt{2 - b^2}, 0 < |b| \leq \sqrt{2}. \end{cases}$$

Таким образом,

$$k^* = \min \left\{ \min_{|b| > \sqrt{2}} \sqrt{\frac{\pi}{2}} b, \min_{0 < |b| < \sqrt{2}} \frac{\sqrt{2\pi}}{b} e^{\frac{b^2}{2} - 1} \right\}.$$

Поскольку $k > 0$, то и $b > 0$. Найдем максимум вспомогательной функции

$$h(b) = \frac{e^{\frac{b^2}{2} - 1}}{b} :$$

$$h'(b) = \frac{1 - b^2}{b^2} e^{\frac{b^2}{2} - 1},$$

следовательно, поскольку $b > 0$, точкой экстремума является $b = 1$.

Получаем оптимум при $a^* = 0$, $b^* = 1$:

$$k^* = \min \left\{ \sqrt{\pi}, \sqrt{\frac{2\pi}{e}} \right\} = \sqrt{\frac{2\pi}{e}}.$$

Докажем, что $a = 0$ — оптимальное значение параметра.

$$\begin{aligned} k^* &= \min_{a,b} \max_x \left(\frac{\sqrt{\pi}}{\sqrt{2}b} e^{-\frac{x^2}{2}} ((x-a)^2 + b^2) \right) = \\ &= \min_a \left\{ \min_{b > \sqrt{2}} \frac{p_1(x)}{p_2(x)} \Big|_{x=0}, \min_{0 < b \leq \sqrt{2}} \frac{p_1(x)}{p_2(x)} \Big|_{x=\pm\sqrt{2-b^2}} \right\} > \\ &> \min_a \left\{ \min_{b > \sqrt{2}} \frac{\sqrt{\pi}}{\sqrt{2}b} (a^2 + b^2), \min_{0 < b \leq \sqrt{2}} \left(\sqrt{2-b^2} + |a| \right) \right\} \quad (6) \end{aligned}$$

Минимум выражения достигается при $a = 0$.

Иллюстрация работы построенного датчика, использующая Python функцию `scipy.stats.probplot`, представлена на Рис. (14). На оси ординат откладываются точки выборки, на оси абсцисс — квантили стандартного нормального распределения. Прямой линии соответствует "точное" нормальное распределение, наилучшим образом приближающее, в смысле указанных осей, значения выборки. Видно, что полученная с помощью датчика Фон-Неймана выборка следует стандартному нормальному распределению.

Возьмем далее случайную величину $\xi \sim N(\mu, \sigma^2)$. Ее функция распределения

$$F_\xi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Введем замену переменной $s = \frac{t-\mu}{\sigma}$. Тогда

$$F_\xi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{s^2}{2}} ds = F\left(\frac{x-\mu}{\sigma}\right)$$

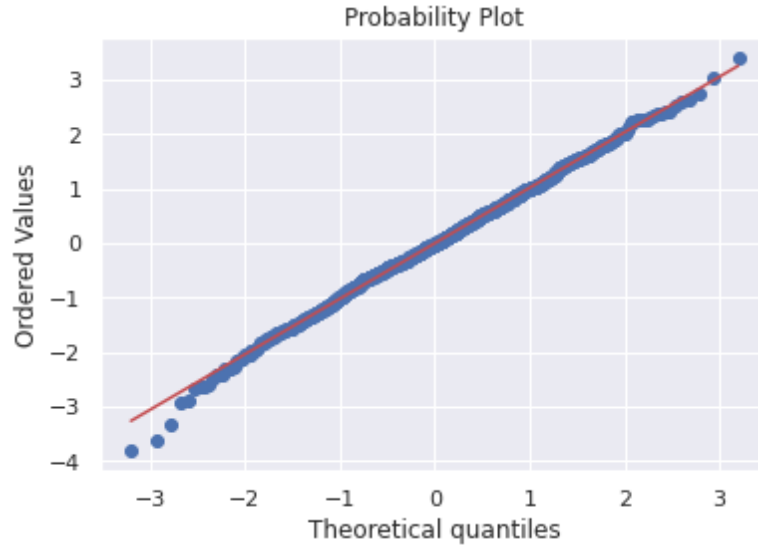


Рис. 14: Демонстрация совпадения построенного с помощью метода Фон-Неймана распределения со стандартным нормальным при размере выборки $n = 1000$.

