STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Ans- option (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

Ans- option (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

Ans- option (b); Modeling bounded count data

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- option (d); All of the mentioned

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans- option (c); Poisson

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

b) False

Ans- option (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

Ans- option (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

Ans- option (a); 0

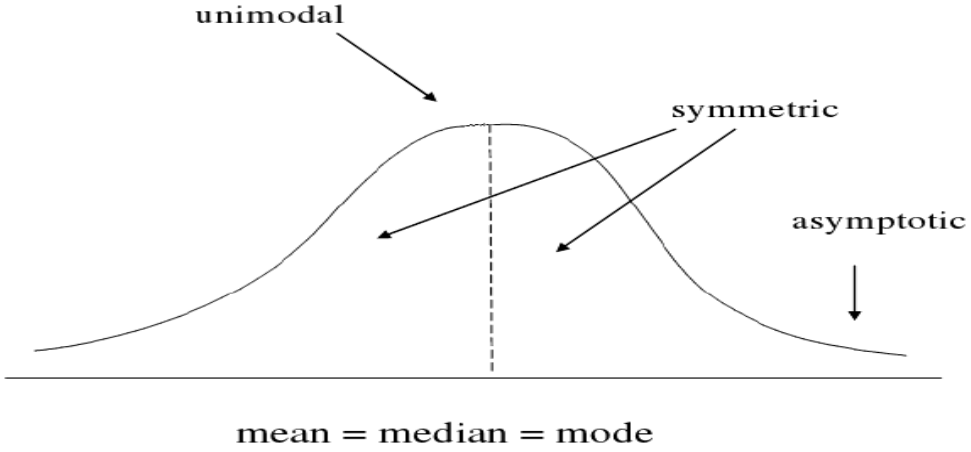
9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Ans- option (c); Outliers cannot conform to the regression relationship

WORKSHEET

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

Ans- A normal distribution is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean. The further a data point is from the mean, the less likely it is to occur. There are many things, such as intelligence, height, and blood pressure, that naturally follow a normal distribution.



The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean

The normal distribution is produced by the normal density function, $p(x) = e^{-(x-\mu)2/2\sigma 2/\sigma}\sqrt{2\pi}$. In this exponential function e is the constant 2.71828…, is the mean, and σ is the standard deviation

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.
Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

- Deleting the Missing values (Delete entire row and column)
- Imputing the Missing Values

Imputing the Missing Value

- Replacing With Arbitrary Value
- Replacing With Mean
- Replacing With Mode
- Replacing With Median
- Replacing with next value – Backward fill
- Missing values can also be imputed using interpolation
- Impute the Most Frequent Value
- Impute the Value “missing”, which treats it as a Separate Category
- Nearest Neighbors Imputations (KNNImputer**)**

12. **What is A/B testing?**

Ans- A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

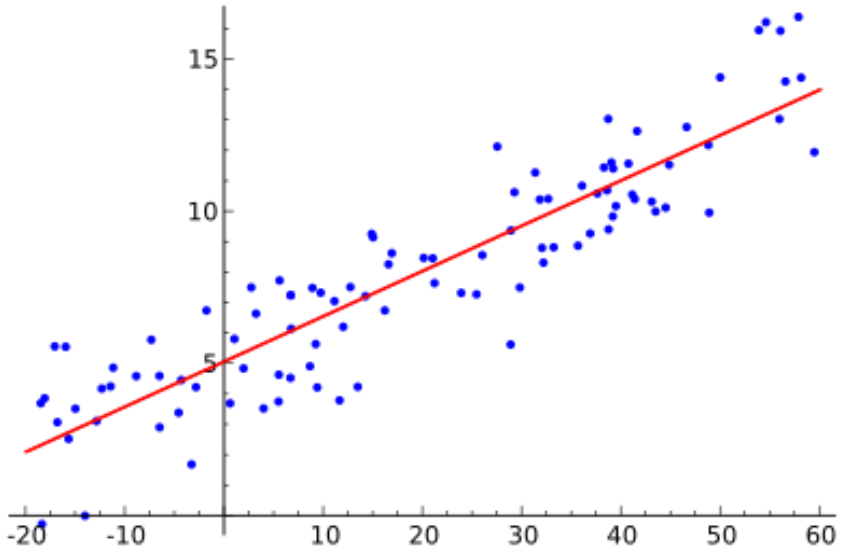
An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

Ans:- Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out

14. What is linear regression in statistics?

Ans- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.



simple linear regression, which has one independent variable

15. What are the various branches of statistics?

Ans-The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.