Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
(ФГАОУ ВО СПбПУ)
Институт промышленного менеджмента, экономики и торговли
Высшая инженерно-экономическая школа

ЛАБОРАТОРНАЯ РАБОТА №1

По дисциплине «Многомерный статистический анализ»

Построение и обоснование модели закона распределения исследуемой случайной величины – вариант 13
(семестр 2)

Студент
группы
3740105/20101 ____________________
подпись, дата К.С. Малышева.

Оценка выполненной студентом работы:

Преподаватель,
Доцент, канд.эк.наук ____________________
подпись, дата Л.В. Павлова

Санкт-Петербург – 2023

Ход работы:

1) Найти выборочные характеристики исследуемой с.в. : выборочное среднее, выборочная дисперсия, выборочные коэффициенты асимметрии и эксцесса.

2) Построить э.ф.р. и нормированную гистограмму (гистограмма - красивая! без провалов и "неровностей").

3) По э.ф.р построить (в одних и тех же координатных осях) доверительные полосы для теор. функции распределения (т.ф.р.) с доверительными вероятностями 0.90 и 0.95.

4) После анализа выборочных характеристик и вида гистограммы выдвинуть (осознанно!) гипотезу (или гипотезы) о виде распределения исследуемой с.в.

5) Проверить гипотезу (гипотезы) о виде распределения на основе критерия хи-квадрат Фишера. В отчете должно присутствовать определение критерия Фишера и описание его применения для конкретного случая (случаев).

6) После того, как принято решение о виде распределения, найти МП-оценки параметров распределения с. в.

7) С этими оценками построить гипотетические теоретические кривые : ф.р.и плотность вероятности. Наложить эти кривые на э.ф.р. и нормированную гистограмму, соответственно.

8) Привести анализ полученных результатов.

№1. Найти выборочные характеристики исследуемой с.в.

Импортируем библиотеки для анализа и данные из варианта 13.

```
Ввод [16]: #Импорт библиотек
import numpy as np
import scipy.stats as ss
from scipy.stats import gamma
from scipy.stats import uniform
import statistics as st
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from statsmodels.distributions.empirical_distribution import ECDF
import random
from scipy.optimize import minimize
```

```
Ввод [2]: with open("C:/Users/simpleuser/Desktop/парсинг/Number_13.txt", "r") as f:
    data = [float(i) for s in f for i in s.split()]
    data = np.array(sorted(data))

len(data)

Out[2]: 60
```

Также рассчитаем минимальное и максимальное значения наших данных и интервал их распределения.

```
print(max(data)-min(data))
print('Максимальное значение', max(data))
print('Минимальное значение', min(data))
```

```
0.99136225778
Максимальное значение 0.99170372
Минимальное значение 0.00034146222
```

Далее по имеющимся данным рассчитываем выборочные характеристики.

```
print('Выборочное среднее',np.mean(data))
print ('Несмещенная выборочная дисперсия', np.var(data, ddof = 0))
print ('Смещенная дисперсия', np.var(data, ddof = 1))
print('Коэффициент асимметрии',ss.skew(data))
print('Коэффициент эксцесса',ss.kurtosis(data))
print('Среднеквадратическое отклонение', st.sqrt(np.var(data, ddof = 0)))
```

```
Выборочное среднее 0.49342269055366667
Несмещенная выборочная дисперсия 0.08001044966562554
Смещенная дисперсия 0.08136655898199208
Коэффициент асимметрии 0.07642724198797152
Коэффициент эксцесса -1.1809048272440879
Среднеквадратическое отклонение 0.2828611844449951
```

На основании коэффициентов асимметрии и эксцесса уже сейчас можно предположить, что распределение выборки является равномерным, так как эксцесс у равномерного распределения составляет -1,2, что близко с нашим значением в -1.18, а выборочный коэффициент асимметрии около 0 и составляет 0.076.

