|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **Министерство науки и высшего образования Российской Федерации**  Калужский филиал  федерального государственного бюджетного  образовательного учреждения высшего образования  ***«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»***  ***(КФ МГТУ им. Н.Э. Баумана)*** |

|  |  |
| --- | --- |
| **ФАКУЛЬТЕТ** | **ИУК «Информатика и управление»** |
| **КАФЕДРА** | **ИУК4 «Программное обеспечение ЭВМ,** |
| **информационные технологии»** | |

**Практическое занятие №2**

**«Графическое представление статистических данных, выборочные числовые характеристики на основе большой выборки»**

**ДИСЦИПЛИНА: «Методы обработки информации»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил: студент гр. ИУК4-72Б | |  |  | ( | Сафронов Н.С. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |
| Проверил: | |  |  | ( | Никитенко У.В. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |

|  |  |
| --- | --- |
| Дата сдачи (защиты):  Результаты сдачи (защиты): | |
|  | - Балльная оценка:  - Оценка: |

Калуга, 2023

**Цель работы:** овладение приёмами первичной обработки большой выборки, выдвижение гипотезы о законе распределения гeнеральной совокупности.

**Вариант 14**

**Постановка задачи**

Для обработки преподавателем выдается случайных чисел.

Эти числа хранятся в файле TestNN.csv.

1. Выборка подвергается обработке и оформляется в виде таблицы.

2. Графические характеристики выборки – строим гистограмму и полигон приведенных частот. Выдвигаем гипотезу о виде плотности вероятности генерального распределения.

3. Находим выборочные характеристики положения и рассеивания.

4. Для сравнения с гистограммой и полигоном приведенных частот на одном чертеже постройте графики гистограммной оценки плотности вероятности параметрической oцeнки плотности вероятности , и усредненную ядерную оценку плотности вероятности .

5. Значения оценок плотности вероятности в средних точках промежутков группированного статистического ряда оформите в виде таблицы.

6. Проанализируйте близость оценок по средним квадратическим отклонениям и от .

**Листинг программы**

import argparse

import csv

import numpy as np

import prettytable

import matplotlib.pyplot as plt

import statistics as st

from scipy.stats import gaussian\_kde

if \_\_name\_\_ == "\_\_main\_\_":

parser = argparse.ArgumentParser()

parser.add\_argument("-file") or "./data/Test14.csv"

args = parser.parse\_args()

file = args.file

points = []

with open(file, newline='') as csvfile:

reader = csv.reader(csvfile, delimiter=' ', quotechar='|')

for row in reader:

points.append(float("".join(row)))

points.sort()

min\_point = points[0]

max\_point = points[-1]

points\_range = max\_point - min\_point

print(f"Размах выборки: {points\_range:.2f}")

num\_bins = 1 + int(np.ceil(np.log2(len(points))))

print(f"Количество интервалов: {num\_bins}")

step = points\_range / num\_bins

print(f"Длина интервала: {step:.2f}")

\_bins = []

for i in range(num\_bins):

current\_min = min\_point + i \* step

current\_max = min\_point + (i + 1) \* step

current\_range = (current\_min, current\_max)

count = len(

list(

filter(

lambda x: current\_min <= x <= current\_max, points

)

)

)

current\_average = (current\_max + current\_min) / 2.

\_bins.append(

{

"average": round(current\_average, 4),

"minimum": round(current\_min, 4),

"maximum": round(current\_max, 4),

"count": round(count, 4)

}

)

table = prettytable.PrettyTable()

table.field\_names = [

"Номер промежутка", "a\_{i-1}", "a\_i", "n\_i",

"Средняя точка промежутка"

]

index = 1

for bin in \_bins:

table.add\_row([index := index + 1, bin["minimum"], bin["maximum"],

bin["count"], bin["average"]])

print(table)

print(f"Выборочное среднее: {np.mean(points):.2f}")

print(f"Медиана: {np.median(points):.2f}")

print(f"Мода: {st.mode(points):.2f}")

print(f"Размах выборки: {max(points) - min(points):.2f}")

print(f"Выборочная дисперсия: {np.var(points):.2f}")

print(f"Стандартное отклонение выборки: {np.sqrt(np.var(points)):.2f}")

print(f"Коэффициент вариации: "

f"{np.sqrt(np.var(points)) / np.mean(points) / 100:.2f}")

plt.hist(

points, color="grey", edgecolor="black", bins=num\_bins, range=(

min\_point,

max\_point), alpha=0.5, density=True, label="Гистограмма"

