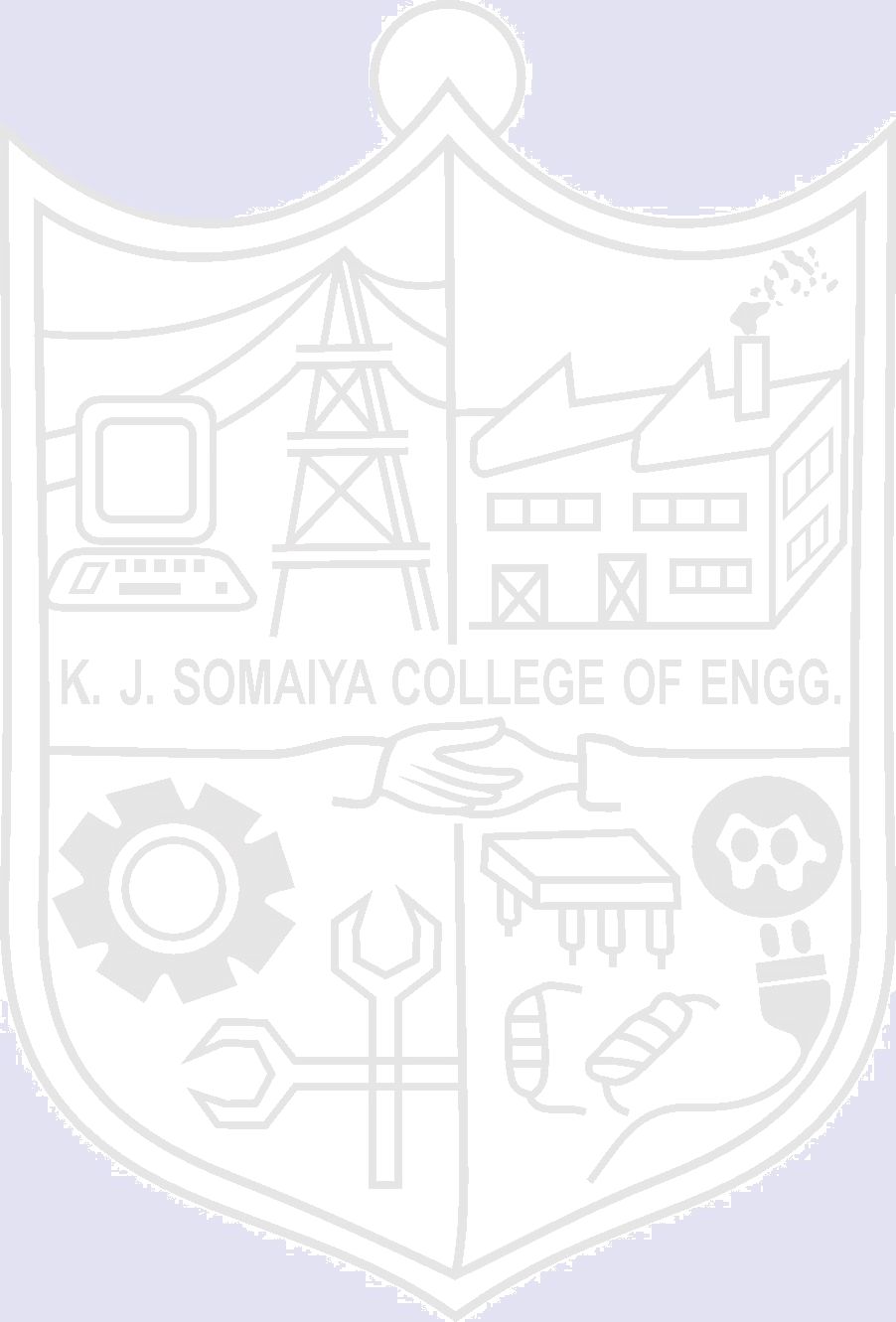


# Batch:A3 Roll No.: 16010421075 Experiment No. 2

**Aim**: To find measures of central tendency and variability of data of data using statistical analysis tool.

**Resources needed:** Any free and Open online statistical analysis tools

# Theory:

Mathematically central tendency means measuring the centre or distribution of location of values of a data set. It gives an idea of the average value of the data in the data set and also an indication of how widely the values are spread in the data set. That in turn helps in evaluating the chances of a new input fitting into the existing data set and hence probability of success.

There are three main measures of central tendency:

* Mean - It is the Average value of the data which is a division of sum of the values with the number of values.
* Median - It is the middle value in distribution when the values are arranged in ascending or descending order.
* Mode - It is the most commonly occurring value in a distribution.

Variability describes how far apart data points lie from each other and from the center of a distribution. Along with measures of [central tendency](https://www.scribbr.com/statistics/central-tendency/), measures of variability give you [descriptive statistics](https://www.scribbr.com/statistics/descriptive-statistics/) that summarize your data. Variability is also referred to as spread, scatter or dispersion. It is most commonly measured with the following:

* + [Range](https://www.scribbr.com/statistics/range/): the difference between the highest and lowest values
  + [Interquartile range](https://www.scribbr.com/statistics/interquartile-range/): the range of the middle half of a distribution
  + [Standard deviation](https://www.scribbr.com/statistics/standard-deviation/): average distance from the mean
  + [Variance](https://www.scribbr.com/statistics/variance/): average of squared distances from the mean
  + Coefficient of variation: ratio of the standard deviation to the mean.

# Activity:

1. Use any programming language to compute mean mode and median of sample data provided by user or applied to any database downloaded from internet? Do not use pre-defined library functions to compute.

import pandas as pd

def get\_median(list1):

    # sort the list

    ls\_sorted = list1.sort()

    # find the median

    if len(list1) % 2 != 0:

        # total number of values are odd

        # subtract 1 since indexing starts at 0

        m = int((len(list1)+1)/2 - 1)

        return list1[m]

    else:

        m1 = int(len(list1)/2 - 1)

        m2 = int(len(list1)/2)

        return (list1[m1]+list1[m2])/2

def mode(list1):

    # dictionary to keep count of each value

    counts = {}

    # iterate through the list

    for item in list1:

        if item in counts:

            counts[item] += 1

        else:

            counts[item] = 1

    # get the keys with the max counts

    return [key for key in counts.keys() if counts[key] == max(counts.values())]

def variance(list1):

    global mean

    variance=0

    for i in list1:

        variance=variance+(i-mean)\*\*2

    return (variance/len(list1))

def standard\_dev(list1):

    global variance

    std=variance(list1)\*\*0.5

    return (std)

df= pd.read\_csv(r'ds\_salaries.csv')

df.head()

salary=df['salary']

s=0

list=df['salary'].tolist()

for i in range(0,21):

    s = salary[i]+s

mean = s/20

print(f"Mean: {mean}")

list1=[]

for i in range(0,20):

    list1.append(list[i])

print(f"Median: {get\_median(list1)}")

print(f"Mode: {mode(list1)}")

print(f"Variance: {variance(list1)}")

print(f"Standard Deviation: {standard\_dev(list1)}")

