

STATISTICS WORKSHEET-1

1. TRUE
2. B
3. B
4. D
5. C
6. False
7. B
8. A
9. C

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

Answer: The Normal Distribution is defined by the probability density function for a continuous random variable in a system.

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

Answer: One way of handling missing values is the deletion of the rows or columns having null values. If any columns have more than half of the values as null then you can drop the entire column. In the same way, rows can also be dropped if having one or more columns values as null..

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.

And how would you choose that estimate? 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.

13. What is A/B testing?

Answer: A/B testing is the process of comparing two variations of a page element, usually by testing users' response to variant A vs. variant B and concluding which of the two variants is more effective.

14. 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 a terrible 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.

14. What is linear regression in statistics?

Answer: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they—indicated by the magnitude and sign of the beta estimates—impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula $y = c + b \cdot x$, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15. What are the various branches of statistics?

Answer: Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics

- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics are used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.

