STATISTICS WORKSHEET-1 Answer Sheet

Q : 1 : Ans : a) True

Q : 2 : Ans : a) Central Limit Theorem

Q : 3 : Ans : b) Modeling bounded count data

Q : 4 : Ans : d) All of the mentioned

Q : 5 : Ans : c) Poisson

Q : 6 : Ans : b) False

Q : 7 : Ans : b) Hypothesis

Q : 8 : Ans : a) 0

Q : 9 : Ans : c) Outliers cannot conform to the regression relationship

Q : 10 : Ans : Nomal Distribution :

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**. In graphical form, the normal distribution appears as a "bell curve".

Q : 11 : Ans :

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 list wise sequence most of the time. Depending on why and how much data is gone, list wise 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.

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.
- When list wise 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 list wise 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.

Furthermore, standard errors are underestimated by all single imputation approaches. Because the imputed observations are estimates, their values have a random error associated with them. However, your programme is unaware of this when you enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values.

And, while imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above–hot deck and stochastic regression–work as the imputation method.

The multiple estimates varied significantly because these two approaches contain a random component. This reintroduces some variance that your program can account for in order to provide reliable standard error estimates for your model.

About 20 years ago, multiple imputation was a big advance in statistics. It eliminates many (but not all) difficulties with missing data and, when done correctly, leads to unbiased parameter estimations and accurate standard errors.

Q : 12 : Ans :

A/B testing (also known as **bucket testing** or **split-run testing**) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

Q : 13 : Ans :

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. 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.

Q : 14 : Ans :

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.

Q : 15 : Ans :

There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics.**

STATISTICS WORKSHEET-1 Answer Sheet

Data Collection : The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection

Descriptive Statistics : Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Inferential Statistics : Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).