МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені Тараса Шевченка
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
Кафедра програмних систем і технологій

Дисципліна
«Ймовірнісні основи програмної інженерії»

Лабораторна робота № 2

Виконав:	Якубець М. В.	**Перевірила**:	Вечерковська А. С.
Група	ІПЗ-22	Дата перевірки	
Форма навчання	денна	Оцінка	
Спеціальність	121		

2022

Назва: Лінійне перетворення та Графічне зображення даних.

Мета: Навчитись використовувати на практиці набуті знання про лінійні перетворення та графічне зображення даних.

Постановка задачі:

1. Знайдіть Q_1, Q_3 та P_{90}.
2. Знайдіть середнє та стандартне відхилення цих оцінок.
3. Через незадоволення низькими оцінками викладач вирішив використати шкалу форми y = ax + b, щоб відредагувати оцінки. Він хотів, щоб середнє значення масштабних оцінок становило 95, а оцінка 100, щоб залишалася рівною 100.
4. Показати дані за допомогою діаграми "стовбур – листя".
5. Відобразити дані за допомогою коробкового графіка.
6. Зробити висновок.

Математична модель:

DEFINITIONS

Concept 1

Know the k^{th} percentile of a set of discrete data

- The k^{th} *percentile*, P_k, is a value that splits the data into two parts Part 1 consisting of N_1 numbers that are less than P_k and part 2 consisting of N_2 numbers that are greater than P_k. The ratio $N_1 : N_2$ is $\dfrac{k}{100-k}$.

Concept 2

Know the three quartiles of a set of discrete data

- The 25^{th} percentile is called the *first* or *lower quartile* and denoted by Q_1.
- The 50^{th} percentile is called the *second* or *middle quartile* Q_2. It is also the median of the data.
- The *third* or *upper quartile* Q_3 is the 75^{th} percentile.

The k^{th} percentiles, the lower quartile, and the upper quartile of a data set of size N are sometimes referred to, respectively, as $\frac{k}{100}(N+1)^{\text{th}}$, $\frac{1}{4}(N+1)^{\text{th}}$, and $\frac{3}{4}(N+1)^{\text{th}}$ terms of the data.

$$\text{Var}(X) = \frac{1}{N}\sum_{x\in X}\left(f_x \cdot x^2\right) - \left(\bar{x}\right)^2, \text{ where } N = \sum_{x\in X} f_x .$$

VARIANCE AND STANDARD DEVIA

Concept 6

Alternative formula for the variance

- The *variance of a sample*, also called *unbiased variance*, is given by:

$$s_x^2 = \frac{\sum_{x \in X} f_i \cdot (x_i - \bar{x})^2}{N-1}$$

Concept 7

Know the definition of the variance of a sample

- And hence the *standard deviation of the sample* becomes:

$$s_x = \sqrt{s_x^2(x)}$$

STANDARDIZED SCORES (Z-SCORES)

Concept 5

Know the transformation that standardizes a set of data

The linear transformation $z = \frac{x - x}{\sigma}$ that associates to each point x_i in a set of data a point z_i in another set, standardizes the distribution given by X. The value of z indicates how many standard deviations an observation is away from the mean. It is called the z-score of this datum. It is left as an exercise to show that the mean of the distribution z is 0 while its standard deviation is 1.

LINEAR TRANSFORMATION

Concept 1

Know the formula for the mean after a linear transformation

The mean, variance, and standard deviation are the most commonly used measures to extract useful information from data. Some of their properties are discussed in this section.

A set of data X is said to be linearly transformed into a set Y if the elements of X are mapped onto the elements of Y by the relation $y = ax + b \in Y$, where a and b are real numbers.

The mean and standard deviation of Y are calculated as follows:

$$\bar{y} = \frac{\sum_{y \in Y} f_y \cdot y}{\sum_{y \in Y} f_y} = \frac{\sum_{x \in X} f_x \cdot (ax+b)}{\sum_{y \in X} f_x} = \frac{\sum_{x \in X} f_x \cdot (ax) + \sum_{x \in X} f_x \cdot (b)}{\sum_{y \in X} f_x} = \frac{a \sum_{x \in X} f_x x + b \sum_{x \in X} f_x}{\sum_{y \in X} f_x}$$

$$= a \frac{\sum_{x \in X} f_x x}{\sum_{y \in X} f_x} + b \frac{\sum_{x \in X} f_x}{\sum_{y \in X} f_x}$$

Hence, $\bar{y} = a \cdot \bar{x} + b$.

In a similar way, one can show that $\mathrm{Var}(Y) = a^2 \mathrm{Var}(X)$ and $\sigma_y = |a| \sigma_x$.

Concept 2

Know the formulas for the variance and the standard deviation after a linear transformation

EXAMPLE 2

Concept 3

Display data using a stem-and-leaf plot

A survey concerning the duration of telephone calls was conducted. Twenty calls were chosen at random. The durations of these calls, to the nearest minute, are listed below.

8, 25, 4, 32, 29, 41, 11, 21, 44, 5, 26, 16, 34, 23, 12, 37, 22, 18, 26, 23

Display the data using a stem-and-leaf plot.

Solution

The data in ascending order:

4, 5, 8, 11, 12, 16, 18, 21, 22, 23, 23, 25, 26, 26, 29, 32, 34, 37, 41, 44.

