STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

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

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

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

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

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

7. 1. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

In statistics Normal Distribution is a type of continuous probability distribution for a real valued random variable. The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters : the mean and the standard deviation.

The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the “bell curve”.

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

Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another strategy - Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values.

There are prevalent methods: Mean imputation, Substitution, Hot deck imputation, Cold deck imputation, Regression imputation, Stochastic regression imputation, Interpolation and extrapolation, Single or Multiple Imputation & lastly While imputation.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above – Hot deck and Stochastic regression are most recommended imputation methods.

12. What is A/B testing?

A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. It is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single Vector-variable are compared. These values are similar except for one variation which might affect a user’s behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product. Large social media sites like Linkedin, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline theirs services.

Today, A/B tests are being used also for conducting complex experiments on subjects such as network effects when users are offline, how online services affect user actions, and how users influence one another. A/B testing is used by data engineers, marketers, designers, software engineers and entrepreneurs among others. Many positions rely on the data from A/B texts, as they allow companies to understand growth, increase revenue and optimize customer satisfaction.

When conducting A/B testing, the user should evaluate the pros and cons of it to see if it aligns best with the results that they’re hoping for.

13. Is mean imputation of missing data acceptable practice?

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 as a bad practice since it ignores feature correlation. Consider the following scenario: 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?

Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.

A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

15. What are the various branches of statistics?

Statistics is a study of presentation, analysis, collection, interpretation and organization of data. There are two main branches of statistics.

Inferential Statistics – It is used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics – It is used to get brief summary of data. You can have the summary of the data in numerical and graphical form.