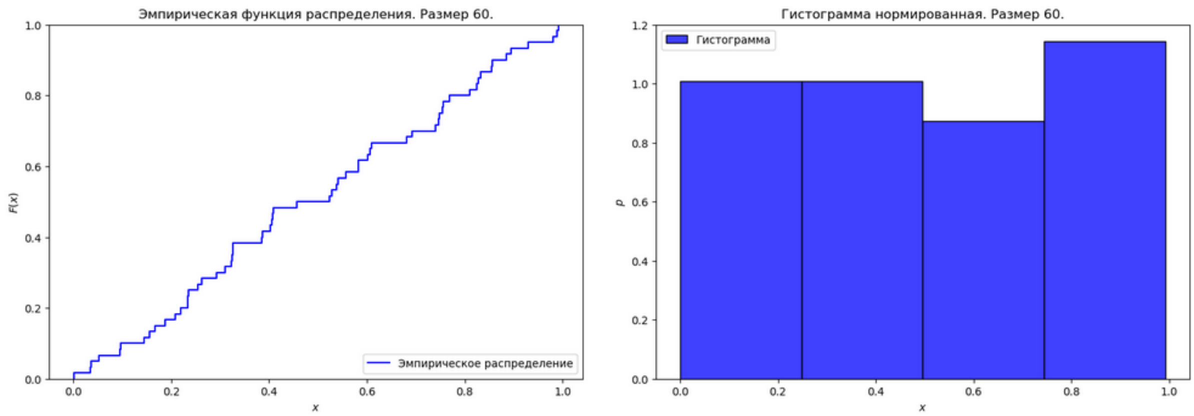
№ 2. Построить э.ф.р. и нормированную гистограмму.

```
fig, axes = plt.subplots(ncols=2, figsize=(16, 6))

sns.ecdfplot(data, label='Эмпирическое распределение', color="blue", ax=axes[0])
axes[0].yaxis.label.set_text('$F(x)$')
axes[0].xaxis.label.set_text('$x$')
axes[0].legend(loc='lower right')
axes[0].title.set_text(f"Эмпирическая функция распределения. Размер {len(data)}.")


sns.histplot(data, stat="density", label='Гистограмма', color="blue", ax=axes[1], bins=4)
axes[1].yaxis.label.set_text('$p$')
axes[1].xaxis.label.set_text('$x$')
axes[1].legend(loc='upper left')
axes[1].title.set_text(f"Гистограмма нормированная. Размер {len(data)}.")

fig.tight_layout(pad=3.0)
plt.show()
```



Исходя из четырех выделенных интервала на гистограмме, сгруппируем всю нашу выборку (все 60 значения на 4 интервала).

```
interval = (max(data)-min(data))/4
one = []
two = []
three = []
four = []

for x in data:
    if x < interval + min(data):
        one.append(x)
    if interval + min(data) <= x < min(data) + interval*2:
        two.append(x)
    if min(data) + interval*2 <= x < min(data) + interval*3:
        three.append(x)
    if min(data) + interval*3 <= x:
        four.append(x)

print(interval)
df = pd.DataFrame(([min(one),max(one)],[min(two),max(two)],[min(three),max(three)],[min(four),max(four)]), columns=['Min', '
df
```

0.247840564445

	Min	Max
0	0.000341	0.234783
1	0.254481	0.456425
2	0.523030	0.740365
3	0.746148	0.991704

```
group = []
groups = []

location = min(data)

for i in range(4):
    group.append(location)
    location += interval
    group.append(location)
    groups.append(group)
    group = []
groups
```

```
[[0.00034146222, 0.248182026665],
 [0.248182026665, 0.49602259111],
 [0.49602259111, 0.743863155555],
 [0.743863155555, 0.9917037200000001]]
```

```
len_of_groups = []
len_of_groups.append(len(one))
len_of_groups.append(len(two))
len_of_groups.append(len(three))
len_of_groups.append(len(four))
len_of_groups
```

```
[15, 15, 13, 17]
```

№ 4. После анализа выборочных характеристик и вида гистограммы выдвинуть гипотезу о виде распределения исследуемой с.в.

По форме гистограммы (количество значений распределено практически равномерно) и по выборочным характеристикам можно выдвинуть гипотезу, что анализируемая выборка относится к классу равномерных распределений.

№5. Проверить гипотезу (гипотезы) о виде распределения на основе критерия хи-квадрат Фишера.

