Statistics Worksheet-1

Q1) True

Q.2) c

Q.3) b

Q.4) d

Q.5) c

Q.6) b

Q.7) b

Q.8) a

Q.9) c

Q.10) Normal Distribution: It is also known as 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. It appears as a bell curve in graphical form. In this the mean is 0 and standard deviation is 1. It has 0 skew and a kurtosis of 3.

Q.11) How do you handle missing data? What imputation techniques do you recommend?
Ans: Most common ways of handling missing data are:

Zero Replacement: replace the missing value with zero irrespective of everything.

Min/Max Replacement: replace the missing value with minimum or maximum value of a feature.

Mean/Median/Mode Replacement: replace missing value with mean or median or most frequent feature value.

Imputation Techniques are:

1. Mean imputation: calculate the mean of the observed values for the variable for all the individuals who are non-missing.
2. Substitution: impute the value from new individual who has not selected to be in the sample.
3. Hot deck imputation: find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.
4. Cold deck imputation: A systematically chosen value from an individual who has similar values on other variables.
5. Regression imputation: the predicted value obtained by regressing the missing variable on other variables.
6. Stochastic regression imputation: the predicated value from the regression plus a random residual value.
7. Interpolation and extrapolation: an estimated value from other observations from the same individual.

Q.12) A/B testing: It 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. It is a hypothetical testing methodology for decision making that estimate population parameters based on sample statistics.

Q.13) Drawbacks:

1. Mean imputation does not preserve the relationships among Variables.
2. It leads to an underestimate of standard errors.

Perhaps, it is really risky.

Q.14) Linear regression in statistics: It is used to predict the value of a variable based on the value of another variable. The variable which we want to predict is called the dependent variable and the variable we are using to predict the other variable's value is called the independent variable. These models are relatively simple and provide an easy to interpret mathematical formula that can generate predictions.

Q.15) Various branches of statistics: There are two main branches of statistics.

1. Descriptive Statistics:- it is part of statistics that deals with presenting the date we have. This can take two basic forms- presenting aspects of the data either visually or numerically. It focuses on collecting, summarizing, and presenting a set of data.

2. Inferential statistics:- it is the aspect that deals with making conclusions about the data.