Санкт-Петербургский Политехнический Университет Петра Великого

Физико-механический институт

Кафедра прикладной математики и вычислительной физики

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

Выполнил студент гр. 5030102/90401: Саськов Л.К.

Преподаватель: к.ф.-м.н., доцент, Павлова Л.В.

Санкт-Петербург

2023

1) Постановка задачи:
   Построить и обосновать модель распределения исследуемой случайной величины.

2) По данному в файле «Number_23.txt» набору чисел были найдены выборочные несмещенные статистики:

Среднее:	1.786
Дисперсия:	4.593
Коэффициент асимметрии:	3.289
Коэффициент эксцесса:	14.126

```
from matplotlib import pyplot as plt
from scipy.stats import kurtosis, skew
import numpy as np

data = open("Number_23.txt")

numbers = data.read().split()

for i in range(len(numbers)):
    num = numbers[i].split('e')
    numbers[i] = float(num[0])*10**int(num[1])

data.close()

numbers = np.array(numbers)
numbers.sort()
```

```
print(f"Mean: {round(np.mean(numbers), 3)}")
# unbiased = несмещенные
print(f"Variance(unbiased): {round(np.var(numbers, ddof=1), 3)}")
print(f"Skew(unbiased): {round(skew(numbers, bias=False), 3)}") # Коэффициент ассиметрии
print(f"Kurtosis(unbiased): {round(kurtosis(numbers, bias=False), 3)}") # Коэффициент эксцесса
```

```
Mean: 1.786
Variance(unbiased): 4.593
Skew(unbiased): 3.289
Kurtosis(unbiased): 14.126
```

3) Построены э.ф.р. и нормированная гистограмма:
Прежде всего выборка сортируется в целях повышения производительности далее описанных алгоритмов.
Гистограмма строится следующим образом:

- Выбираются полуинтервалы. В данной работе используются интервалы равной длины, за исключением последнего – левая граница последнего полуинтервала выбирается таким образом, чтобы избежать “пробелов” столбцов гистограммы.
- Левая граница первого, а точнее нулевого, интервала на малое $\varepsilon = 10^{-15}$ меньше минимального элемента выборки, а правая граница последнего интервала равна максимальному элементу выборки.
- Строятся столбцы гистограммы. Высота каждого столбца согласуется с условием того, что площадь столбца пропорциональна относительной частоте подвыборки, попавшей в полуинтервал, а сумма площадей всех интервалов равна 1:

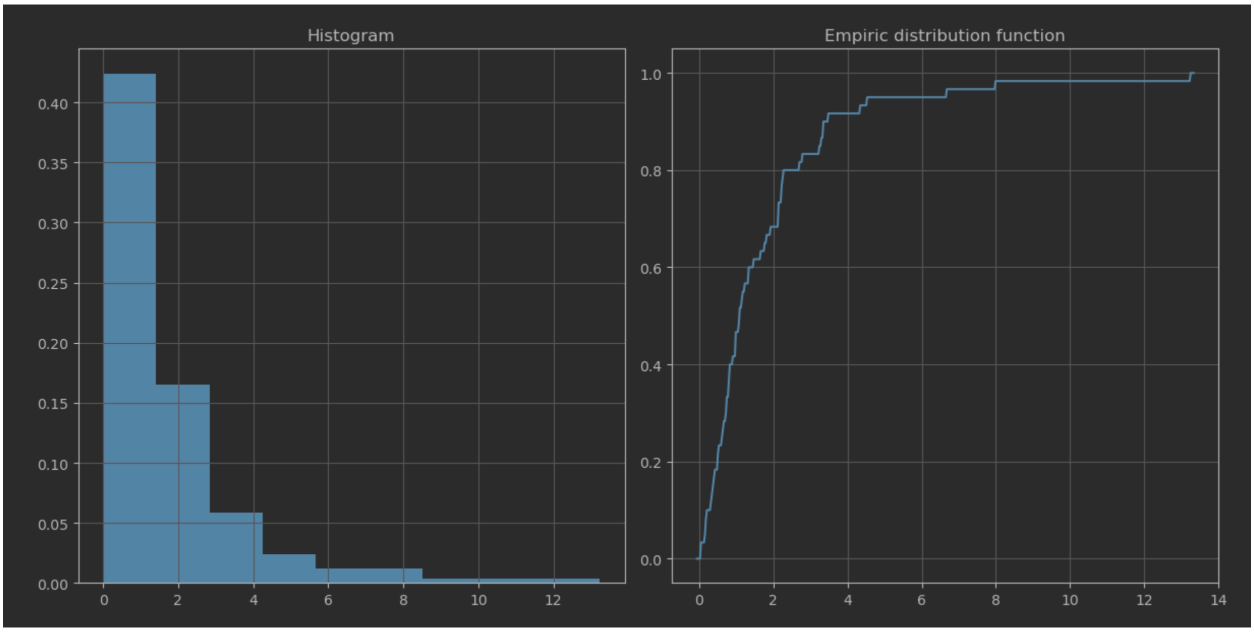
$$h\colon\ S_i = h \cdot (x_{i+1} - x_i) \sim \frac{n_i}{n}, \qquad i = \overline{0, N-1}, \qquad \sum_{i=0}^{N-1} S_i = 1$$