Используя критерий согласия хи-квадрат Фишера, проверим гипотезу о равномерном распределении (в наших интервалах больше, чем 5 значений, и выборка составляет более 50).

$$\mathcal{F}_0 = \{F(t;\theta), \theta \in \Theta\},\ \theta \in \Theta \subset \mathrm{R}^r$$

$$(*)\quad X_n^2 = X_n^2(\nu) = \textstyle\sum_{j=1}^{N} (\nu_j - np_j^0)^2 / (np_j^0)$$

$$p_j^0(\theta) = \mathrm{P}\{\xi \in \Delta_j \mid \mathrm{H}_0\} = \int_{\Delta_j} d\,\mathrm{F}(t;\theta),\ j = 1,\ldots,\mathrm{N},\ \theta - ?$$

$\theta \to \hat{\theta}$ (мультиномиальная оценка максимального правдоподобия)

$X_n^2 = X_n^2(\hat{\theta}) = \sum_{j=1}^{N}(\nu_j - np_j^0(\hat{\theta}))^2/(np_j^0(\hat{\theta}))$

при $n \to \infty$ в условиях гипотезы H_0 стремится

к распределению χ^2 (N–r–1)

(т.е. при $n \to \infty$ $X_n^2 \mid H_0 \approx \chi^2_{(N-r-1)}$)

$\alpha \in (0,1) \to \mathcal{T}_{1\alpha} = \{ t: \ t = T(X) \geq t_\alpha \mid H_0\}$

Критическое значение статистики: $t_\alpha = \chi^2_{1-\alpha, N-r-1}$

Пркатическое применение:

1) $\nu_j \geq 5$, $j = 1, \ldots, N$; $n \geq 50$
2) $\hat{\theta}$ находят, решая задачу:

$$X_n^2(\theta) = \sum_{j=1}^{N}(\nu_j - np_j^0(\theta))^2/(np_j^0(\theta)) \to \min_{\theta \in \Theta}$$

Создадим функцию критерия хи-квадрат и минимизируем ее значения.

```
def f(x):
    loc, scale = x[0], x[1]
    def p_group(left_bound, right_bound):
        # Возвращаем вероятность попасть в отрезок от left_bound до right_bound
        return ss.uniform.cdf(right_bound, loc=loc, scale=scale) - ss.uniform.cdf(left_bound, loc=loc, scale=scale)
    hi = 0
    for index, group in enumerate(groups):
        if index == 0:
            left_bound = float('-inf') # - бесконечность
        else:
            left_bound = group[0]
        if index == len(groups) - 1:
            right_bound = float('inf') # + бесконечность
        else:
            right_bound = group[1]
        p_i = p_group(left_bound, right_bound)
        hi += (len_of_groups[index] - len(data) * p_i)**2 / (len(data) * p_i)

    return hi
```

```
x = [0, 1]
result = minimize(f, [x[0], x[1]])

print(f'Значение Хи-квадрат = {result.fun}')
print(f'Значения формы и масштаба соответственно: {result.x}')
```

```
Значение Хи-квадрат = 0.14276020825224534
Значения формы и масштаба соответственно: [-0.01668663  1.06073424]
```

Таким образом, у нас выходят следующие параметры статистики и распределения.

```
loc1 = result.x[0]
scale1 = result.x[1]
print(loc1,scale1)
```

```
-0.01668663134461831 1.0607342377961857
```

```
X_stat = f([loc1, scale1])
X_stat
```

```
0.14276020825224534
```

Также выведем параметры критического значения статистики.

```
X_crit = ss.chi2.ppf(1-.05, df=4)
X_crit
```

```
9.487729036781154
```

```
X_stat < X_crit
```

```
True
```

Таким образом, значения статистики нашей выборки не попадает в область критических значений статистики хи-квадрат, что означает, что мы не можем отвергнуть нулевую гипотезу о том, что наши данные не принадлежат классу равномерных распределений. То есть наша выдвинутая гипотеза, скорее всего, верна и анализируемая выборка принадлежит равномерному закону распределения.

№6. После того, как принято решение о виде распределения, найти МП- оценки параметров распределения с.в.

С помощью метода максимального правдоподобия найдем оценки параметров равномерного распределения на основе наших данных с целью построения теоретической ф.р. Для этого используем функцию «fit».

```
loc2,scale2 = uniform.fit(data)
print(loc2,scale2)
```

```
0.00034146222 0.99136225778
```

№7. С этими оценками построить гипотетические теоретические кривые: ф.р.и плотность вероятности. Наложить эти кривые на э.ф.р. и нормированную гистограмму, соответственно.

