Санкт-Петербургский Политехнический Университет имени Петра Великого
Физико-Механический Институт

Отчет по лабораторной работе №1
по дисциплине
"Многомерный Статистический Анализ"
Вариант 12

Выполнил студент:
Тырыкин Я. А.
группа 5030102/80401
Преподаватель:
Павлова Л. В.

Санкт-Петербург
2022

Содержание

1 Формулировка задачи

В данной лабораторной работе требуется выполнить построение и обоснование модели распределения исследуемой случайной величины, заданной в виде некоторой выборки. Для данной случайной величины требуется:

- Найти выборочные характеристики: выборочное среднее, выборочную дисперсию, выборочные коэффициенты асимметрии и эксцесса.
- Построить эмпирическую функцию распределения (э. ф. р.) и нормированную гистограмму.
- На основе э. ф. р. построить доверительные полосы для теоретической функции распределения (т. ф. р.) с доверительными вероятностями 0.90 и 0.95.
- После анализа выборочных характеристик выдвинуть гипотезу о виде распределения исследуемой случайной величины и проверить ее на основе критерия χ^2 Фишера.
- Построить МП-оценки параметров случайной величины после того, как было принято решение о виде ее распределения.
- На основании полученных оценок построить гипотетические теоретические кривые т. ф. р. и плотности вероятности.

Будем анализировать данные, представленной выборкой из 60 вещественных чисел, расположенных в файле "Number_12.txt".

2 Построение выборочных параметров распределения

Вычислим следующие 4 выборочных параметра заданной случайной величины:

- Выборочное среднее
- Выборочную дисперсию
- Выборочный коэффициент асимметрии
- Выборочный эксцесс

Вычисление проведем по приведенным ниже формулам:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.618198$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = 2.17317$$

$$A = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n\sigma_x^3} = 1.18902$$

$$E = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n\sigma_x^4} = 1.12451$$

3 Построение э. ф. р. и нормированной гистограммы

Визуально распределение случайной величины, отраженной в исходных данных, можно представить в виде гистограммы и эмпирической функции распределения:

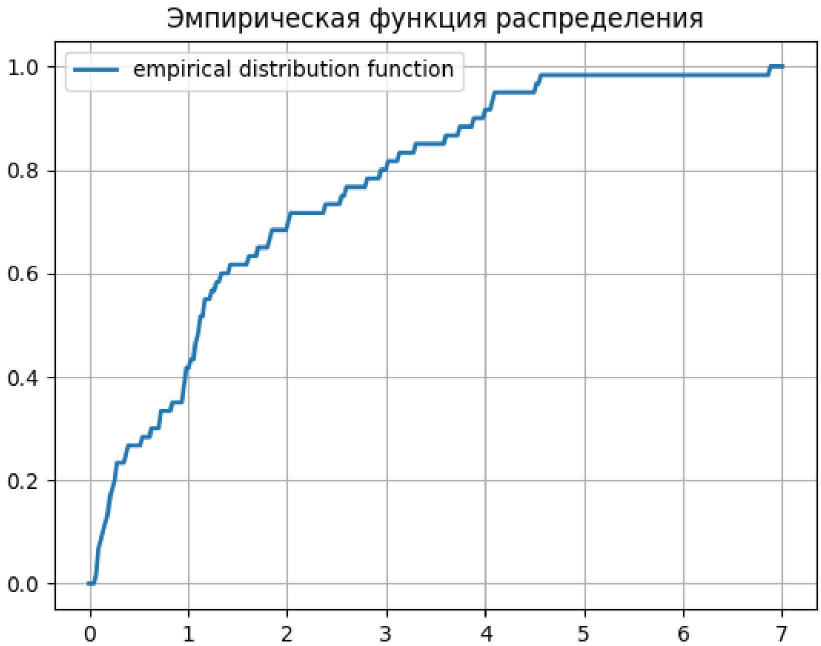

Рис. 1: Эмпирическая функция распределения случайной величины

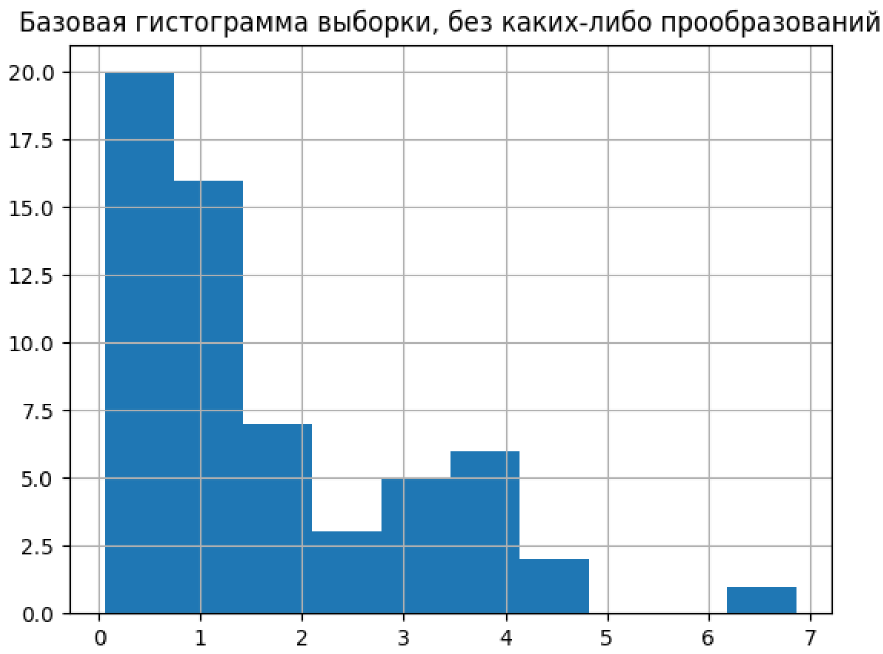

Рис. 2: Гистограмма исходной выборки случайной величины

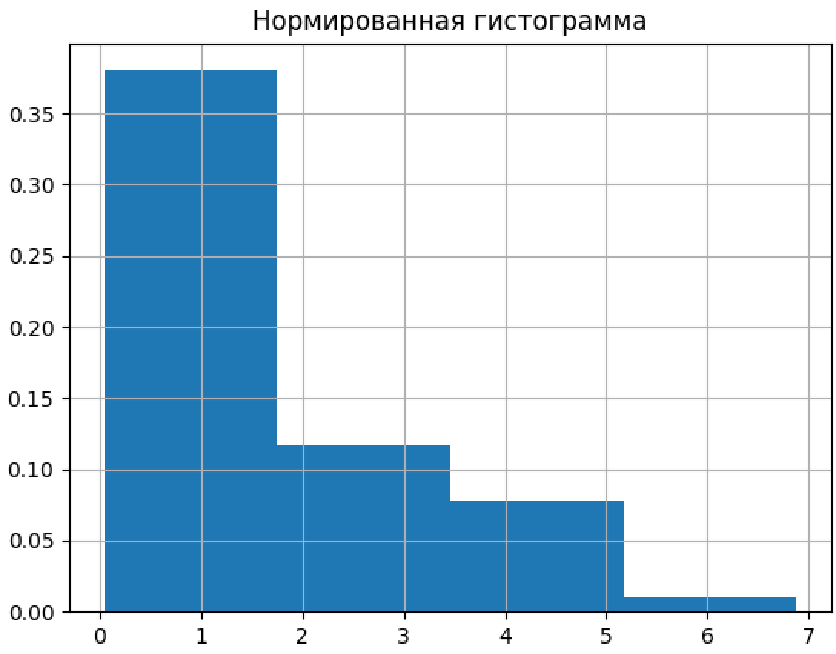

Рис. 3: Нормированная гистограмма случайной величины

4 Построение доверительных полос для теоретической функции распределения

Перед тем, как найти доверительную полосу для теоретической функции распределения, нужно привести формулу для эмпирической функции распределения, построенной в прошлом пункте. В качестве э. ф. р. нашей случайной величины будем пользоваться следующей формулой:

$$\hat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - X_i)$$

где $H(t)$ - функция Хэвисайда, определяемая следующим выражением:

$$H(t) = \begin{cases} 0, t < 0 \\ 1, t \geq 0 \end{cases}$$

Для построения полос, в которых с заданной вероятностью будет находиться наша непрерывная функция распределения случайной величины, воспользуемся теоремой Колмогорова. Данная теорема гласит, что для любого параметра $u > 0$ вероятность того, что значение статистики D_n, построенной на выборке размера n, не превосходит значения $\frac{u}{\sqrt{n}}$ стремится к значению распределения Колмогорова от данного параметра при бесконечном увеличении выборки ($n \to \infty$). Говоря проще, если статистика $\sqrt{n}D_n$ не превышает процентную точку распределения Колмогорова K_α заданного уровня значимости α, то гипотеза H_0 принимается (говорят, что принимается на уровне α).

Данная теорема подходит для нашего исследования, так как значения статистики D_n практически не меняется при $n \geq 20$. Статистика D_n определяется по следующей формуле:

$$D_n = sup\big\{|\hat{F_n}(t) - F(t)|\big\}, t \in R$$

где $\hat{F_n}(x)$ - выборочная функция распределения, построенная на выборке размера n исследуемой случайной величины, $F(t)$ - предполагаемая непрерывная функция распределения той же случайной величины.

Распределение Колмогорова опредляется известной формулой:

$$K(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$$

Математически теорема записывается следующим образом:

$$\forall u > 0 \lim_{n \to \infty} P\{\sqrt{n} D_n \leq u\} = K(u)$$

Преобразуя основное неравенство теоремы Колмогорова, получим вид доверительной полосы, которую нам необходимо построить:

$$max\{0, \hat{F_n}(t) - \frac{u_\gamma}{\sqrt{n}}\} \leq F(t) \leq min\{\hat{F_n}(t) + \frac{u_\gamma}{\sqrt{n}}, 0\}$$

К сожалению, программной реализации, удобной для самостоятельного нахождения квантилей распределения Коломгорова, найти не удалось, поэтому квантили уровней $1 - \gamma_1 = 1 - 0.90 = 0.1, 1 - \gamma_2 = 1 - 0.95 = 0.05$ (u_{γ_1} и u_{γ_2} соответственно) возьмем в виде известных табличных значений: $u_{1-\gamma_1} = 1.22$ и $u_{1-\gamma_2} = 1.36$.

Приведем полученные полосы в виде их визуального построения относительно э. ф. р.:

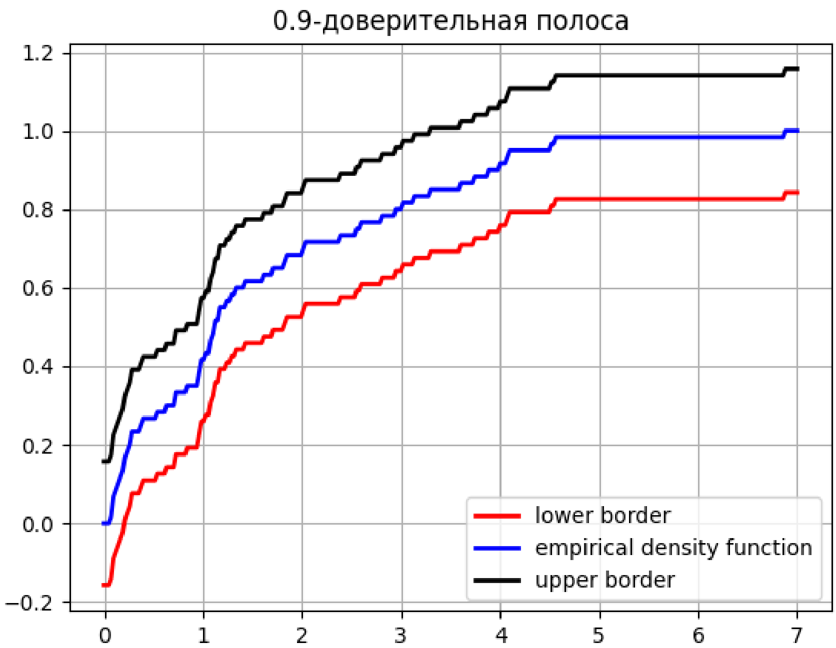

Рис. 4: Построение доверительной полосы с доверительной вероятностью 0.90

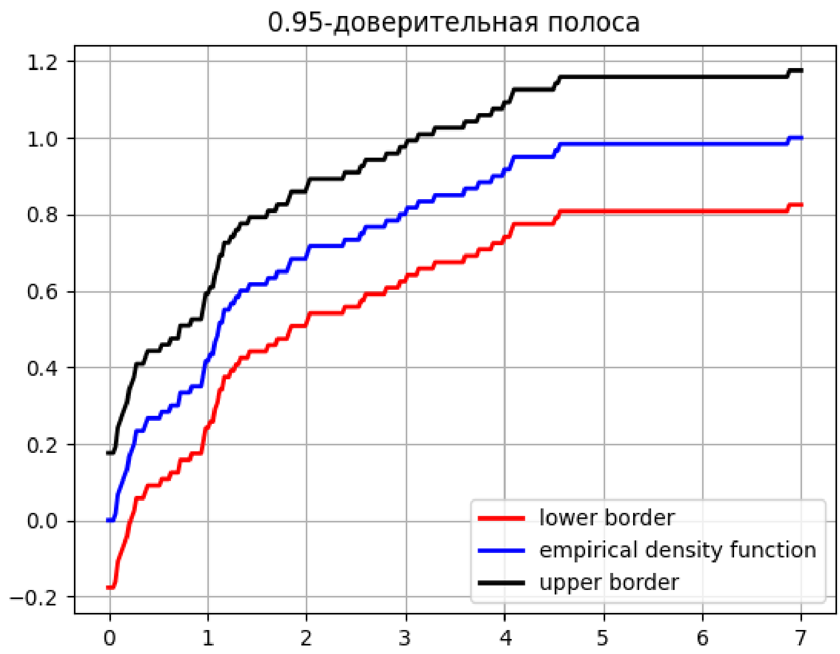

Рис. 5: Построение доверительной полосы с доверительной вероятностью 0.95

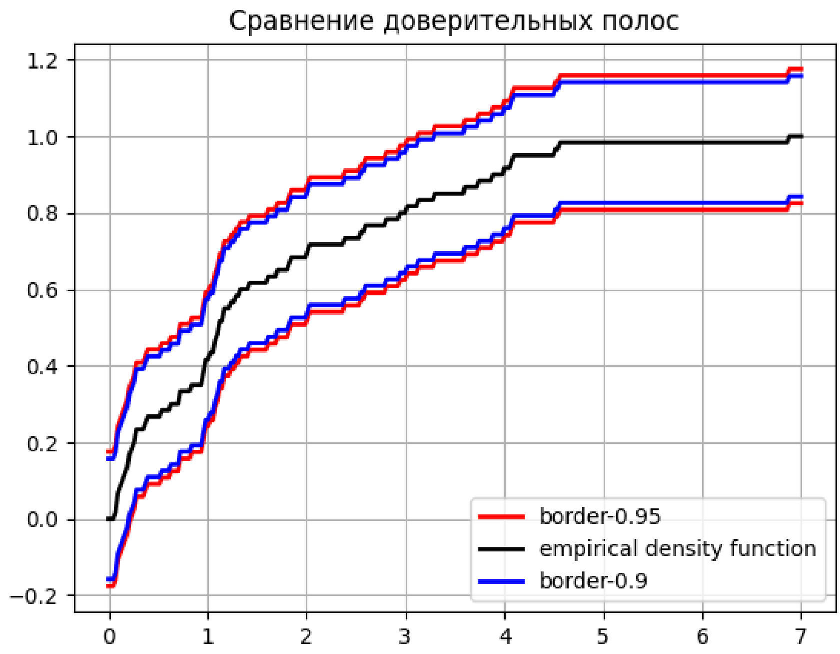

Рис. 6: Сравнение построеных выше доверительных полос

5 Проверка гипотезы о виде распределения рассматриваемой случайной величины

Исходя из того, что выборочный коэффициент асимметрии положителен, и из общего вида нормированной гистограммы и эмпирической функции распределения случайной величины, выдвигаем гипотезу H_0, заключающуюся в том, что наша случайная величина распределена либо по показательному закону с функцией распределения:

$$F(x,\lambda) = \begin{cases} 0, x < 0 \\ \lambda e^{-\lambda x}, x \geq 0 \end{cases}, \lambda > 0$$

либо по закону гамма-распределения (несмотря на то, что с левой части гистограммы нет промежутка убывания, по остальным параметрам данное распределение вполне может подойти) со следующей функцией распределения:

$$F(x,\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \alpha, \beta > 0$$

Данную гипотезу проверим при помощи критерия χ^2-Фишера.

Для этого будем строить статистику X_N^2, зависящую от параметров закона распределения случайной величины, обозначаемых $\theta-$вектором. Статистика вычисляется по следующей формуле:

$$X_n^2(\nu) = \sum_{i=1}^{N} \frac{(\nu_i - Np_i^0)^2}{Np_i^0}$$

где ν_i - элементы вектора частот ν попадания выборочных данных в интервал группировки Δ_j (интервалы группировки $\Delta_1, \Delta_2, ..., \Delta_N$ выбираются так, что покрывают всю область значений исследуемой выборки случайной величины); p_j^0 - элементы вектора реальных частот p^0 попадания случайной величины в интервал Δ_j

Приведем также математические обозначения для векторов частот:

$$\nu = (\nu_1, \nu_2, \ldots, \nu_N), \nu_j = \sum_{i=1}^{N} I(X_i \in \Delta_j), \sum_j \nu_j = N$$

$$p^0 = (p_1^0, p_2^0, \ldots, p_N^0), p_j^0 = P\big\{\xi \in \Delta_j | H_0\big\} = \int_{\Delta_k} dF(t,\theta)$$

В наших экспериментах число интервалов разбиения множества значений выборки случайной величины берется равным 5 и 10 соответственно. Векторы параметров Θ будут иметь размерность 2 (для гамма-распределения) и 1(для показательного распределения) соответственно: $\theta_\Gamma \in R^2, \theta_exp \in R^1$.

Далее вводим уровень значимости $\alpha \in (0,1)$ (в данной работе берется уровень значимости равный 0.05), после чего строится критическое множество $\Im_{1\alpha} = \{t | t = X_n^2(\hat{\theta}) \geq t_\alpha\}$, где $t_\alpha = \chi^2_{1-\alpha, N-r-1}$ - квантиль уровня $1-\alpha$ распределения χ^2 с $N-r-1$ степенями свободы (N - число интервалов, r - размерность вектора параметров θ).

Оценки параметров $\hat{\theta}$ ищем как решение следующей оптимизационной задачи, выраженной через распределение χ^2:

$$\min_{\theta \in \Theta} X_n^2(\theta) | H_0$$

Приведем вычисления статистик на основе распределения χ^2 и квантилей уровня $1-\alpha$ для проверки, попадают ли статистики в критическое множество и можем ли мы принять гипотезы об общем виде распределения генеральной случайной величины, на основе которой построена наша исходная выборка:

- Проверка гипотезы о гамма-распределении, N = 5:
  - $t_\alpha = 7.824046$
  - $\hat{\theta} = \{0.9591, 1.8124\}$
  - $X_N^2(\hat{\theta}) = 5.27837$
  - $t_\alpha > X_N^2(\hat{\theta}) \to X_N^2(\hat{\theta}) \notin \Im_{1\alpha}$
- Проверка гипотезы о гамма-распределении, N = 10:
  - $t_\alpha = 16.6224$
  - $\hat{\theta} = \{1.0191, 1.8395\}$
  - $X_N^2(\hat{\theta}) = 9.08725$
  - $t_\alpha > X_N^2(\hat{\theta}) \to X_N^2(\hat{\theta}) \notin \Im_{1\alpha}$
- Проверка гипотезы об экспоненциальном распределении, N = 5:

- $t_\alpha = 7.814727$
- $\hat{\theta} = 1.7335$
- $X_N^2(\hat{\theta}) = 5.29002$
- $t_\alpha > X_N^2(\hat{\theta}) \rightarrow X_N^2(\hat{\theta}) \notin \Im_{1\alpha}$

- Проверка гипотезы об экспоненциальном распределении, N = 10:

  - $t_\alpha = 15.5073$
  - $\hat{\theta} = 1.87903$
  - $X_N^2(\hat{\theta}) = 9.092895$
  - $t_\alpha > X_N^2(\hat{\theta}) \rightarrow X_N^2(\hat{\theta}) \notin \Im_{1\alpha}$

6 МП-оценки параметров полученных распределений

После получения оценок параметров как показательного, так и гамма-распределения, найдем оценки максимального правдоподобия характеристик данных случайных величин (при помощи встроенных средств пакета **scipy**):

- Параметры показательного распределения:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.8679$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = 3.489$$

$$A = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n\sigma_x^3} = 2.0$$

$$E = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n\sigma_x^4} = 6.0$$

- Параметры гамма-распределения:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.91261$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = 3.8963774$$

$$A = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n\sigma_x^3} = 2.0641$$

$$E = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n\sigma_x^4} = 6.3907$$

7 Сравнительный анализ результатов

В целом, можно видеть, что для всех вариантов выбранных разбиений интервала, в который укладываются значения исходной выборки, гипонеза H_0 была принята, то есть можно говорить, что при определенных параметрах данная случайная величина может быть распределена как по показательном закону, так и подчиняться гамма-распределению.

Также видно, что оценки характеристик распределений, полученных на основе приближенно посчитанных параметров, получились достаточно близки как друг к другу, так и к характеристикам исходной случайной величины.

Также представим визуальный результат в виде распределений с параметрами, полученными в результате проверки гипотез о виде распределения исходной случайной величины. Для наглядности сравним функции распределения и плотности вероятности с эмпирическими зависимостями, построенными перед проверками гипотез:

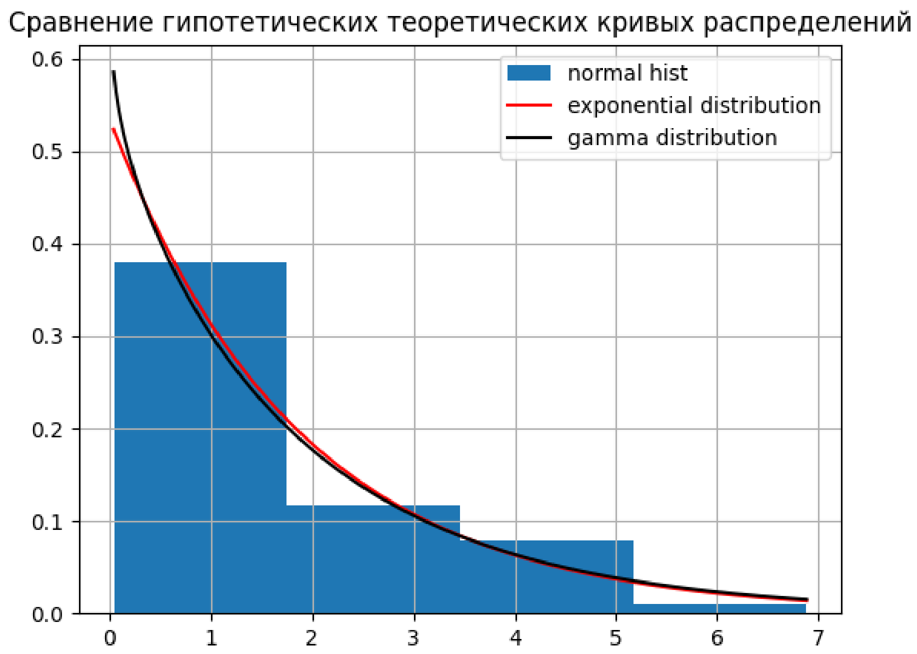

Рис. 7: Сравнение полученных функций распределения с нормированной гистограммой исходной случайной величины

