Statistics worksheet

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

ANSWER : 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?

ANSWER : Central Limit Theorem (CLT)

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

ANSWER : Modeling bounded count data

4. Point out the correct statement

ANSWER :

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

5. Poisson random variables are used to model rates

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

ANSWER : False

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

ANSWER : Hypothesis

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

ANSWER : 0

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

ANSWER : Outliers cannot conform to the regression relationship

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

ANSWER : We say the distribution of data is normal when it is symmetric around the mean,when +-68% of the the data lies within one standard deviation of the mean, 95% within two standard deviation of the mean and 99.8% of it within three standard deviations of the mean.

normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

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

ANSWER : Types of Missing Data

1. Missing Completely At Random (MCAR)

2. Missing At Random (MAR)
3. Not Missing At Random (NMAR)

Common Methods

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations
3. Random forest

12. What is A/B testing?

ANSWER: 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.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above case, you may divide the products into two parts – A and B. Here A will remain unchanged while you

make significant changes in B's packaging. Now, on the basis of the response from customer groups. who used A and B respectively, you try to decide which is performing better.

who used A and B respectively, you try to decide which is performing better.

13. Is mean imputation of missing data acceptable practice?

ANSWER: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation.

For an example:

We have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the 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 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?

ANSWER: 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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

15. What are the various branches of statistics?

ANSWER : Two branches, **descriptive statistics and inferential statistics**, comprise the field of statistics.