Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій.

Лабораторна Робота №1

Теорія ймовірності та математична статистика

Тема: **Описова статистика**

Виконали
студенти групи ІА – 11:
Юрченко Владислав, Момот Аркадій
Спускан Дарина, Старовойтов Вадим

Перевірив:
Цимбал С.

Київ 2023

Тема: Описова статистика

Мета: Навчитись розраховувати числові характеристики вибірки програмними засобами.

Теоретичні відомості

Множину однорідних об’єктів називають статистичною сукупністю. Вибірковою сукупністю (вибіркою) називають сукупність випадково взятих об’єктів із статистичної сукупності.

Генеральною називають сукупність об’єктів, з яких зроблено вибірку. Об’ємом сукупності (вибіркової або генеральної) називають кількість об’єктів цієї сукупності.

Полігоном частот вибірки називають ламану з вершинами в точках (xi , ni).

Полігоном відносних частот вибірки називають ламану з вершинами в точках (xi , ni /n). Полігони частот є аналогами щільності ймовірностей.

Гістограмою частот називають ступінчасту фігуру, яка складається з прямокутників, основами яких є інтервали варіант довжиною h = xi – xi-1, а висоти дорівнюють ni /h.

Нехай x1, x2, ... , xn – спостереження (значення величини X) у вибірці з об’ємом n.

Тоді вибіркове середнє можна знайти за формулою: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Медіана вибірки, що має n відсортованих значень, визначається як центральне значення, якщо n непарне число, або як середнє значення двох центральних значень, якщо n парне число.

Мода вибірки визначається як значення, що зустрічається найчастіше у вибірці. Вибірка даних може мати більше однієї моди, і в цьому випадку вона називається мультимодальною вибіркою. Якщо x1 , x2 , ... , xn є вибіркою з об’ємом n, тоді

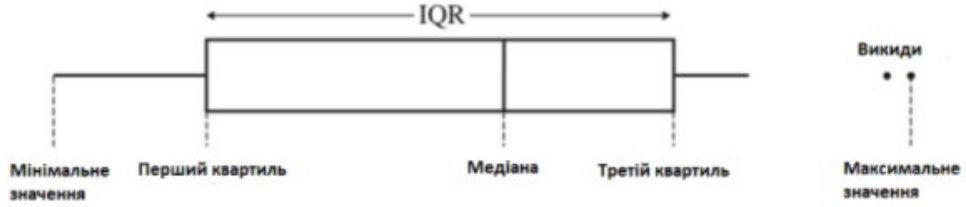
вибіркова дисперсія розраховується за формулою: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ або

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}{n-1}.$$

Значення $s = \sqrt{s^2}$ називається вибірковим середнім квадратичним відхиленням.

Діаграма розмаху або коробкова діаграма – це схематичне представлення положення даних, включаючи найменші та найбільші значення, нижню та верхню чверть вибірки (нижній та верхній квартилі), медіану та статистичні викиди.

Виглядає діаграма наступним чином:



Діаграма Парето – це діаграма, де категорії розташовані в порядку зменшення частоти. Діаграма Парето використовується для контролю якості та вдосконалення процесів з метою визначення декількох основних причин більшості проблем та першочергового їх вирішення.

Кругова діаграма або секторна діаграма створюється діленням кола на сектори, де розмір сектора відображає відносну частоту категорії у відсотках.

ХІД РОБОТИ

Завдання

Формула розрахунку варіанта: No варіанта^1 = 120 % No групи^2 + Nогрупи×No команди^3

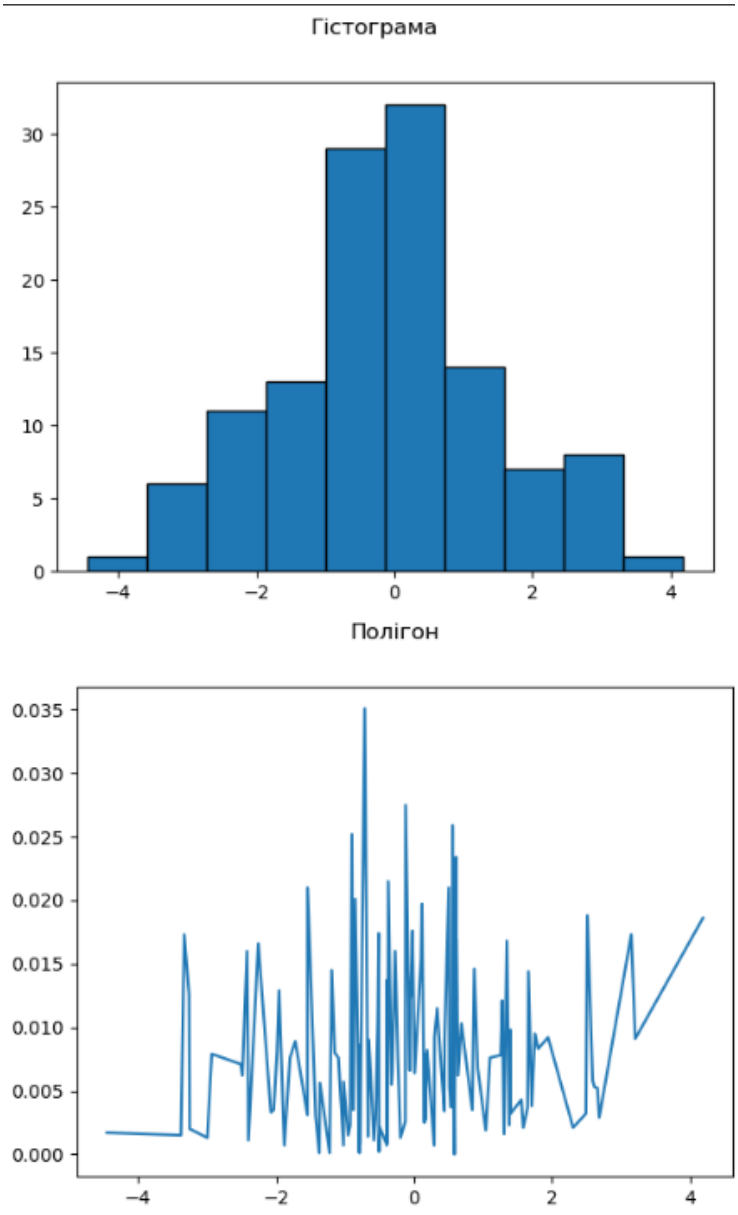
1. Згенерувати вибірку об’ємом n (див. варіанти завдань нижче) з нормальної популяції. Можна використати онлайн генератор. Значення математичних сподівань обрати самостійно.

```python
# * Оголошення варіанту та початкових змінних
variant = 120 % 11 + 11 * 3
n = 122
sigma = 1.7

# * Генерація вибірки об'ємом n
np.random.seed(42)
sample = np.random.normal(0, sigma, n).round(4)

# * Генерація ймовірностей
math_expectations = np.random.dirichlet(np.ones(n), size=1).round(4)

# * Створення словника для спрощення подальшої роботи
sample_dict = dict(zip(sample, math_expectations[0]))

od = collections.OrderedDict(sorted(sample_dict.items()))
odv = dict(sorted(od.items(), key=lambda item: item[1], reverse=True))
```

2. Написати програму, що:

a. Будує полігон та гістограму чистот

```
# * Полігон
plt.plot(od.keys(), od.values())
plt.suptitle("Полігон")
#
# # * Гістограма
fig = plt.figure()
ax = fig.add_subplot (111)
ax.hist (sample, edgecolor='black')
plt.suptitle("Гістограма")
```



