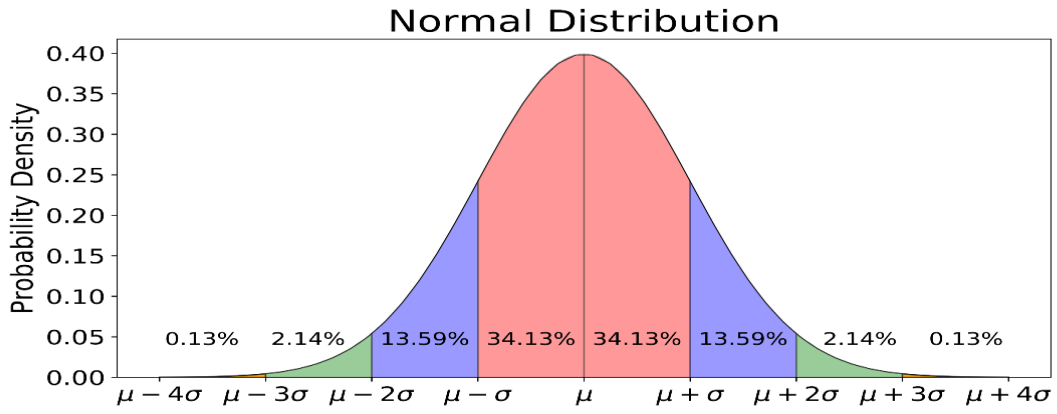
Statistics Worksheet 1

Question NO	Answers
01	a
02	a
03	c
04	d
05	c
06	b
07	b
08	a
09	c

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

Ans - Normal distribution is most popular is type of continuous probability distribution. Normal distribution is symmetric about its mean. i.e half of the data is at less than the mean and other half data is greater than the mean. That's why graphical representation of normal distribution looks like a Bell-shaped curve as shown in fig below. In normal distribution reading near the mean has more frequency of occurrence than that of reading far from mean.



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.
Examples of normal distribution are – average height, marks scored by a class of students etc.

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 statistical way of comparing two or more versions to determine which version performs better and wheatear difference between two versions is significant or not. No matter how good we design a website or a page when it comes to attracting customers one constantly needs to optimise. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. At first a null hypothesis is made and sample data is chosen. Then sample data is tested with conducting A/B testing and we compare if we could reject the null hypothesis or not. 2 Sample t test is most commonly used hypothesis test by data scientists it applied to compare the average difference between two groups.

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 attempts to model the relationship between two or more continuous variables by fitting a linear equation to observed data. Variables is considered to be an independent variable (features), and the other is considered to be a dependent variable(columns). Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting. It shows the extent of the impact of multiple independent variables on dependent variable.

The equation for linear regression is Y=a+b*X+e where, a = intercept, b = slope of line and e is the error term.
The above equation is used to find or draw a best fit line with minimum variance.

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 -

1 - Measures of central tendency
2 - Measures of variability

To easily understand the analysed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.
1 – 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
2 – Measure of Variability - The measure of variability help 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.

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 –

1. Regression analysis
2. Analysis of variance (ANOVA)
3. Analysis of covariance (ANCOVA)
4. Statistical significance (t-test)
5. Correlation analysis

The equation of the Linear Regression is: $Y=a+b*X + e$

It shows the extent of the impact of multiple independent variables on the dependent variable.

Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting