Statistics Worksheet -1

Q.1. Bernoulli random variables take (only) the values 1 and 0

a) True

b) False

Solution : a) True

Q.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

Solution : a) Central Limit Theorem because it states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.

Q.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

Solution : b) Modeling bounded count data because Poisson distribution is used for modeling unbounded count data

Q.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

Solution: d) All of the mentioned

Q.5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Solution: c) Poisson

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

a) True

b) False

Solution : b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Solution : b) The Hypothesis testing method is concerned with making decisions using data.

Q.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

Solution : a) 0

Q.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

Solution : c) Outliers cannot conform to the regression relationship

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

Solution : Normal Distribution , also known as Gaussian, Gauss, or Laplace–Gauss distribution is a type of continuous probability distribution , that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. For Example, Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

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

Solution : Best techniques to handle missing data**.**

1. Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings.

2. Use regression analysis to systematically eliminate data

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Data scientists can use data imputation techniques

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data.

Q.12. What is A/B testing?

Solution : A/B tests, also known as split tests, is one of the simplest forms of a randomized controlled experiment. allow you to compare 2 versions of something to learn which is more effective.

Q.13. Is mean imputation of missing data acceptable practice?

Solution : The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q.14. What is linear regression in statistics?

Solution : Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable. The variable we are using to predict the other variable's value is called the independent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Q.15. What are the various branches of statistics?

Solution : There are two branches of Statistics.

Descriptive Statistics :

Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

Inferential Statistics :

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.