1.What is the purpose of descriptive statistics?

Descriptive statistics provide a summary view of the main characteristics of a dataset, such as central tendency (mean, median, mode), variability (range, standard deviation), and distribution (skewness, kurtosis). They enable comparisons between different groups or sets of data, aiding in decision-making processes or hypothesis testing.

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

To calculate the mean, you add up all the values in the dataset and then divide by the number of values. For example, if you have the numbers 2, 4, 6, and 20, the mean is (2 + 4 + 6 + 20) / 4 = 32 / 4 = 8.

The median is the middle value when the data is arranged in ascending or descending order. If there is an odd number of observations, the median is the middle value. If there's an even number, it's the average of the two middle values.

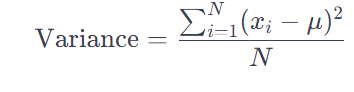
The mode is the value that appears most frequently in the data.

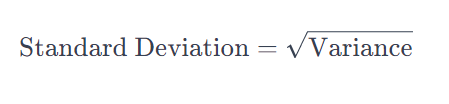
3.How do you interpret the standard deviation of a dataset?

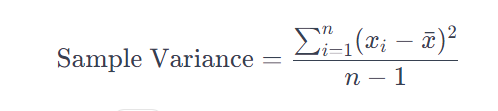
It indicates how much individual values deviate or differ from the mean (average) of the dataset.

Find the mean (*μ*) of the dataset.

Calculate the squared differences between each data point and the mean. Sum up these squared differences and divide by the total number of data points (*N* for a population).







4.Describe the concept of skewness in statistics.

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

Positive Skewness (Right Skewed) When the tail on the right-hand side of the distribution is longer than the left-hand side. The mean is typically greater than the median, and the bulk of the data is concentrated on the left side.

Negative Skewness (Left Skewed) When the tail on the left-hand side of the distribution is longer or fatter than the right-hand side. The mean is usually less than the median, and the majority of the data is concentrated on the right side.

**Inferential Statistics:**

5.What is the main goal of inferential statistics?

Inferential statistics allows us to make predictions or estimate characteristics of a population based on sample data. For example, using a sample to predict election outcomes for an entire country.

6.Explain the difference between a population and a sample.

The population refers to the entire group that you're interested in studying or analyzing. It includes all the individuals, items etc.

A sample is a subset or a smaller group selected from the larger population. The goal is to use the sample to make inferences or generalizations about the larger population.

7.What is a confidence interval, and how is it useful in inferential statistics? Define p-value

It offers a more informative view than a point estimate (single value) by showing the range of values within which the true parameter is likely to fall. This aids in making informed decisions and drawing conclusions about the population

p-value

p-value in statistics represents the probability of observing a result (or one more extreme) assuming that the null hypothesis is true

The p-value helps in decision-making regarding whether to reject the null hypothesis or not. If the p-value is less than the chosen significance level, it provides evidence to reject the null hypothesis in favor of the alternative hypothesis, suggesting that there is a statistically significant effect or relationship

8.What are the techniques if inferential statistics?

1 Hypothesis Testing: This technique involves making decisions using sample data to determine whether there's enough evidence to reject or fail to reject a null hypothesis about a population parameter. Common tests include t-tests, chi-square tests, ANOVA, z-tests, etc

Cross validation

Partitioning the Data: The dataset is divided into k equal-sized subsets or "folds."

Training and Testing: The model is trained on k-1 folds (training set) and evaluated on the remaining fold (test set). This process is repeated k times, with each fold serving as the test set exactly once and the remaining k-1 folds as the training set.

Performance Evaluation: The performance metric (e.g., accuracy, mean squared error, etc.) is calculated for each iteration of training and testing. The final evaluation is often the average performance across all k folds

Population inference

Population inference refers to the process of drawing conclusions or making inferences about a population based on information collected from a sample taken from that population.

Sampling: Selecting a representative sample from the population of interest is crucial. The sample should ideally reflect the diversity and characteristics of the entire population.

Data Collection and Analysis: Data is collected from the sample and analyzed using various statistical techniques. Descriptive statistics summarize the sample data, while inferential statistics help draw conclusions or make predictions about the population parameters (such as means, proportions, correlations, etc.).

Generalization: The findings from the sample are generalized to make statements or predictions about the entire population. This involves using the sample data to estimate population parameters or to test hypotheses about the population.

Uncertainty and Confidence: In population inference, there's always some level of uncertainty due to sampling variability. Confidence intervals and hypothesis testing are used to quantify this uncertainty and provide a measure of how confident we can be in our conclusions about the population.