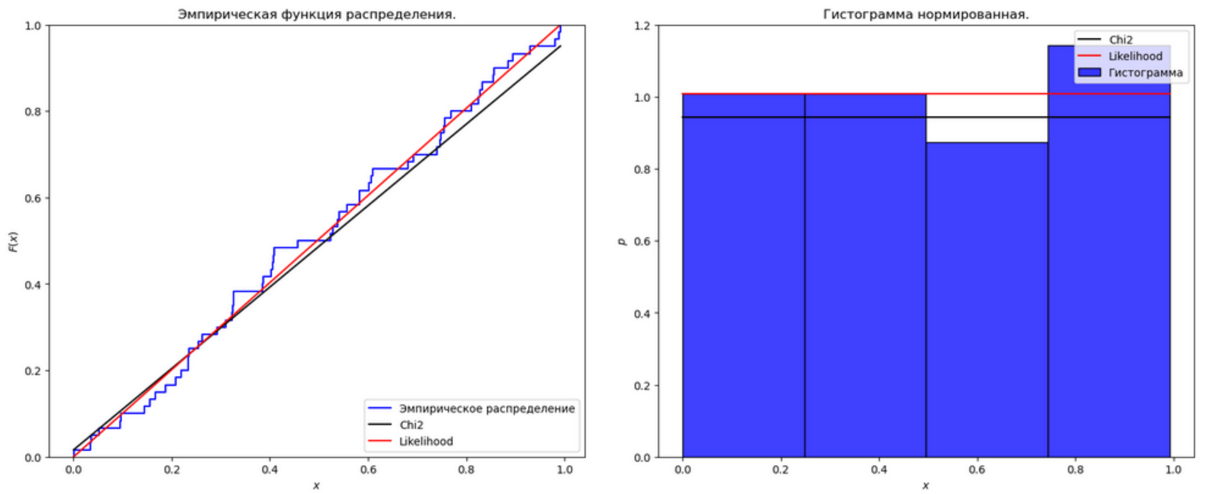
Отобразим на графиках распределение наших данных, кривые по МП-оценкам и в результате минимизации хи-квадрат.

```
loc1 = -0.01668663134461831
scale1 = 1.0607342377961857
loc2 = 0.00034146222
scale2 = 0.99136225778
fig, axes = plt.subplots(ncols=2, figsize=(16, 7))

sns.ecdfplot(data, label='Эмпирическое распределение', color="blue", ax=axes[0])
x_linspace = np.linspace(min(data), max(data), 1000)
sns.lineplot(x=x_linspace, y=ss.uniform.cdf(x_linspace, loc=loc1, scale=scale1), label='Chi2', color='black', ax=axes[0])
sns.lineplot(x=x_linspace, y=ss.uniform.cdf(x_linspace, loc=loc2, scale=scale2), label='Likelihood', color='red', ax=axes[0]
axes[0].yaxis.label.set_text('$F(x)$')
axes[0].xaxis.label.set_text('$x$')
axes[0].legend(loc='lower right')
axes[0].title.set_text(f"Эмпирическая функция распределения.")

sns.histplot(data, stat="density", label='Гистограмма', color="blue", ax=axes[1], bins=4)
sns.lineplot(x=x_linspace, y=ss.uniform.pdf(x_linspace, loc=loc1, scale=scale1), label='Chi2', color='black', ax=axes[1])
sns.lineplot(x=x_linspace, y=ss.uniform.pdf(x_linspace, loc=loc2, scale=scale2), label='Likelihood', color='red', ax=axes[1]
axes[1].yaxis.label.set_text('$p$')
axes[1].xaxis.label.set_text('$x$')
axes[1].legend(loc='upper right')
axes[1].title.set_text(f"Гистограмма нормированная.")

fig.tight_layout(pad=3.0)
plt.show()
```



Таким образом, теоретическая кривая на основании МП-оценок близка к эмпирическому распределению.

№8. Привести анализ полученных результатов.

В работе была выдвинута гипотеза о равномерном распределении. Данная гипотеза была проверена с помощью критерии хи-квадрат Фишера, значения которой показали, что статистика критерия исследуемой выборки не попадает в критическую область критических значений, и обозначает, что наша гипотеза верна. Построив теоретическую функцию равномерного распределения на основании МП-оценок и отобразив ее на графике с эмпирической ф.р., можно сделать вывод о том, что теоретические кривые отображают общую тенденцию распределения нашей анализируемой выборки.