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 генеральной 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 $f_{\Gamma}(x)$ параметрической оценки плотности вероятности $f_{\Pi}(x)$, и усредненную ядерную оценку плотности вероятности $f_{\text{уя}}(x)$.

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

6. Проанализируйте близость оценок по средним квадратическим отклонениям $f_{\text{уя}}(x)$ и $f_{\Pi}(x)$ от $f_{\Gamma}(x)$.

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

```
Размах выборки: 3.25
Количество интервалов: 9
Длина интервала: 0.36
```

Рисунок 1 – Параметры построения гистограммы

Номер промежутка	a_{i-1}	a_i	n_i	Средняя точка промежутка
2	-1.538	-1.1768	10	-1.3574
3	-1.1768	-0.8157	16	-0.9963
4	-0.8157	-0.4546	13	-0.6352
5	-0.4546	-0.0935	20	-0.274
6	-0.0935	0.2677	22	0.0871
7	0.2677	0.6288	20	0.4482
8	0.6288	0.9899	5	0.8093
9	0.9899	1.351	19	1.1705
10	1.351	1.7121	10	1.5316

Рисунок 2 – Результат обработки выборки

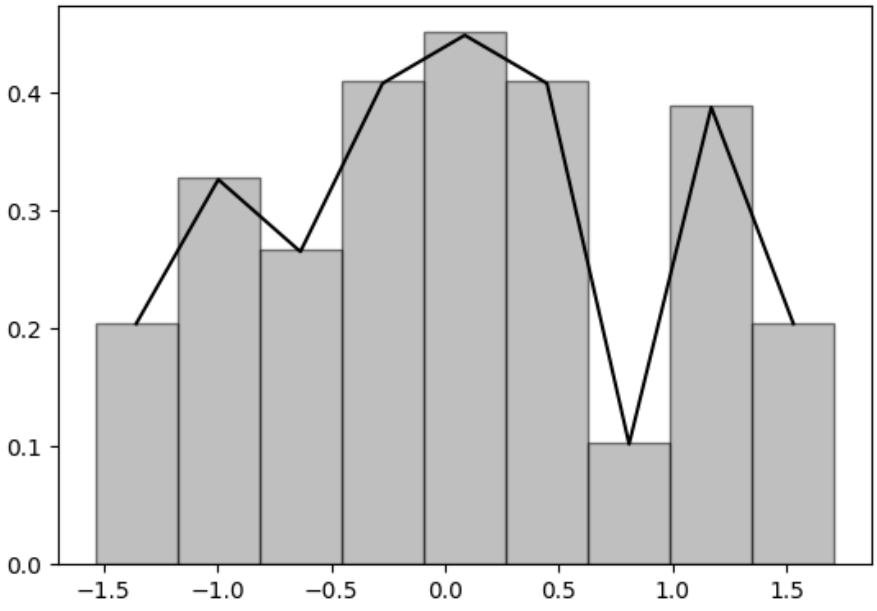

Рисунок 3 – Построенные гистограмма и полигон частот

```
Выборочное среднее: 0.07
Медиана: 0.05
Мода: -1.54
Размах выборки: 3.25
Выборочная дисперсия: 0.74
Стандартное отклонение выборки: 0.86
Коэффициент вариации: 0.12
```

Рисунок 4 – Выборочные характеристики положения и рассеяния

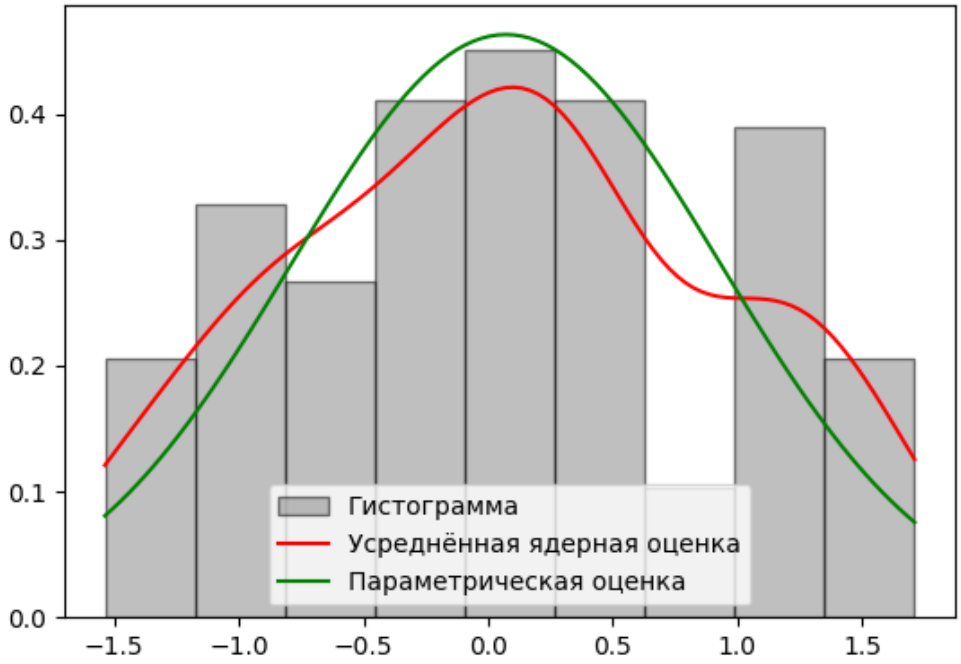

Рисунок 5 – Параметрическая и усреднённая ядерная оценки выборки

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
-1.3574	10	0.1481	0.1685	0.1169	0.0004	0.001
-0.9963	16	0.237	0.2556	0.2146	0.0003	0.0005
-0.6352	13	0.1926	0.3149	0.3306	0.015	0.019
-0.274	20	0.2963	0.3752	0.4272	0.0062	0.0171
0.0871	22	0.3259	0.4212	0.463	0.0091	0.0188
0.4482	20	0.2963	0.3592	0.421	0.004	0.0156
0.8093	5	0.0741	0.2637	0.3211	0.0359	0.061
1.1705	19	0.2815	0.2508	0.2054	0.0009	0.0058
1.5316	10	0.1481	0.184	0.1102	0.0013	0.0014

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 генеральной совокупности.