где $F(x)$ — функция стандартного нормального распределения.

Таким образом, квантили различных распределений связаны между собой линейно, что означает, что любую нормальную случайную величину $\xi \sim N(\mu, \sigma^2)$ можно представить в виде $\xi = \sigma\eta + \mu$, где $\eta \sim N(0, 1)$, а прямая в функции probplot будет прямой со сдвигом μ и с коэффициентом наклона σ .

Сравнение времени работы

На Рис. (15) приведен график сравнения скорости работы датчика стандартного нормального распределения с моделированием случайных величин парами и датчика, построенного методом Фон-Неймана.

Задание 5

Формулировка задания

1. Пусть $X_i \sim N(\mu, \sigma^2)$. Убедиться эмпирически в справедливости ЗБЧ и ЦПТ, т.е. исследовать поведение суммы S_n и эмпирического распределения величины

$$\sqrt{n} \left(\frac{S_n}{n} - a \right).$$

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

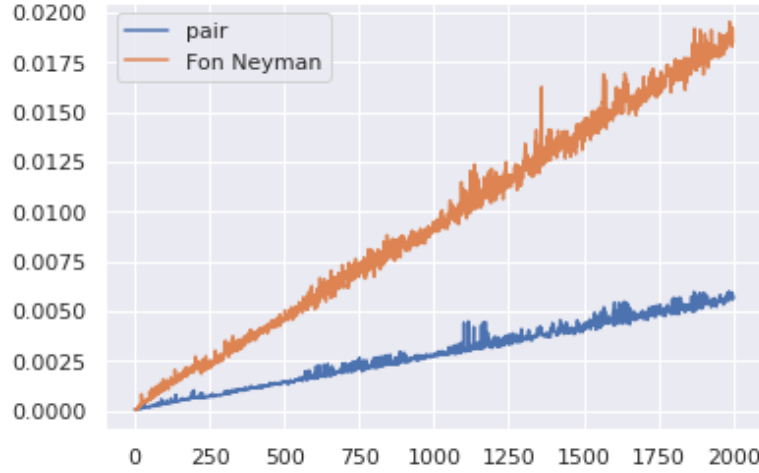


Рис. 15: Зависимость времени моделирования от размера генерируемой выборки.

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

Закон больших чисел и центральная предельная теорема для нормального распределения

Пусть $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Исследуем поведение суммы $\frac{S_n}{n}$ и эмпирического распределения величины

$$\sqrt{n} \left(\frac{S_n}{n} - \mu \right).$$

Теорема 9 (Закон больших чисел). Пусть X_1, X_2, \dots — независимые одинаково распределенные случайные величины, $\mathbb{E}X_i = \mu$, $\forall i \in \mathbb{N}$, $|\mu| < \infty$, $S_n = X_1 + \dots + X_n$. Тогда $\frac{S_n}{n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \mu$, т. е.

$$\forall \varepsilon > 0 \quad \mathbb{P} \left(\left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \xrightarrow[n \rightarrow \infty]{} 0.$$

Теорема 10 (Центральная предельная теорема). Пусть X_1, X_2, \dots — независимые одинаково распределенные случайные величины, $0 < \mathbb{E}X_i^2 < \infty$, $\forall i \in \mathbb{N}$, $S_n = X_1 + \dots + X_n$. Тогда

$$\mathbb{P} \left(\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathbb{D}S_n}} \leq x \right) \xrightarrow[n \rightarrow \infty]{} \Phi(x), \quad x \in (R),$$

где $\Phi(x)$ — функция стандартного нормального распределения:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Доказательство этих теорем представлено в [3]. На рисунке (16) представлена иллюстрация сходимости из закона больших чисел. На рисунке (17) в свою очередь проиллюстрирована сходимость из центральной предельной теоремы.

