A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

Group A.

Experiment - 03.

- Title of the Experiment: DESCRIPTIVE Statistics. - Measures of Central Tendency and variability.

- Problem Statement: Perform the following operations on any open source dataset (eg. 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 eg. if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by 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 detail like percentile, mean, standard deviation, etc. of the species of 'Iris-setosa', 'Iris-versicolor' of iris.csv. dataset.

- Objective of the experiment -

Students should be able to perform the statistical operations using python on any open source dataset.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

- **Pre-requisite:**
  - Basic of python programming.
  - Concept of statistics such as mean, median, minimum, maximum, standard deviation, etc.

- **THEORY :**

I] **Summary Statistics :-**

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

There are two branches of statistics -

(1). Descriptive statistics - measure that describes the data.

(2). Inferential statistics - Using random sample of data to make inferences about it.

1] **Descriptive statistics :-**

Summarising data at hand to certain numbers like mean, median, etc.

It does not involve any generalisation of inference beyond what is available.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

* Commonly used measures:-
  1. Measures of central tendency.
  2. Measures of dispersion (or variablity).

(i). Measures of central tendency -
It is one number summary of the data that typically describes the center of data. This one number summary is of three types.

(a) Mean: Mean is defined as the ratio of the sum of all the observations in the data to total number of observations. This is also known as average.

$$\text{Mean} = \frac{\Sigma x}{N}$$

Σx → Sum of all observation
N → Total observations.

(b). Median:- Median is the point which divides the entire data into two equal halves. Median is calculated by arranging data ascending or descending order.

- If no. of observation is odd, the median is given by middle observation.
- If it is even then median is the mean of middle two observations.

Eg- 11, 15, 16, 17, 17, 18, 19, 21, 21, 23.

No. of observation = 10

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

$$= \boxed{17.5}$$

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

(c). **Mode**: Mode is the number which has maximum frequency in the entire data set or in other words, mode is the number that appears maximum no. of times.

- If there is only one number that appears that maximum amount then data has one mode or is unimodal.
- If there is two modes then data is Bimodal.
- If mode is more than two observation then data is multimodal.

(ii). **Measures of Dispersion** (or variability) –

- It describes the spread of the data around the central value (or the measures of central tendency).

1. Absolute Deviation from Mean :-
describes variation in dataset, in the sense that it tells the average absolute distance of each data point in the set. It is calculated as :

$$\text{Mean Absolute deviation} = \text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$$

2. **Variance**: It measures how far are the data points spread out from the mean. A high variance indicates that data points are spread widely and small means dataset is closer to mean.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Standard Deviation :-

The square root of variance is called the standard deviation. It is calculated as -

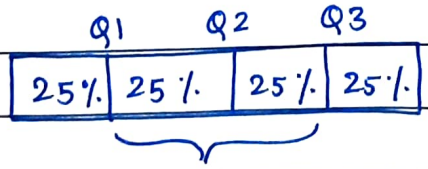
$$S.D = \sqrt{\text{Variance}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Range: It is the difference between the Maximum value and the minimum value in data set.

$$\text{Range} = \text{Maximum} - \text{Minimum}.$$

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

- 25% of data points lie below Q1 and 75% lie above it.
- 50% of data points lie below Q2 and 50% lie above it.
- 75% lie below Q3, and 25% above it.



