Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Отчет по лабораторной работе №1 по дисциплине «Многомерный статистический анализ»

Выполнила студентка группы 5030102/90401: О. А. Ковалёва

Преподаватель: Л. В. Павлова

Санкт-Петербург
2023

Оглавление

Цель

По заданной выборке построить и обосновать модель закона распределения исследуемой случайной величины.

Ход работы

0. Подготовим данные к работе с ними.

Считаем данные из файла и перейдем от научной записи чисел к естественному формату.

```python
def to_natural_format(num_str: str):
    mantissa, exponent = num_str.split("e")
    return float(mantissa) * 10 ** int(exponent)
```

```python
with open("Number_11.txt") as file:
    scientific_nums = np.array(file.read().split())
    sample = np.full_like(scientific_nums, 0, dtype=float)
    for i in range(scientific_nums.size):
        sample[i] = to_natural_format(scientific_nums[i])

n = sample.size
print(sample)
```

Посмотрим на полученную выборку:

```
[3.6413331  2.1894662  7.0672141  0.83373732 3.0708073  5.332609
 9.3817367  2.513246   4.1662498  1.9058249  1.2388978  3.7832059
 1.3678309  1.1036777  0.31161687 2.2967959  3.1524509  0.58616777
 1.4951332  5.8103628  4.2427576  3.8029665  6.0873556  1.2273975
 1.979971   3.8500065  7.836419   4.3650879  1.9621465  1.8550813
 7.4784227  1.9693777  2.3578568  1.4985633  2.3604825  2.7425887
 1.8175576  3.3539797  1.3733023  3.670338   1.4584889  0.54154897
 0.4366424  2.8968079  0.35526998 1.6729069  2.0402258  2.4476081
 2.4117325  2.9465026  0.17806556 1.3887417  2.6011661  1.1286504
 1.8054993  0.90979188 1.7414732  2.0375143  2.1737967  1.9149072 ]
```

1. Находим выборочные характеристики исследуемой случайной величины.

Значения выборочных характеристик получим с помощью функций, реализованных по формулам, и библиотечных. Сравним полученные результаты.

Используем формулы для вычисления выборочных характеристик:

Выборочное среднее:

$$\overline{x_n} = \sum_{i=1}^{n} x_i$$

Неисправленная выборочная дисперсия:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$$

Выборочный коэффициент асимметрии:

$$skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^3}{(\sigma^2)^{\frac{3}{2}}}$$

Выборочный коэффициент эксцесса:

$$kurt = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^4}{(\sigma^2)^2} - 3$$

Полученные результаты:

```
sample mean: 2.6694560541666665 / 2.669456054166666
sample variance: 3.6881836517453213 / 3.6881836517453213
sample skewness: 1.4866686622370526 / 1.4866686622370537
sample kurtoses: 2.1279393830448443 / 2.1279393830448488
```

Видим, что значения, вычисленные по приведенным формулам(слева), совпадают со значениями, полученными библиотечными функциями(справа). Прежде результаты были различны по причине использования формулы несмещённой дисперсии, когда в библиотечных функциях применяется смещённая.

2. Строим эмпирическую функцию распределения и нормированную гистограмму.

Эмпирическая функция распределения задается равенством:

$$F_э(x) = \frac{m_x}{n},$$

Где m_x – число элементов выборки, меньше x.

Вычислим эмпирическую функцию распределения на множестве значений x из интервала, на котором задана выборка.

```
edf = lambda grid, smpl: np.array([len(sample[smpl < x]) / n for x in grid])
sorted_sample = np.sort(sample)
grid = np.linspace(int(sorted_sample[0]), ceil(sorted_sample[-1]), 1000)

plt.figure(figsize=(10,8))
plt.plot(grid, edf(grid, sample))
plt.title('Empirical distribution function', fontsize=16)
plt.grid(ls=':')
plt.xlabel('x', fontsize=15)
plt.ylabel('F(x)', fontsize=15)
plt.show()
```

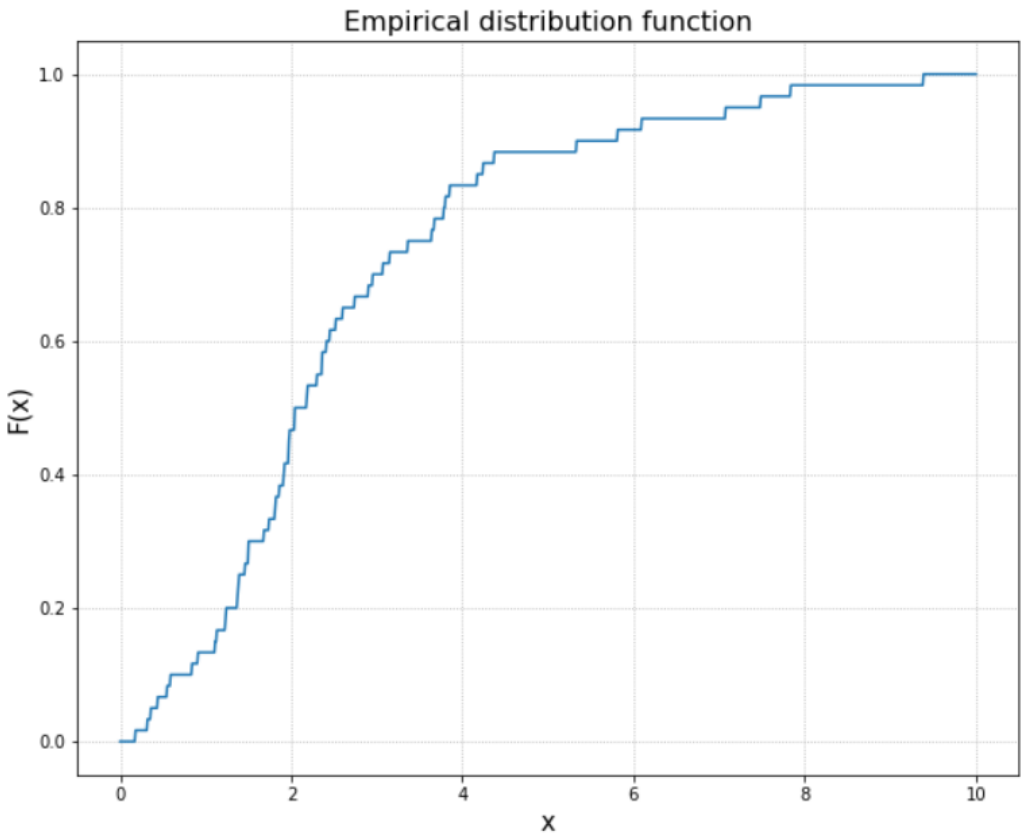

Рисунок 1. Эмпирическая функция распределения случайной величины.

Посмотрим на отсортированную в порядке возрастания выборку:

```
[0.17806556 0.31161687 0.35526998 0.4366424  0.54154897 0.58616777
 0.83373732 0.90979188 1.1036777  1.1286504  1.2273975  1.2388978
 1.3678309  1.3733023  1.3887417  1.4584889  1.4951332  1.4985633
 1.6729069  1.7414732  1.8054993  1.8175576  1.8550813  1.9058249
 1.9149072  1.9621465  1.9693777  1.979971   2.0375143  2.0402258
 2.1737967  2.1894662  2.2967959  2.3578568  2.3604825  2.4117325
 2.4476081  2.513246   2.6011661  2.7425887  2.8968079  2.9465026
 3.0708073  3.1524509  3.3539797  3.6413331  3.670338   3.7832059
 3.8029665  3.8500065  4.1662498  4.2427576  4.3650879  5.332609
 5.8103628  6.0873556  7.0672141  7.4784227  7.836419   9.3817367 ]
```

Для построения гистограммы подберем такое разбиение на интервалы, чтобы по нему выполнялись условия применения критерия хи-квадрат: частоты попадания элементов выборки в интервалы > 5.

```
borders:  [0.17806556 1.2273975  2.01879979 3.3539797  6.0873556  9.3817367 ]
num of intervals:  5
counts:  [10 18 16 11  5]
```

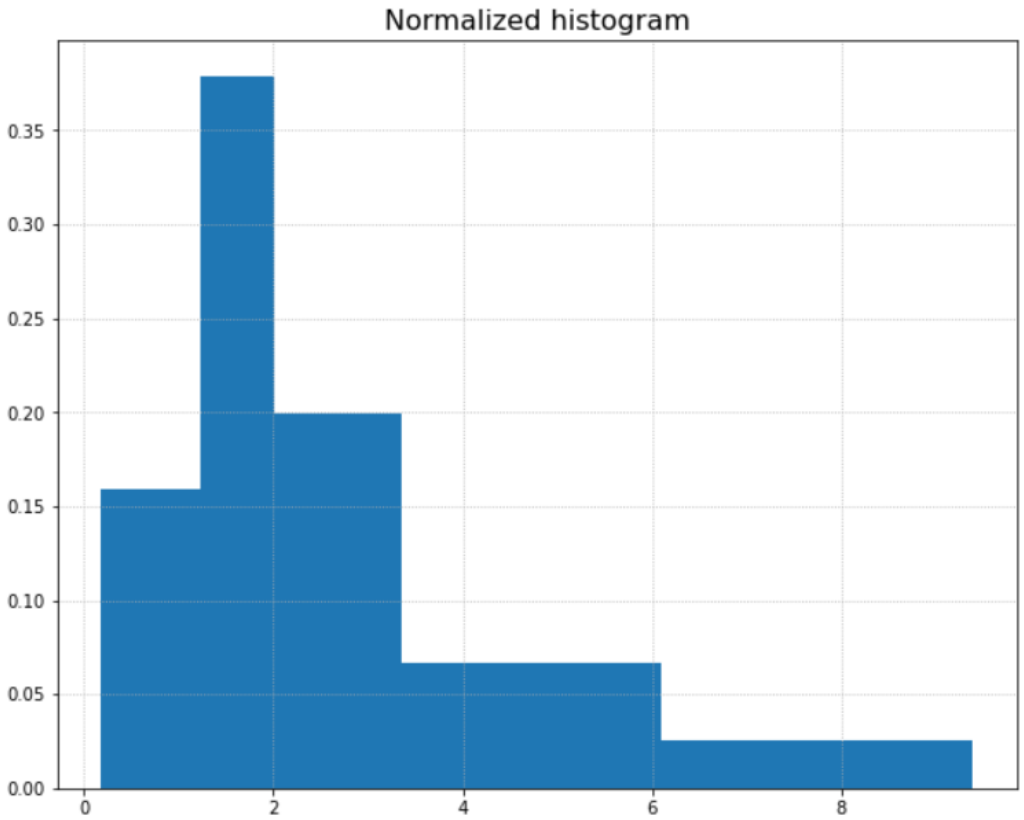

Рисунок 2. Нормированная гистограмма случайной величины

3. Строим доверительные полосы для теоретической функции распределения с доверительными вероятностями 0.90 и 0.95.

Введем статистику $D_n(X) = \sup|F_n(t) - f(t)|, t \in \mathbb{R}^1$. Применим теорему Колмогорова для определения границ, в которых с заданной вероятностью γ находится $F(t)$. Согласно теореме, при любом $u > 0$ $P\{\sqrt{n}D_n(X) \leq u\} \to K(u)$ при $n \to \infty$, где $K(u)$ – функция распределения Колмогорова. Найдем значения квантиля уровня γ распределения Колмогорова с помощью библиотечной функции.

Тогда при $n \to \infty$: $P\{\sqrt{n}D_n(X) \leq u_\gamma\} =$

$$= P\{F_n(t) - \frac{u_\gamma}{\sqrt{n}} \leq F(t) \leq F_n(t) + \frac{u_\gamma}{\sqrt{n}}\} \to \gamma.$$

И тогда область, определяемая границами $max\{0, F_n(t) - \frac{u_\gamma}{\sqrt{n}}\}$ и $min\{F_n(t) + \frac{u_\gamma}{\sqrt{n}}, 1\}$, $\forall t$, является асимптотической γ-доверительной полосой для генеральной функции распределения $F(t)$.

```
plt.figure(figsize=(10,8))
plt.plot(grid, empirical_df(grid, sorted_sample), color='b', label='Empirical distribution function')

for gamma, clr in zip([0.9, 0.95], ['r', 'g']):
    u_k = sps.kstwobign.ppf(gamma)
    print(f"gamma = {gamma}, ", f"u_k = {u_k}")
    L = np.array([max(0, F - u_k / np.sqrt(n)) for F in edf(grid, sample)])
    R = np.array([min(F + u_k / np.sqrt(n), 1) for F in edf(grid, sample)])
    plt.plot(grid, L, color=clr, label=r"$\gamma$" + f" = {gamma}")
    plt.plot(grid, R, color=clr)

plt.grid()
plt.legend(prop={'size': 15})
plt.xlabel('x', fontsize=15)
plt.ylabel('F(x)', fontsize=15)
plt.show()
```

```
gamma = 0.9,  u_k = 1.2238478702170825
gamma = 0.95,  u_k = 1.3580986393225505
```

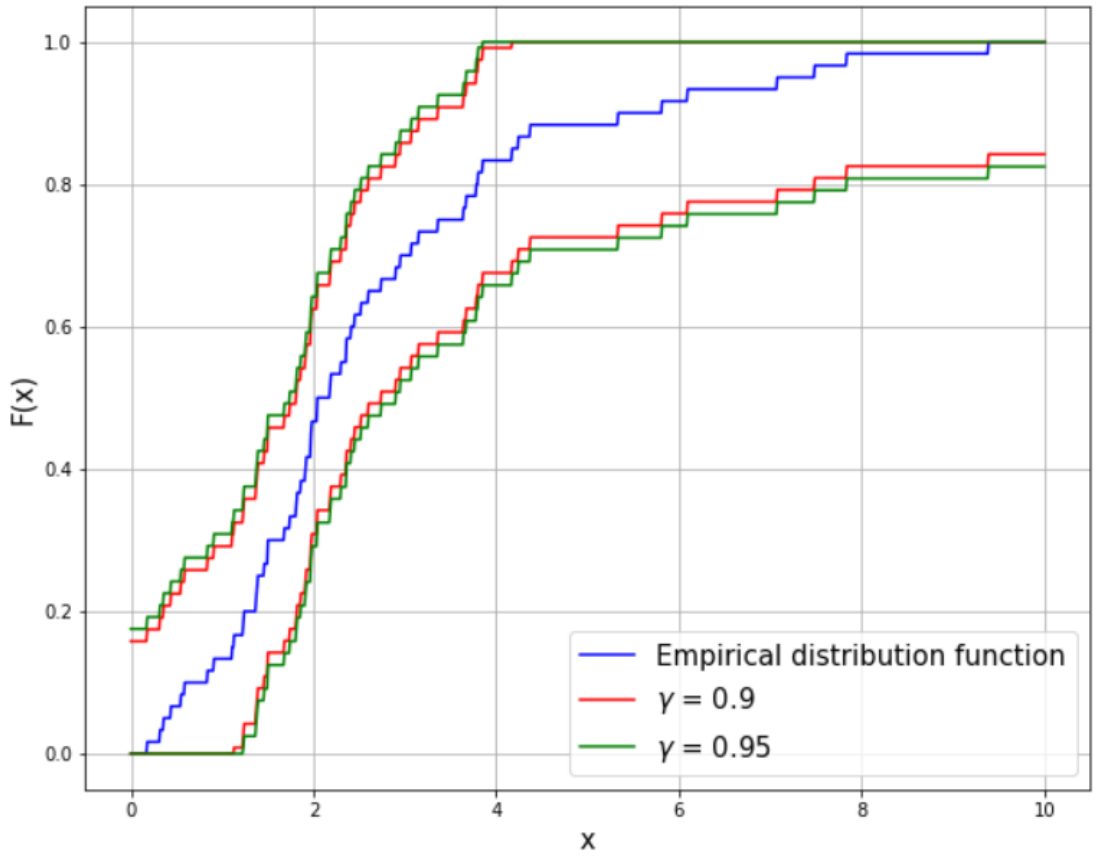

Рисунок 3. Доверительные полосы для теоретической функции распределения случайной величины.

4. Выдвигаем гипотезу о виде распределения случайной величины.

Для определения закона распределения проанализируем полученные результаты:

- Случайная величина имеет только положительные значения.
- Коэффициент асимметрии положительный
- Коэффициент эксцесса положительный
- Эмпирическая функция распределения и гистограмма по форме напоминают функцию распределения и плотность вероятности гамма-распределения и распределения хи-квадрат.

Таким образом, выдвигаем гипотезы:

$$H_{01} = \{\text{Случайная величина имеет гамма} - \text{распределение}\}$$

$$H_{02} = \{\text{Случайная величина имеет распределение хи} - \text{квадрат}\}$$

5. Проверяем гипотезы о виде распределения на основе критерия хи-квадрат Фишера.

