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

Answer:- (a)

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

Answer:- (a)

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

Answer :- (B)

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.

Answer:- (d)

5. ______ random variables are used to model rates.

(a) Empirical

(b) Binomial

(c) Poisson

(d) All of the mentioned.

Answer:- (c)

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

(a) True

(b) False

Answer:- (b)

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

(a) Probability

(b) Hypothesis

(c) Causal

(d) None of the mentioned

Answer:- (b)

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

(a) 0 (b) 5 (c) 1 (d) 10

Answer:- (a)

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

Answer:- (c)

Q (10 – 15) are subjective answer type questions, Answer them in your own words briefly.

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

Answer:- The normal distribution, also called the Gaussian distribution. Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than the data far from the mean. The normal distribution is the most commonly seen continuous distribution in nature. Just as the binomial distribution, every event is independent from one another. In the normal distribution the mean, medium and mode all line up such that the center of the distribution is the mean. Because of this, exactly half of the results fall to either side of the mean. The normal distribution is also identifiable by its bell shape and may sometimes be referred to as a bell curve.

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

Answer:- Missing data is an inevitable part of the process. As data researchers, we pour a lot of resources, time and energy into making sure the data set is as accurate as possible. However, data inevitably goes missing. Missing data is a huge problem for data analysis because it distorts findings. According to data scientists, there are three types of missing data:-

(1) Missing completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern.
(2) Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data.
(3) Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

Imputation techniques are:-

a. Predict missing values.
b. Substitute a value such as mean
c. Ignore the records with missing values (Many tools ignore records with missing values. When the percentage of records with missing values is small).
d. Use deletion methods to eliminate missing data.

12. What is A/B testing?

Answer: - A/B testing is also known as split testing or bucket testing. It allows decision makers to choose the best design for a website by looking at analytics results obtained with two possible alternatives A and B. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal**.** A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why certain elements of their experiences impact user behavior. In A/B testing, **A** refers to 'control' or the original testing variable. Whereas **B** refers to 'variation' or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

Answer:- True, the process to replace null values in a dataset with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Disadvantages of mean imputations:-

- It reduces the variance of the imputed variables.
- It shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- It does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Answer:- L**inear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable, denoted as *x*, is regarded as the predictor, features or independent variable. The other variable, denoted as *y*, is regarded as the response, outcome, label or dependent variable. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) Does a set of independent variables do a good job in predicting an outcome (dependent) variable?
(2) Which variables in particular are significant predictors of the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula :

$$y = m*x + c$$

where, y = estimated dependent variable score.
c = constant.
m = regression coefficient.
x = score on the independent variable.

15. What are the various branches of statistics?

Answer:- There are two real branches of statistics: Descriptive and Inferential Statistics

Descriptive Statistics:- The term “descriptive statistics” refers to the analysis, summary, and presentation of findings related to a data set derived from a sample or entire population. Descriptive statistics comprises two main categories – Measures of Central tendency, and Measures of dispersion. Although descriptive statistics may provide information regarding a data set, they do not allow for conclusions to be made based on the data analysis but rather provide a description of the data being analyzed.

Inferential Statistics:- Inferential statistics enables one to make descriptions of data and draw inferences and conclusions from the respective data. Through inferential statistics, an individual can conclude what a population may think or how it’s been affected by taking sample data. Inferential statistics is mainly used to derive estimates about a large group (or population) and draw conclusions on the data based on hypotheses testing methods. Inferential statistics uses sample data because it is more cost-effective and less tedious than collecting data from an entire population. It allows one to come to reasonable assumptions about the larger population based on a sample’s characteristics. Sampling methods need to be unbiased and random for statistical conclusions and inferences to be validated.