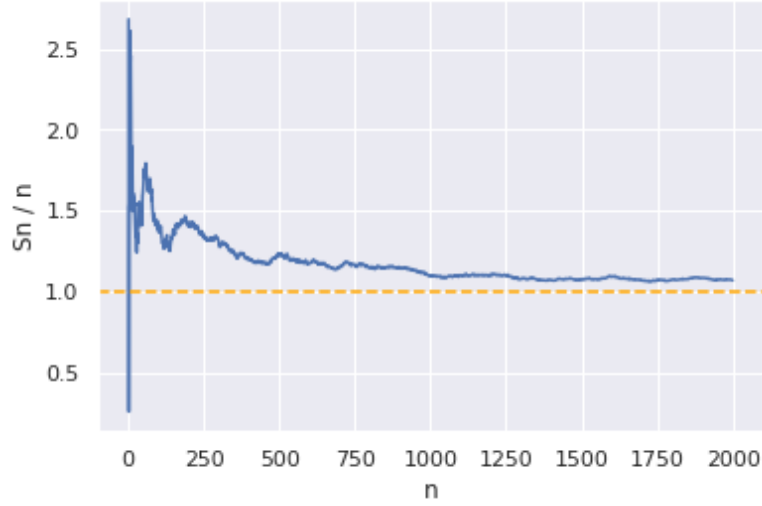


Рис. 16: Иллюстрация сходимости среднего S_n/n к матожиданию $\mu = 1$ при увеличении n .

Доверительные интервалы для матожидания и дисперсии

Построим доверительные интервалы для матожидания и дисперсии, считая их неизвестными. Случайная величина

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

где

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

— уже встречавшаяся ранее, несмещенная оценка дисперсии, имеет распределение Стьюдента с $n - 1$ степенями свободы. Тогда, в силу симметрии распределения, получим:

$$\mathbb{P} \left(-t_{1-\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

или же

$$\mathbb{P} \left(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right) = 1 - \alpha,$$

что непосредственно дает доверительный интервал для матожидания с уровнем доверия $1 - \alpha$.

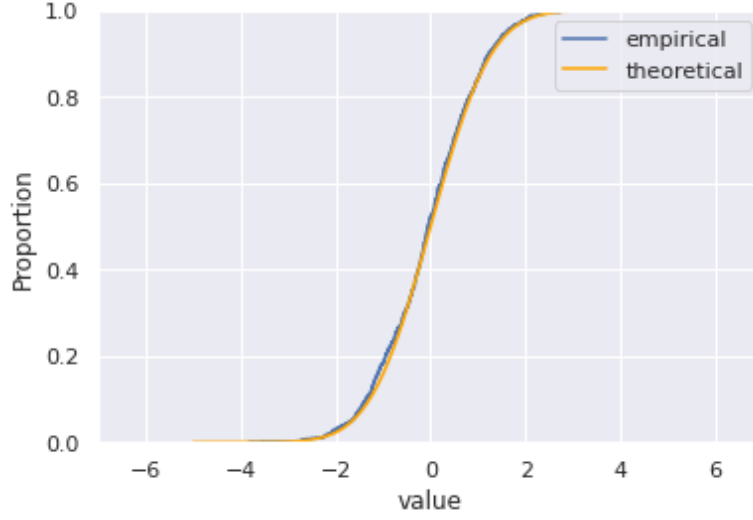


Рис. 17: Эмпирическая функция распределения величины $\frac{S_n - \mu^* n}{\sigma \sqrt{n}}$ при $n = 1000$ и теоретическая функция распределения стандартного нормального распределения.

Для построения доверительного интервала для дисперсии рассмотрим случайную величину

$$H = \frac{(n-1)S^2}{\sigma^2},$$

имеющую распределение хи-квадрат χ_{n-1}^2 с $n-1$ степенями свободы. В данном случае распределение уже не обладает свойством симметрии, поэтому будем брать $\frac{\alpha}{2}$ и $1 - \frac{\alpha}{2}$ квантили. В таком случае получаем:

$$\mathbb{P}\left(\chi_{\frac{\alpha}{2}}^2 \leq H \leq \chi_{1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

или

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2}\right) = 1 - \alpha,$$

откуда немедленно получаем доверительный интервал для дисперсии с уровнем доверия $1 - \alpha$.