Критерий Фишера с применением критерия согласия χ^2 используется для проверки сложной гипотезы о принадлежности неизвестной функции распределения исследуемой случайной величины заданному семейству функций распределения. Пусть $\mathcal{F} = \{F(x;\theta), \theta \in \Theta\}$ – заданное параметрическое семейство функций распределения и $X = (X_1, \dots, X_n)$ – выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения. Требуется проверить гипотезу H_0: $\mathcal{L}(\xi) \in \mathcal{F}$.

Сгруппируем выборку по интервалам, полученным в пункте 2: вектор $\nu = (\nu_1, \dots, \nu_N)$ – соответствующий вектор частот попадания элементов выборки в интервалы группировки. Составим статистику X_n^2:

$$\mathrm{X}_n^2 = \mathrm{X}_n^2(\theta) = \sum_{i=1}^{N} \frac{\left(\nu_i - np_i(\theta)\right)^2}{np_i(\theta)},$$

где $p_i(\theta) = P(\xi \in \Delta_i | H_0) = \int_{x_{i-1}}^{x_i} dF(x;\theta) = F(x_i;\,\theta) - F(x_{i-1};\,\theta)$, $i = 1, \dots, N$.

Для исключения из статистики неопределенности, связанной с неизвестным параметром θ, используем его оценку. Таким образом, статистика будет представлять собой функцию только от выборки:

$$\widehat{\theta_n} = \hat{\theta}_n(X)$$

$$\widehat{\mathrm{X}_n^2} = \mathrm{X}_n^2\left(\widehat{\theta_n}\right) = \sum_{i=1}^{N} \frac{\left(\nu_i - np_i\left(\widehat{\theta_n}\right)\right)^2}{np_i\left(\widehat{\theta_n}\right)}$$

Однако, в данном случае величины $p_i\left(\widehat{\theta_n}\right)$ являются случайными, так как представляют собой функцию от выборки. Фишер показал, что существуют методы оценивания параметра θ, при которых предельное распределение –

распределение хи-квадрат с числом степеней свободы $N-1-r$, где r – размерность параметра θ. Используем мультиномиальную оценку максимального правдоподобия и найдем $\hat{\theta}$, решая задачу минимизации: $\min_{\theta\in\Theta} \mathrm{X}_n^2(\theta) = \min_{\theta\in\Theta} \sum_{i=1}^{N} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}$.

Таким образом, по теореме Фишера:

$$\mathcal{L}\left(\widehat{\mathrm{X}_n^2} | H_0\right) \to \chi^2_{N-r-1}$$

Для заданного уровня значимости $\alpha = 0.05$ определим по таблицам распределения χ^2_{N-r-1} значения $(1-\alpha)$-квантили $\chi^2_{1-\alpha,\ N-r-1}$ и сравним с ним найденное значение $\widehat{\mathrm{X}_n^2}$. Гипотеза H_0 отвергается, если $\widehat{\mathrm{X}_n^2} \geq \chi^2_{1-\alpha,\ N-r-1}$.

В нашем случае имеем гамма-распределение с двумя параметрами, соответствующее критическое значение статистики – 5.99. Хи-квадрат – однопараметрическое распределение, критическое значение – 7.81.

Реализуем функции, вычисляющие значение статистики по параметрам распределений – для однопараметрического и двухпараметрического семейства.

```
def chi_2_value_R_2(theoretical_df, _counts, _bins, shape_param, scale_param):
    chi_2 = 0
    for j in range(len(_counts)):
        p_j = theoretical_df(_bins[j + 1], shape_param, scale_param) - theoretical_df(_bins[j], shape_param,
          scale_param)
        v_j = _counts[j]
        try:
            chi_2 += (v_j - n * p_j) ** 2 / (n * p_j)
        except Exception:
            print(f"p_j is zero = {p_j}")

    return chi_2
```

```python
def chi_2_value_R_1(theoretical_df, _counts, _bins, deg_of_freedom):
    chi_2 = 0
    for j in range(len(_counts)):
        p_j = theoretical_df(_bins[j + 1], deg_of_freedom) - theoretical_df(_bins[j], deg_of_freedom)
        v_j = _counts[j]
        try:
            chi_2 += (v_j - n * p_j) ** 2 / (n * p_j)
        except Exception:
            print(f"p_j is zero = {p_j}")

    return chi_2
```

И функцию, вычисляющую мультиномиальную оценку максимального правдоподобия параметра θ:

```python
bnds_1 = Bounds(0, np.inf)
bnds_2 = Bounds([0, 0], [np.inf, np.inf])
chi2_2 = 5.99
chi2_3 = 7.81

def multi_max_likelihood_est(theoretical_df, _counts, _bins, init_value, dimension):
    if dimension == 1:
        chi_2 = lambda param: chi_2_value_R_1(theoretical_df, _counts, _bins, param)
        opt_res = minimize(chi_2, init_value, bounds=bnds_1, method='TNC')
        print(f"chi2 value: {opt_res['fun'][0]:.2f} < {chi2_3} ?   ", opt_res['fun'][0] < chi2_3,
         "\nmultinomial maximum likelihood estimation: ", opt_res['x'])
    else:
        chi_2 = lambda params: chi_2_value_R_2(theoretical_df, _counts, _bins, params[0], params[1])
        opt_res = minimize(chi_2, init_value, bounds=bnds_2, method='TNC')
        print(f"chi2 value: {opt_res['fun']:.2f} < {chi2_2} ?   " , opt_res['fun'] < chi2_2, "\nmultinomial
         maximum likelihood estimation: ", opt_res['x'])
    return opt_res['x']
```

Применив, получим следующие результаты:

```
H0 = {Gamma distribution}
chi2 value: 4.53 < 5.99 ?    True
multinomial maximum likelihood estimation:  [2.17415495 1.28206221]

H0 = {Chi-square distribution}
chi2 value: 7.22 < 7.81 ?    True
multinomial maximum likelihood estimation:  [3.00182637]
```

В соответствии с критерием согласия χ^2 обе гипотезы H_0 принимаются, так как полученные значения статистик лежат вне критических областей.

6. Находим оценки максимального правдоподобия параметров распределения случайной величины.

Будем считать, что случайная величина имеет гамма-распределение, так как полученное значение статистики находится дальше от критической области.

Для гамма-распределения оценка максимального правдоподобия параметра формы:

$$\alpha \approx \frac{3-s+\sqrt{(s-3)^2+24s}}{12s}, \qquad \text{где } s = \ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Тогда по вычисленному параметру формы можем получить параметр масштаба:

$$\beta = \frac{1}{\alpha n}\sum_{i=1}^{n} x_i$$

```python
def shape_param_MLE(sample):
    s = log(np.mean(sample)) - sum(log(x) for x in sample) / n
    k = (3 - s + ((s - 3) ** 2 + 24 * s) ** 0.5) / (12 * s)
    return k

def scale_param_MLE(sample, shape_param_mle):
    theta = np.mean(sample) / shape_param_mle
    return theta
```

Получим оценки:

```
shape parameter: 2.038669509625593
scale parameter: 1.3094108886029883
```

Применяя библиотечные функции:

```
shape parameter: 1.8868928690464348
scale parameter: 1.3811703878501465
```

Видим, что полученные оценки близки.

7. Строим гипотетические теоретические кривые.

По полученным оценкам максимального правдоподобия и приближенным оценкам максимального правдоподобия построим функции распределения и плотности вероятности.

```
plt.figure(figsize=(10,8))
plt.plot(grid, edf(grid, sorted_sample), color='b', label='Empirical distribution function')
plt.plot(grid, gamma.cdf(grid, a=a_mle, scale=scale_mle), color='g', label=f'Gamma({a_mle:.4f}, {scale_mle:.4f}) distribution function')
plt.plot(grid, gamma.cdf(grid, a=a, scale=scale), color='c', label=f'Gamma({a:.4f}, {scale:.4f}) distribution function')
plt.grid(ls=':')
plt.xlabel('x', fontsize=15)
plt.ylabel('F(x)', fontsize=15)
plt.legend(prop={'size': 15})
plt.title(f'Gamma distribution function', fontsize=16)
plt.show()
```

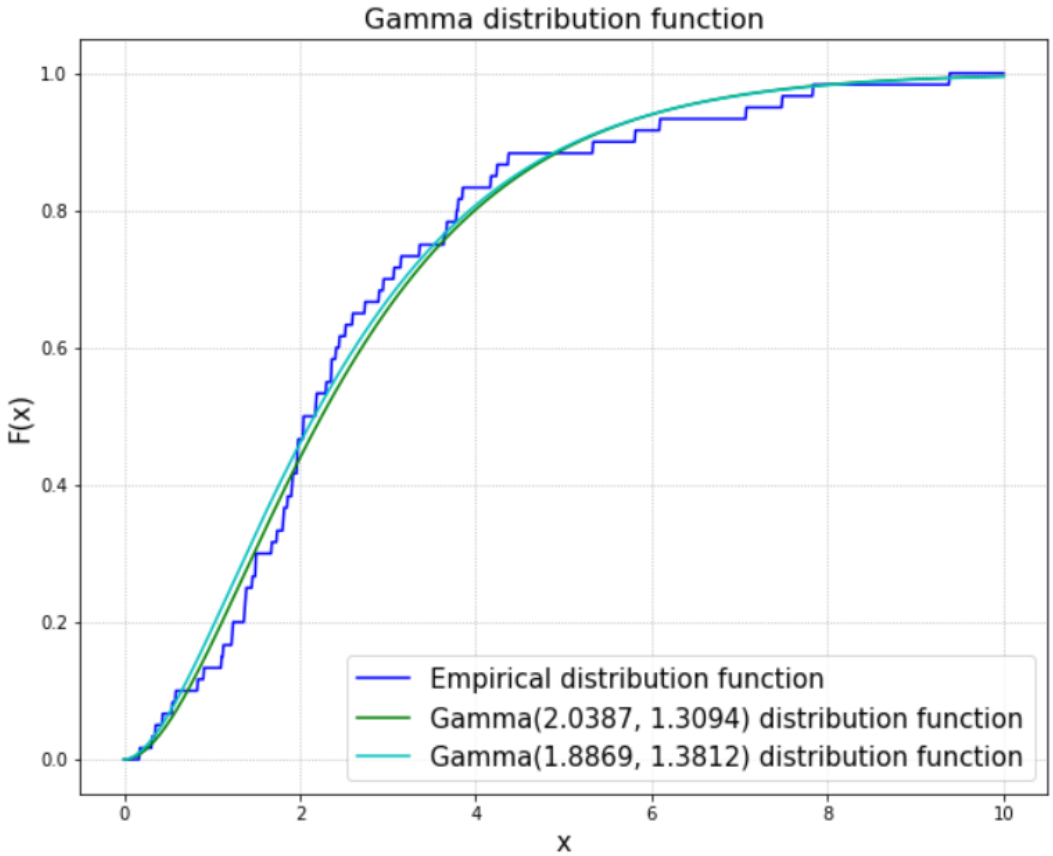

Рисунок 4. Эмпирическая и гипотетические теоретические функции распределения по двум оценкам параметров.

```
beauty_density = np.array([0.17806556, 0.9, 2.22332581, 3.24595594, 5.29121619, 7.33647645, 8.35910657,
 9.3817367])

plt.figure(figsize=(10,8))
plt.hist(sample, bins=beauty_density, density=True)
plt.plot(grid, gamma.pdf(grid, a=a_mle, scale=scale_mle), color='r', label=f'Gamma({a_mle:.4f}, {scale_mle:
 .4f}) probability density function')
plt.plot(grid, gamma.pdf(grid, a=a, scale=scale), color='orange', label=f'Gamma({a:.4f}, {scale:.4f})
 probability density function')
plt.grid(ls=':')
plt.legend(prop={'size': 15})
plt.title(f'Gamma probability density function', fontsize=16)
plt.show()
```