)

centers = [bin["average"] for bin in \_bins]

\_bins\_plot = [bin["count"] / 49 for bin in \_bins]

plt.plot(centers, \_bins\_plot, color="black")

plt.show()

def normal\_distribution(x):

return 1 / np.sqrt(2 \* np.pi) / np.sqrt(np.var(points)) \* (

np.exp(-1 / 2 \* (

(x - np.mean(points)) / np.sqrt(np.var(points))) \*\* 2

)

)

kde = gaussian\_kde(points)

xs = np.linspace(min\_point, max\_point, 100)

plt.hist(

points, color="grey", edgecolor="black", bins=num\_bins, range=(

min\_point,

max\_point), alpha=0.5, density=True, label="Гистограмма"

)

plt.plot(xs, [kde(x) for x in xs], color="red",

label="Усреднённая ядерная оценка")

plt.plot(xs, [normal\_distribution(x) for x in xs], color="green",

label="Параметрическая оценка")

plt.legend()

plt.show()

table = prettytable.PrettyTable()

table.field\_names = [

"z\_i", "n\_i", "f\_Г(x)", "f\_УЯ(x)", "f\_П(x)", "(f\_УЯ(x)-f\_Г(x))^2",

"(f\_П(x)-f\_Г(x))^2",

]

index = 1

for bin in \_bins:

current\_average = bin["average"]

current\_count = bin["count"]

current\_histogram = round(current\_count / len(points) / 0.5, 4)

current\_kde = round(float(kde(current\_average)), 4)

current\_parametric = round(normal\_distribution(current\_average), 4)

diff\_kde = round((current\_kde - current\_histogram) \*\* 2, 4)

diff\_parametric = round((current\_parametric - current\_histogram) \*\* 2, 4)

table.add\_row([

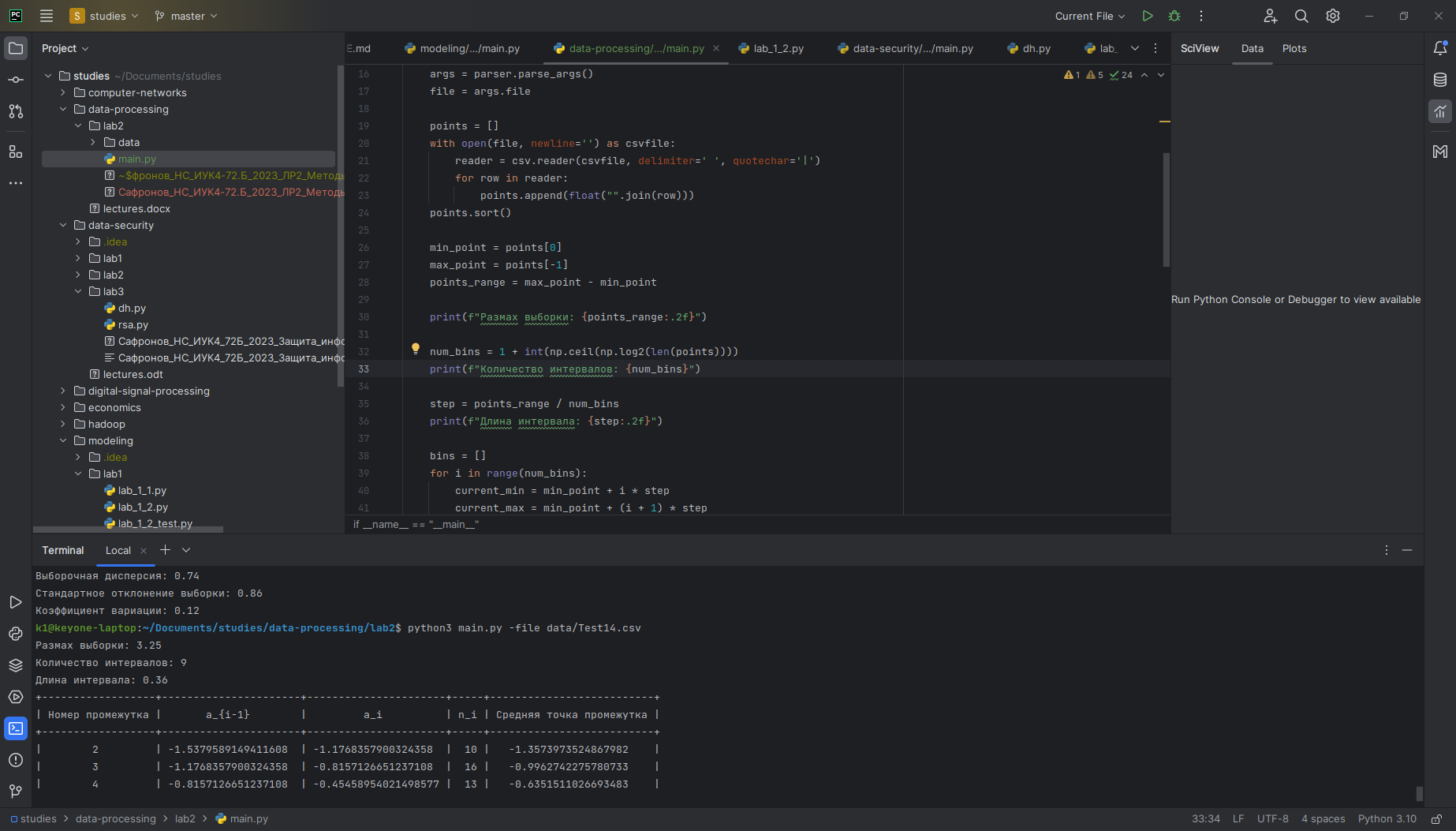
current\_average, current\_count, current\_histogram,