где x_{i+1}, x_i – правая и левая границы полуинтервала, h – высота столбца, n_i – количество элементов выборки, попавших в интервал $(x_i, x_{i+1}]$, n – размер выборки.

Эмпирическая функция распределения в каждой точке рассчитывается как относительная частота элементов, расположенных на числовой прямой левее аргумента:

$$\hat{F}(x) = \frac{n_i}{n}, \qquad \forall x_j < x, j = \overline{0, \iota}$$

```python
def emp_distribution_func(x: np.ndarray, selection: np.ndarray):
    f_arr = []
    size = len(selection)
    for i in x:
        freq = (len(selection[selection < i])) / size
        f_arr.append(freq)
    return np.array(f_arr)

left = min(numbers)-0.1
right = max(numbers)+0.1
x = np.linspace(left, right, 500)

f, ax = plt.subplots(1, 2, figsize=(12, 6))

b = list(np.linspace(0, 8.5, 7))
b.extend([max(numbers)])
print(b)
ax[0].hist(numbers, density=True, bins=b)
ax[0].grid()
ax[0].set_title('Histogram')

ax[1].plot(x, emp_distribution_func(x, numbers))
ax[1].grid()
ax[1].set_title('Empiric distribution function')

f.tight_layout()
```


Histogram
Empiric distribution function
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
1.0
0.8
0.6
0.4
0.2
0.0
14

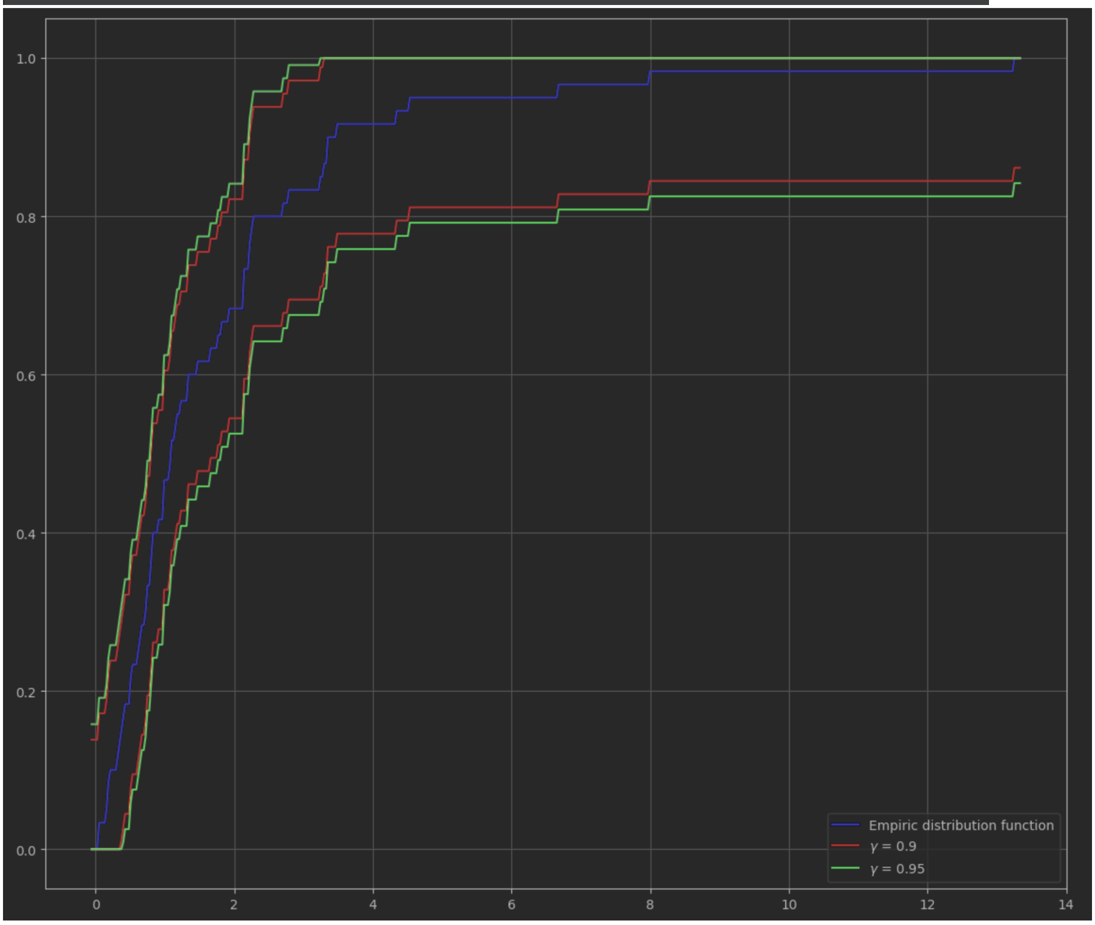
4) По эмпирической функции распределения были построены доверительные полосы для теоретической функции распределения (т.ф.р.) с доверительными вероятностями $\gamma = 0.90$ и $\gamma = 0.95$.