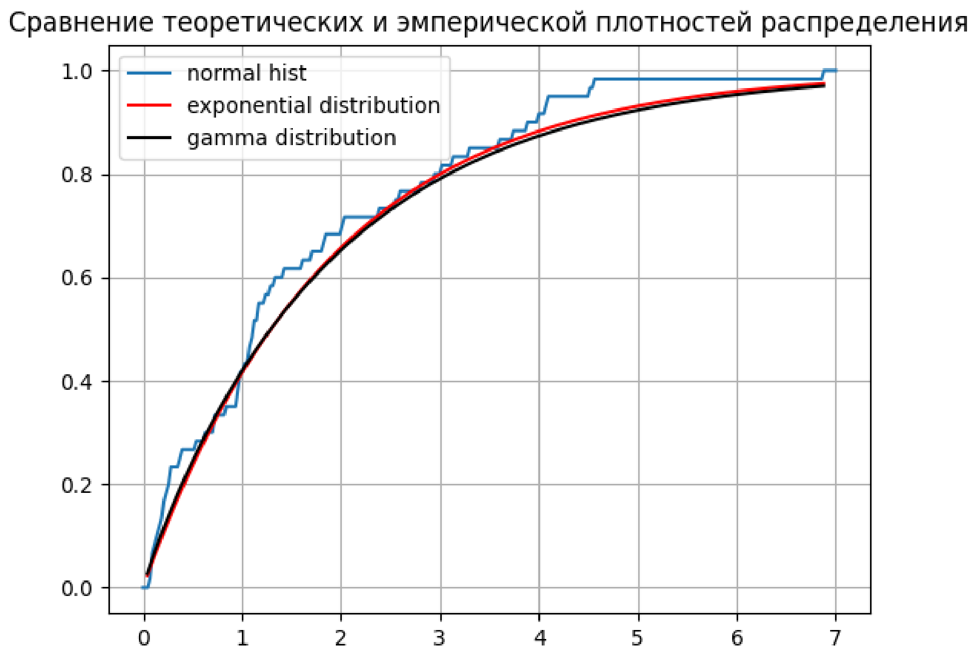

Рис. 8: Сравнение теоретических функций плотности вероятности с эмпирической

8 Модульная структура программы

Построение эмпирической функции распределения, минимизация статистики χ^2, а также визуализация результатов исследований производились в среде разработки PyCharm на языке программирования Python версии 3.10.1 с применением таких фреймворков, как **numpy**, **scipy**, **matplotlib** .

Репозиторий с кодом данной лабораторной работы находится по **ссылке**.

9 Заключение

Из полученных графиков и численных результатов видно, что исходная случайная величина достаточно близка по своим фенотипическим свойствам (если можно так выразиться) как к показательному, так и к гамма-распределению. В целом, можно сказать, что критерий оценки χ^2 Фишера является серьезным математическим инструментом, который качественно определяет вид распределения случайной величины. Даже с учетом того, что саму гипотезу выдвигает исследователь на основании выборочных характеристик исходной выборки и что метод является достаточно громоздким в плане вычислений (решение задачи безусловной минимизации - нетривиальная и ресурсоемкая задача).

В дальнейшем можно улучшить данный метод путем выбора другого алгоритма оптимизации (например, можно взять метод сопряженных направлений или что-нибудь стохастическое) либо путем доуточнения значений оценок параметров распределения (например, можно запустить метод БФГШ на меньшей точности, чем допустимый машинный эпсилон, а затем уточнить оптимум, применив метод Ньютона).