Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра системного анализа

Отчёт по практикуму

«Стохастический анализ и моделирование»

Студент 415 группы

Н. Ю. Заварзин

Руководитель практикума

аспирант В. А. Сливинский

Москва, 2023

Содержание

Задание 1

1.1

Определение 1. *Схемой Бернулли с заданной вероятностью p называется эксперимент, состоящий из серии испытаний, удовлетворяющих следующим условиям:*

1. *Отсутствие взаимного влияния.*
2. *Воспроизводимость. Однородные испытания проводятся в сходных, аналогичных условиях (не одинаковых, иначе результат был бы один и тот же).*
3. *Существует признак, который реализуется ("успех" с вероятностью p) или не реализуется ("неуспех" с вероятностью $q = 1 - p$) в испытании. Признак может быть отнесён к любому из испытаний (в силу их однородности).*

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.*

Случайную величину X, имеющую распределение Бернулли, можно сэмплировать, беря логическое значение $u < p$, где u — равномерно распределенная на $[0, 1]$ случайная величина. Ведь при таком подходе $\mathbb{P}(X = 1) = \mathbb{P}(u < p) = p$, $\mathbb{P}(X = 0) = \mathbb{P}(u >= p) = q$.

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

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

Случайную величину X, имеющую биномиальное распределение можно представить в виде суммы n независимых, одинаково распределенных бернуллиевских случайных величин

$$X = \sum_{i=1}^{n} Y_i. \tag{2}$$

Так как

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i = k\right) = C_n^k \cdot \mathbb{P}(Y = 1)^k \cdot \mathbb{P}(Y = 0)^{n-k} = C_n^k p^k (1 - p)^{n-k},$$

где C_n^k — число сочетаний k "успехов" в n испытаниях Бернулли, $Y \sim Y_i$.

1.2

Определение 4. *Случайная величина X, равная количеству неудач до появления первого успеха в схеме Бернулли с параметром p, имеет геометрическое распределение с параметром p ($X \sim Geom(p)$),*

$$\mathbb{P}(X = k) = p(1-p)^k. \tag{3}$$

Для моделирования X, имеющей геометрическое распределение с параметром p, будем сэмплировать случайную величину $Y \sim Ber(p)$, до появления первого значения 1 и возвращать число встреченных нулей. В таком случае, аналогично тому, как это было сделано для биномиального распределения, можно показать корректность используемого построения.

Свойство 1. *Если $Y \sim Geom(p)$, то $\mathbb{P}(Y > m+n \mid Y \geqslant m) = \mathbb{P}(Y > n)$, для любых целых неотрицательных m и n. Это называется свойством отсутствия памяти (количество "неудач" в прошлом не влияет на число будущих).*

Доказательство.

$$\mathbb{P}(Y > m+n \mid Y \geqslant m) = \frac{\mathbb{P}(Y > m+n)}{\mathbb{P}(Y \geqslant m)} = \frac{1 - \sum_{k=0}^{m+n} p(1-p)^k}{1 - \sum_{k=0}^{m-1} p(1-p)^k},$$

покажем, что эта вероятность совпадает с $\mathbb{P}(Y > n) = 1 - \sum_{k=0}^{n} p(1-p)^k$, или, что

$$1 - \sum_{k=0}^{m+n} p(1-p)^k = \left(1 - \sum_{k=0}^{m-1} p(1-p)^k\right)\left(1 - \sum_{k=0}^{n} p(1-p)^k\right) \Leftrightarrow$$

$$\Leftrightarrow 1 - \left(1 - \frac{p(1-p)^{m+n+1}}{1-(1-p)}\right) = \left(1 - \left(1 - \frac{p(1-p)^m}{1-(1-p)}\right)\right)\left(1 - \left(1 - \frac{p(1-p)^{n+1}}{1-(1-p)}\right)\right) \Leftrightarrow$$

$$\Leftrightarrow (1-p)^{m+n+1} = (1-p)^m (1-p)^{n+1}.$$

□

1.3

Рассмотрим процесс игры в орлянку, для этого построим последовательность случайных величин X_i таких, что

$$X_i = \begin{cases} 1, & p = 0.5, \\ -1, & p = 0.5. \end{cases}$$

$Y(i)$ примем равным $\dfrac{X_1 + \cdots + X_i}{\sqrt{n}}$, $i = \overline{1, n}$

Теорема 1. (***Центральная предельная теорема.***) (*док-во см. в* [2])
Пусть $X_1, X_2, \ldots$ *— независимые, одинаково распределенные случайные величины, с конечными* $\mathbb{E}X_i = a$ *и* $\mathbb{D}X_i = \sigma^2$*. Тогда*

$$\mathbb{P}\left(\frac{S_n - na}{\sigma\sqrt{n}} < z\right) \underset{n\to\infty}{\to} \Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{z} e^{-\frac{y^2}{2}}\,\mathrm{d}y.$$

Отметим, что математическое ожидание X_i равно нулю, дисперсия равна 1, а значит по сформулированной выше теореме, в пределе для $Y(n)$ мы получим стандартное нормальное распределение.

Задание 2

2.1

Определение 5. *Распределение называется сингулярным, если оно сосредоточено на континуальном множестве с нулевой мерой Лебега.*

Моделировать канторову случайную величину стандартным методом через обращение функции распределения крайне проблематично, в силу сложности аналитического нахождения обратной функции. Поэтому перейдём от равномерного сэмплирования на $[0, 1]$ по oy к равномерному сэмплированию элементов множества кантора (считаем, что между ними есть взаимнооднозначное соответствие из-за монотонности лестницы и её возрастания исключительно на рассматриваемом множестве). Для того, чтобы понять как нам генерировать элементы канторова множества вспомним процесс его построения.

