STATISTICS WORKSHEET -1(ANSWERS)

Q1) option [a]

Q2) option [a]

Q3) option [b]

Q4 option [d]

Q5) option [c]

Q6) option [b]

Q7) option [b]

Q8) option [a]

Q9) option [c]

Q10) What do you understand by the term Normal Distribution?

Ans: The normal distribution is the most widely known and used of all distributions, Because the normal distribution approximates many natural phenomena so well, it has helped into a standard of reference for many probability problems. 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 graph form, normal distribution will appear as a "bell curve".

In a Normal distribution the mean is zero and the standard deviation is 1.It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution. The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution itself is definitely not normal.

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

Ans: There are a lot of techniques to treat missing value. According to me think the best way to organize some of the most commonly used methods, if I use SAS to implement it –

- Ignore the records with missing value

- Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- Substitute a value such as mean
  When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.
- Predict missing values.

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

- Logistic Regression
- Discriminant Regression
- Markov Chain Monte Carlo (MCMC)

- Predict missing values-Multiple Imputations
  Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required **statistical assumptions** for multiple imputation:

1. Whether the data is missing at random (MAR).
2. Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).

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.

For instance, let's say we own a company and want to increase the sales of our product. Here, either we can use random experiments, or we can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, we may divide the products into two parts – A and B. Here A will remain unchanged while we make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, we try to decide which is performing better.

A/B Testing is a widely used concept in most industries nowadays, and data scientists are at the forefront of implementing it. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

Q13) Is mean imputation of missing data acceptable practice?

Ans: Yes, mean imputation of missing data is acceptable practice because imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing, also by imputing the mean, we are able to keep our sample size up to the full sample size.

Q14) What is Linear regression in statistics?

Ans: In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) 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.

Naming the Variables: There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are: determining the strength of predictors, forecastingan effect, and trend forecasting.

Q15) What are the various branches of statistics?

Ans: Statistics plays a main role in the field of research. It helps us in the collection, analysis and presentation of data. In this blog post we will try to learn about the two main branches of statistics that is descriptive and inferential statistics. **Statistics** is concerned with developing and studying different methods for collecting, analyzing and presenting the empirical data.

The field of statistics is composed of two broad categories- Descriptive and inferential statistics. Both of them give us different insights about the data. One alone doesn't not help us much to understand the complete picture of our data but using both of them together gives us a powerful tool for description and prediction.

Descriptive Statistics:It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.Data can be summarized and represented in an accurate way using charts, tables and graphs.

For example: We have marks of 1000 students and we may be interested in the overall performance of those students and the distribution as well as the spread of marks. Descriptive statistics provides us the tools to define our data in a most understandable and appropriate way.

Inferential Statistics: It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.

For example: Suppose we are interested in the exam marks of all the students in India. But it is not feasible to measure the exam marks of all the students in India. So now we will measure the marks of a smaller sample of students, for example 1000 students. This sample will now represent the large population of Indian students. We would consider this sample for our statistical study for studying the population from which it's deduced.