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

a)True
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

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

e) 0
f) 5
g) 1
h) 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.

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

Ans.) Also called Gaussian distribution, the most common distribution function for independent, randomly generated variables. The measures of central tendency (mean, mode and median) are exactly the same in a normal distribution.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 graph form, normal distribution will appear as a bell curve. Data is symmetrically distributed with no skew.

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

Ans.) The first step in handling missing values is to look at the data carefully and find out all the missing values.The following code shows the total number of missing values in each column.It also shows the total number of missing values in entire data set.Lets take train dataset.

```
import pandas as pd
train_df = pd.read_csv("train.csv")
#Find the missing values from each column
train_df.isnull().sum()
#Find the total number of missing values from the entire dataset
train_df.isnull().sum().sum()
```

Some Imputation techniques……

1) **Replacing With Arbitrary Value**: If you can make an educated guess about the missing value then you can replace it with some arbitrary value using the “fill na” method
2) **Replacing With Mean** : This is the most common method of imputing missing values of numeric columns.If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.Then we will use ‘fillna’method for imputing the columns with the mean of respective column values.
3) **Replacing With Mode** : Mode is the most frequently occurring value. It is used in the case of categorical features.Then the ‘fillna’ method is used for imputing the categorical columns
4) **Replacing With Median :** Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.You can use ‘fillna’ method for imputing the column values.
5) **Simple imputer**
6) **KNN imputer**
7) **Iterative imputer**

Q.12What is A/B testing?

Ans.) A/B testing is basically statistical hypothesis testing, or statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

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

Ans.) 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 .

Q.14 What is linear regression in statistics?

Ans.) 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 predictor 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 :

$$y = c + mx,$$

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

Q.15 What are the various branches of statistics?

Ans.) **DESCRIPTIVE STATISTICS**

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, Kurtosis and skewness.

INFERENTIAL STATISTICS

Inferential statistics help you come to conclusions and make predictions based on your data.When you have collected data from a sample, you can use inferential statistics to understand the larger population from which the sample is taken.Inferential statistics have two main uses:

- making estimates about populations
- testing hypothesis to draw conclusions about populations