При построении канторова множества C на отрезке $[0, 1]$ мы выбрасываем из него интервалы $(\frac{1}{3}, \frac{2}{3})$, $(\frac{1}{9}, \frac{2}{9})$, $(\frac{7}{9}, \frac{8}{9})$, $\ldots$ В итоге получаем замкнутое множество C (как пересечение замкнутых). Оно получается из отрезка $[0, 1]$ выбрасыванием счётного числа интервалов. Из построения получаем, что канторово множество состоит из точек, в записи которых в троичной системе счисления нет единиц. Поэтому элементы этого множества можно сэмплировать в виде

$$X = \sum_{i=1}^{\infty} \frac{2y_i}{3^i},$$

где $y_i \sim Ber(0.5)$.

Беря во внимание конечность машинной точности вычислений, элементы будем генери-

ровать как

$$X = \sum_{i=1}^{n} \frac{2y_i}{3^i},$$

тут n выбирается исходя из заданной точности ε следующим образом:

$$\sum_{k=n}^{\infty} \frac{2y_k}{3^k} \leqslant \sum_{k=n}^{\infty} \frac{2}{3^k} = \frac{\frac{2}{3^n}}{1-\frac{1}{3}} = \frac{1}{3^{n-1}} < \varepsilon \Leftrightarrow$$

$$\Leftrightarrow n > \log_3\left(\frac{1}{\varepsilon}\right) + 1.$$

Используем критерий Колмогорова для проверки корректности работы датчика. За гипотезу H_0 примем случай совпадения наблюдаемой функции распределения $F_n(x)$ с теоретической $F(x)$, за H_1 —отрицание H_0. Также определим статистику Колмогорова

$$T_n = \sqrt{n} D_n = \sqrt{n} \cdot \sup_{x \in [0,1]} |F_n(x) - F(x)|.$$

Для проверки гипотезы через известное распределение статистики Колмогорова $F_k(T)$ и заданный уровень значимости α ($\mathbb{P}(H_1 \mid H_0)$) рассчитаем $p_{value} = 1 - F_k(T_n)$, сравним его с α:

$$p_{value} \vee \alpha,$$

здесь знак $\leqslant$ будет говорить в пользу отклонения гипотезы H_0.

В программной реализации p_{value} будем определять с задаваемой пользователем точностью ε. Для этого рассмотрим саму

$$F_k(T) = 1 + 2\sum_{s=1}^{\infty} (-1)^s e^{-2s^2T^2}$$

и найдём остаток её ряда, дающий вклад $< \varepsilon$.

$$2\sum_{s=n}^{\infty} (-1)^s e^{-2s^2T^2} \leqslant 2\sum_{s=n}^{\infty} e^{-2s^2T^2} \leqslant 2\sum_{s=n}^{\infty} e^{-s^2T^2} \leqslant$$

$$\leqslant \left\{ -s^2T^2 < -s \Leftrightarrow 1 < sT^2 \Leftrightarrow s > \frac{1}{T^2} \right\} \leqslant$$

$$\leqslant 2\sum_{s=n}^{\infty} e^{-s} = \frac{2e^{-n}}{1-\frac{1}{e}} = \frac{2}{e^{n-1}(e-1)} \Rightarrow$$

$$\Rightarrow \frac{2}{e^{n-1}(e-1)} < \varepsilon \Leftrightarrow \frac{2}{\varepsilon(e-1)} \leqslant e^{n-1} \Leftrightarrow \ln\left(\frac{2}{\varepsilon(e-1)}\right) < n-1 \Rightarrow$$

$$\Rightarrow n > \max\left\{\ln\left(\frac{2}{\varepsilon(e-1)}\right) + 1,\ \frac{1}{T^2}\right\} \tag{4}$$

Теорема 2. (***Теорема Колмогорова***) *Пусть $X_1, \ldots, X_n, \ldots$ — бесконечная выборка из распределения, задаваемого непрерывной функцией $F(x)$. Пусть $F_n(x)$ — выборочная функция распределения, построенная на первых n элементах выборки. Тогда*

$$\sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to K$$

по распределению при $n \to +\infty$, где K — случайная величина, имеющая распределение Колмогорова.

Доказательство представлено в [6]. Результаты проверки гипотезы с $\alpha = 5\%$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^3	10^3	0.954
10^3	10^4	0.9542
10^4	10^3	0.951
10^4	10^4	0.9529

2.2

Свойство симметричности канторовых случайных величин относительно $\frac{1}{2}$ эквивалентно совпадению распределений для X и $1 - X$.

$$1 - X = 1 - \sum_{i=1}^{\infty} \frac{2y_i}{3^i} = \sum_{i=1}^{\infty} \frac{2(1-y_i)}{3^i} = \sum_{i=1}^{\infty} \frac{2z_i}{3^i},$$

здесь $z_i \sim Ber(0.5)$, а значит распределения действительно идентичны.

Проверим свойство самоподобия относительно деления на 3 для канторовых случайных величин, то есть, что условное распределение X при условии $X \in [0, \frac{1}{3}]$ совпадает с распределением $\frac{X}{3}$.

$$\frac{X}{3} = \frac{1}{3} \cdot \sum_{i=1}^{\infty} \frac{2y_i}{3^i} = \sum_{i=1}^{\infty} \frac{2y_i}{3^{i+1}} = \sum_{i=2}^{\infty} \frac{2y_i}{3^i} = X,$$

ведь $y_1 = 0 \Leftrightarrow X \in [0, \frac{1}{3}]$.

