**My Name – Vaishali Ramdas Salve**

**Date- 26/12/2023**

**Batch No- DS2311**

**Sir- Shwetank Mishra**

**Assignment Name – Statistics worksheet – 1**

**Q1 to Q9 have only one correct answer. Choose the correct to answer your question.**

1. Bernoulli random variables take ( only ) the value. 1 and 0
2. Ttue
3. Which of the following theorem states that the distribution of iid variables, property normalization becomes that of a standard normal as the sample size increase?

c) Central Limit Theorem

3. Which of the following is incorrect with respect to use of poussing distribution

B) Modeling bounded count data

4. Point out the correct statement

D) All of the mentioned

5. ------- Random variables are used to model rates.

C) Poisson

6. 10.Usually replacing the standard error by the its estimated value does change the CLT.

B) False

7. Which of the following testing is concerned with making decisions using data?

B) Hypothesis

8. 4.Normalized data are centered with making decision using data?

A) 0

9. Which of the following statement is incorrect with respect to outliers?

C) Outliers can not conform to the regression relationship

**Q10 and Q15 are subjected answer type question , Answer them in your own words brie fly.**

**10. what do you understand by the term normal Distribution?**

A normal distribution , also known as a Gaussian distribution ,is a symmetric probability distribution characterized by its bell – shaped curve . in this distribution , the majorty of data points cluster around the mean, and the probabilities for values farther from the mean decrease symmetrically. The standard deviation . measures the spread of the distribution. Many natural phenomena, such as human height or tedt scored, exhit a normal distribution.

1. Deletion:

. Listwise Deletion: Removing entire cases with missing values.

2 . Imputation:

. Mean/ Median/ Mode Imputation : Filling missing values with the mean, median , or model of the observed values.

1. Forward or Back wared Fill: using the value before or after the missing value in a time series.
2. Interpolation: Estimating missing values based on existing data points.
3. Multiple Imputaton: Generating multiple data based with imputed values to account for uncertainty.

The choice of imputation depends on the nature of the data and potential impact on analysis.

12. what is A/B testing

A/B testing is a statistical method used in marketing used in marketing , product development , and other fields to compare two versions of a variable, typically labeled as A and B. the goal is to determine which version performs better in term of a predefined metric , such as click-through rates, conversion rates, or user engagement.

In an A/B test , a randomly selected group of individuals is exposed to version a, while another group is exposed to version b.

13. Is mean imputation of missing data acceptable practice?

**Pros:**

1. Preservation of Sample Size: Mean imputation allows you to retain the full sample size for analysis.
2. Easy Implementation: It is straightforward to implement and does not require complex computations.

**Cons:**

1. Distortion of variability : Mean Imputation can underestimate the true variability in the data since it artificially reduces variance.
2. Bias Introduction : If missing data is not completely random, mean imputation may introduce bias, especially if missingness is related to the variable being imputed.
3. Inappropriate for Categorical Data: Mean imputation is not suitable for categorical data.

**Considerations:**

. Use with caution : Its acceptable in certain scenarios, but caution is advised, especially when missingness is non-random or substantial.

. Explore Alternatives : Depending on the context, other imputation method like multiple imputation or model-based imputation may be more appropriate.

14. What is linear Regression in statistics?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variable by fitting a linear equation to observed data. The goal is to find the best –fitting line, that minimizes the sum of the observes and predicted valves.

The linear regression equation is typically expressed as:

***Y = Bo + B1 X1 + B2 X2 + . . . + Bn Xn + E***

Where:

. Y is the dependent variable,

. X1,X2,. . . ,Xn are independent variables,

. B0 is the y- intercept,

. B1 B2,. . . ,Bn are the coefficient representing the relationship’s strength and direction,

. E is the error term accounting for unobserved factors.

Linear regression is widely used for prediction and understanding the relationship between variables. It assumes a linear relationship , which means the change in the dependent variable is proportional to changes in the independent variables.

15. What are the various branches of statistics?

1. Descriptive Statistics:

. Involves method for summarizing and describing the main features of a dataset. Measures such as mean, median , and standard deviation fall under this category.

2 . Inferential Statisstics:

. Focuses on making prediction or inferences about a population based on a sample of data. This includes hypothesis testing and confidence intervals.

3 . Biostatistics:

. Applies statistics method to biological and medical data. It plays a crucial role in clinical trials, epidemiology, and genetics.

4 . Econometrics:

. Utilizes statistical methods to analyze economic data. It help economists test theories, forecast trends, and make policy recommendations.

5 . Actuarial science:

. Involves the application of statistical method to assess risk in insurance, finance and pension industries.

6 . Spatial statistics:

. Focuses on analyzing spatial and geographical dats.

7 . Multivariate statistics:

. Deals with the analysis of dataset with multiple variables. Techniques such as factor analysis and principal component analysis fall under this branch.

These branches highlight the diverse application of statistics across various field , reflecting its importance in making informed decisions based on data.