Построение доверительных полос обосновывается теоремой Дворецкого-Кифера-Вольфовица. Границы полос выражаются следующим образом:

$$L(x) = \max\{\hat{F}(x) - \epsilon_n, 0\}$$
$$U(x) = \min\{\hat{F}(x) + \epsilon_n, 1\}$$
$$\epsilon_n = \sqrt{\frac{1}{2n}\ln\frac{2}{1-\gamma}}$$

где γ – доверительная вероятность, n – размер выборки.

```python
epsilon = lambda gamma: np.sqrt(np.log(1 / (1 - gamma)) / 2 / len(numbers))
L = lambda gamma: np.array([max(F - epsilon(gamma), 0) for F in emp_distr(x)])
R = lambda gamma: np.array([min(F + epsilon(gamma), 1) for F in emp_distr(x)])

f, ax = plt.subplots(1, 1, figsize=(14, 12))

ax.plot(x, emp_distr(x), color='b', label='Empiric distribution function')

for _, vargamma, clr in zip(range(2), [0.9, 0.95], ['r', 'g']):
    ax.plot(x, L(vargamma), color=clr, label=r"$\gamma$" + f" = {vargamma}")
    ax.plot(x, R(vargamma), color=clr)

ax.grid()
ax.legend()
```



5) После анализа гистограммы, эмпирической функции распределения и выборочных статистик была выдвинута гипотеза о принадлежности распределения случайной величины семейству гамма-распределения:

$$f_X(x) = \frac{1}{\Gamma(k)\theta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\theta}}$$

Параметры:

$$\alpha, \theta > 0$$

6) Гипотеза о виде распределения проверяется на основе критерия хи-квадрат Фишера

Разбивая допустимый для экспоненциального и гамма распределения (от нуля до бесконечности) интервал на 5 интервалов в ходе проверки гипотезы хи-квадрат методом Фишера, получаем величины,

$$\mathrm{X}_{\mathrm{n}}^2 = \mathrm{X}_{\mathrm{n}}^2(\theta) = \sum_{k=1}^{N} \frac{(\nu_k - n * p_k(\theta))^2}{n * p_k(\theta)}, \qquad \theta - \text{вектор параметров}$$

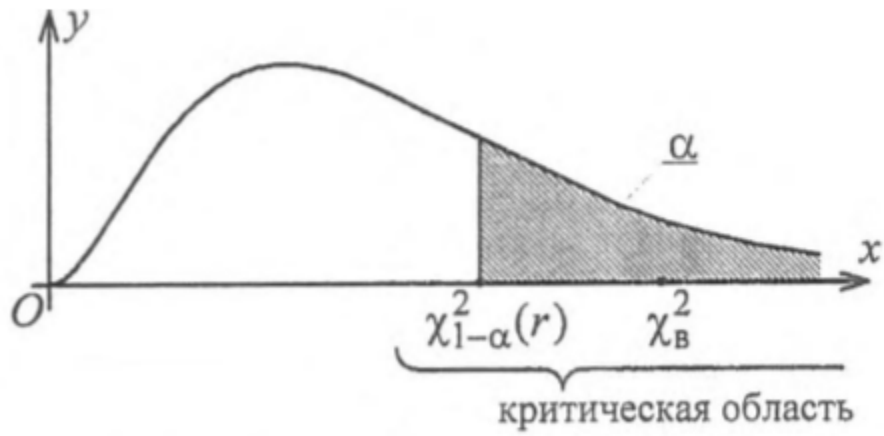
ν_k – количество элементов выборки, попавших в k-ый интервал; $p_k(\theta)$ – теоретическая вероятность попасть в k-ый интервал, $p_k(\theta) = \int_{\Delta_{\mathrm{k}}} dF(t;\ \theta)$