current\_kde, current\_parametric, diff\_kde, diff\_parametric

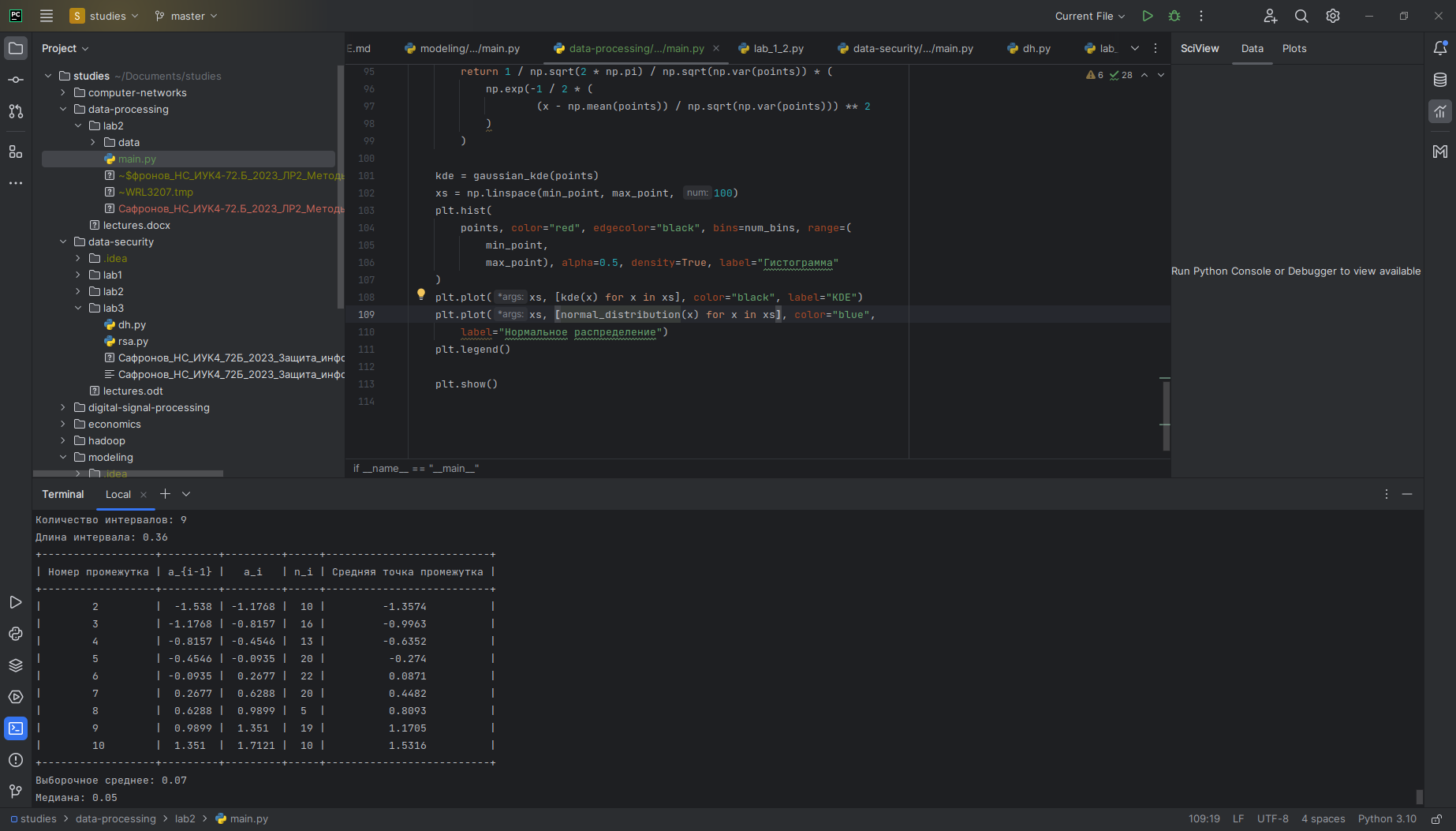
])