Графики доверительных интервалов для математического ожидания и дисперсии в зависимости от размера выборки, по которой они строятся изображены на Рис. (18) и Рис.(19) соответственно.

Закон больших чисел и распределение Коши

Исследуем вопрос справедливости закона больших чисел для последовательностей случайных величин Коши. Эмпирически можно сделать вывод, что в отличие

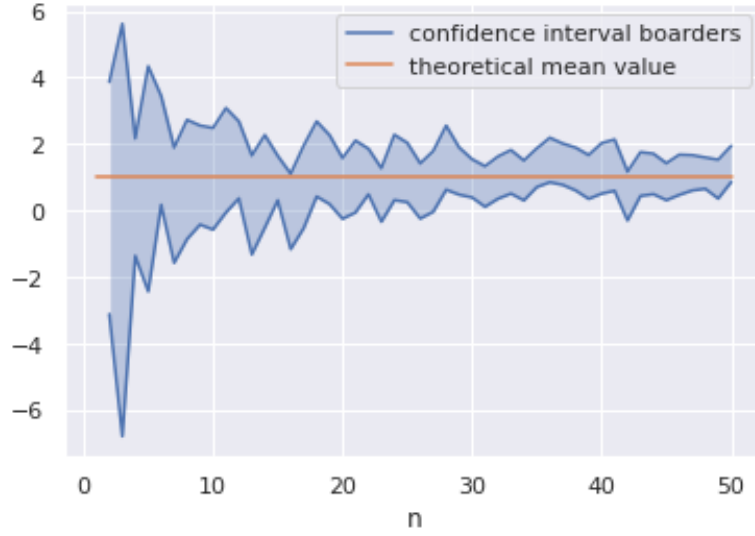


Рис. 18: Эволюция доверительного интервала для математического ожидания $\mu = 1$ при увеличении n с уровнем доверия 95%.

от выборок из нормального распределения, сходимость к математическому ожиданию в данном случае отсутствует. Это можно объяснить тем, что у распределения Коши вовсе нет математического ожидания, а выборочное среднее в выборке из таких случайных величин также будет распределена по закону Коши. Иными словами, если $X_1, \dots, X_n \sim C(a, b)$, то

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim C(a, b).$$

Это свойство доказывается с помощью характеристических функций.

Определение 15. Функция $\phi_X(t) = \mathbb{E} e^{itX}$ вещественного переменного t называется характеристической функцией случайной величины X .

Утверждение 4. Характеристическая функция суммы независимых случайных величин равна произведению функций слагаемых.

Доказательство. Если случайные величины X и Y независимы, то, по свойству математических ожиданий получаем:

$$\phi_{X+Y}(t) = \mathbb{E} e^{it(X+Y)} = \mathbb{E} e^{itX} e^{itY} = \phi_X(t) \phi_Y(t).$$

□

Характеристическая функция однозначно задает распределение, то есть если две случайные величины имеют одинаковые характеристические функции, то их распределения совпадают. В таком случае, рассмотрим характеристическую функцию