$= Q3 - Q1$

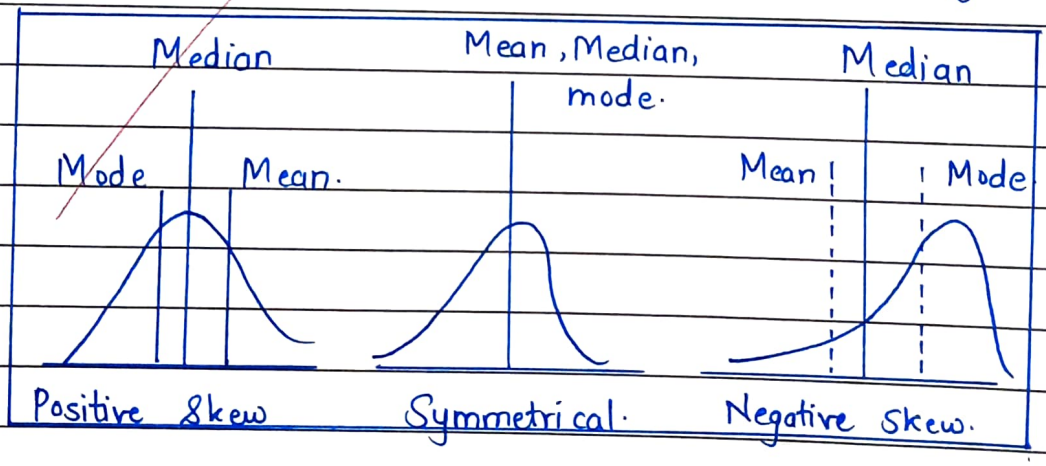
A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

6. Skewness: The measure of assymetry in a probability distribution is defined by skewness. It can be positive, negative or undefined.

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

- Positive Skew: Mean is greater than mode.
- Negative skew: Mean is smaller than mode.

If skewness is zero, then distribution is symmetrical,
If it is negative, the distribution is negatively skewed
if positive, then distribution is positively skewed.



- Python code :-

1. Mean:

To find mean of all columns -

Syntax - df.mean()

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

- To find mean of specific column -

```
df.loc[:, 'Age'].mean()
```

- To find mean row-wise -

```
df.mean(axis=1)[0:4]
```

2. Median : `df.median()`
   - To find median of specific columns -

```
df.loc[:, 'Age'].median()
```

3. Mode : To find mode of all columns.
   - `df.mode()`.

   To find mode of specific columns.
   - `df.loc[:, 'Age'].mode()`.

4. Minimum: `df.min()`

   To find for specific columns.
   - `df.loc[:, 'Age'].min(skipna = False)`.

II. 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 variables.

There are two categories of variables -

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

- Categorical variables - Also known as qualitative variables refers to a characteristics that can't be quantifiable.

* They can be either nominal or ordinal -

(1) Nominal: Nominal variable is one that describes name, label or category without natural order.

(2). Ordinal: Its values are defined by an order relation between different categories.

- Numerical Variables: - Also called as quantitative variable it is a quantifiable characteristic.

* They can be continuous or discrete. -

(1). Continuous variable: A variable is said to be continuous if it can assume an infinite number of real values within a given interval.

For eg: Consider height of student - It cannot be negative or greater than a few meters.

(2). Discrete: As opposed to continuous, it is only a finite number of real values within a given interval.

Eg - Score of student can be in range of 0 to 100 & in decimals too.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

- Algorithm:

1. Import pandas library.
2. The dataset is downloaded from VCI Repository -
3. Assign column names.
4. Load iris.csv into a pandas dataframe.
   iris = pd.read_csv("iris.csv")
5. Load all rows with Iris-setosa species in variable irisSet.
   irisSet = (iris['Species'] == 'Iris-Setosa')
6. To display basic statistical details like percentile, mean, standard deviation, etc. for Iris-setosa use describe
7. Load all rows with Iris-versicolor species in variable irisVer
   irisVer = (iris['Species'] == 'Iris-versicolor')
8. To display basic statistical details like percentile, mean, standard deviation, etc. for irisVersicolor use describe.
   iris[irisVer].describe().
9. Load all in with Iris-virginica species in variable irisVir.
10. Similarly use describe.
11. Then apply all above formulae on each set.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

* # 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.
- measures of variability. (or spread).

Measures of central tendency describe the center of a dataset. It includes the mean, median, mode

Measures of variability or spread describe the dispersion of data within the set and it includes std deviation, variance, min and max.

Thus, we successfully performed these operations on a dataset.

* * * * * * *