STATISTICS WORKSHEET-1

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

a) True

b) False

Answer: - a) True

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) Central Limit Theorem

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) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of 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) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: - c) Poisson

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

a) True

b) False

Answer: - b) False

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) Hypothesis

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) 0

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) Outliers cannot conform to the regression relationship

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

Answer: - Normal Distribution is also called as bell curve. It is a 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 a normal distribution the mean is zero and the standard deviation is 1. It has zero skew. The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. The normal distribution model is motivated by the Central Limit Theorem.

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

Answer: - To handle missing data first I will use the isnull() to identify how many null values are present in a table. Then I will check the percentage of missing data because if more than 50 percent of the data in particular row is missing then we will need to accumulate proper information for that particular column else deleting the entire column would be logical instead of treating it manually and giving incorrect data for the model to be trained and tested upon. If there are very few data missing then depending on the data can use mean, median and mode options to fill the correct data. The imputation techniques that I will be using are mean imputation, simple imputer, iterative imputer and knn imputer.

12. What is A/B testing?

Answer: - A/B testing also known as split testing. An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

Answer: - Mean imputation is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e., continuous as well as categorical data and so it makes advantageous over other imputations.

There are some limitations too: -

1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random, the estimate of the mean remains unbiased.
2. Mean Imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

Answer:- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form $Y = mx + c$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when $x = 0$).

Types of linear regression: -

1. Simple linear regression
2. Multiple linear regressions
3. Logistic regression

4. Ordinal regression
5. Multinomial regression

15. What are the various branches of statistics?

Answer: -The two branches of statistics are descriptive statistics and inferential statistics. All these branches of statistics follow a specific scientific approach which makes them equally essential to every statistics student.

1. **Descriptive Statistics:** Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations.

Descriptive statistics can be categorized into

a) Measures of central tendency
b) Measures of variability

To easily understand the analyzed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

a) Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

- **Mean**

  This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.

- **Median**

  This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score which is at the center of the arranged sample.

- **Mode**

  This is the frequently occurring value in a given set of scores.

b) Measures of Variability

The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

2. **Inferential Statistics:** Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.

   The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis