

Statistics

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

Ans. 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?

Ans. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. b) Modeling bounded count data

4. Point out the correct statement.

Ans. d) All of the mentioned

5. _____ random variables are used to model rates.

Ans. c) Poisson

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

Ans. b) False

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

Ans. b) Hypothesis

8. 4. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

Ans. a) 0

9. Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers cannot conform to the regression relationship

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

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.

A normal distribution is symmetrical around its center. Which means the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution. Normal distributions are continuous and have tails that are asymptotic, which means that they approach but never touch the x-axis. The center of a normal distribution is located at its peak, and 50% of the data lies above the mean, while 50% lies below. It follows that the mean, median, and mode are all equal in a normal distribution.

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

Ans. To handle missing data, we need to analyse 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:

1. Deleting the Missing values
2. Imputing the Missing Values

I would recommend KNN as it can be much more accurate than the mean, median or most frequent imputation methods.

12. What is A/B testing?

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

13. Is mean imputation of missing data acceptable practice?

Ans. The process of replacing null values with the mean value of the data is known as mean imputation.

Mean imputation is not considered as acceptable practice since it ignores feature correlation.

Also, it decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans. Linear regression is the most widely used statistical technique. It is a technique to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predictions about data.

15. What are the various branches of statistics?

Ans. The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data (central tendency, variability, and distribution of sample data.), and inferential statistics, which uses those properties to test hypotheses and draw conclusions.