It is displayed in a stem-and-leaf plot as follows:

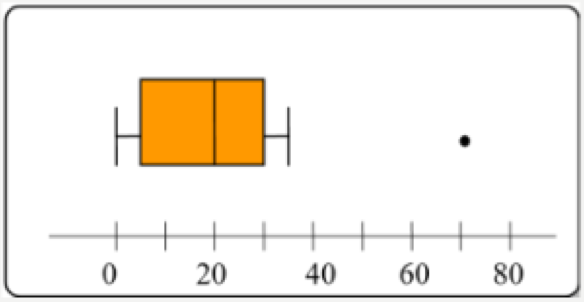
Stem	Leaf
0	4 5 8
1	1 2 6 8
2	1 2 3 3 5 6 6 9
3	2 4 7
4	1 4

Key: 2| 3 means 23

A quick look at the above stem-and-leaf plot shows that the minimum duration of a telephone call is 4 minutes and the maximum is 44. It also tells us that there are more calls between 20 and 30 minutes than any other 10-minute interval.

EXAMPLE 12. *SOLUTION*

The minimum distance is 0 km and the maximum distance is 70 km. The median is the 43rd term which is 20, the lower quartile is the 86/4 = 21.5th term which is 5 km, and the upper quartile is the 0.75×86 = 64.5th term which is 30 km.



IQR = 30 – 5 = 25

To determine the outliers: 1.5×25 = 37.5. There are no items that are more than 37.5 to the left of 5 but there is 1 item that is more than 37.5 to the right of 30. Therefore, 70 is the only outlier.

Код алгоритму:

```
from functools import reduce
from math import sqrt

import numpy as np

from utils import convert_list_items_to_type
```

```
def calculate_q_or_p(list_: list[int], fraction: float) ->
float:
    index = int(fraction * (len(list_) + 1)) - 1
    return list_[index] + fraction * (list_[index + 1] -
list_[index])


def calculate_mean(list_: list[int]) -> float:
    return reduce(lambda x, y: x + y, list_) / len(list_)


def _calculate_numerator_dispersion(list_: list[int],
mean: float) -> float:
    list_ = map(lambda x: (x - mean) ** 2, list_)
    return reduce(lambda x, y: x + y, list_)


def calculate_average_square_deviation(list_: list[int],
mean: float) -> float:
    numerator = _calculate_numerator_dispersion(list_,
mean)
    return sqrt(numerator / (len(list_) - 1))


def calculate_standard_deviation(list_: list[int], mean:
float) -> float:
    numerator = _calculate_numerator_dispersion(list_,
mean)
    return sqrt(numerator / len(list_))


def calculate_z_score(integer: int, mean: float,
standard_deviation: float) -> float:
    return (integer - mean) / standard_deviation


def __get_a_and_b(mean: float, to: int) -> np.ndarray:
    coefficients = [[100, 1], [mean, 1]]
    answers = [100, to]
    return np.linalg.solve(coefficients, answers)
```

```
def get_rearranged_list_for_teacher(list_: list[int],
mean: float) -> list[float]:
    a, b = _get_a_and_b(mean, 95)
    return [round(value * a + b, 2) for value in list_]


def create_stem_and_leaf_data(list_: list[int]) ->
dict[str, list[str]]:
    stem_and_leaf_data = {str(k): [] for k in range(10)}
    for value in convert_list_items_to_type(str, list_):
        stem, leaf = value[0], value[1:]
        if len(value) == 1:
            stem, leaf = "0", value
        stem_and_leaf_data[stem].append(leaf)
    return stem_and_leaf_data
```

Випробування алгоритму:

Задача №1:

$Q_1 = 62.75$

$Q_3 = 93.75$

$P_{90} = 99.5$

Задача №2:

Середнє квадратичне відхилення = 18.1

Стандартне відхилення = -1.99

Задача №3:

Відредаговані оцінки = [88.37, 92.64, 93.22, 93.41, 94.19, 94.19, 96.9, 98.06, 99.03, 100.0]

Задача №4:

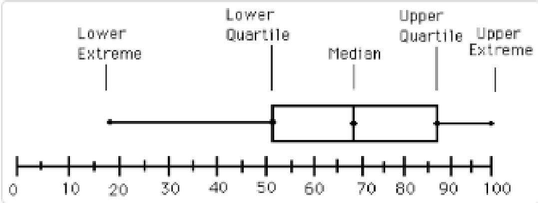
Діаграма стовбур-листя

Stem Leaf

```
 0 |
 1 | 00
 2 |
 3 |
 4 | 0
 5 |
 6 | 2 5 6
 7 | 0 0
 8 | 4
 9 | 0 5
```

Quartile calculator Q1, Q3

For quartiles Q1, Q3 calculation, please enter numerical data separated with comma (or space, tab, semicolon, or newline). For example: -235.4 -303.8 838.9 271.2 903.7 269.6 596.4 285.8 632.0 383.9 508.2 144.6 769.6



40, 65, 62, 70, 100, 90, 66, 70, 95, 84

Recalculate Reset

*For low count distributions, there is no universal agreement on selecting the quartile values (divide the ordered data set into two halves and then next halving...). If there are even number of data points, all methods give the same results.

Calculation:

Statistical file:
{40, 65, 62, 70, 100, 90, 66, 70, 95, 84}

Quartile Q1: 64.25
Quartile Q2: 70
Quartile Q3: 91.25

Other statistical characteristics:
Average (mean): μ=74.2
Absolute deviation: 144.4
Mean deviation: 14.44
Minimum: 40
Maximum: 100
Variance: 294.96
Standard deviation σ=17.174399552823
Corrected sample standard deviation s=18.103406677566

Z-score: {-1.9913, -0.5357, -0.7104, -0.2446, 1.5022, 0.92, -0.4775, -0.2446, 1.2111, 0.5706}
Count items: 10

Висновок: Навчився використовувати на практиці набуті знання про лінійні перетворення та графічне зображення даних. Перевірив зв’язок між квартилями, перцентилями та модою. Отримав досвід побудови діаграми стовбур-листя та коробкового графіка. Виявив, що чим більше даних, тим більше квартилі та перцентилі.