STATISTICS WORKSHEET-1

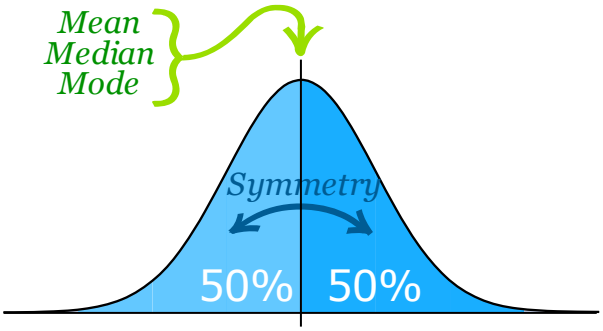
Question Numbers.	Answers.
01	A
02	A
03	B
04	D
05	C
06	A
07	B
08	A
09	C

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

Ans.

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.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, median & mode are all the same as shown in figure.



In data science it is ideal to fit a normally disturbed data for model building. In above figure **'μ'** is the mean of the data and x-axis indicates the density of data. The red, blue and green areas are the standard deviations indicated by **'σ'** for the mean.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

Some important points about Normal Distribution:

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

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

In a data set it is likely to have a missing data or NaNs. Generally, this data is filled by calculating and filling with mean of the data if data is continuous and filling with mode if the data is discrete. There are various data imputation techniques used depending upon the type of data and missing values:

- Univariate imputation uses statistics (mean, median) of the same feature or same column to find the missing data.
- Multivariate imputation uses entire data feature set available to fill the missing data.
- K-Nearest neighbour imputation - This imputer utilizes the k-Nearest Neighbours method to replace the missing values in the datasets with the mean value from the parameter 'n-neighbours' nearest neighbours found in the training set. By default, it uses a Euclidean distance metric to impute the missing values.

12. 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 you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which 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 customers buying your product, while the **sample** refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?
Ans.

Mean manipulation is not considered as an ideal practice when it comes to fill multiple missing values. Mean manipulation ignores the variance of the feature. It shrinks the standard errors which invalidates most hypothesis tests and the calculation of confidence interval. Also, mean variation does not preserve the relationships between the variables such as corelations.

14. What is linear regression in statistics?
Ans.

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?

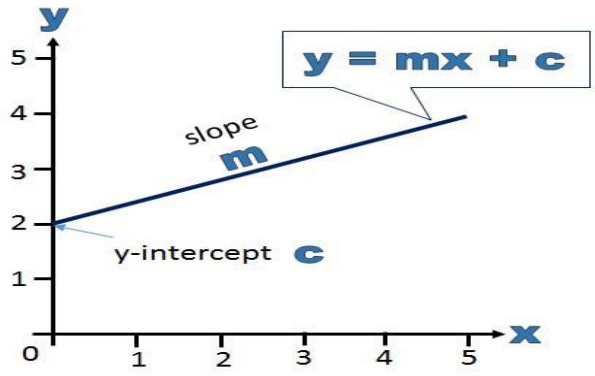
These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

$$y = mx + c$$

here,

y = estimated dependent variable score,
m = regression coefficient (slop of the line),
x = score on the independent variable (features),
c = Interception constant



Example: Predicting stock prices of the company based on company annual performance, customer reputation, future projects of company and etc.

15. What are the various branches of statistics?
Ans.

Statistics is classified into two branches:

1. Descriptive statistics.
2. Inferential Statistics.

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set.

Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into:

i. Measure of central tendency: Measures of central tendency specifically help the statisticians to estimate the centre of values distribution. These measures of tendency are – Mean, mode and median
ii. Measure of Variability: The measure of variability helps statisticians to analyse the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

To easily understand the analysed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population.

Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyse data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.

The different types of calculation of inferential statistics include –

i. Regression analysis
ii. Analysis of variance (ANOVA)
iii. Analysis of covariance (ANCOVA)
iv. Statistical significance (t-test)
v. Correlation analysis.