

**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. Provide an example of a research question and how it can be converted into a testable hypothesis **(6mks)**
2. **Identify the Research Problem: Recognize and define the problem or topic of interest. Example: Investigating the impact of a new diet on weight loss.**
3. **Conduct a Literature Review: Review existing literature to understand what has already been studied and identify gaps. Example: Finding that there is limited research on the specific diet's long-term effects.**
4. **Define the Research Question: Narrow down the focus to a specific, researchable question.**
5. **Specify Variables: Determine the independent and dependent variables. Example: Independent variable - type of diet (ketogenic vs. low-fat), Dependent variable - weight loss.**
6. **Formulate the Hypothesis: Create a clear and testable hypothesis based on the research question. Example: "Individuals on the ketogenic diet will lose more weight over six months compared to those on a low-fat diet."**
7. **Test the Hypothesis: Design an experiment or study to test the hypothesis. Ensure that it is measurable and that the hypothesis can be supported or refuted by data.**

**Example:**

**-Research Question: Does caffeine improve cognitive performance in adults?**

**- Hypothesis: Adults who consume caffeine will perform better on cognitive tests than those who do not consume caffeine**.

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

**- Correlation Coefficient: Used to determine the strength and direction of the relationship between two continuous variables. Example: Examining the relationship between hours of study and exam scores. If the correlation coefficient is close to +1, it indicates a strong positive relationship.**

**- Linear Regression: Used to predict the value of a dependent variable based on the value of one or more independent variables. Example: Predicting exam scores based on the number of hours studied. The regression equation can help estimate the expected exam score for a given number of study hours.**

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

**Univariate Outliers- Outliers in a single variable.**

**Multivariate Outliers-Outliers in the context of multiple variables.**

**Contextual Outliers- Values that are considered outliers within a specific context or environment.**

**Consequences**

**1. Distorted Statistical Results: Outliers can skew measures such as the mean and standard deviation, leading to inaccurate conclusions.**

**2. Misleading Patterns: They can obscure true patterns in the data, potentially leading to incorrect inferences.**

**3. Reduced Model Accuracy: In regression analysis, outliers can lead to models that do not accurately represent the data, affecting predictions and decisions based on the model.**

## 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)**

**-Null Hypothesis -the type of water source has no effect on the duration of symptoms experienced.**

**-Alternative Hypothesis- The type of water source affects the duration of symptoms experienced.**

**Variables**

**- Dependent Variable: Duration of symptoms.**

**- Independent Variable: Water source**.

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)**

**Null Hypothesis - Provides a statement of no effect or no difference**

**Alternative Hypothesis - Indicates the presence of an effect or difference**

**They ensure a rigorous and unbiased approach to testing and interpreting results.**

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

**Linear regression can be applied to see the relationship between the independent variable which is age and dependent variable which is duration of symptoms. It will determine if age affects the duration of symptoms Does old age mean longer duration while young age mean lesser.**

## 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)**

**1. High-Spending Urban Professionals**

**Attributes: Age-25-45**

**Purchase History- High total amount spent**

**Location- Urban areas**

**High frequency of purchasing**

**Product Category -Electronics and Apparel**

**Rationale-This segment likely consists of professionals with disposable income who frequently shop online for convenience. They are significant contributors to the platform's revenue and may respond well to marketing campaigns focused on premium products, new technology releases, and fashion trends.**

**2.**

**Budget-Conscious Rural Families**

**Attributes; -Age- 30-50**

**Location-Rural**

**Less frequent at purchasing things**

**Preferred Product Category- Beauty, Home Goods, and Groceries**

**Rationale: This segment is likely more price-sensitive and values practical, everyday items. They may benefit from targeted campaigns featuring discounts, bulk buying options, and promotions on essential goods. Marketing strategies could highlight the convenience and cost-effectiveness of using the platform**.

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

**Purchasing History and Frequency of buying products.**

**This is reasonable to hypothesize that customers who purchase more frequently also tend to have a higher total amount spent. Analyzing this relationship can help in understanding purchasing behavior patterns, such as whether high-frequency shoppers are also high spenders, which can inform inventory and promotional strategies.**

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)**

**1. Gender and Preferred Product Category**

**-This test can help determine if there is an association between a customer's gender and their preferred product category. Understanding this relationship can guide the platform in creating gender-specific marketing strategies and tailoring product recommendations to better align with customer preferences.**

**2. Location and Preferred Product Category**

**-This test can reveal if there is a significant relationship between a customer's location (Urban, Suburban, Rural) and their preferred product category. Insights from this analysis can aid in regional marketing efforts, stock management, and localization of the shopping experience to cater to the preferences of customers from different area.**

## 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)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

**Hypotheses**

**- Null Hypothesis (H0): There is no association between the severity of the condition and blood group.**

**- Alternative Hypothesis (H1): There is an association between the severity of the condition and blood group.**

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)**

**Means= bar{X} = \frac{17 + 15 + 30 + 45 + 11 + 32 + 56}{7} = \frac{206}{7} =29.43**

**bar{Y} = \frac{55 + 56 + 62 + 23 + 33 + 63 + 59}{7} = \frac{351}{7} = 50.14**

**Standard Deviaton**

**| Student | Age (X) | Weight (Y) | \(X - \bar{X}\) | \(Y - \bar{Y}\) | \((X - \bar{X})(Y - \bar{Y})\) |**

**-60.39 +-84.53 +6.76 +-422.51+315.93+33.04 +235.64 = 23.945**

**Sum of squared deviations**

**sum (X - \bar{X})^2 = (-12.43)^2 + (-14.43)^2 + (0.57)^2 + (15.57)^2 + (-18.43)^2 + (2.57)^2 + (26.57)^2 = 1514.57**

**sum (Y - \bar{Y})^2 = (4.86)^2 + (5.86)^2 + (11.86)^2 + (-27.14)^2 + (-17.14)^2 + (12.86)^2 + (8.86)^2 = 1377.94**

**Pearson correlation coefficient**

**r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{23.94}{\sqrt{1514.57 \times 1377.94}} = 0.015**

**The Pearson correlation coefficient is 0015,showing a very weak positive linear relationship between age and weight**.

1. 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)**

**Means**

**bar{X} = \frac{12 + 13 + 14 + 15 + 7 + 9 + 11}{7} = \frac{81}{7} =11.57**

**bar{Y} = \frac{55 + 56 + 62 + 70 + 75 + 85 + 80}{7} = \frac{483}{7}= 69.00**

**Sum of products of deviations**

**sum (X - \bar{X})(Y - \bar{Y}) = -6.02 - 18.59 - 17.01 + 3.43 - 27.42 - 41.09 - 6.29 = -113.99**

**Squared deviations**

**sum (X - \bar{X})^2 = (0.43)^2 + (1.43)^2 + (2.43)^2 + (3.43)^2 + (-4.57)^2 + (-2.57)^2 + (-0.57)^2 = 47.43**

**sum (Y - \bar{Y})^2 = (-14.00)^2 + (-13.00)^2 + (-7.00)^2 + (1.00)^2 + (6.00)^2 + (16.00)^2 + (11.00)^2 = 788.00**

**Pearson =-0569**

1. Predict y when x is 9 **(4mks)**

**Y = 83.23 - 1.23 \times 9 = 83.23 - 11.07 = 72.16**