Реализуем численно критерий Смирнова и убедимся в корректности работы программы,

посредством проверки указанных выше свойств (так как теоретически доказана их правильность, то частота принятия критерия должна быть близка к 1 - α).

Действовать будем как и в случае с критерием Колмогорова, только статистику возьмём

$$T_{nm} = \sqrt{\frac{nm}{n+m}} D_{nm} = \sqrt{\frac{nm}{n+m}} \cdot \sup_{x \in [0,1]} |F_n(x) - F_m(x)|,$$

тут $F_n(x)$, $F_m(x)$ — эмпирические функции распределения рассматриваемых выборок, а n и m — их длины.

Теорема 3. (***Теорема Смирнова***) *Пусть $F_n^1(x)$ и $F_m^2(x)$ — эмпирические функции распределения с объёмами выборок n и m соответственно случайной величины Y. Тогда, если $F(x) \in C^1(\mathbb{R})$, то*

$$\lim_{n,m \to +\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}} D_{nm} \leqslant t \right) = K(t) = 1 + 2 \sum_{s=1}^{\infty} (-1)^s e^{-2s^2 t^2}, \quad \forall t > 0,$$

где $D_{nm} = \sup\limits_{x \in \mathbb{R}} |F_n^1(x) - F_m^2(x)|$.

Доказательство представлено в [3]. Результаты проверки, отталкиваясь от свойства симметричности при $\alpha = 5\%$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^3	10^3	0.959
10^3	10^4	0.9588
10^4	10^3	0.945
10^4	10^4	0.9557

Результаты проверки, отталкиваясь от свойства самоподобия при $\alpha = 5\%$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^3	10^3	0.96
10^3	10^4	0.9596
10^4	10^3	0.946
10^4	10^4	0.9578

2.3

Математическое ожидание канторовой случайной величины:

$$\mathbb{E}X = \mathbb{E}\sum_{i=1}^{\infty}\frac{2y_i}{3^i} = \sum_{i=1}^{\infty}\frac{2\mathbb{E}y_i}{3^i} = \sum_{i=1}^{\infty}\frac{1}{3^i} = \frac{\frac{1}{3}}{1-\frac{1}{3}} = \frac{1}{2}.$$

Дисперсия:

$$\mathbb{D}X = \mathbb{D}\sum_{i=1}^{\infty}\frac{2y_i}{3^i} = \sum_{i=1}^{\infty}\frac{4\mathbb{D}y_i}{3^{2i}} = \sum_{i=1}^{\infty}\frac{1}{3^{2i}} = \frac{\frac{1}{9}}{1-\frac{1}{9}} = \frac{1}{8},$$

пользовались выше тем, что для бернуллиевских y_i известны $\mathbb{E}y_i = \frac{1}{2}$, $\mathbb{D}y_i = \frac{1}{4}$.

Задание 3

3.1

Определение 6. *Случайная величина X имеет экспоненциальное распределение с параметром $\lambda > 0$, если её функция распределения имеет вид:*

$$F_x(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geqslant 0, \\ 0, t < 0. \end{cases}$$

Теорема 4. *Пусть функция $F(x)$ непрерывна и монотонно возрастает на $\mathbb{R}$, $\lim\limits_{x\to-\infty} F(x) = 0$, $\lim\limits_{x\to+\infty} F(x) = 1$, случайная величина $Y \sim \mathbb{U}[0,1]$ распределение, то случайная величина $X = F^{-1}(Y)$ имеет функцию распределения $F_x(x) = F(x)$. (док-во см. в [3]).*

Воспользуемся этой теоремой для построения датчика экспоненциального распределения:

$$F_x(t) = 1 - e^{-\lambda t} \Rightarrow F_x^{-1}(y) = -\frac{1}{\lambda}\ln(1-y)$$

определяя y как реализацию некоторой случайной величины $Y \sim \mathbb{U}[0,1]$, мы попадём в условие теоремы 4, следовательно, получим искомый способ моделирования.

Свойство 2. *Если $Y \sim Exp(\lambda)$, то $\mathbb{P}(Y > m+n \mid Y \geqslant m) = \mathbb{P}(Y > n)$, для любых неотрицательных a и b. Это называется свойством отсутствия памяти для экспоненциального распределения.*

Доказательство.

$$\mathbb{P}(Y > a + b \mid Y \geqslant a) = \frac{\mathbb{P}(Y > a + b)}{\mathbb{P}(Y \geqslant b)} = \frac{1 - (1 - e^{-\lambda(a+b)})}{1 - (1 - e^{-\lambda b})},$$

покажем, что эта вероятность совпадает с $\mathbb{P}(Y > a) = 1 - (1 - e^{-\lambda a})$, или, что

$$1 - (1 - e^{-\lambda(a+b)}) = (1 - (1 - e^{-\lambda b}))(1 - (1 - e^{-\lambda a})) \Leftrightarrow$$

$$\Leftrightarrow e^{-\lambda(a+b)} = e^{-\lambda b} e^{-\lambda a}.$$

□

Теорема 5. *Пусть $X_1, \ldots, X_n$ независимые случайные величины, и $X_i \sim Exp(\lambda_i)$. Тогда $Y = \min\limits_{i=1,\ldots,n} X_i \sim Exp(\sum_{i=1}^n \lambda_i)$.*

Доказательство.

$$\mathbb{P}(Y < t) = 1 - \mathbb{P}(Y \geqslant t) = 1 - \prod_{i=1}^{n} \mathbb{P}(X_i \geqslant t) = 1 - \prod_{i=1}^{n} (1 - F_{X_i}(t)) = 1 - \prod_{i=1}^{n} e^{-\lambda_i t} = 1 - e^{-(\sum_{i=1}^n \lambda_i)t}$$

□

3.2

Определение 7. *Случайная величина X имеет распределение Пуассона с параметром $\lambda > 0$, если*

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

Строить датчик будем на основе следующей теоремы.

Теорема 6. *Пусть $\eta_1, \ldots, \eta_n, \ldots$ независимые случайные величины, такие, что $\eta_i \sim Exp(1)$. Если*

$$X = \max\left\{ k \;\middle|\; \sum_{i=1}^{k} \eta_i < \lambda \right\},$$

то случайная величина $X \sim Poiss(\lambda)$. В случае, $\eta_1 \geqslant \lambda$ полагаем $X = 0$.

Доказательство можно найти в [2].

Таким образом, мы будем сэмплировать $Exp(1)$, пока их сумма не превзойдёт заданного параметра λ и возвращать число генераций минус 1.

3.3

Теорема 7. *Пусть случайная величина* $X \sim Binom(n, p)$. *Пусть* $np = \lambda = \text{const}$. *Тогда при* $n \to \infty$

$$\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Доказательство данной теоремы можно найти в [2].

С помощью критерия хи-квадрат Пирсона убедимся в корректности работы построенного датчика. Для этого определим статистику

$$\chi_n^2 = n \sum_{i=1}^{k} \frac{\left(\frac{n_i}{n} - p_i\right)^2}{p_i},$$

в которой p_i задают вероятности предполагаемого распределения (H_0), в нашем случае это $Poiss(\lambda)$, n_i — число встреченных в выборке значений i-го номинала, k — количество различных номиналов значений в совокупности наблюдений. Получается, что разность $(\frac{n_i}{n} - p_i)$ как бы находит отклонение эмпирической вероятности наблюдать некоторое i-е значение от теоретической.

Также введём функцию плотности распределения статистики χ^2:

$$g(s) = \frac{1}{2^{\frac{r}{2}} \cdot \Gamma(\frac{r}{2})} s^{\frac{r}{2}-1} e^{-\frac{s}{2}}, \tag{5}$$

с $r = k - 1$ — числом степеней свободы.

Проверяемую гипотезу H_0 будем отвергать в случае, если p_{value} окажется меньше α. Так как критерий имеет правостороннюю критическую область, то

$$p_{value} = \int_{\chi_n^2}^{\infty} g(s)\,\mathrm{d}s.$$

Примем c равным $\dfrac{1}{2^{\frac{r}{2}} \cdot \Gamma(\frac{r}{2})}$ и найдём p_{value} с точностью ε, для этого определим такое l, что

$$c \cdot \int_{l}^{\infty} s^{\frac{r}{2}-1} e^{-\frac{s}{2}}\,\mathrm{d}s < \varepsilon.$$

Обозначим для удобства $k = \frac{r}{2} - 1$, тогда нам необходимо показать, что

$$\int_l^\infty s^k e^{-\frac{s}{2}}\,\mathrm{d}s < \frac{\varepsilon}{c}.$$

Взяв $l : l^k < e^{\frac{l}{4}}$, мы придём к

$$\int_l^\infty s^k e^{-\frac{s}{2}}\,\mathrm{d}s < \int_l^\infty e^{-\frac{s}{4}}\,\mathrm{d}s = -4e^{-\frac{s}{4}}\Big|_l^\infty = 4e^{-\frac{l}{4}} < \frac{\varepsilon}{c} \Leftrightarrow$$

$$\Leftrightarrow -l < 4\ln\left(\frac{\varepsilon}{4c}\right) \Leftrightarrow l > -4\ln\left(\frac{\varepsilon}{4c}\right).$$

Теперь поподробнее рассмотрим неравенство

$$l^k < e^{\frac{l}{4}}. \tag{6}$$

Воспользуемся известным фактом, что степенная функция растёт медленнее показательной с аргументом > 1. Поэтому, если мы найдём некоторое l^*, для которого неравенство (6) выполняется, то оно будет справедливо и при больших значениях. Преобразуем неравенство для удобства следующим образом:

$$l^k < e^{\frac{l}{4}} \Leftrightarrow 4k\ln l < l.$$

Нетрудно проверить, что $l^* = \max\{(4k)^2, e\}$ подходит.

- Если $e \geqslant (4k)^2$.

  Тогда очевидно выполняется

  $$4k\ln(e) = 4k < e.$$

- Если $e < (4k)^2$.

  $$4k\ln(4k)^2 < (4k)^2 \Leftrightarrow 2\ln(4k) < 4k,$$

  что выполнено в силу совокупности двух фактов: указанного выше о сравнении скорости роста двух типов функций и второго: $2\ln(s) < s$ при $s = 1$.

Поэтому l положим равной $\max\left\{e, (4k)^2, -4\ln\left(\frac{\varepsilon}{4c}\right)\right\}$.

Результаты проверки с $\alpha = 5\%$, $\varepsilon = 10^{-10}$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^3	10^3	0.954
10^3	10^4	0.9552
10^4	10^3	0.954
10^4	10^4	0.959

3.4

Построим датчик стандартного нормального распределения методом моделирования случайных величин парами с переходом в полярные координаты. Для этого рассмотрим случайные величины $\xi \sim \mathcal{N}(0,1)$, $\eta \sim \mathcal{N}(0,1)$. Совместная функция распределения имеет вид:

$$\mathbb{P}(\xi < x, \eta < y) = \frac{1}{2\pi} \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} e^{-\frac{(x_1^2+x_2^2)}{2}} \,\mathrm{d}x_1 \mathrm{d}x_2 = \frac{1}{2\pi} \iint\limits_{\substack{r\cos\phi \,<\, x \\ r\sin\phi \,<\, y}} r e^{-\frac{(r^2)}{2}} \,\mathrm{d}r\mathrm{d}\phi = \frac{1}{4\pi} \iint\limits_{\substack{r\cos\phi \,<\, x \\ r\sin\phi \,<\, y}} e^{-\frac{w}{2}} \,\mathrm{d}w\mathrm{d}\phi.$$

Получили общую функцию распределения случайных величин $w \sim Exp(\frac{1}{2})$, $\phi \sim U[0, 2\pi]$. Таким образом совместное распределение ξ и η совпадает с таковым у $\sqrt{w}\cos\phi$, $\sqrt{w}\sin\phi$. Поэтому случайные величины ξ и η можно сэмплировать как:

$$\xi = \sqrt{w}\cos\phi, \quad \eta = \sqrt{w}\sin\phi.$$

Проверим равенство нулю математических ожиданий $\mathbb{E}\sqrt{w}\cos\phi$ и $\mathbb{E}\sqrt{w}\sin\phi$ при помощи t-критерия Стьюдента, статистика которого имеет вид

$$T_n = \sqrt{n}\frac{\overline{X} - m}{s_X},$$

здесь $\overline{X}$ — выборочное среднее, m — предполагаемое значение математического ожидания (в нашем случае 0), s_X — корень из выборочной дисперсии. Распределение данной статистики для этого критерия задаётся как

$$F(T) = \int\limits_{-\infty}^{T} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\cdot\Gamma(\frac{\nu}{2})} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \,\mathrm{d}x,$$

где ν — число степеней свободы (для рассматриваемого случай $\nu = n - 1$). Обозначим $c = \dfrac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \cdot \Gamma(\frac{\nu}{2})}$. Проверяемую гипотезу будем отвергать, если

$$p^*_{value} = \int\limits_{|T|}^{+\infty} c \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x < \frac{\alpha}{2}$$

в силу симметрии t-распределения. Вычислять значение p^*_{value} будем с точностью ε, для этого найдём l такое, что

$$\int\limits_{l}^{+\infty} c \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x < \varepsilon.$$

Заметим, что

$$\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \left(\frac{\nu + x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \left(\frac{\nu}{\nu + x^2}\right)^{\frac{\nu+1}{2}} < \left(\frac{\nu}{x^2}\right)^{\nu},$$

считаем $\nu \geqslant 1$, то есть $n \geqslant 2$. Тогда получим

$$\int\limits_{l}^{+\infty} \frac{\nu^\nu}{x^{2\nu}} \mathrm{d}x < \frac{\varepsilon}{c} \Leftrightarrow \nu^\nu \cdot \frac{x^{-(2\nu-1)}}{-(2\nu-1)}\Bigg|_{l}^{+\infty} < \frac{\varepsilon}{c} \Leftrightarrow \nu^\nu \frac{l^{-(2\nu-1)}}{2\nu-1} < \frac{\varepsilon}{c} \Leftrightarrow l > \sqrt[(2\nu-1)]{\frac{\nu^\nu \cdot c}{(2\nu-1)\varepsilon}}.$$

А значит, достаточно определять

$$p^*_{value} = \int\limits_{|T|}^{l} c \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x \vee \frac{\alpha}{2}.$$

Результаты проверки с $\alpha = 5\%$, $\varepsilon = 10^{-10}$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^2	10^2	0.96
10^2	10^3	0.956
10^2	10^4	0.9476

Далее сформулируем критерий Фишера, равенства дисперсий и проверим его на смоделированных нами случайных величинах ξ и η. За гипотезу H_0 примем $\sigma^2_\xi = \sigma^2_\eta$ для независимых нормальных случайных величин. Используемая статистика

$$T_{nm} = \frac{S^2_\xi}{S^2_\eta},$$

здесь S^2_ξ, S^2_η — выборочные дисперсии (несмещённые). Пусть в первой выборке n элементов,

а во второй — m, тогда, если проверяемая гипотеза верна, то T_n имеет F-распределение со степенями свободы $(n-1)$ и $(m-1)$. Кроме того, так как критерий двусторонний, то для принятия H_0 станем проверять неравенство

$$F_{1-\frac{\alpha}{2}}(n-1, m-1) > T_{nm} > F_{\frac{\alpha}{2}}(n-1, m-1).$$

В силу сложности получения аналитической оценки значения квантилей F-распределения, для их определения воспользуемся библиотечной функцией. Ниже показаны результаты применения критерия с $\alpha = 5\%$:

Обьём выборки	Количество испытаний	Частота принятия гипотезы
10^3	10^3	0.955
10^3	10^4	0.9524
10^4	10^3	0.941
10^4	10^4	0.9507

Задание 4

4.1

Определение 8. *Случайная величина ξ имеет распределение Коши с параметрами сдвига x_0 и масштаба $\gamma > 0$, если её функция распределения имеет вид*

$$F_\xi(x) = \frac{1}{\pi}\operatorname{arctg}\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}.$$

Датчик распределение Коши легко построить через:

$$F_\xi^{-1}(y) = x_0 + \gamma \operatorname{tg}\left(\pi\left(y - \frac{1}{2}\right)\right).$$

Воспользовавшись теоремой 4 получим, что если $Y \sim U[0,1]$, то $X = F_\xi^{-1}(Y)$ имеет распределение Коши.

