

**INTRODUCTION TO DATA ANALYTICS MODULE 1 REVIEW**

**INSTRUCTOR: DR. Mursi**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Total marks (50)**

**Due date: 26th Feb 2023**

**Instructions:**

1. **Attempt all the questions**
2. **Paste your output below each question**

Question 1 (15mks)

1. **Describe the steps involved in formulating a research hypothesis.**

STEPS INVOLVED IN FORMULATING A RESEARCH HYPOTHESIS

1. Writing a research question that needs to be answered. The question should be specific and researchable.

Example: Does ones blood group increase their susceptibility to HIV infection?

1. Conduct research on past studies concerning the topic to enable you form informed assumptions about th research findings.
2. Identify the independent and dependent variables. In this case the blood group is dependent while HIV infection is independent.
3. Formulating the hypothesis : involves writing the answer to the research question. Formulate the alternative hypothesis and the null hypothesis.
4. Make your hypothesis testable by ensuring that it has relevant variables, specific groups being studied and predicted outcome.
5. Test the hypothesis: a study research can be carried and data collected for HIV positive patients and their blood groups. Then the H0 can be tested

**Provide an example of a research question and how it can be converted into a testable hypothesis (6mks)**

Research question: Is there an association between ABO blood group and the risk of HIV infection?

H1: There is an association between HIV infection and blood group.

Ho: There is no association between HIV infection and blood group.

1. Using examples describe when to use correlation coefficient and linear regression? **(4mks)**

Correlation coefficient is used to investigate whether changes in one variable are associated with changes in other variables. The correlation coefficient measures the strength and direction of the linear relationship between two variables (denoted as x and y)

For example if students in Kenya are making subject selection in form 2. Teachers like to use scores in mathematics to guide learners in choosing Physics subjects.

The Pearson correlation coefficient can be used to investigate whether scores in mathematics are related to physics.

Linear regression is used to describe the relationships between a set of independent variables and the dependent variable. It predicts the value of one variable (dependent variable) based on the value of another (independent variable).

Example one can predict a students score in Physics based on their mathematics results.

1. **Differentiate the different types outliers in data analysis and identify potential consequences of outliers (5mks).**

Outliers are values that significantly deviate from the majority of data points in a dataset.

They can be either **true outliers** (representing natural variations) or **other outliers** (resulting from errors or anomalies).

**Types of Outliers:**

**Contextual Outliers (Conditional Anomalies)**

These outliers depend on specific contexts or conditions.

Example: Suppose you measure running times, but accidentally start the timer midway through someone’s sprint. This erroneous data point is a contextual outlier.

**Global Outliers (Point Anomalies)**

These are extreme values that stand out from the entire dataset.

Example: Imagine measuring 100-meter running times for college students. A few exceptionally fast or slow times represent global outliers due to various influencing factors.

**Collective Outliers**

These occur when a group of data points deviates together eg.Unusuallylow scores in an exam following a long school holiday, collectively forms outliers.

**CONSEQUENCES OF OUTLIERS:**

1. **Impact on Statistical Analyses:** Outliers can skew statistical results, affecting measures like means, standard deviations, and correlations. Hypothesis tests may yield misleading conclusions if outliers are not addressed.
2. **Modeling Distortion:** Outliers can significantly impact predictive models (e.g., linear regression). Models may become less accurate or less robust.
3. **Data Integrity and Interpretation:** Incorrectly handled outliers can compromise data integrity. Interpretation of results becomes challenging if outliers are not properly manage

## Question 2 (9mks)

1. A group of Biostatistics students were tasked with investigating a recent outbreak of waterborne disease in a particular region. They have collected data on various factors that may be related to the spread of the disease. **(9mks)**

**Variables:**

1. Age (Numeric): Age of the affected individuals.
2. Symptom Onset Date (Date): Date when symptoms first appeared.
3. Location (Categorical): Categorized as Urban, Suburban, or Rural.
4. Water Source (Categorical): Source of water supply, such as Municipal, Well, or Spring.
5. Duration of Symptoms (Numeric): Number of days the individuals experienced symptoms.
6. Household Size (Numeric): Number of people in the affected individuals' households.
7. **In the context of the outbreak investigation, provide an example of a null hypothesis and an alternative hypothesis related to one of the variables and identify both dependent and independent Variables (e.g., water source) (4mks)**

H1: The source of water is associated with the location

H0: The source of water is not associated with the location

1. **Why is it important to have both null and alternative hypotheses in a hypothesis test? How do they complement each other in the decision-making process? (2mks)**
   1. Together, the null and alternative hypotheses cover **all possible outcomes**.
   2. They are **exhaustive**, meaning that one of them must be true (there’s no middle ground).
   3. They are also **mutually exclusive**, ensuring that only one hypothesis can hold at a time.