print(table)

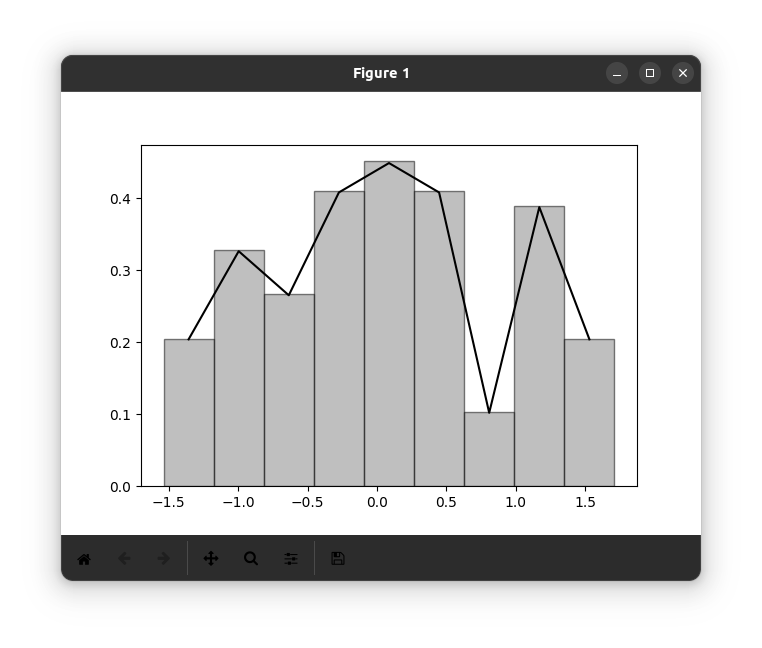
**Результаты выполнения программы**

****

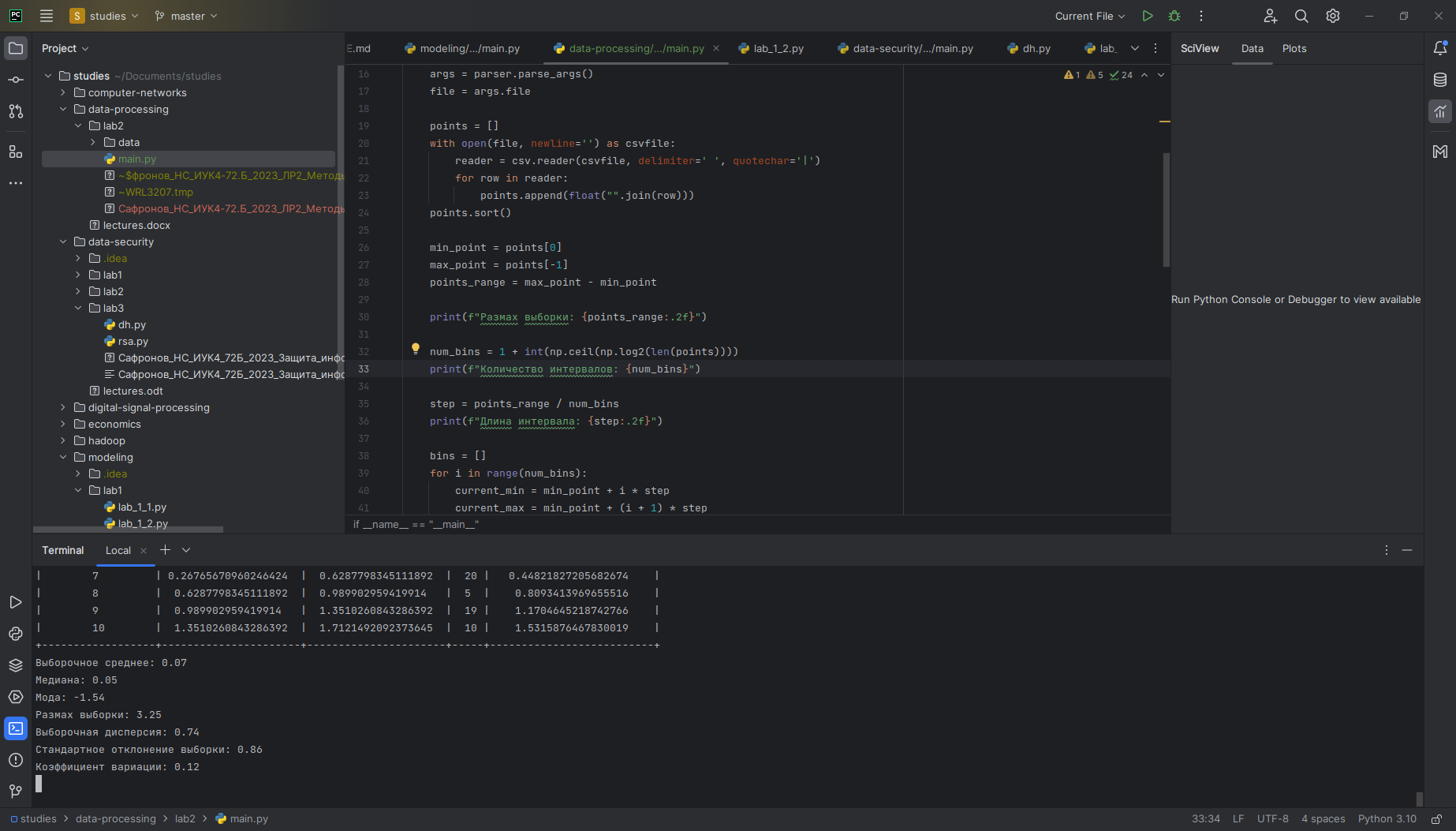
