Assignment No - 03 (Group A)

Problem statement:

Perform the following operations on any open source dataset (e.g., data.csv) 1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variables. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable. 2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris- versicolor' of iris.csv dataset. Provide the codes with outputs and explain everything that you do in this step.

Pre-requisite

1. Basic of Python Programming
2. Concept of statistics such as mean, median, minimum, maximum, standard deviation etc

Objective

Students should be able to perform the Statistical operations using Python on any open source dataset.

Software and Hardware requirements:-

1. **Operating system:** Linux- Ubuntu 16.04 to 17.10, or Windows 7 to 10,
2. **RAM-** 2GB RAM (4GB preferable)
3. IDE :-Anaconda Jupiter Notebook / pycharm /Visual Studio

Theory

What is Statistics?

Statistics is the science of collecting data and analysing them to infer proportions (sample) that are representative of the population.

In other words, statistics is interpreting data in order to make predictions for the population.

Branches of Statistics: There are two branches of Statistics.

1. DESCRIPTIVE STATISTICS:

Descriptive Statistics is a statistics or a measure that describes the data.

2. INFERENTIAL STATISTICS:

Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

1. DESCRIPTIVE STATISTICS:

Descriptive Statistics is summarising the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier.

It does not involve any generalisation or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. Measures of Central Tendency
2. Measures of Dispersion (or Variability)

1. Measures of Central Tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the centre of the data. This one number summary is of three types

1. Mean:

Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread. Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

$$Mean = \frac{17+16+21+18+15+17+21+19+11+23}{10} = \frac{178}{10} = 17.8$$

2. Median:

Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

If the number of observations is odd, the median is given by the middle observation in the sorted form.

If the number of observations are even, median is given by the mean of the two middle observations in the sorted form.

An important point to note is that the order of the data (ascending or descending) does not affect the median. To calculate Median, let's arrange the data in ascending order. 11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17+18}{2} = 17.5$$

3.Mode:

Mode is the number which has the maximum frequency in the entire data set, or in other words, mode is the number that appears the maximum number of times. A data can have one or more than one mode.

If there is only one number that appears the maximum number of times, the data has one mode, and is called Uni-modal.

If there are two numbers that appear the maximum number of times, the data has two modes, and is called Bi-modal.

If there are more than two numbers that appear the maximum number of times, the data has more than two modes, and is called Multi-modal.

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

Mode is given by the number that occurs the maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

2. Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1. Absolute Deviation from Mean —

The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describes the variation in the data set, in the sense that it tells the average absolute distance of each data point in the set.

It is calculated as

$$Mean\ Absolute\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \bar{X}\right|$$

2. Variance —

Variance measures how far are data points spread out from the mean.

A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as

$$Variance = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

3. Standard Deviation —

The square root of Variance is called the Standard Deviation.

It is calculated as

$$Std\ Deviation = \sqrt{Variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

4. Range —

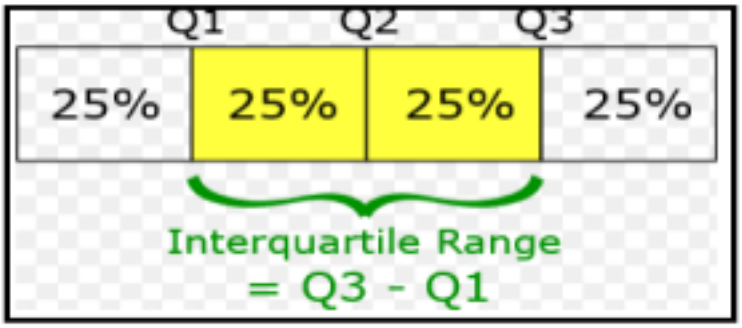
Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

Range = Maximum – Minimum

5. Quartiles

Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

- 25% of the data points lie below Q1 and 75% lie above it.
- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
- 75% of the data points lie below Q3 and 25% lie above it



6. Skewness —

The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

$$Skewness = \frac{3(Mean - Median)}{Std\ Deviation}$$

Positive Skew —

This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.

