STATISTICS WORKSHEET-1

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

a) True
b) False

Ans. a) True , Bernoulli random variables take (only) the value 1 and 0.

Q.2 Which of the following theorm 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

Ans. a) Central Limit Theorem

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

Ans. b) Modeling bounded 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

Ans. 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

Ans. c) Poisson

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

a) True

b) False

Ans. b) False

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

a) Probability

b)Hypothesis

c) Casual

d) None of the mentioned

Ans. b) Hypothesis

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

Ans. 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

Ans. c) Outliers cannot conform to the regression relationship

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

Ans. Normal distribution is the probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurance than data far from the mean. It is also called Gaussian distribution. In graphical form the normal distribution appears as a “bell curve”.

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

Ans. Missing data reduces the statistical power of the analysis, which can distort the validity of the results. When dealing with missing data, we can use two primary methods to solve the error: imputation or the removal of data.

When data is missing, it may make sense to delete data. Instead of deletion, data scientists have multiple solutions to impute the value of missing data.

There are two types of imputation –

a) Single Imputation:- When there are a small number of missing observations we can calculate the mean or median of the existing observations but when there are many missing variables, mean, median results can result in a loss of variation in the data. This method does not use time series characteristics or depend on the relationship between the variables.

b) Multiple Imputation :- Multiple imputations is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point , the missing values are exchanged for values that encompass the natural variability and uncertinity of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.Multiple Imputationa can produce statistically valid results even when there is a small sample size or a large amount of missing data.

Q.12 What is A/B testing?

Ans. A/B also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable are shown two different segmentsof website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

A/B is one of the component of the overaching process of Conversion Rate Optimization(CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behaviour, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections etc.

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

Ans. Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.
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 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 increases 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?
Ans. Linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative. For example, it can be used to quantify the relative impact of age, gender, and diet (the predictor variables) on height (the outcome variable).

Q.15 What are various branches of statistics?
Ans. Descriptive statistics and inferential statistics are the two main branches of statistics. Both the statistics branches are used in scientific data analysis and are equally significant for a statistics student.

a) Descriptive statistics ;- It deals with the presentation and collection of data. Descriptive statistics can be categorized into:-
   (i) Measures of central tendency to examine the value distribution center.
   (ii) Measure of variability which helps the statisticians analyze the distribution from a particular data set.
   b) inferential statistics:- Inferential statistics are statistical techniques to utilize data from a sample to conclude, predict the behaviour of a given population, and make judgement or decisions.