WORKSHEET

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

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?
c) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

4. Point out the correct statement.
d) All of the mentioned

5. _____ random variables are used to model rates.
c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
b) False

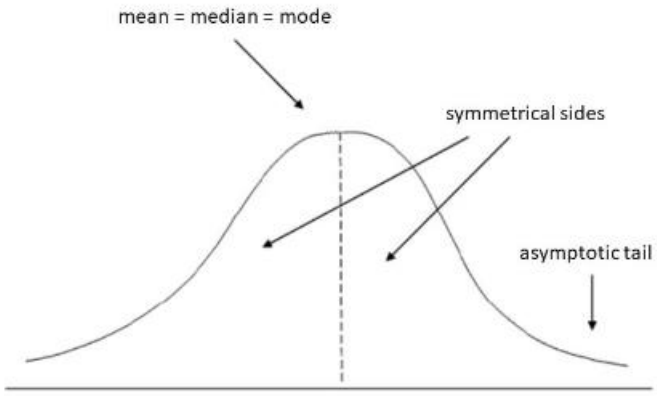
7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
Ans.: Normal Distribution is continuous distribution in nature. In this every event is independent from one another. In this distribution the center of the distribution is mean and all mean, median and mode are line up in such a way that the center is mean.
Here half of the results are fall on the either side of the mean forming bell shape curve; also called bell curve.



The Normal Distribution provides the probability of the value in a particular range for a given experiment.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans.: Missing data indicates that while getting or capturing the information from the data source, required data is missing or wrongly updated in system for analysis. In this the conditions may arise that few data is not required or irrelevant; for such cases we can remove the data which is not required. At few places were data is wrongly updated, we need to replace/put the data by using imputation methods; because doing the data analysis with missing can affect the outcome or result. There are various types of missing data handling. Data can be missing at 1.Missing At Random(MAR) 2. Missing completely At Random(MAR) 3. Missing Not At Random(MCAR) Depending upon the conditions, observation, and requirement missing data can be filled for analysis.
Data can be deleted with below methods:
1.List wise 2.Pair wise and 3.Dropping variables. Here depending upon the data availability and the data required for analysis is considered. If the large amount of data available few thing can be dropped which will not affect the outcome/result of the analysis.

Imputation methods for replacing data are as below;
1. Mean Median and Mode
2. Time Series Specific Method
3. LOCF and NOCB method
4. Linear Interpol
5. Multiple Imputation
6. KNN(K Nearest Neighbor) method

Multiple Imputation technique is good for imputation. In this instead of putting/replace one single value at each place missing values are exchanged for values that incorporate the natural variability and uncertainty of the right value. Using imputed data the process is repeated to make multiple imputed data sets. Then each set is analyzed using standard analytical procedure and results are combined to get overall result. This will give best result in case of small data sample or large data missing.

12. What is A/B testing?

Ans.:A/B testing is a statistical way of comparing two or more versions, such as version A and version B. To determine not only which version perform better but also to understand if a difference between two versions is statistically significant.

A/B tests are used to optimize marketing campaigns, increase conversation etc.

13. Is mean imputation of missing data acceptable practice?

Ans.: Mean imputation of missing data is not a acceptable practice.

The reasons are as bellow;

1).It reduces the variance of the imputed variable.

2).It does not preserve the relation between variables like correlation.

3).Since the imputed values are at mean it shrinks the standard errors which affects the hypothesis tests and confidence interval calculations.

14. What is linear regression in statistics?

Ans.: Linear Regression is a statistical model used for predictive analysis. In this there are basically two variables independent variable and dependent variable. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15. What are the various branches of statistics?

Ans.: There are two main branches of statistics, i) Descriptive statistics ii) Inferential statistics.

i)Descriptive statistics- If data can be described without any statistical tools then it is called descriptive statistics . ex, marks in class , height of student.

Descriptive statistics include mean (average), variance, skewness, and kurtosis.

ii) Inferential statistics: If data is too big then then we use inferential statistics. This statistical techniques allow us to utilize data from a sample to conclude, predict the behavior of a given population, and make judgments or decisions.