STATISTICS WORKSHEET-1

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

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Ans1- a) True
Ans2 – a) Central Limit Theorem
Ans3 – b) Modeling bounded count data
Ans4- d) all of the mentioned
Ans5- c) Poisson
Ans6- b) False
Ans7- b) Hypothesis
Ans8 – a) 0
Ans9 – c) Outliers cannot conform to the regression relationship.

Q10. What do you understand by the term Normal Distribution?

Ans. - Normal distribution is also known as Gaussian Distribution / bell curve. In this majority of data point cluster around mean and here mean =median = mode and shape of Normal distribution is perfectly symmetrical like bell shaped. Ex- Height, weight, test scores generally have normal distribution.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. – Missing data occurs when there is no data stored for a variable in a dataset. Handling of missing data is crucial for the accuracy of analysis. We can use various imputation technique for the missing values but it depends on nature of dataset and analysis being conducted.

Imputation techniques

1 – for basic imputation techniques we can use mean or mode or median to fill missing values

2- for advanced imputation techniques we use KNN(k-nearest neighbours), Time series imputation, Regression imputation etc

the choices of technique depends on type of data set. For ex—

Mean- is used where missing data completely believed to be random and it follows approximately normal distribution.
Median – is preferred when distribution is not normal or outlier is present.
Mode- is preferred for categorical or nominal variables.
KNN- is effective when there is local similarity among observations and KNN can be applied for numerical and categorical variables both.
Time-series imputation- can be used for time series structure, imputing missing values based on historical data.
Regression imputation- can be used for the dataset where there are strong relationship between the variables with missing data and other variables.

Thus every data set need specific imputation and some data set also need domain specific . So we recommend to use imputation based on data requirement.

Q12. What is A/B testing?

Ans- A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

EX- If we have a product and we run two adds for the product then we can compare which brings more customers or higher traffic on the website. We can use hypothesis testing for this by taking Ho→ No significant difference between both and H1→ one brings more traffic than others

Q13.Is mean imputation of missing data acceptable practice?

Ans – Mean imputation of missing data is simple and widely used technique but nature of data and goals of analysis dfecide its acceptability

Benefits-

- Simple to implement
- Allow to retain same number of observation.
- It works good when data is commonly random.

Limitations-

- It can introduce bias if data is not random.
- It can underestimate variability of the imputed data.
- It can lead to incorrect estimate of correlation.

Yes, mean imputation of missing data is generally an acceptable practice, but it can introduce biases and potentially impact the accuracy of your analysis.

Q14.What is linear regression in statistics?

Ans. Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms is a statistical method that is used to model and quantify the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best fits the observed data, allowing prediction and understanding of how changes in the independent variables affect the dependent variable.

Q15. What are the various branches of statistics?

Ans. The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.