**Рисунок 1 –** Параметры построения гистограммы



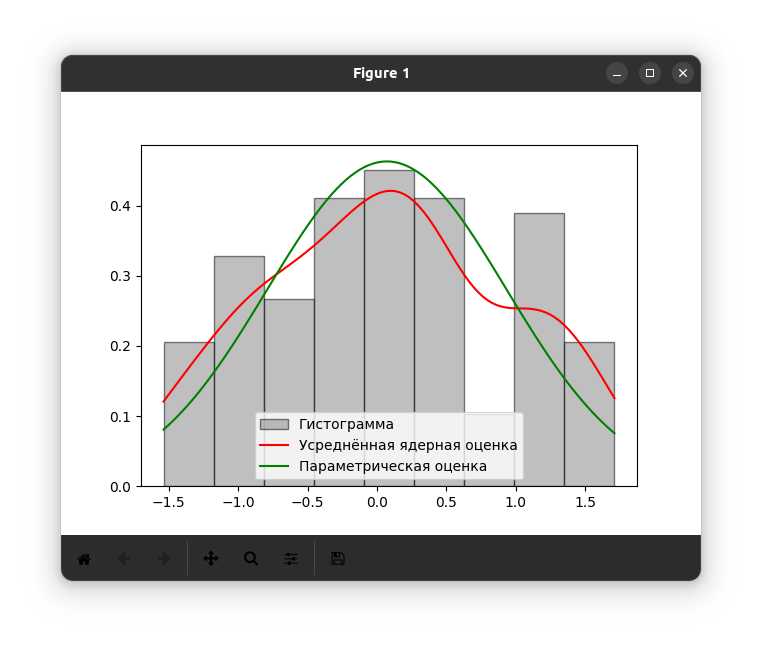
**Рисунок 2 –** Результат обработки выборки