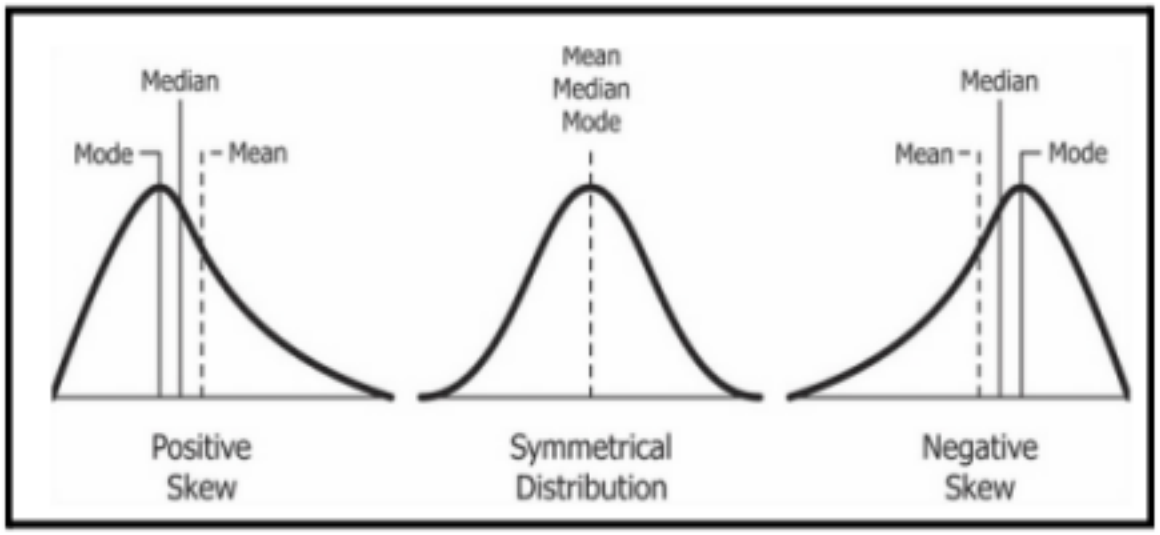
Negative Skew —

This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

The most commonly used method of calculating Skewness is

If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution

is Negatively Skewed and if it is positive, it is Positively Skewe



Types of Variables:

A variable is a characteristic that can be measured and that can assume different values.

Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables.

Variables may be classified into two main categories:

- Categorical and
- Numeric.

Each category is then classified in two subcategories: nominal or ordinal for categorical variables, discrete or continuous for numeric variables.

Categorical variables:-Ordinal Variable:-

An ordinal variable is a variable whose values are defined by an order relation between the different categories.

In the following table, the variable “behaviour” is ordinal because the category “Excellent” is better than the category “Very good,” which is better than the category “Good,” etc.

There is some natural ordering, but it is limited since we do not know by how much “Excellent” behaviour is better than “Very good” behaviour.

Student behaviour ranking	
Behaviour	**Number of students**
Excellent	5
Very good	12
Good	10
Bad	2
Very bad	1

Categorical variables:-Numerical Variables:-

A numeric variable (also called quantitative variable) is a quantifiable characteristic whose values are numbers (except numbers which are codes standing up for categories). Numeric variables may be either continuous or discrete.

Continuous variable:-

A variable is said to be continuous if it can assume an infinite number of real values within a given interval. For instance, consider the height of a student. The height can't take any values.

It can't be negative and it can't be higher than three metres. But between 0 and 3, the number of possible values is theoretically infinite. A student may be 1.6321748755 … metres tall.

Categorical variables:-Discrete variables:-

As opposed to a continuous variable, a discrete variable can assume only a finite number of real values within a given interval.

An example of a discrete variable would be the score given by a judge to a gymnast in competition: the range is 0 to 10 and the score is always given to one decimal (e.g. a score of 8.5)

Conclusion:

Descriptive statistics summarises or describes the characteristics of a data set.

Descriptive statistics consists of two basic categories of measures:

● Measures of central tendency and measures of variability (or spread). Measures of central tendency describe the centre of a data set. It includes the mean, median, and mode.

Measures of variability or spread describe the dispersion of data within the set and it includes standard deviation, variance, minimum and maximum variables.