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

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

a. True

b. False

Answer: 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

Answer: b) Hypothesis

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

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: A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.

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

Answer: 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 common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10. Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value. This has all of the benefits of regression imputation plus the random component's benefits. The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Single or Multiple Imputation

- **Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.**
- **The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.**
- **It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.**
- **When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.**
- **Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.**
- **The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.**

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. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

Answer: Yes, 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 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.

15. What are the various branches of statistics?

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