STATISTICS – WORKSHEET 1: ANSWERS

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, 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.

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

Answer:

There are mainly two ways that we can deal with missing values in a dataset:

- If there are only a few rows with missing values or if a column has an overwhelming number of missing values, we can simply drop them from the dataset without running into the risk of losing too much information.

- An alternative approach to handling missing values is via imputation. Imputation is the process whereby the missing values in a dataset are replaced with some substituted values.

There are three types of different imputation techniques:

- **Simple Imputer:** Simple imputer follows a univariate approach to imputing missing values i.e. it only takes a single feature into consideration. Some of the most common uses of simple imputer are: a) Mean, b) Median, c) Mode.

- **Iterative Imputer:** Iterative imputer is an example of a multivariate approach to imputation. It models the missing values in a column by using information from the other columns in a dataset.

- **KNN Imputer:** KNN Imputer which is another multivariate imputation technique. KNN Imputer scans our dataframe for k nearest observations to the row with missing value.

12. What is A/B testing?

Answer: A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

13. Is mean imputation of missing data acceptable practice?

Answer: YES, Mean imputation is typically considered terrible practice since it ignores feature correlation (The process of replacing null values in a data collection with the data's mean is known as mean imputation).

14. What is linear regression in statistics?

Answer: Linear regression establishes the linear relationship between two variables based on a line of best fit.

15. What are the various branches of statistics?

Types of Statistics:

- **Descriptive Statistics:** Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.

- **Inferential Statistics:** Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

** ** **