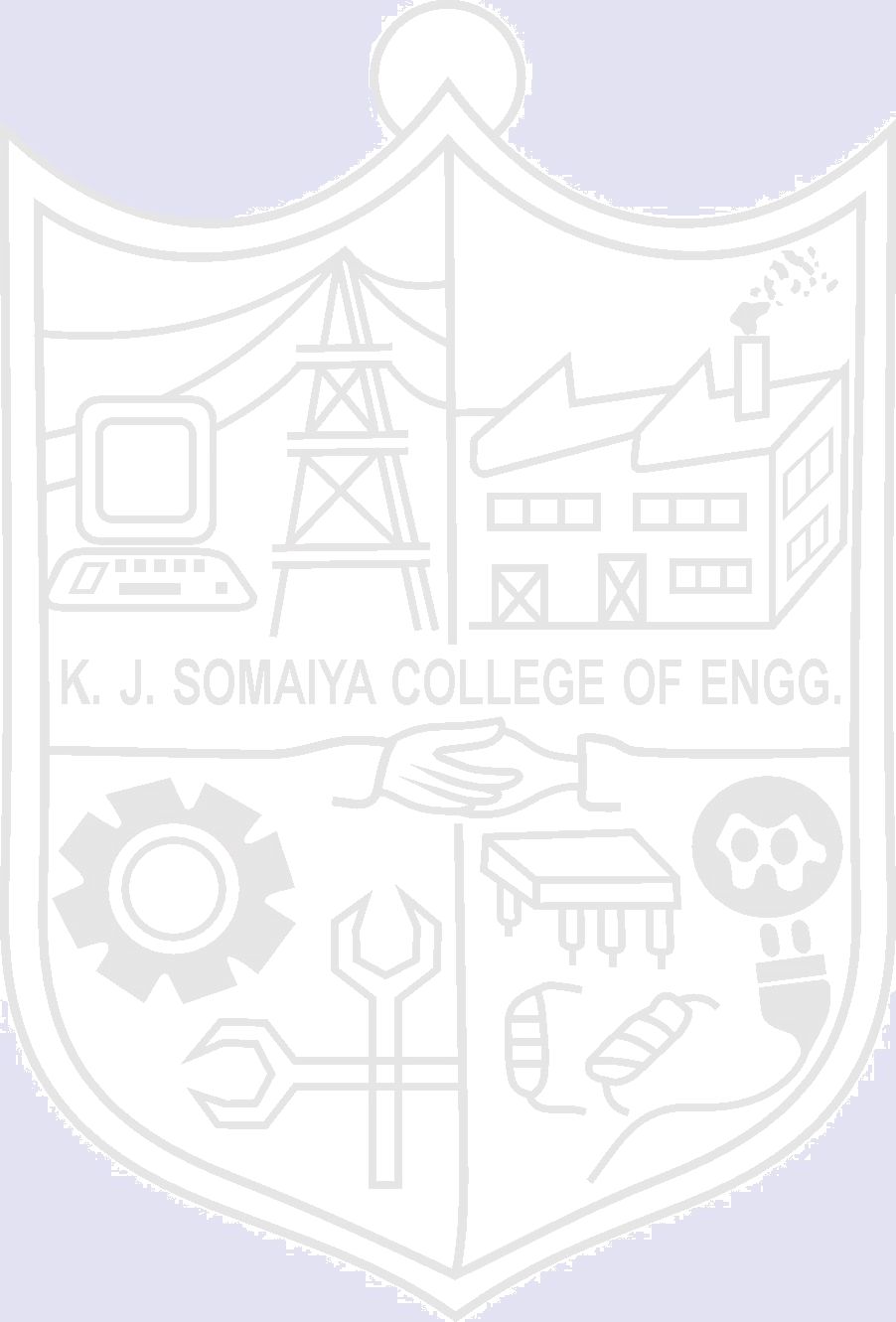
1. Compute mean, standard deviation, mean deviation, coefficient of variation and standard error on sample data of your choice? Use any programing language of your choice.

# Results: Results must be submitted online in a separate doc/docx file for the above activity containing the program code along with Input and Output mentioned below the program. Attempt post lab questions in the same file.

# 

# 

**Questions:**

1. What are the various applications of central tendency and variability of data?

Real life applications include:

• Explicating poverty line

• Business

• Grouping data on climate change

• Buying a property

1. What are the outlier’s data? What are the different ways to find out it? Give suitable example with its effect on central tendency and variability of data?

Outliers are data points that are far from other data points. In other words, they’re unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. Different ways to find out outlier:

• Sorting data

• Graphing data

• Calculating z-score

• Calculating interquartile range

Outliers affect mean but have a very little effect on median and mode. A median would be an appropriate measure of central tendency in such a case. If the outlier lies to the left of the graph, it is causing the mean to decrease, but if the outlier is present on the right side, it increases the mean. This conflict no longer provides a good representation of data. Outliers affect on variance and standard deviation of a data distribution. With extreme outliers, the distribution is skewed in the direction of the outliers making it difficult to analyze the data.

**Outcomes: CO2:** Comprehend descriptive and proximity measures of data

# Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

# References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
2. S.C. Gupta , V. K. Kapoor Fundamentals of mathematical statistics Sultan Chand and Sons 2014