****

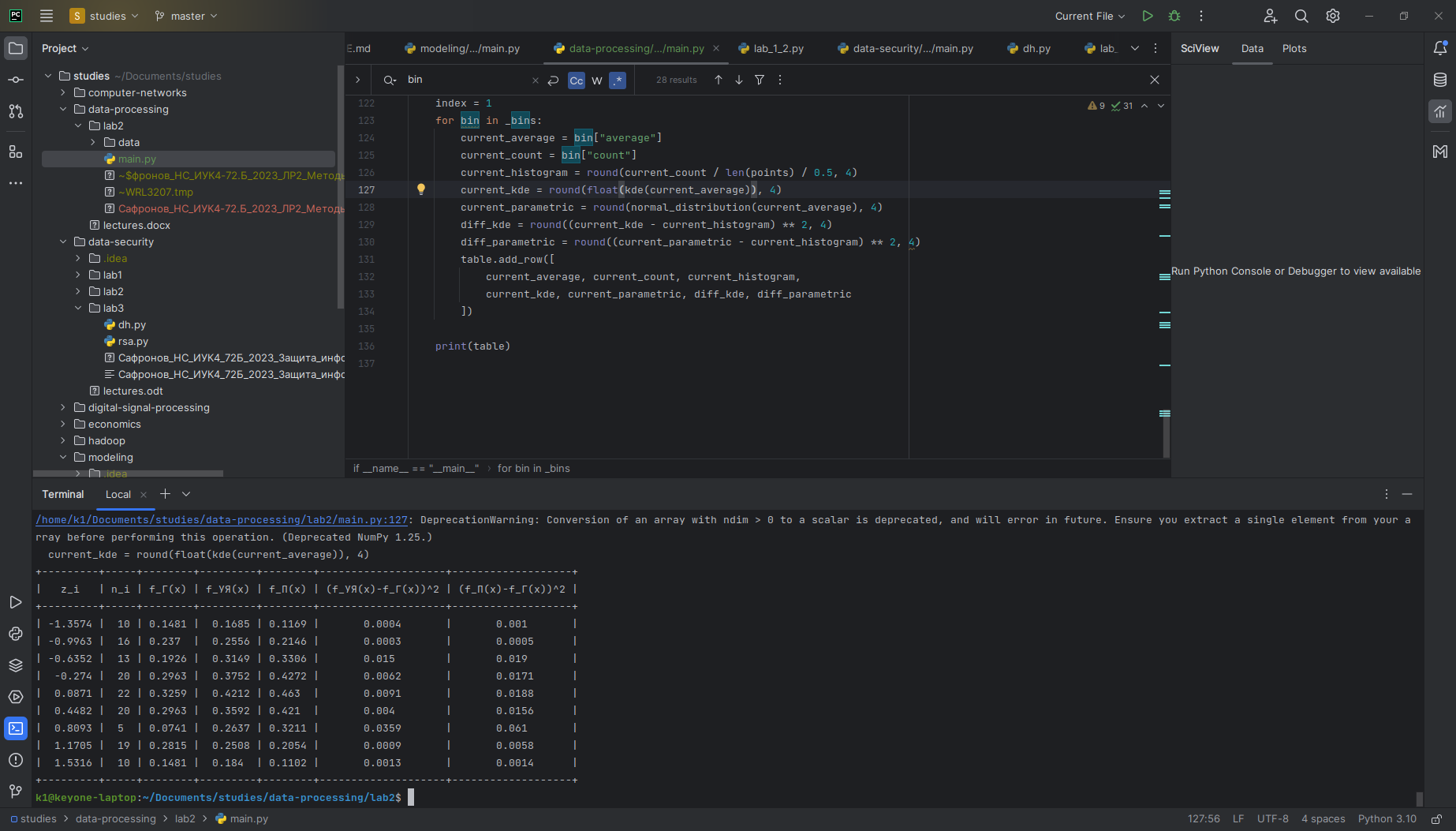
**Рисунок 3 –** Построенные гистограмма и полигон частот



**Рисунок 4 –** Выборочные характеристики положения и рассеяния



**Рисунок 5 –** Параметрическая и усреднённая ядерная оценки выборки



**Рисунок 6 –** Значения плотностей вероятности в средних точках интервалов

Усреднённая ядерная оценка находиться ближе к гистограммной оценке плотности, поскольку в отличие от параметрической она является ассиметричной.

**Вывод:** в ходе лабораторной работы были изучены приёмы первичной обработки большой выборки для выдвижения гипотезы о законе распределения гeнеральной совокупности.