4.2

Пусть заданы плотности двух распределений $f_1(x)$, $f_2(x)$. Суть метода фон Неймана заключается в том, что бы определив константу k: $kf_1(x) \geqslant f_2(x)\ \forall x$ сгенерировать ξ по сле-

дующему алгоритму:

1. Сэмплируется элемент y соответствующий $f_1(x)$.
2. Вычисляется $R = \dfrac{f_2(y)}{kf_1(y)}$.
3. Генерируется $\eta(y) \sim Ber(R)$ и если она равна 1, то значение x возвращаем как результат генерации случайного элемента, имеющего плотность распределения $f_2(x)$, иначе повторяем заново.

В нашем случае:

$$f_1(x) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2},\ f_2(x) = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}}.$$

Обоснуем метод фон Неймана.

Теорема 8. *пусть P и Q — вероятностные меры, заданные на измеримом пространстве $(\mathcal{X}, \mathcal{A})$, причём $P(A) < kQ(A)$ для всех $A \in \mathcal{A}$ для некоторого вещественного числа k. Пусть ν — бернуллиевская случайная величина с параметром $\dfrac{\mathrm{d}P}{k\mathrm{d}Q}$. Тогда*

$$\mathbb{P}(X \in A \mid \nu = 1) = P(A).$$

Доказательство.

$$\mathbb{P}(X \in A \mid \nu = 1) = \frac{\mathbb{P}(X \in A,\ \nu = 1)}{\mathbb{P}(\nu = 1)} = \frac{\int_A \mathbb{P}(X = x,\ \nu = 1)\,\mathrm{d}Q(x)}{\int_\Omega \mathbb{P}(X = x,\ \nu = 1)\,\mathrm{d}Q(x)} = \frac{\int_A \dfrac{\mathrm{d}P}{k\mathrm{d}Q}\,\mathrm{d}Q(x)}{\int_\Omega \dfrac{\mathrm{d}P}{k\mathrm{d}Q}\,\mathrm{d}Q(x)} = P(A).$$

□

Функция `normplot` сопоставляет квантилям рассматриваемой выборки квантили нормального распределения. Таким образом, если выборка соответствует нормальному распределению, то мы будем наблюдать линейную зависимость на графике, а в случае распределения, отличного от нормального, получим квантили, которые сильно отличаются от таковых у нормального.

4.3

Скорость моделирования стандартного нормального распределения методом фон Неймана сильно ниже, чем у метода пар. Убедимся в этом эмпирически, результаты на рис.

(будет в чистовой версии).

Задание 5

5.1

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

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

Теорема 9. (*Закон больших чисел*) *Пусть $X_1, X_2, \ldots$ — независимые, одинаково распределенные случайные величины, $\mathbb{E}X_i = \mu\ \forall i \in \mathbb{N}$, $|\mu| < \infty$, $\mathbb{D}X_i \leqslant c$, $S_n = X_1 + \cdots + X_n$. Тогда $\dfrac{S_n}{n} \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mu$, т. е.*

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \underset{n\to\infty}{\to} 0.$$

Теорема 10. (*Центральная предельная теорема*) *Пусть $X_1, X_2, \ldots$ — независимые, одинаково распределенные случайные величины, с конечными $\mathbb{E}X_i = a$ и $\mathbb{D}X_i = \sigma^2$. Тогда*

$$\mathbb{P}\left(\frac{S_n - na}{\sigma\sqrt{n}} < z\right) \underset{n\to\infty}{\to} \Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{z} e^{-\frac{y^2}{2}}\, \mathrm{d}y.$$

Доказательство данных теорем представлено в [2].

5.2

Доверительный интервал для среднего будем строить отталкиваясь от t-распределения, для дисперсии — χ^2-распределения.

- Для неизвестного среднего μ при неизвестной дисперсии σ^2:
  по теореме Фишера ([3]) для независимой выборки из нормального распределения справедливо:
  $$\frac{\sqrt{n}(\overline{x} - \mu)}{\sigma} \sim \mathcal{N}(0,1),\ \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 \sim \chi^2_{(n-1)}.$$
  Следовательно отношение
  $$\frac{\sqrt{n}(\overline{x} - \mu)}{\sqrt{\frac{1}{n-1} \cdot \sum\limits_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{\sqrt{n}(\overline{x} - \mu)}{s}$$

есть не что иное, как распределение Стьюдента с $(n-1)$ степенью свободы. Тогда с вероятностью γ (рассматриваемая надёжность) выполнено неравенство

$$t^{(n-1)}_{\frac{1-\gamma}{2}} < \frac{\sqrt{n}(\overline{x}-\mu)}{s} < t^{(n-1)}_{\frac{1+\gamma}{2}},$$

тут t^n_α — квантиль t-распределения с n степенями свободы. Откуда легко получить, пользуясь симметрией распределения:

$$\overline{x} - \frac{s}{\sqrt{n}} t^{(n-1)}_{\frac{1+\gamma}{2}} < \mu < \overline{x} + \frac{s}{\sqrt{n}} t^{(n-1)}_{\frac{1+\gamma}{2}}.$$

