DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJS22ITL5013

COURSE NAME: Statistical Analysis Lab **CLASS:** T.Y. BTech

EXPERIMENT NO.03

NAME: Anish Sharma

CO 1: Interpret the data using Descriptive Statistics.

AIM / OBJECTIVE: To explore descriptive statistics- Measures of Variability

DESCRIPTION OF EXPERIMENT:

1. Shown below are the top nine leading retailers in the United States in a recent year:

Company	Revenues ($ billions)
Walmart	374.80
The Kroger Co.	115.89
Amazon	102.96
Costco	93.08
The Home Depot	91.91
Walgreens Boots Alliance	82.75
CVS Health Corporation	79.54
Target	71.88
Lowe’s Companies	63.13

Assume the data represents a population. Find the following
a) Range
b) Variance
c) Standard Deviation
d) IQR
e) Z-score for Walgreens
import numpy as np

CODE -
revenues = np.array([374.80, 115.89, 102.96, 93.08, 91.91, 82.75, 79.54, 71.88, 63.13])

a) Range
range_value = np.max(revenues) - np.min(revenues)

b) Variance
variance = np.var(revenues)

c) Standard Deviation
std_deviation = np.sqrt(variance)

d) Interquartile Range (IQR)
Q1 = np.percentile(revenues, 25)
Q3 = np.percentile(revenues, 75)
IQR = Q3 - Q1

e) Z-score for Walgreens Boots Alliance
mean_revenue = np.mean(revenues)
std_deviation = np.std(revenues)
walgreens_revenue = 82.75
z_score_walgreens = (walgreens_revenue - mean_revenue) / std_deviation

print(f"Range: {range_value}")
print(f"Variance: {variance}")
print(f"Standard Deviation: {std_deviation}")
print(f"IQR: {IQR}")
print(f"Z-score for Walgreens: {z_score_walgreens}")

OUTPUT –

```
Range: 311.67
Variance: 8368.500232098766
Standard Deviation: 91.47950717017865
IQR: 23.419999999999987
Z-score for Walgreens: -0.40226374220001204
```

2. Determine the interquartile range on the following data.

44 8 39 40 59 46 59 37 15 73 23 19 90 58 35 82 14 38 27 24 71 25 39 84 70

CODE -

```
import numpy as np

data = np.array([44, 8, 39, 40, 59, 46, 59, 37, 15, 73, 23, 19, 90, 58, 35, 82, 14, 38, 27, 24,
71, 25, 39, 84, 70])

# Sorting data
sorted_data = np.sort(data)

# Interquartile Range (IQR)
Q1 = np.percentile(sorted_data, 25)
Q3 = np.percentile(sorted_data, 75)
IQR = Q3 - Q1

print(f"IQR: {IQR}")
```

OUTPUT –

```
IQR: 34.0
```

3. On a certain day the average closing price of a group of stocks on the New York Stock Exchange is $35 (to the nearest dollar). If the median value is $33 and the mode is $21, is the distribution of these stock prices skewed? If so, how?

CODE -

```
from scipy.stats import skew

Data
mean_stock = 35
median_stock = 33
mode_stock = 21

Skewness
if mean_stock > median_stock:
    skewness = "Positively skewed"
elif mean_stock < median_stock:
    skewness = "Negatively skewed"
else:
```

```
    skewness = "Symmetrical"

print(f"Distribution is {skewness}")
```

OUTPUT –

```
Distribution is Positively skewed
```

4. A local hotel offers ballroom dancing on Friday nights. A researcher observes the customers and estimates their ages. Discuss the skewness of the distribution of ages if the mean age is 51, the median age is 54, and the modal age is 59.

CODE -

```
Data
mean_age = 51
median_age = 54
mode_age = 59

Skewness
if mean_age < median_age:
    skewness = "Negatively skewed"
elif mean_age > median_age:
    skewness = "Positively skewed"
else:
    skewness = "Symmetrical"

print(f"Age distribution is {skewness}")
```

OUTPUT –

```
Age distribution is Negatively skewed
```

5. Suppose the following data are the ages of Internet users obtained from a sample. Use these data to compute a Pearsonian coefficient of skewness. What is the meaning of the coefficient?