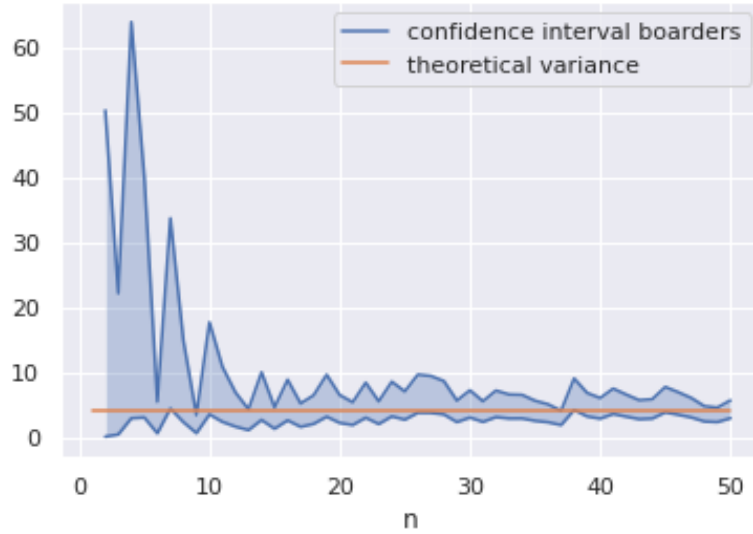


Рис. 19: Эволюция доверительного интервала для дисперсии $\sigma^2 = 4$ при увеличении n с уровнем доверия 90%.

выборочного среднего и учтем при этом, что для $X_i \sim C(a, b)$, $\phi_{X_i} = e^{ait-b|t|}$.

$$\psi_{\bar{X}}(t) = \phi_{X_1 + \dots + X_n} \left(\frac{t}{n} \right) = \left(\phi_{X_1} \left(\frac{t}{n} \right) \right)^n = \left(e^{\frac{ait-b|t|}{n}} \right)^n = \phi_{X_1}(t).$$

То есть получаем, что $\bar{X} \sim C(a, b)$. Отсутствие сходимости выборочного среднего проиллюстрировано на рисунке (20). На рисунке (21) проиллюстрировано совпадение распределений Коши и выборочного среднего для выборки из распределения Коши (так называемого свойства устойчивости).

Задание 6

Формулировка задания

1. Посчитать интеграл

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{e^{-\left(x_1^2 + \dots + x_{10}^2 + \frac{1}{2^7 \cdot x_1^2 \cdot \dots \cdot x_{10}^2}\right)}}{x_1^2 \cdot \dots \cdot x_{10}^2} dx_1 \dots dx_{10}$$

- методом Монте-Карло
 - методом квадратур, сводя задачу к вычислению собственного интеграла Римана
2. Для каждого случая оценить точность вычислений.

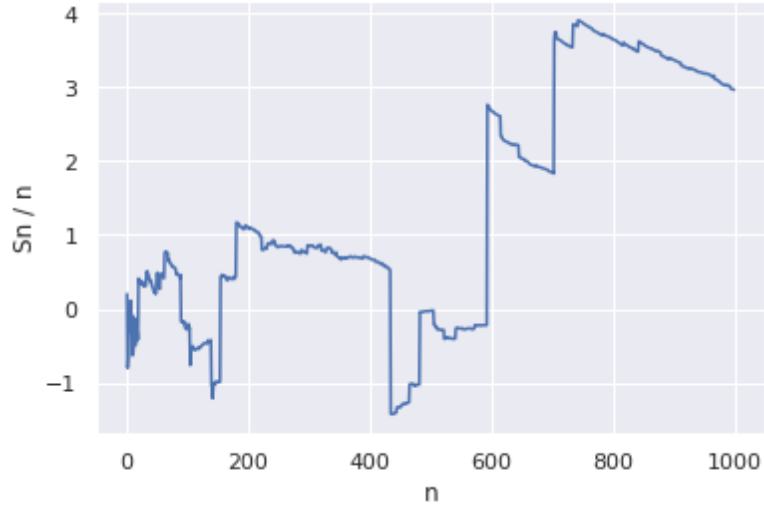


Рис. 20: Поведение среднего значения в выборке из распределений Коши с ростом размера выборки.

Метод Монте-Карло

Перепишем интеграл

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{e^{-\left(x_1^2 + \dots + x_{10}^2 + \frac{1}{27 \cdot x_1^2 \dots x_{10}^2}\right)}}{x_1^2 \cdot \dots \cdot x_{10}^2} dx_1 \dots dx_{10}$$

в виде

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_{10}) g(x_1, \dots, x_{10}) dx_1 \dots dx_{10},$$

где

$$f(x) = \sqrt{\pi^{10}} \cdot \frac{e^{-\frac{1}{27 \cdot x_1^2 \dots x_{10}^2}}}{x_1^2 \cdot \dots \cdot x_{10}^2}, \quad g(x) = \frac{1}{\sqrt{\pi^{10}}} \cdot e^{-(x_1^2 + \dots + x_{10}^2)}.$$