b. розраховує вибіркове середнє, медіану, моду, вибіркові дисперсію та середньоквадратичне відхилення заданої вибірки (написати власні реалізації розрахунків відповідних характеристик);

```python
# Вибіркове середнє
def calc_sample_mean(distribution_dict: dict) -> float:
    sample_mean = 0
    for v, probability in distribution_dict.items():
        sample_mean += v
    sample_mean *= 1 / len(distribution_dict)
    return sample_mean
```

```python
# Медіана
def calc_median(distribution_dict: dict) -> float:
    sorted_x = sorted(distribution_dict.keys())
    length = len(sorted_x)
    if length % 2 == 0:
        median = (sorted_x[round(length / 2)] + sorted_x[round(length / 2) + 1]
    else:
        median = sorted_x[round(length / 2)]
    return median
```

```python
# Мода
def calc_mode(distribution_dict: dict) -> list:
    mode = []
    list_numbers = {}

    for k in distribution_dict.keys():
        values = list(distribution_dict.keys())
        counter = values.count(k)
        list_numbers.update({k: counter})

    max_counter = max(list_numbers.values())
    for k, v in list_numbers.items():
        if v == max_counter:
            mode.append(k)

    return mode
```

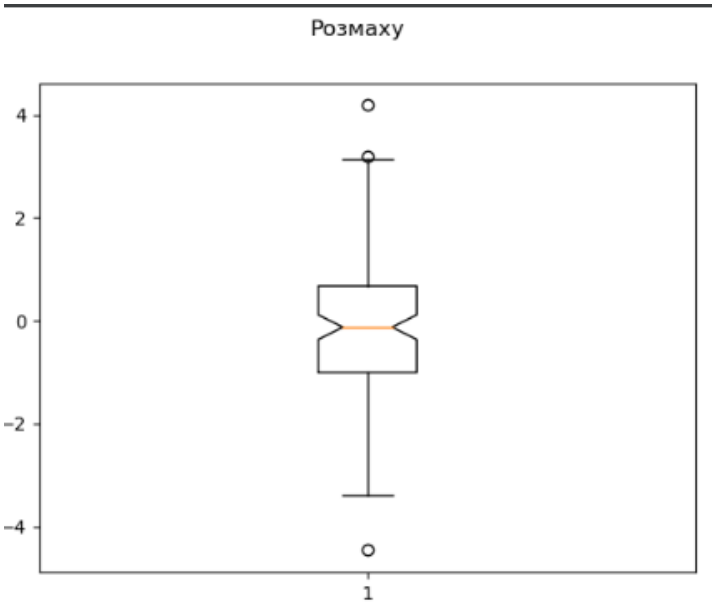
```python
# Дисперсія
def calc_variance(distribution_dict: dict) -> float:
    sample_mean = calc_sample_mean(distribution_dict)
    variance = 0
    for x in distribution_dict.keys():
        variance += (x - sample_mean) ** 2
    variance /= len(distribution_dict) - 1
    return variance
```

```python
# Середньоквадратичне відхилення
def calc_standard_deviation(distribution_dict: dict) -> float:
    variance = calc_variance(distribution_dict)
    return math.sqrt(variance)
```

```
Вибіркове середнє:-0.014029670000000063
Медіана:-0.09165
Мода:[0.8444, -0.235, 1.1011, 2.5892, -0.3981, -0.398, 2.6847, 1.3046,
Дисперсія:2.468113722295759
Середньоквадратичне відхилення:1.5710231450541265
```

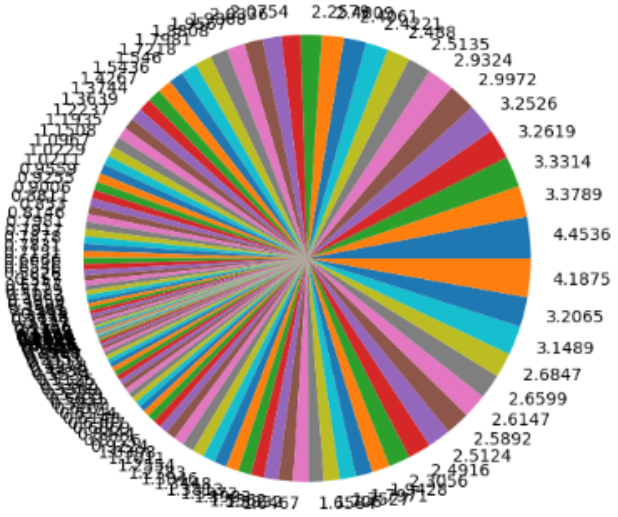
c. Будує діаграму розмаху, Парето та кругову

```python
# * Розмаху
plt.boxplot(od.keys(), od.values())
plt.suptitle("Розмаху")
```



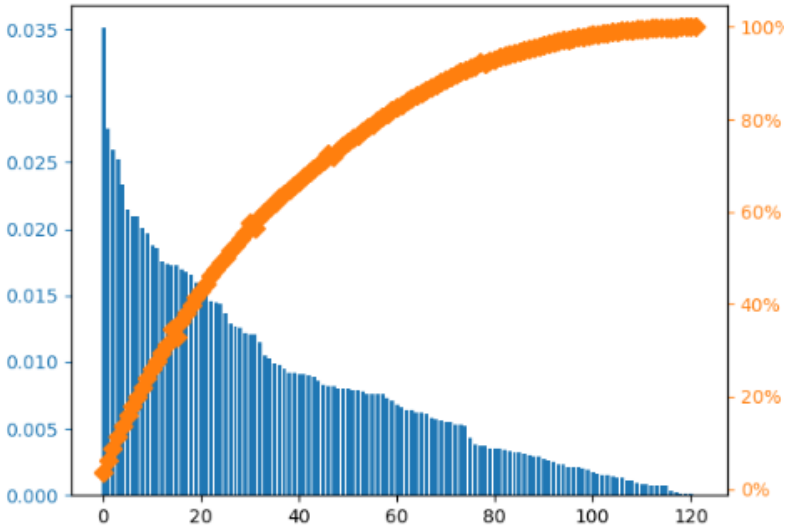

```python
# * Кругова
# * Створення списку для подальшої роботи
data = []
for key, value in od.items():
    if key >= 0:
        data.append((key, value))
    else:
        data.append((-key, value))

# * Створення кругової діаграми з абсолютними значеннями
plt.figure(figsize=(7, 6))
plt.pie([x[0] for x in data], labels=[str(x[0]) for x in data])
```

```python
# * Функція для зворотнього відображення значень на графіку
def backmapping(value):
    for x in data:
        if x[0] == value:
            if key < 0:
                return -x[0]
            else:
                return x[0]


# * Додавання зворотнього відображення на графіку
plt.gca().set_aspect('equal')
# plt.gca().legend(
#     [f"{backmapping(v)} ({v:.2f})" for v in [x[0] for x in data]],
#     title="Values",
#     loc='upper right'
# )
plt.suptitle("Кругова")

plt.show()
```



```python
# * Парета
df = pd.DataFrame.from_dict({'value': odv.values()})

df = df.sort_values(by='value', ascending=False)
df["cumpercentage"] = df["value"].cumsum()/df["value"].sum()*100


fig, ax = plt.subplots()
ax.bar(df.index, df["value"], color="C0")
ax2 = ax.twinx()
ax2.plot(df.index, df["cumpercentage"], color="C1", marker="D", ms=7)
ax2.yaxis.set_major_formatter(PercentFormatter())

ax.tick_params(axis="y", colors="C0")
ax2.tick_params(axis="y", colors="C1")
plt.suptitle("Парета")
```

Парета



d.

x	probability
0.8444	0.0035
-0.235	0.0076
1.1011	0.0076
2.5892	0.0058
-0.3981	0.0007
-0.398	0.0137
2.6847	0.0029
1.3046	0.0016
-0.7981	0.0003
0.9224	0.0068
-0.7878	0.0086
-0.7917	0.0001
0.4113	0.0055
-3.2526	0.002
-2.9324	0.0079
-0.9559	0.0015
-1.7218	0.0089
0.5342	0.0037

3.

Висновок: виконуючи лабораторну роботу ми навчились розраховувати числові характеристики вибірки програмними засобами.