Которые в предельном случае распределены по закону хи-квадрат с $(N - r - 1)$ степенями свободы, где N – число интервалов, r – число параметров:

$$\mathrm{X}_{\mathrm{n}}^2 \sim^{n\to\infty} \chi^2(N - r - 1)$$

Выражение $\mathrm{X}_{\mathrm{n}}^2(\theta) = \sum_{k=1}^{N} \frac{\left(\nu_k - n * p_k(\theta)\right)^2}{n * p_k(\theta)}$ минимизируется, вычисляются оптимальные значения вектора параметров θ^*

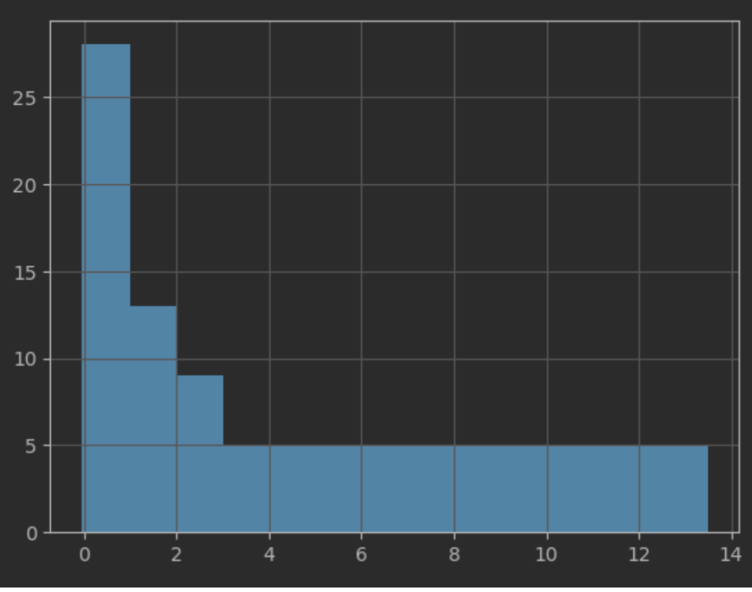
Далее строится критическая область:



Если вычисленное значение $\mathrm{X}_{\mathrm{n}}^2(\theta^*)$ меньше $\chi^2_{1-\alpha}(N - r - 1)$, то гипотеза H_0 принимается, в ином случае $\mathrm{X}_{\mathrm{n}}^2$ попадает в критическую область – нулевая гипотеза отвергается.

Следует отметить, что при разделении точек на 5 интервалов, в каждом интервале располагается минимум 5 точек (границы интервалов: [-0.05700837, 1, 2, 3, 4, 13.5]):

```
[-0.05700837, 1, 2, 3, 4, 13.5]
```



Функция вычисляющая хи-квадрат:

```python
def chi2_value_big(cdf2check, borders, sigma: float, mu: float, nums, logging: bool = False):
    N = len(nums)
    if logging:
        print(f"borders {borders}")
        print(f"sample size: {N}")
    res = 0
    for i in range(len(borders)-1):
        p_k = cdf2check(borders[i+1], sigma, mu) - cdf2check(borders[i], sigma, mu)
        v_k = len([num for num in nums if borders[i] < num and num < borders[i+1]])
        if logging:
            print(f"curr borders: {borders[i], borders[i+1]}")
            print(f"v_k: {v_k}, p_k: {p_k}")
        try:
            res += (v_k - N*p_k)**2 / (N*p_k)
        except Exception:
            print(f"potential zero = {(N*p_k)}")

    return res
```

Проверка гипотезы:

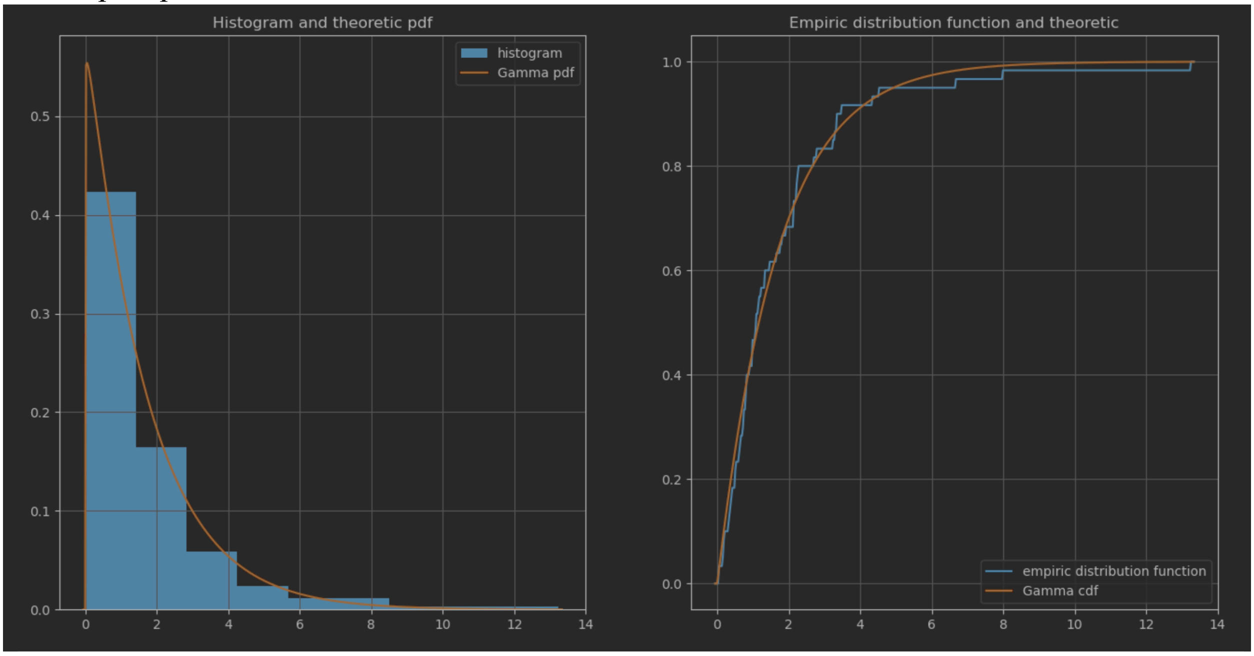
```python
from scipy.optimize import minimize

chi_2 = 6.0
borders = [-0.05700837, 1, 2, 3, 4, 13.5]
theta_gamma = []
print("gamma")
chi2 = lambda x: chi2_value_big(theoretic_gamma, sigma=x[0], mu=x[1], nums=numbers, logging=False)
result = minimize(chi2, np.array([2, 1.5]), method='TNC', tol=1e-15)
print(f"Value: {result['fun']}, min: {result['x']}")
theta_gamma.append(result['x'])
print(result['fun'] < chi_2)
```

```
gamma
Value: 0.47810585500676295, min: [1.02717783 1.61637804]
True
```

Гипотеза принимается, так как вычисленное значения меньше табличного значения – 0,95 квантиля.

7) В результате минимизации выражения $\theta^*: X_n^2(\theta^*) = \min(X_n^2(\theta)|H_0), \theta \in \Theta$
Возьмем оценку параметра для интервалов образованных 5 точками и построим теоретические функции плотности и распределения:

Гамма распределение:



8) В результате исследования были получены выборочные статистики, построена эмпирическая функция распределения и гистограмма. Построены доверительные полосы э.ф.р. Была предложена гипотеза о принадлежности распределения случайной величины семейству экспоненциального распределения. Далее гипотеза была принята на основании критерия хи-квадрат Фишера.
Была получена оценка параметров гамма распределения:

$\alpha = 1.027, \theta = 1.616$ для гамма распределения

Посчитаем оценки среднего и дисперсии, коэффициента асимметрии, эксцесса для полученного распределения:

```
mean, var, skew, kurt = gamma.stats(a=theta_gamma[0][0], scale=theta_gamma[0][1], moments='mvsk')
print(f"{mean}, {var}, {skew}, {kurt}")
```

```
1.6603076950622853, 2.6836848970519087, 1.9733638857560365, 5.841247538409246
```

$$M = \alpha\,\theta = 1.66$$

$$D = \alpha\,\theta^2 = 2.68$$

При выборочных оценках из данной нам выборки:

$$M = 1.786$$

$$D = 4.593$$

Видим, что значения для среднего и дисперсии близки.