WEEKLY ASSIGNMENT – 3

Descriptive Statistics:

1.What is the purpose of descriptive statistics?

Descriptive Statistics describes the characteristics of a data set. It is a simple technique to describe, show and summarize data in a meaningful way. The primary purpose of descriptive statistics is to characterize and summarize the key features of a dataset. This involves presenting the data in a meaningful and easily understandable manner.

Descriptive statistics serve two main functions:

1. **Providing Basic Information:** Descriptive statistics offer fundamental information about the variables within a dataset. This includes measures such as central tendency (mean, median, mode) and measures of variability (range, standard deviation), which give an overview of the typical values and the extent to which values deviate from the average.
2. **Highlighting Relationships:** Descriptive statistics can reveal potential relationships between variables. Correlation coefficients and scatterplots are examples of tools used to assess the degree and direction of relationships between different variables. This helps in identifying patterns and tendencies within the data.

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

The 3 most common measures of central tendency are the mean, median and mode.

- **Mean** is the measurement of central tendency that represents the average value of the dataset. Mean is calculated by adding all the values in the dataset and dividing by the total number of values.
- **Median** is the measure of central tendency that represents the middle value of the dataset when the data are arranged in order (either ascending or descending).
- **Mode** is a measure of central tendency that is used to represent the most frequently occurring value in a dataset.

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

The **Standard deviation** (SD) is a single number that summarizes the variability in a dataset. It represents the typical distance between each data point and the mean.

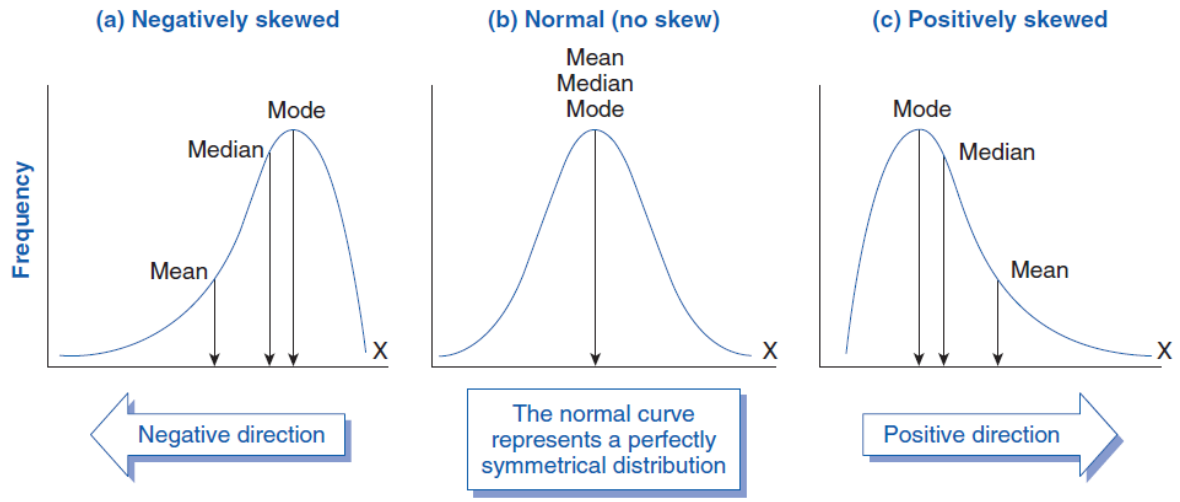
- A smaller SD implies that data points tend to be closer to the mean, indicating a more consistent and tightly clustered dataset.
- Conversely, a higher SD suggests a greater divergence of values from the mean, signifying increased variability and a propensity for more distant or extreme data points.

In essence, the standard deviation provides valuable insights into the degree of uniformity or scatter in the dataset.

4.Describe the concept of skewness in statistics.

Skewness in statistics quantifies the degree of asymmetry or departure from a symmetrical distribution within a dataset. It gauges how the distribution of variable measurements deviates from a perfect symmetry around its mean. Visualizing skewness on a bell curve reveals a lack of balance in

the distribution, with data points clustering more on one side of the median than the other. As skewness increases, indicating departure from a normal distribution, reliance on the mean as a central tendency measure diminishes in accuracy.



Inferential Statistics:

5.What is the main goal of inferential statistics?

Inferential statistics is a branch of statistics that makes the use of various analytical tools to extrapolate information from a smaller sample to make predictions and draw conclusions about a larger population. The goal of inferential statistics is to use the sample data to extract insights and conclusions about the population that the sample represents. To do this, inferential statistics uses various techniques, such as hypothesis testing, regression analysis, analysis of variance (ANOVA), etc.

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

A **population** is the entire group that you want to draw conclusions about. A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population. In research, a population doesn't always refer to people.

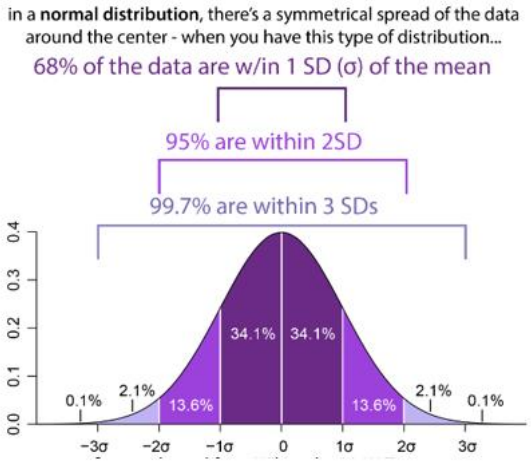
7.What is a confidence interval, and how is it useful in inferential statistics?

The **confidence interval** (CI) is a statistical concept that provides a range of values within which we can be reasonably certain that the true value of a population parameter lies.

By convention, the confidence level is usually set at 95%. If you have a 95% confidence interval for the mean of a variable, it means that you are 95% confident that the true population mean falls within the

calculated interval. The width of the confidence interval is influenced by factors such as the sample size and the variability of the data.

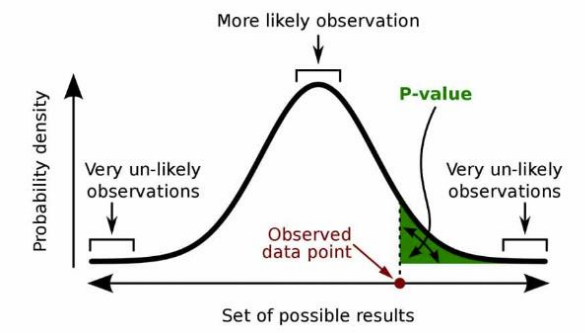
In the context of hypothesis testing and parameter estimation, a 95% confidence level is commonly used, but other confidence levels such as 90%, 99%, or even higher can be chosen depending on the specific requirements and considerations of the analysis.

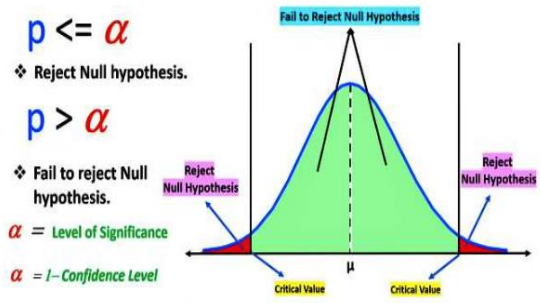


8.Define p-value

A **p-value** is a statistical measurement used to validate a hypothesis against observed data. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.

P-values are calculated from the deviation between the observed value and a chosen reference value(mean), given the probability distribution of the statistic, with a greater difference between the two values corresponding to a lower p-value.

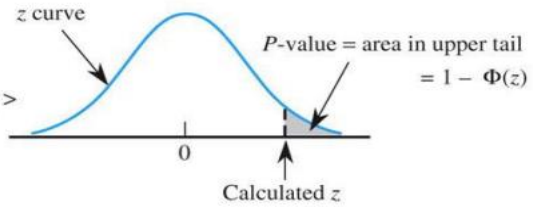




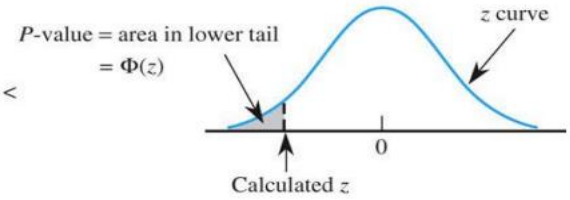
The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant.

The calculation for a p-value varies based on the type of test performed. The three test types describe the location on the probability distribution curve: **lower-tailed test, upper-tailed test, or two-tailed test.**

1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