**Worksheet\_Statistics\_1**

Solution:1 A

Solution:2 A

Solution:3 B

Solution:4 D

Solution:5 C

Solution:6 B

Solution:7 B

Solution:8 A

Solution:9 B

**Solution:10** 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.

**Solution:11** I have three options when dealing with missing data. The most obvious and by far the easiest option, is to simply ignore any observations that have missing values.

Another approach is to impute the missing values. This involves using statistical or machine learning models to make educated guesses based about the values of the missing data.

A third approach is to use analysis methods that are specifically designed to deal with missing values, such as latent class analysis.

I would recommend these techniques to impute the missing data:

1. **Generalized Imputation:**

 Here we take the average of the entire feature and impute that value for the missing values.

2. **Simple Case Imputation:**

Here the mean is calculated by keeping in the specific groups. For example, here the specific species is taken into consideration and it’s grouped and the mean is calculated. That mean is imputed to its respective group’s missing value.

3. **K-NEAREST NEIGHBOR IMPUTATION:**

It can be done in two ways; one is by using the K-nearest Neighbour Algorithm and the second one is by using the K-means Clustering Algorithm. They are quite similar in functionality. In K-Nearest Neighbours, the missing values are found by calculating the K number of observations that are most similar to the instance in question.

**Solution:12** A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB 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.

**Solution:13** Yes, mean imputation of missing data is acceptable practice. So if the data are missing completely at random, the estimate of the mean remains unbiased. That’s a good thing.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That’s good too.

**Solution:14** 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\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

**Solution:15** There are two branches of Statistics

1. Descriptive Statistics
2. Inferential Statistics
3. **Descriptive Statistics**

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

1. **Inferential Statistics**

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.