- Для неизвестной дисперсии σ^2 при неизвестном среднем μ:
  из сказанного выше
  $$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 \sim \chi^2_{(n-1)},$$
  а значит,
  $$h^{(n-1)}_{\varepsilon_1} < \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 < h^{(n-1)}_{1-\varepsilon_2}$$
  наш искомый интервал, если взять $\varepsilon_1 + \varepsilon_2 = 1-\gamma$. h^n_α — это квантиль хи-квадрат распределения с n степенями свободы. Из представленного неравенства получаем:
  $$\frac{(n-1)s^2}{h^{(n-1)}_{1-\varepsilon_2}} < \sigma^2 < \frac{(n-1)s^2}{h^{(n-1)}_{\varepsilon_1}}.$$

5.3

Пусть $X_i \sim C(a,b)\ \forall i \in \mathbb{N}$. Рассмотрим график, по нему видно, что $\frac{S_n}{n}$ не имеет предела (ЗБЧ не выполнен), что неудивительно, ведь у случайной величины, имеющей распределение Коши отсутствует математическое ожидание

$$\mathbb{E}X_i = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{bx}{(x-a)^2+b^2}\,\mathrm{d}x = \frac{b}{2\pi}\ln((x-a)^2+b^2)\Big|_{-\infty}^{+\infty} - \frac{a}{\pi}\operatorname{arctg}\left(\frac{a-x}{b}\right)\Bigg|_{-\infty}^{+\infty} = \infty - \infty.$$

Покажем, что закон распределения сумм $\frac{S_n}{n}$ есть $C(a,b)$. Для этого Вспомним понятие характеристической функции, которая задаётся как $\mathbb{E}e^{itX}$, а также тот факт, что существует взаимно однозначное соответствие между характеристическими функциями и функциями распределения, а они в свою очередь однозначно связаны со случайными величинами.

Непосредственно для X_i:

$$\mathbb{E}e^{itX_i} = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{b}{(x-a)^2+b^2} e^{itx}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \sqrt{\frac{2}{\pi}}\cdot\frac{b}{(x-a)^2+b^2}\cdot e^{itx}\,\mathrm{d}x.$$

Заметим, что выше представлено обратное преобразование Фурье к $\sqrt{\frac{2}{\pi}}\cdot\frac{b}{(x-a)^2+b^2}$. Поэтому вспомним основные свойства этого преобразования, приняв за $F(\xi)$ — образ рассматриваемой функции $f(t)$:

$$f(at) \;\to\; \frac{1}{a}F\left(\frac{\xi}{a}\right),$$

$$e^{-|t|} \;\to\; \sqrt{\frac{2}{\pi}}\cdot\frac{1}{1+\xi^2},$$

$$e^{itx_0}f(t) \;\to\; F(\xi - x_0).$$

Следовательно, легко проверить, что

$$e^{ita-|t|b} \;\to\; \sqrt{\frac{2}{\pi}}\cdot\frac{b}{(x-a)^2+b^2}$$

а значит — искомый прообраз. Отметим два важных свойства характеристической функции, которыми мы воспользуемся в дальнейшем:

1. Если x_1 и x_2 — независимы, а $f_{x_1}(t)$, $f_{x_2}(t)$ – их характеристические функции, то

$$f_{x_1+x_2}(t) = f_{x_1}(t)\cdot f_{x_2}(t).$$

2. Кроме того

$$f_{ax+b}(t) = e^{itb}f_x(at).$$

Отталкиваясь от озвученного выше

$$f_{S_n}(t) = \prod_{j=1}^{n} f_{x_j}(t) = \prod_{j=1}^{n} e^{ita-|t|b} = e^{n(ita-|t|b)} \Rightarrow$$

$$\Rightarrow f_{\frac{S_n}{n}}(t) = f_{S_n}\left(\frac{t}{n}\right) = e^{(ita-|t|b)}.$$

Поэтому среднее арифметическое случайных величин, имеющих одно и то же распределение Коши, распределено идентичным образом.

Задание 6

6.1

Представим интеграл

$$\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\cdots\int\limits_{-\infty}^{+\infty}\frac{e^{-\left(x_1^2+\ldots+x_{10}^2+\frac{1}{2^7\cdot x_1^2\cdot\ldots\cdot x_{10}^2}\right)}}{x_1^2\cdot\ldots\cdot x_{10}^2}\,\mathrm{d}x_1\ldots\,\mathrm{d}x_{10}$$

в виде

$$\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\cdots\int\limits_{-\infty}^{+\infty}\sqrt{\pi^{10}}\frac{e^{-\left(\frac{1}{2^7\cdot x_1^2\cdot\ldots\cdot x_{10}^2}\right)}}{x_1^2\cdot\ldots\cdot x_{10}^2}\cdot\frac{1}{\sqrt{\pi^{10}}}e^{-\left(x_1^2+\ldots+x_{10}^2\right)}\,\mathrm{d}x_1\ldots\,\mathrm{d}x_{10}.$$

Примем

$$f(x)=\sqrt{\pi^{10}}\cdot\frac{e^{-\left(\frac{1}{2^7\cdot x_1^2\cdot\ldots\cdot x_{10}^2}\right)}}{x_1^2\cdot\ldots\cdot x_{10}^2},\ g(x)=\frac{1}{\sqrt{\pi^{10}}}e^{-\left(x_1^2+\ldots+x_{10}^2\right)}.$$

