Unit 04
Data And Analysis

EXERCISE

Multiple Choice Questions (MCQs)

MCQ	1	2	3	4	5	6	7	8	9	10
Answer	A	C	B	B	B	C	A	A	D	D

Short Questions

Q1. List out the parameters and statistics from given statements.

a) **Average length of height of a giraffe.**
b) **Average weight of watermelon**
c) **There are 430 doctors in a hospital.**
d) **Average age of students of 6th class in a school is 12 years.**

Parameters

Parameters refers to the ***entire*** population.

a) Average length of height of a giraffe
b) Average weight of watermelon
c) There are 430 doctors in a hospital
d) Average age of students of 6th class in a school is 12 years.

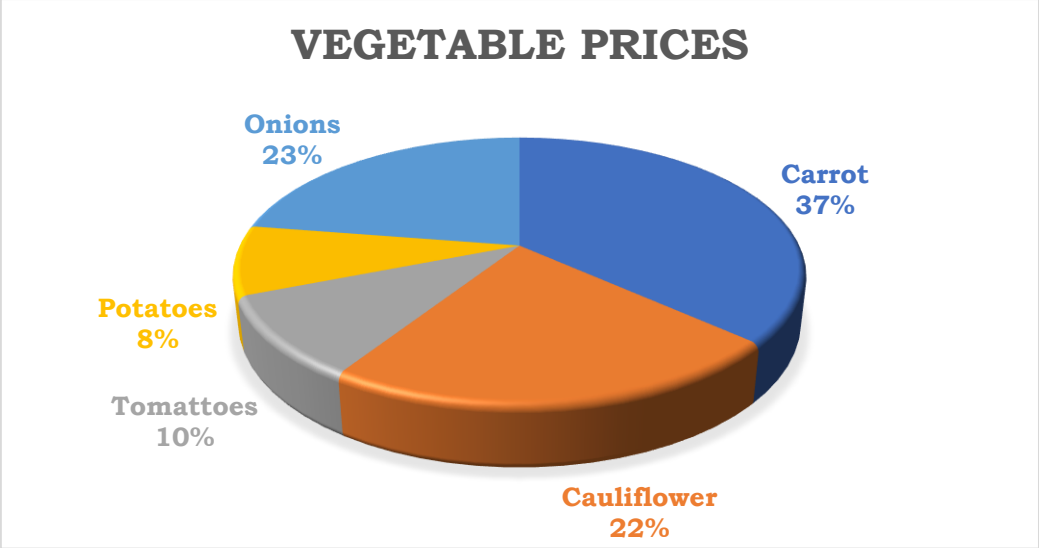
All these statements refer to the *entire* population therefore, they are all parameters.

Q2. If you want to make a report regarding the products exported from Pakistan in last five years, how libraries can help you to collect data? Write steps.

Using Libraries for Data Collection

- Use online libraries to access specialized trade databases like Trade Map for export data.
- Use library catalogs to find books, reports, or research papers on Pakistan's export history over the last five years.
- Seek help from librarians to locate specific trade-related sources and historical data archives.

Q3. Make a pie chart of vegetable prices in the market. Consider five to ten vegetables.



Q4. Enlist steps to represent monthly temperatures of a Pakistani city in 2023 from January till December using a line graph.

Steps:

- Get monthly temperatures of a Pakistani city in 2023 from Jan to Dec.
- Write this Python code to create a line graph.

```
import numpy as np

import matplotlib.pyplot as plt

x=np.array(["Jan", "Feb", "Mar", "Apr", "May", "Jun", "Aug", "Sep",
"Oct", "Nov", "Dec"])

y=np.array([15,19,22,25,28,33,36,27,24,21,17,10])

plt.plot(x,y,'o-b',label='Temperatures')

plt.xlabel("Month")

plt.ylabel("Temperature")

plt.legend()

plt.show()
```

- Save the file as ***"temp.py"***
- Run the program. The line chart will appear.

Long Questions

Q1. Sketch primary data collection methods in the context of disease outbreak like seasonal flu.

Primary Data Collection Methods For Seasonal Flu

The primary data collection methods in the context of disease outbreak like seasonal flue are the following.

1. Surveys and Questionnaires

The surveys can be distributed to individuals in affected areas to collect data about flu symptoms, vaccination history and healthcare utilization.

2. Interviews

The interviews can be conducted with patients and health workers to understand flu symptoms, timeline and contact with others.

3. Observations

The observation method is used to monitor and record real time data. You can observe and document the number of patients come in clinics with flu symptoms.

4. Medical Record Review

This method is used to analyze patient records for trends and patterns about the flu. The hospital record can be reviewed to identify the number of flu cases, severity and outcomes.

5. Field Reports

Field reports can be collected from public health workers about the spread of flu, vaccination rates and community response.

6. Focus Groups

The qualitative data can be collected from a group of individuals. Focus groups can be held with community members to discuss their concerns and experiences with the flu.

Q2. Argue about the use of statistical modeling techniques. Highlight all techniques discussed in this unit.

Statistical Modeling Techniques

Statistical modeling techniques are used for gathering data. There are two categories of statistical modeling methods in data analysis.

- Supervised learning
- Unsupervised learning

1. Supervised Learning

In supervised learning model, the algorithm uses a labeled dataset for learning. The answer key is used by the algorithm to determine the accuracy of data.

Supervised learning techniques in statistical modeling include ***regression model*** and ***classification model***.

Regression Model

If the result, label or outcome of a model is continuous value then it is called *regression*. The most common regression model is linear model. A *linear regression model* is a mathematical equation that allows us to predict a response for a given predictor value.

