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 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

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 known as the Gaussian distribution, 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 graphical form, the normal distribution appears as a "bell curve".

Importance of normal distribution

The normal distribution is one of the most important probability distributions for independent random variables for three main reasons.

1. Normal distribution describes the distribution of values for many natural phenomena in a wide range of areas, including biology, physical science, mathematics, finance and economics. It can also represent these random variables accurately.

2. the normal distribution is important because it can be used to approximate other types of probability distribution, such as binomial, hypergeometric, inverse (or negative) hypergeometric, negative binomial and Poisson distribution.

3. normal distribution is the key idea behind the central limit theorem, or CLT, which states that averages calculated from independent, identically distributed random variables have approximately normal distributions. This is true regardless of the type of distribution from which the variables are sampled, as long as it has finite variance.

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

Answer: Mostly used strategies to handle missing values in the dataset -

1. Deleting Rows with missing values.
2. Impute missing values for continuous variable.
3. Impute missing values for categorical variable.
4. Other Imputation Methods.
5. Using Algorithms that support missing values.
6. Prediction of missing values.

Imputation techniques -

Next or Previous Value.
K Nearest Neighbors.
Maximum or Minimum Value.
Missing Value Prediction.
Most Frequent Value.
Average or Linear Interpolation.
Median Value.
Fixed Value.

12. What is A/B testing?

Answer: A/B testing is a methodology for comparing two versions of a webpage or app against each other to determine which one performs better. Also known as split testing, refer to a randomized experimentation process wherein two or more version of a variable are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Is mean imputation of missing data acceptable practice?

Answer: Yes, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

14. What is linear regression in statistics?

Answer: Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

15. What are the various branches of statistics?

Answer: The two main branches of statistics are descriptive statistics and inferential statistics.

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.