

**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)**

**Answer:** STEP 1: Coming up with a question, for example does exercise affect weight loss?

STEP2:Youare supposed to identify the variables; Ind ependent is like testing by doing exercise while Dependent is by observing if there will be loss of weight.

STEP 3: Formulating null hypothesis; in this case exercise does not affect loss of weight.

STEP 4: Formulating alternative hypothesis; In this step there is a relationship between doing exercise and losing of weight. When one does, an exercise will lose compared to those who do not do exercise.

STEP 5: Making it testable; if one does an exercise frequently his /her results of losing weight will be excellent compared to those who do not do exercise.

STEP 6: Predict the outcome; In our case is that those who does exercise will have lost a lot of weight compared to those who don’t.

STEP 7: Test the hypothesis-It’s about analyzing the Hypothesis to determine if we reject or fail to reject.

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

**Answer:** Correlation Coefficient: Used to assess the strength and direction of the linear relationship between two variables, such as height and weight.

Linear Regression: Employed to model the relationship between variables and make predictions, such as predicting sales revenue based on advertising spending

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

**Answer:**1. Collective outliers: are clusters or groups of observations that collectively deviate from the overall pattern of the dataset.

Consequences Misleading interpretation of data patterns if collective outliers are not properly identified and analysed

2. Global outliers: are data points that deviate significantly from the rest of the dataset across all variables. Consequences, Misrepresentation of central tendency Measure.

3. Contextual outliers: are observations that are unusual only within a specific subset or context of the data. Consequences: Influence on subgroup analysis, leading to incorrect conclusions.

## 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)**

**Answer:**

Null hypothesis (H0): There is no significant difference in the mean duration of symptoms between individuals who use municipal water, well water, and spring water as their primary water source.

Alternative hypothesis (H1): There is a significant difference in the mean duration of symptoms between individuals who use municipal water, well water, and spring water as their primary water source.

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)**

**Answer:**

The null hypothesis provides a baseline assumption that there is no effect or difference in the population being studied. It serves as the starting point for statistical testing. The alternative hypothesis, on the other hand, suggests that there is an effect or difference present in the population.

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

**Answer:**

Linear regression could be applied to this dataset to examine the relationship between age (independent variable) and the duration of symptoms (dependent variable)

## 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)**

**Answer:** Age and Preferred Product Category: Segmenting customers based on age and preferred product category allows targeted marketing campaigns tailored to different age groups' interests.

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

**Answer:** Purchase History and Frequency of Purchases: These variables might have a linear relationship as customers who spend more are likely to purchase 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)**

**Answer:** Location and Preferred Product Category: Analyzing if location influences product preferences helps tailor promotions to regional interests.

## 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)**

**Answer:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |  |  | A=0.05 |  |
|  |  |  |  |  |  |  |  |  |  |
| **Expected** |  |  |  |  |  |  |  |  |  |
| BG | O | A | B | AB | TOTAL |  |  | DF=(R-1)(C-1) | 3 |
| SEVERE | 47.66667 | 43.69444 | 10.69444 | 7.944444 | 110 |  |  | CV | 7.814728 |
| MODERATE | 108.3333 | 99.30556 | 24.30556 | 18.05556 | 250 |  |  |  |  |
| TOTAL | 156 | 143 | 35 | 26 | 360 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **CHI SQUARE** |  |  |  |  |  |  |  |  |  |
| BG | O | A | B | AB | TOTAL |  |  |  |  |
| SEVERE | 0.2331 | 0.312372 | 0.045094 | 0.140249 | 0.730815 |  |  |  |  |
| MODERATE | 0.102564 | 0.137444 | 0.019841 | 0.061709 | 0.321558 |  |  |  |  |
| TOTAL | 0.335664 | 0.449816 | 0.064935 | 0.201958 | 1.052373 |  |  |  |  |

**It can be associated.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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)**

**Answer**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| student | Age | Weight (Kgs) | xD1:G15y | x^2 | y^2 |
| 1 | 17 | 55 | 935 | 289 | 3025 |
| 2 | 15 | 56 | 840 | 225 | 3136 |
| 3 | 30 | 62 | 1860 | 900 | 3844 |
| 4 | 45 | 23 | 1035 | 2025 | 529 |
| 5 | 11 | 33 | 363 | 121 | 1089 |
| 6 | 32 | 63 | 2016 | 1024 | 3969 |
| 7 | 56 | 59 | 3304 | 3136 | 3481 |
|  | 206 | 351 | 10353 | 7720 | 19073 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| num | 72471 | 72306 | 165 |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| den | 11604 | 10310 |  |  |  |
|  | 107.7219 | 101.5382 | 10937.88 |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| r | 0.015085 |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| there is no correlation between the age and weight |  |  |  | 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)**
2. Predict y when x is 9 **(4mks)**

**Answer:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Study hours (X) | Exam scores (Y) | xy | x^2 | y^2 |  |  |
| 12 | 55 | 660 | 144 | 3025 |  |  |
| 13 | 56 | 728 | 169 | 3136 |  |  |
| 14 | 62 | 868 | 196 | 3844 |  |  |
| 15 | 70 | 1050 | 225 | 4900 |  |  |
| 7 | 75 | 525 | 49 | 5625 |  |  |
| 9 | 85 | 765 | 81 | 7225 |  |  |
| 11 | 80 | 880 | 121 | 6400 |  |  |
| 81 | 483 | 5476 | 985 |  | sum |  |
| 11.57143 | 69 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| b | -113 | -2.36826 |  |  |  |  |
|  | 47.71429 |  |  |  |  |  |
|  |  |  |  |  | r | -0.58604 |
| a | 96.40419 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Y | 94.03593 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| y | 75.08982 | prediction |  |  |  |  |