Classification Model

If the result, label or outcome of a model is a discrete value then it will be called *classification*. For example, if we predict that a certain employee will get raise in salary or not then it is classification.

2. Unsupervised Learning

In unsupervised learning model, the algorithm is given unlabeled data and attempts to extract features and determine patterns independently. There are two methods of unsupervised learning.

- Clustering algorithm
- Association rules

Clustering

Clustering means group of data items with maximum similarities. *For example*, to reduce churn rate of customer, a telecom company analyzes the usage of customers and divide them into three clusters:

- Customers with long call duration.
- Customers with heavy internet usage.
- Customers with short calls and average Internet usage.

Now the company provides attractive packages to all these kinds of users to retain them.

Association

Association means how likely is to do second action if first action is done. For example, in a supermarket:

- *Customer 1* buys bread, milk, tea and tissues.
- *Customer 2* buys bread, milk, coffee and eggs.

Now if *customer 3* buys bread, then there are more chances that he will also buy milk. There is an *association* between milk and bread.

K-means Clustering

This algorithm combines a specified number of data points into specific groupings based on similarities.

Q3. Compare linear regression and classification. Emphasize on their respective roles in statistical modeling.

Linear Regression Model

The most common regression model is linear model. A *linear regression model* is a mathematical equation that allows us to predict a response for a given predictor value.

The variable which is used for prediction is called *independent variable* (**x**) and the variable to be predicted based on the value of variable **x** is called *dependent variable* (**y**).

In simple linear regression the equation of the line is **y=mx+b** where **m** is the slope and **b** is intercept.

Intercept is the point where the graph line meets the y-axis. It is also called *y-intercept*.

Slope is calculated by the term *rise/run* where ***rise*** is the distance from x-axis to y-intercept and ***run*** is the distance between y-axis and y-intercept.

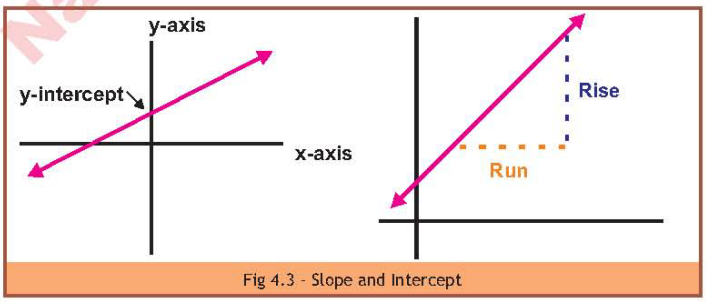


Fig 4.3 - Slope and Intercept

Classification Model

If the result, label or outcome of a model is a discrete value then it will be called *classification*. For example, if we predict that a certain employee will get increase in salary or not then it is classification. But if we make a prediction that how much salary increase an employee will get based on some statistical data then it is called *regression model*.

Q5. Defend either of supervised learning and unsupervised learning. Give reasons for your preference to the other.

Supervise Learning vs Unsupervised Learning

I prefer *supervised learning* over unsupervised learning because it's more straightforward and accurate for many tasks. Here's why:

Better Predictions

Supervised learning uses labeled data, meaning the model knows the correct answers during training. This makes it much better at making accurate predictions.

Clear Goals

In supervised learning, the goal is simple: predict the right answer. Since we know what the outcome should be, it's easier to measure how well the model is performing. Unsupervised learning doesn't have clear goals, so it's harder to evaluate its success.

Easy to Understand

Supervised learning models are often easier to interpret, especially simpler ones like decision trees. This makes it easier for people to trust and understand the decisions, which is important in fields like healthcare or finance.

Real-World Use

Many practical applications, like customer churn prediction, product recommendations, or image classification, are best handled by supervised learning because it provides more reliable and usable results.

Why I Prefer Supervised Over Unsupervised Learning?

While unsupervised learning is useful for exploring patterns and finding clusters in data, it's harder to interpret and act on its results. Supervised learning is more direct, producing clear, accurate predictions that are easier to use in real-world situations.

Q6. Write a Python code to generate a dataset with variables where $y=x^2+2x$. Fit scatter plot and box plot on this data.

Scatter Plot For $y=x^2+2x$

```
1. import numpy as np
2. import matplotlib as plt
3. x=np.linspace(-15, 30, 50)
4. y=(x*x)+(2*x)
5. plt.figure(figsize=(10,6))
6. plt.scatter(x, y, color='red', label= 'y=x^2+2x')
7. plt.title('Scatter Plot')
8. plt.xlabel('x')
9. plt.ylabel('y')
10. plt.grid(True)
11. plt.legend()
12. plt.show()
```

Box Plot For $y=x^2+2x$

```
1. import numpy as np
2. import matplotlib as plt
3. x=np.linspace(-15, 30, 100)
4. y=(x*x)+(2*x)
5. plt.figure(figsize=(10,6))
6. plt.boxplot(y)
7. plt.title('Box Plot')
8. plt.xlabel('y')
9. plt.show()
```

Q7. Relate some real-world examples (other than Airbnb, Facebook and YouTube) where data science was used to improve marketing strategies and enhance the business.

Real-World Examples of Data Science

Here are some examples of companies using data science to improve their marketing strategies and enhance the business.

1. Netflix

It recommends shows and movies based on what people like to watch, keeping users engaged and helping them find new content.

2. Amazon

It suggests products based on shopping habits, which boosts sales. About 35% of their sales come from these recommendations.

3. Spotify

It creates personalized music playlists by analyzing listening habits, making users stay on the app longer.

4. Coca-Cola

It uses customer data to predict what people want and sends them personalized offers, improving engagement.

5. Zara

Tracks fashion trends in real time and adjusts its clothing lines and marketing to match what customers are currently into, keeping sales high.