Statistics Worksheet 1 Solution

A1. (a) True

A2. (a) The Central Limit Theorem

A3. (b) Modeling bounded count data

A4. (d) All of the mentioned

A5. (c) Poisson distribution

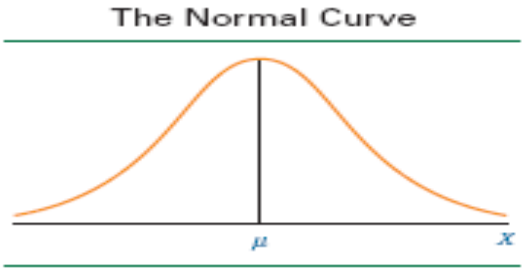
A6. (b) False

A7. (b) Hypothesis

A8. (a) 0

A9. (c) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?
Ans.:



Normal distribution is also known as the Gaussian distribution. It 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 graph form, normal distribution will appear as a bell curve.

It is symmetrical in Nature. Each half of the distribution is the mirror image of the other half. The Normal Distribution exhibits the following characteristics:

a) It is a continuous distribution.
b) It is a symmetrical distribution about its mean.
c) It is asymptotic to the horizontal axis.
d) It is unimodal.
e) Area under the curve is 1.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans.: The concept of missing data is implied as the data which is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results.

When dealing with missing data, there are two primary methods that can be used to solve the error:

a) Imputation: The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an ineffective model.

b) The removal of data: It is used when data is missing at random then related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis.

Recommendation of Imputation techniques: there are multiple solutions to impute the value of missing data depending on the reason 'why the data are missing'? hence recommending only one technique to be used in every case will increase the chances of getting a ineffective model.
The examples of single imputation methods for replacing missing data are as follows

Mean, Median and Mode
This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, the mean or median of the existing observations is calculated and replaced with the missing observation. However, when there are many missing variables, mean or median can result in a loss of variation in the data. This method does not use time-series characteristics or dependency on the relationship between the variables.

Time-Series Specific Methods
Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)
These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Linear Interpolation
Linear interpolation is used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.
When dealing with missing data, this method should be used in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

Seasonal Adjustment with Linear Interpolation
When the data shows trend and seasonality characteristics both, seasonal adjustment with linear interpolation is used. First, the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another is performed and then complete data smoothing with linear interpolation is done.

K Nearest Neighbors
In this method, a distance measure for k neighbors is used and the average is used to impute an estimate. The number of nearest neighbors and the distance metric is selected. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

Linear Regression
Several predictors of the variable with missing values are identified using a correlation matrix. The best predictors are selected and used as independent variables in a regression equation. The variable with missing data is used as the dependent variable. Cases with complete data for the predictor variables are used to generate the regression equation, the equation is then used to predict missing values for incomplete cases. In an iterative process, values for the missing variable are inserted and then all cases are used to predict the dependent variable. These steps are repeated until there is little difference between the predicted values from one step to the next that is they converge.

Q12. What is A/B testing?

Ans.: A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

In this method the products are divided into two parts – A and B. Here A will remain unchanged while significant changes are being made in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, it is decided which part is performing better.
It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the products, while the sample refers to the number of products chosen to participate in the test.

Q13. Is mean imputation of missing data acceptable practice?
Ans.: Mean Imputation of missing data is a bad practice in general because:

a) It does not take correlation between the variables into account.
b) Mean reduces the variance of the data in turn a small variance leads to the narrow confidence interval in the probability distribution.

Q14. What is linear regression in statistics?
Ans.: Linear regression is a basic and commonly used predictive analysis technique. It is used to examine two things:

a) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

b) 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:

$$y = c + b*x$$

Where

y = estimated dependent variable score

c = constant

b = regression coefficient

x = score on the independent variable.

The regression's dependent variables are called by many names as:

a) An outcome variable
b) Criterion variable
c) Endogenous variable
d) Regressand

The independent variables can be called as:

a) Exogenous variables
b) Predictor variables
c) Regressors.

Three major uses for regression analysis are:

a) Determining the strength of predictors
b) Forecasting an effect
c) Trend forecasting.

Types of Linear Regression

a) Simple linear regression: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
b) Multiple linear regression: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
c) Logistic regression: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
d) Ordinal regression: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
e) Multinomial regression: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
f) Discriminant analysis: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Q15. What are the various branches of statistics?

Ans.: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are used in scientific analysis of data and are equally important.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is the first part of a statistical analysis.

Inferential Statistics

Inferential statistics, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample are done with the help of inferential statistics.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other.