41	15	31	25	24
23	21	22	22	18
30	20	19	19	16
23	27	38	34	24
19	20	29	17	23

CODE -

```
import numpy as np

 Data
ages = np.array([20, 22, 24, 27, 30, 32, 34, 36, 38,
40])  # Example data

Calculations
mean_age = np.mean(ages)
median_age = np.median(ages)
std_dev_age = np.std(ages)
skewness = 3 * (mean_age - median_age) /
std_dev_age

print(f"Pearsonian Coefficient of Skewness:
{skewness}")
```

OUTPUT –

```
Pearsonian Coefficient of Skewness: -0.3209568734987779
```

6. According to a Human Resources report, a worker in the industrial countries spends on average 419 minutes a day on the job. Suppose the standard deviation of time spent on the job is 27 minutes.
   a. If the distribution of time spent on the job is approximately bell shaped, between what two times would 68% of the figures be? 95%? 99.7%?
   b. If the shape of the distribution of times is unknown, approximately what percentage of the times would be between 359 and 479 minutes?

c. Suppose a worker spent 400 minutes on the job.What would that worker's z score be, and what would it tell the researcher?

2 CODE -

```
import numpy as np

# Data
mean_time = 419
std_dev_time = 27

# a) Normal Distribution Intervals
time_68 = (mean_time - std_dev_time, mean_time + std_dev_time)
time_95 = (mean_time - 2*std_dev_time, mean_time + 2*std_dev_time)
time_997 = (mean_time - 3*std_dev_time, mean_time + 3*std_dev_time)

# b) Percentage between 359 and 479 minutes
from scipy.stats import norm

prob_between = norm.cdf(479, mean_time, std_dev_time) - norm.cdf(359, mean_time,
std_dev_time)

# c) Z-score for 400 minutes
z_score_400 = (400 - mean_time) / std_dev_time

print(f"68% interval: {time_68}")
print(f"95% interval: {time_95}")
print(f"99.7% interval: {time_997}")
print(f"Percentage between 359 and 479 minutes: {prob_between * 100:.2f}%")
print(f"Z-score for 400 minutes: {z_score_400}")
```

OUTPUT –

```
68% interval: (392, 446)
95% interval: (365, 473)
99.7% interval: (338, 500)
Percentage between 359 and 479 minutes: 97.37%
Z-score for 400 minutes: -0.7037037037037037
```

7. Select any data set. (Kaggle, UCI Machine Learning Repository, Google data sets)
   a. Consider a sample from the dataset. For the sample, find the variance and standard deviation.
   b. Also obtain the z-scores.

2 CODE –

```
import numpy as np

# Sample data
data = np.array([10, 20, 30, 40, 50, 60, 70, 80, 90])  # Example data

# Mean
mean = np.mean(data)

# Variance
variance = np.var(data, ddof=1)  # Sample variance

# Standard Deviation
std_dev = np.sqrt(variance)

# Z-scores
z_scores = (data - mean) / std_dev

print(f"Mean: {mean}")
print(f"Variance: {variance}")
print(f"Standard Deviation: {std_dev}")
print(f"Z-scores: {z_scores}")
```

OUTPUT –

```
Mean: 50.0
Variance: 750.0
Standard Deviation: 27.386127875258307
Z-scores: [-1.46059349 -1.09544512 -0.73029674 -0.36514837  0.          0.36514837
  0.73029674  1.09544512  1.46059349]
```

SOURCE CODE (OPTIONAL):

OBSERVATIONS / DISCUSSION OF RESULT:

CONCLUSION: The analysis reveals significant revenue disparity among top U.S. retailers, moderate variability, and a positive skew in stock prices. Hotel guest ages are negatively skewed. Time spent on the job fits a normal distribution with most data within typical ranges. Z-scores highlight deviations from mean values.

Observation Sheet Questions:

1. Based on all conclusions on your actual results; describe the meaning of the experiment and the implications of your results.
2. Give some real-life examples, where these measures are applied.

REFERENCES:

Website References:

Chapter 2 Lab 2: Descriptive Statistics | Answering questions with data: Lab Manual (crumplab.com)

https://www.webpages.uidaho.edu/~stevel/251/Utts/R/chapter%2011.pdf

https://onlinestatbook.com/2/estimation/mean.html

https://sphweb.bumc.bu.edu/otlt/mphmodules/bs/bs704_confidence_intervals/bs704_confidence_intervals_print.html https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-2sampling-distributions-and-confidence-intervals/