Отметим, что $g(x)$ — плотность многомерного вектора $(x_1,\ldots,x_{10})$, если $x_i\sim\mathcal{N}(0,\frac{1}{2})$. Тогда исходный интеграл можно записать как

$$I=\mathbb{E}f(x_1,\ldots,x_{10}).$$

Воспользуемся законом больших чисел. Для этого рассмотрим выборку

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

Придём к

$$\frac{S_n}{n}=\frac{1}{n}\sum_{i=1}^{n}f(x^i)\underset{n\to\infty}{\to}I.$$

Оценим погрешность метода Монте-Карло.

$$\mathbb{P}\left(\left|\frac{S_n}{n}-I\right|<\varepsilon\right)=\mathbb{P}\left(\left|\frac{S_n-nI}{ns}\right|<\frac{\varepsilon}{s}\right)=\mathbb{P}\left(-\frac{\varepsilon}{s}<\frac{S_n-nI}{ns}<\frac{\varepsilon}{s}\right)=$$

$$=\mathbb{P}\left(-\sqrt{n}\frac{\varepsilon}{s}<\sqrt{n}\frac{S_n-nI}{ns}<\sqrt{n}\frac{\varepsilon}{s}\right)=\Phi\left(\sqrt{n}\frac{\varepsilon}{s}\right)-\Phi\left(-\sqrt{n}\frac{\varepsilon}{s}\right)=2\Phi\left(\sqrt{n}\frac{\varepsilon}{s}\right)-1=\gamma.$$

Где ε — погрешность вычисления I, γ — вероятность, с которой мы будем иметь расхождения с I в пределах ε, $\Phi(x)$ — функция распределения нормальной случайной величины, s —

несмещённая выборочная дисперсия. Напомню, что

$$\Phi(x) = \int\limits_{-\infty}^{x} e^{-\frac{y^2}{2}}\,\mathrm{d}y.$$

Поясним, каким образом будет происходить оценка погрешности. Изначально нам задаётся коэффициент доверия γ, отталкиваясь от него решаем задачу

$$2\Phi\left(x_{\frac{1+\gamma}{2}}\right) - 1 = \gamma,$$

находя при этом $x_{\frac{1+\gamma}{2}}$, откуда получаем

$$\varepsilon = \frac{s \cdot x_{\frac{1+\gamma}{2}}}{\sqrt{n}}.$$

Стоит отметить, что конечность математического ожидания и дисперсии многомерной случайной величины $f(x)$ (можно доопределить до непрерывной $\Rightarrow$ борелевское отображение $\Rightarrow$ сл. в.) следует из её ограниченности. А значит, можем воспользоваться ЦПТ, при этом опираясь на состоятельность в оценке эталонной дисперсии выборочной.
Таблица испытаний с $\gamma = 0.95$:

Обьём выборки	Результат	Погрешность	Время работы
10^4	140.886	19.516	10.8 ms
10^5	128.210	7.187	76.3 ms
10^6	121.768	4.241	777 ms
10^7	124.991	0.708	10.6 s

6.2

Сведём задачу к собственному интегралу Римана заменой

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

Получим

$$I = \left(\frac{\pi}{2}\right)^{10} \cdot \int\limits_{-1}^{1} \dots \int\limits_{-1}^{1} \frac{\exp\left\{-\left(\sum_{i=1}^{10} \operatorname{tg}^2(\frac{\pi}{2}t_i) + \frac{1}{2^7 \cdot \prod_{i=1}^{10} \operatorname{tg}^2(\frac{\pi}{2}t_i)}\right)\right\}}{\prod_{i=1}^{10} \sin^2(\frac{\pi}{2}t_i)}\,\mathrm{d}t_1 \dots \mathrm{d}t_{10}.$$

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

$$-1 = t_0 < t_1 < \ldots < t_N = 1,\ t_i = -1 + i\frac{2}{N},\ i = \overline{1, N}.$$

Обозначим через $f(y_1, \ldots, y_{10})$ подынтегральную функцию интеграла I. Будем использовать метод срединных прямоугольников. Для этого нам необходимо выбрать середины нашего разбиения:

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

Тогда наш интеграл приближённо можно посчитать следующим образом:

$$I_N = \left(\frac{\pi}{2}\right)^{10} \sum_{i_1=1}^{N} \ldots \sum_{i_{10}=1}^{N} f(y_1, \ldots, y_{10}).$$

Оценка погрешности метода прямоугольников на равномерной сетке имеет вид:

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

Приведем таблицу зависимости результата от количества точек разбиения отрезка:

N	Результат	Время работы
4	0.087	18.3 s
6	272.603	18min 49s
8	183.489	7h 15min 33s

Вывод: метод Монте–Карло работает намного эффективнее по скорости и точности, чем метод квадратур.

Список литературы

[1] Смирнов С. Н. Лекции по курсу «Стохастический анализ и моделирование», 2023–2024.

[2] Ширяев А. Н. Вероятность, Наука. М.: 1989.

[3] Востриков И. В. Лекции по курсу «Теория идентификации», 2008.

[4] Кропачёва Н. Ю., Тихомиров А. С. Моделирование случайных величин: метод указания, НовГУ им. Ярослава Мудрого, 2004.

[5] Колмогоров А. Н. Избранные труды, в 6 томах. Том 2. Теория вероятностей и математическая статистика. М., Математический институт им. В. А. Стеклова РАН.

[6] Феллер В. Введение в теорию вероятностей и её приложения, в 2-х томах. Т.1, М., Мир, 1984.