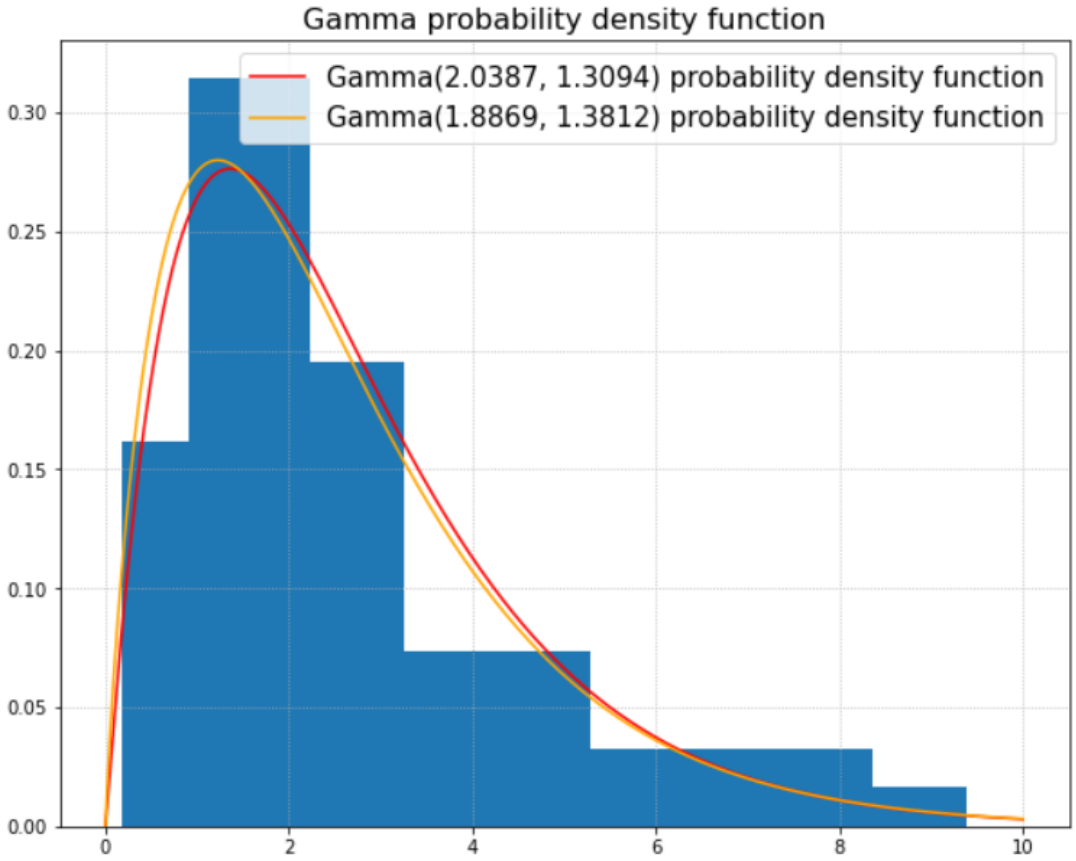

Рисунок 5. Гистограмма и гипотетические теоретические функции плотности вероятности по двум оценкам параметров.

Выводы

По полученным графикам можно сделать вывод о том, что построенные модели закона распределения достаточно хорошо отражают характеристики исходной случайной величины. По виду кривых, соответствующих двум

разным оценкам, можно сказать, что более подходящей является модель, построенная по параметрам, полученным встроенной функцией. Однако, сравнивая характеристики полученных гамма-распределений с выборочными, видим, что некоторые характеристики гамма(1.89, 1.38)-распределения находятся ближе к ним, а некоторые расходятся еще больше, чем для гамма(2.04, 1.31)-распределения:

```
ОМП, полученные по примерным формулам: (2.04, 1.31)
gamma mean: 2.669456054166666 / sample mean: 2.6694560541666665
gamma var: 3.495414823973001 / sample var: 3.6881836517453213
gamma skewness: 1.4007369396618037 / sample skewness: 1.4866686622370526
gamma kurtoses: 2.9430959611996728 / sample kurtoses: 2.1279393830448443

ОМП, полученные встроенной функцией: (1.89, 1.38)
gamma mean: 2.60612055577254 / sample mean: 2.6694560541666665
gamma var: 3.5994965388005986 / sample var: 3.6881836517453213
gamma skewness: 1.4559832342248524 / sample skewness: 1.4866686622370526
gamma kurtoses: 3.179830767515792 / sample kurtoses: 2.1279393830448443
```

Критерий Фишера, используемый для проверки сложной гипотезы, показал себя как достаточно точный метод принятия решения о виде распределения. Важным является корректное разбиение выборки на интервалы, так как при малом числе интервалов по гистограмме можно было бы судить об экспоненциальном распределении, хотя выборочные характеристики далеки от экспоненциальных. Но все же гипотеза об экспоненциальном распределении была проверена с помощью критерия Фишера и, как и ожидалось, отвергнута.