1. **What is the purpose of descriptive statistics?**

The aim of descriptive statistics is to describe and summarize the main features of a dataset without making inferences about a larger population. It mainly summarizes the characteristics and distribution of a dataset such as mean, median, mode, variance, standard deviation etc. It can be presented using tables and data visualization methods like box plots, histogram.

1. **Can you explain the difference between mean, median, and mode?**

**Mean** - The arithmetic average of a set of values. It is calculated by summing up all the values and dividing by the number of observations.

Xmean = ∑X / N

N - Number of observations, X - Individual Value

**Median** - The middle value of a dataset when it is sorted in ascending or descending order. If there is an even number of observations, the median is the average of the two middle values.

**Mode** - The value that occurs most frequently in a dataset. Identify the value with the highest frequency. These measures help summarize the central location of the data and provide insights into the typical or representative value in a dataset.

1. **How do you interpret the standard deviation of a dataset?**

**Standard Deviation** - The square root of the variance. It provides a measure of the average distance of each data point from the mean.

√ ∑ (X – Mean) 2 / n

X - Individual Numbers, mean - avg of individual numbers, n - size

1. **Describe the concept of skewness in statistics.**

There are two types of skewness in distribution of population,

**Left Skewness (Negatively Skewed)** – Left skewed or negatively skewed or left tailed distribution is a distribution where the left tail is longer than the right tail. The majority of the data points are on the right side. The mean is less than median.

**Right Skeweness (Positively Skewed)** – Right skewed or positively skewed or right tailed distribution where the right tail is longer than the left tail. The majority of the data points are on the left side. The mean is greater than median.

1. **Inferential Statistics: What is the main goal of inferential statistics?**

Inferential statistics involves making inferences or predictions about a population based on a sample of data. It allows formulating and testing hypothesis for a sample and generalizing the results to a broader population parameters.

1. **Explain the difference between a population and a sample.**

Population means the whole data or the individuals while sample is the subset of the population which is a sample data or sample individuals from the whole data.

1. **What is a confidence interval, and how is it useful in inferential statistics?**

The confidence level is the probability that the true parameter lies within the confidence interval. Commonly used confidence levels include 90%, 95%, and 99%. A 95% confidence level implies that if we were to calculate the confidence interval from many random samples, we would expect about 95% of those intervals to contain the true parameter.

1. **Define p-value**

The p-value is the probability of obtaining observed data or more extreme results when the null hypothesis is true.

Null Hypothesis – denoted by H0, is a tentative assumption about a population parameter.

Alternative Hypothesis – denoted by Ha, is the opposite of what is stated in the null hypothesis.

1. **Inferential Statistics Techniques**
2. Population Inference – We check the performance of a model. Model performance metrics by taking the subset of the data. From that, calculating the metrics and check the prediction. So basically from the subset of the data gets the model performance of the entire dataset.
3. Hypothesis Testing – Mere assumptions or suppositions which are to be proved or disproved. Used to determine whether a statement abou t the value of a population parameter should or should not be rejected. It draw conclusion based on the population of data from sample data.

Null Hypothesis – denoted by H0, is a tentative assumption about a population parameter.

Alternative Hypothesis – denoted by Ha, is the opposite of what is stated in the null hypothesis

1. Cross Validation – Partitions data into testing and training, allows inferring the model’s performance based on unseen data. It solves the issues like overfitting and underfitting by evaluating model on multiple divisions of the data. Like example K Fold cross validation.