In summary, the null hypothesis provides a baseline assumption, while the alternative hypothesis represents the researcher’s hypothesis about an effect or relationship in the population. By comparing sample data to these hypotheses,

1. **Describe a scenario where linear regression could be applied to this dataset (3mks)**

Linear regression is used to describe the relationships between a set of independent variables and the dependent variable. It predicts the value of one variable (dependent variable) based on the value of another (independent variable).

In this case it can be used to describe the relationship between age (independent variable) and duration of symptoms (dependent variable) in a patient. It can also predict the duration of symptoms for a given age

## Question 3 (10mks)

**Case Study: Customer Segmentation for an E-commerce Platform**

**A leading e-commerce platform is seeking to refine its marketing strategies to target different customer segments effectively. They have gathered a dataset containing various attributes related to customer behavior and demographics.**

**Attributes**:

1. **Age (Numeric)**: Age of the customer in years.
2. **Gender (Categorical)**: Categorized as Male, Female, or Non-binary.
3. **Purchase History (Numeric)**: Total amount spent by the customer on the platform.
4. **Location (Categorical)**: Customer's location, categorized as Urban, Suburban, or Rural.
5. **Frequency of Purchases (Numeric)**: Number of purchases made by the customer.
6. **Preferred Product Category (Categorical)**: Customer's preferred product category, such as Electronics, Apparel, Beauty, etc.
7. **The e-commerce platform intends to target specific customer segments for personalized marketing campaigns. Based on the attributes provided, suggest two potential customer segments and explain why you chose them. (4mks)**

Frequency of purchases and preferred product category will show me what the customer wants and their purchasing power.

1. **Identify a pair of variables in the dataset that could potentially have a linear relationship. Explain why you chose these variables** **(3mks)**

Frequency of purchases and preferred product have a linear relationship because a customer will buy a product more often because they prefer it over another.

Example: a customer may prefer Elliots bread over Festive, United or Supa loaf brands hence purchase it more frequently.

1. Identify two categorical variables from the dataset that could be used for a Chi-square test of independence. Explain why you chose these variables. **(3mks)**

Gender and preferred product category.

Chi-square test of independence is a statistical method used to test whether two categorical variables are related or independent. Gender can be categorized as male/female. Preferred product is also categorized into beauty, electronics etc

## Question 4 (16mks)

1. The severity of a disease and blood group were studied in a research project. The findings are given in the following table, known as the contingency table. *Can the severity of the condition and blood group be associated*? Conclude on Hypothesis at 5% level of significance **(6mks) chi sq**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Condition/BG | O | A | B | AB | Total |
| Severe | 51 | 40 | 10 | 9 | 110 |
| Moderate | 105 | 103 | 25 | 17 | 250 |
| Total | 156 | 143 | 35 | 26 | 360 |

1. Below Table shows the Age and Weight of 7 students.

|  |  |  |
| --- | --- | --- |
| Student | Age | Weight (Kgs) |
| 1 | 17 | 55 |
| 2 | 15 | 56 |
| 3 | 30 | 62 |
| 4 | 45 | 23 |
| 5 | 11 | 33 |
| 6 | 32 | 63 |
| 7 | 56 | 59 |

Calculate;

1. Calculate coefficient correlation and make a conclusion **(3mks)**
2. Suppose you have a dataset representing the relationship between hours of study (x) and exam scores (y) for a group of students

|  |  |
| --- | --- |
| Study hours (X) | Exam scores (Y) |
| 12 | 55 |
| 13 | 56 |
| 14 | 62 |
| 15 | 70 |
| 7 | 75 |
| 9 | 85 |
| 11 | 80 |

Calculate;

1. Correlation r **(3mks)**
2. Predict y when x is 9 **(4mks) regression**