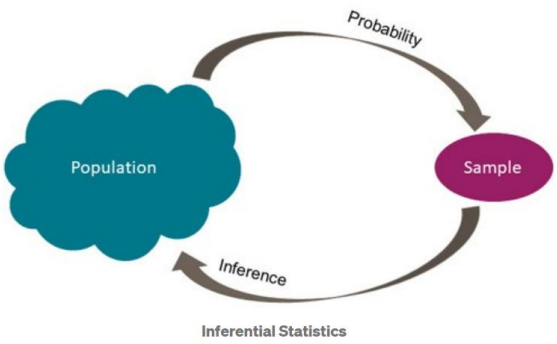
ASSIGNMENT 2

1) What is the difference between inferential statistics and descriptive statistics?

DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Descriptive Statistics result from gathering data from a body, group, or population and reaching conclusions only about that group.	Inferential Statistics are generated from the process of gathering sample data from a group, body, or population and reaching conclusions about the larger group from which the sample was drawn.
Organize and present data in a purely factual way.	Help us to make estimates and predict future outcomes.
Present final results visually, using tables, charts, or graphs.	Present final results in the form of probabilities.
Draw conclusions based on known data and limited to a sample or population having small size.	Draw conclusions that go beyond the available data and attempts to reach the conclusion about the population.
Use measures like central tendency, distribution, and variance.	Use techniques like hypothesis testing, confidence intervals, and regression and correlation analysis.
It is used to describe a situation.	It is used to explain the chance of occurrence of an event.

2) What is the difference between population and sample in inferential statistics?

Inferential Statistics make decisions or predictions about a population based on the sample data.



As the name defines, population means large set of data. It refers to the collections of all elements with similar characteristics. For example, Population of a country includes all people currently within that country. It's a finite but potentially large list of members. Population includes each and every unit of the group and focus on identifying the characteristics.

Sample is the subset of population. It draws one or more observations from the population. Sampling is a process of selecting the sample from the population.

3) Most Common characteristics used in descriptive statistics?

- The Measure of Central Tendency - The measures of central tendency are used to show the center of the data set. The central tendency is estimated using the mean, median and mode.
    - Mean is the average of all the data.
    - Median is the middle of the entire data set.
    - Mode indicates the most commonly occurring value in the data set.
- The Measure of Spread - The objective of measure of dispersion or variation is to identify the extent to which the entire data set is spread from the central tendency – specifically mean. he commonly used estimates are range, standard deviation, and variance.
    - The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.
    - The standard deviation (s) is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.
    - The variance is the average of squared deviations from the mean. Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

4) How to calculate Range and Interquartile Range?

Range

The range of a dataset is the difference between the largest and smallest values in that dataset. For example, in the two datasets below, dataset 1 has a range of 38 – 20 = 18 while dataset 2 has a range of 52-11 = 41.

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

Interquartile Range

The interquartile range is the middle half of the data.

We can find the interquartile range or IQR in four simple steps:

- Order the data from least to greatest
- Find the median
- Calculate the median of both the lower and upper half of the data
- The IQR is the difference between the upper and lower medians

- Let's sort an example data set with an odd number of values into ascending order.

  {Odd data set}: 9, 3, 2, 5, 6, 11, 4, 3, 2

  {Odd data set (ascending)}: 2, 2, 3, 3, 4, 5, 6, 9, 11

- First, we will find the median of a set with an odd number of values. Cross out values until you find the centermost point

  The median of the odd valued data set is four.

- Once we have found the median of the entire set, we can find the medians of the upper and lower portions of the data. If the data set has an odd number of values, we will omit the median or centermost value of the set. Afterwards, we will find the individual medians for the upper and lower portions of the data.

  {Odd data set}: 2, 2, 3, 3, 4, 5, 6, 9, 11

  Omit the centermost value.

  {Odd data set}: 2, 2, 3, 3| 5, 6, 9, 11

  Find the median of the lower portion.

  The median of the lower portion is 2.5

  Find the median of the upper portion.

  The median of the upper portion is 7.5

- Last, we need to calculate the difference of the upper and lower medians by subtracting the lower median from the upper median. This value equals the IQR.

  Let's find the IQR of the odd data set.

  IQR of the odd data set=7.5-2.5

  IQR=5

5) How is the statistical significance of an insight assessed?

Statistical significance is the claim that the results or observations from an experiment are due to an underlying cause, rather than chance. Researchers conduct hypothesis testing to determine statistical significance. The hypothesis is a researcher's theory or belief about something before testing their theory. It is also referred to as the alternative hypothesis.

The alternative hypothesis contrasts with the null hypothesis. The null hypothesis is that the researcher's theory is not true, and there is no underlying cause present during an experiment. If testing shows that the researcher's theory is true, we reject the null hypothesis, and the alternative hypothesis is validated.

Integral to hypothesis testing is the concept of a "p-value." The p-value is the probability that the observations in testing a hypothesis result from random chance instead of an underlying cause. A higher p-value indicates the higher likelihood the observations were due to change. A lower p-value indicates the higher likelihood the observations were due to a cause theorized in the hypothesis.

Practical Example
Consider a company that wants to test the theory that its stock being mentioned on a certain business television show attracts a statistically significant number of new investors. It may construct an experiment where they arrange to get their stock mentioned on the show every other Tuesday for three months. The company can then compare the number of new investors from Tuesdays on the show with the number of new investors from Tuesdays not on the show.

The number of new investors on Tuesdays when their stock is not mentioned would be taken as the baseline average of new investors. If the average number of new investors on Tuesdays when their stock is mentioned is substantially higher than the baseline average, the company can conclude that the result is statistically significant and that it is to the company's advantage to arrange for it to be mentioned.