Заметим, что $g(x)$ является совместной плотностью распределения набора независимых случайных величин, имеющих нормальное распределение с параметрами 0 и $\frac{1}{2}$:

$$x = (x_1, \dots, x_{10}), \quad x_i \sim \mathcal{N}\left(0, \frac{1}{2}\right).$$

Тогда интеграл (1) можно записать в виде:

$$I = \mathbb{E}f(x_1, \dots, x_{10}), \quad x_i \sim \mathcal{N}\left(0, \frac{1}{2}\right).$$

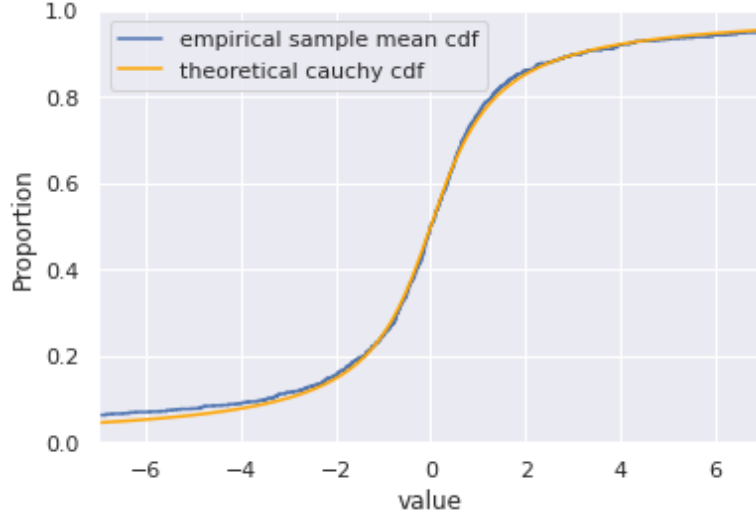


Рис. 21: Иллюстрация устойчивости распределения Коши: $\bar{X} \sim C(a, b)$. $a = 0$, $b = 1$

Рассмотрим выборку

$$x^i = (x_1^i, \dots, x_{10}^i), \quad x_k^i \sim \mathcal{N}\left(0, \frac{1}{2}\right), \quad k = \overline{1, 10}, \quad i = \overline{1, n}.$$

Согласно ЗБЧ выборочное среднее будет стремиться к математическому ожиданию, то есть:

$$\bar{f} = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n f(x^i) \xrightarrow{n \rightarrow \infty} I.$$

Оценим погрешность метода Монте-Карло с помощью центральной предельной теоремы:

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S_n}{n} - I\right| < \varepsilon\right) &= \mathbb{P}\left(\left|\frac{S_n - nI}{n}\right| < \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_n - nI}{\sigma\sqrt{n}}\right| < \frac{\sqrt{n}}{\sigma}\varepsilon\right) = \\ &= \mathbb{P}\left(-\frac{\sqrt{n}}{\sigma}\varepsilon < \frac{S_n - nI}{\sigma\sqrt{n}} < \frac{\sqrt{n}}{\sigma}\varepsilon\right) = \Phi_0\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) - \Phi_0\left(-\frac{\sqrt{n}}{\sigma}\varepsilon\right) = \\ &= \Phi_0\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) - \left(1 - \Phi_0\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right)\right) = 2\Phi_0\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) - 1 = 2\Phi_0(x_p) - 1 = 1 - \alpha, \quad (7) \end{aligned}$$

где

- $\Phi_0(x)$ — функция Лапласа или функция ошибок:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt,$$

- $x_p = \frac{\sqrt{n}}{\sigma}\varepsilon$ — квантиль уровня p . Из (7) видно, что в данном случае $p = 1 - \frac{\alpha}{2}$
- α — уровень значимости.

Погрешность ε для соответствующего уровня значимости $\alpha = 2 - 2\Phi_0(x_p)$ связана с x_p соотношением:

$$\varepsilon = \frac{\sigma x_p}{\sqrt{n}}.$$

Значение $\sigma > 0$ используем как значение выборочной дисперсии:

$$\sigma = \frac{1}{n} \sum_{i=1}^n f^2(x_i) - \left(\frac{1}{n} \sum_{i=1}^n f(x_i) \right)^2.$$

В качестве уровня значимости возьмем $\alpha = 0.05$:

$$\mathbb{P} \left(\left| \frac{S_n}{n} - I \right| < \varepsilon \right) = 1 - \alpha = 0.95.$$

Ниже приведена таблица зависимости вычисленных значений интеграла и полученной погрешности при разном количестве испытаний:

Число испытаний	Результат	Погрешность	Время работы
10^2	107.6762	79.4061	0.0005
10^3	139.7728	31.0156	0.00474
10^4	108.3251	19.266	0.0078
10^5	123.9544	6.9206	0.0956
10^6	124.5817	2.3284	0.3685
10^7	124.78	0.7101	3.0683
10^8	124.7298	0.2214	16.911
10^9	124.789	0.071	186.962
10^{10}	124.8231	0.0221	1436.962

Метод квадратур

Сведем задачу к вычислению собственного интеграла Римана. Для этого сделаем следующую замену переменных:

$$x_i = \operatorname{tg} \left(\frac{\pi}{2} t_i \right), t_i \in [0; 1].$$

Таким образом, по методу прямоугольников исходный интеграл приблизится значением:

$$I = \left(\frac{\pi}{2} \right)^{10} \int_{-1}^1 \dots \int_{-1}^1 \frac{\exp \left\{ - \left(\sum_{k=1}^{10} \operatorname{tg} \left(\frac{\pi}{2} t_k \right)^2 + \frac{1}{2^7 \cdot \prod_{k=1}^{10} \operatorname{tg} \left(\frac{\pi}{2} t_k \right)^2} \right) \right\}}{\prod_{k=1}^{10} \operatorname{tg} \left(\frac{\pi}{2} t_k \right)^2 \cdot \prod_{k=1}^{10} \cos \left(\frac{\pi}{2} t_k \right)^2} dt_1 \dots dt_{10}.$$

Проведём равномерное разбиение отрезка $[-1, 1]$ на N частей:

$$-1 = t_0 < t_1 < \dots < t_N = 1, \quad t_i = -1 + i \cdot \frac{1 - (-1)}{N} = i \cdot \frac{2}{N} - 1.$$

Обозначим через $f(t_1, \dots, t_{10})$ подынтегральную функцию интеграла I . Будем использовать метод средних прямоугольников. Для этого нам необходимо выбрать середины нашего разбиения:

$$y_i = \frac{t_i + t_{i-1}}{2}, \quad i = \overline{1, N}.$$

Тогда наш интеграл приближённо можно посчитать следующим образом:

$$I_N = \left(\frac{\pi}{N}\right)^{10} \sum_{i_1=1}^N \dots \sum_{i_{10}=1}^N f(y_{i_1}, \dots, y_{i_{10}}).$$

Оценка погрешности метода прямоугольников на равномерной сетке имеет следующий вид:

$$\varepsilon = \frac{h^2}{24}(b-a) \sum_{i,j=1}^{10} \max |f''_{x_i, x_j}| = \frac{1}{6N^2} \sum_{i,j=1}^{10} \max |f''_{x_i, x_j}|.$$

Приведем таблицу зависимости результата от количества точек разбиения отрезка:

N	Результат	Время работы
3	0.086797	0.114195
4	272.6029	0.363554
5	183.4886	4.940286
6	116.3903	45.398591
7	120.6386	283.993222

Список литературы

- [1] Лагутин М. Б. *Наглядная математическая статистика*, Бином. М.: 2009.
- [2] Кропачёва Н. Ю., Тихомиров А. С. *Моделирование случайных величин*, НовГУ им. Ярослава Мудрого. Великий Новгород: 2004.
- [3] Ширяев А. Н. *Вероятность*, МЦНМО. М.: 2007.