**Experiment No.2**

**Title:** Measuring Central Tendency and variabilityof the Data

**Batch: A4 Roll No.:1914078 Experiment No.:1**

**Aim**: 1.Find out measures of central tendency of data using statistical analysis

2. Find out measures of variability of data using statistical analysis

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:**Any statistical analysis tools and any programming language

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

Mathematically central tendency means measuring the center 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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Procedure / Approach /Algorithm / Activity Diagram:**

1. Write a programming code with a procedure to compute mean mode and median of sample data without using readymade function?
2. Write a programmingcode with a procedure to find out mean, standard deviation, mean deviation, coefficient of variation and standard error on sample data of your choice?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Results: (Program printout with output / Document printout as per the format)**

**Using data of number of deaths per year due to earthquakes**

from math import sqrt

from pprint import pprint

def get\_desc(arr):

    values = list(arr.values())

    values.sort()

    n = len(values)

    total = sum(values) # total

    mean = total/n # mean

    # median

    if n%2 == 1:

        median =  values[n//2+1]

    else:

        median =  (values[n//2] + values[n//2+1])/2

    # mode

    count = {}

    for i in values:

        if i not in count:

            count[i] = 0

        else:

            count[i] += 1

    a = sorted(count.items(), key = lambda x: x[1])

    mode = a[-1][0]

    if a[-1][1] == 0:

        mode = "None"

    # standard deviation

    diff\_square = 0

    for i in values:

        diff\_square += (mean - i)\*\*2

    diff\_square /= n

    std = sqrt(diff\_square)

    # mean deviation

    diff = 0

    for i in values:

        diff += abs(mean-i)

    mean\_dev = diff/n

    # co-efficient of variation

    cov = std/mean

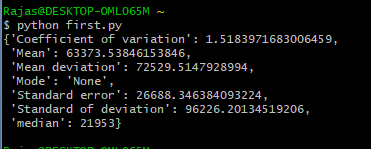
    # standard error

    st\_error = std/sqrt(n)

    return {"Mean": mean, "median": median, "Mode": mode, "Standard of deviation": std, "Mean deviation": mean\_dev, "Coefficient of variation": cov, "Standard error": st\_error}

arr = {'2000': 231,'2001': 21357,'2002': 11685,'2003':33819,'2004': 228802,'2005': 88003,'2006': 6605,'2007': 712,'2008': 88011,'2009': 1790,'2010': 320120,'2011': 21953,'2012': 768}

pprint(get\_desc(arr))

****

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

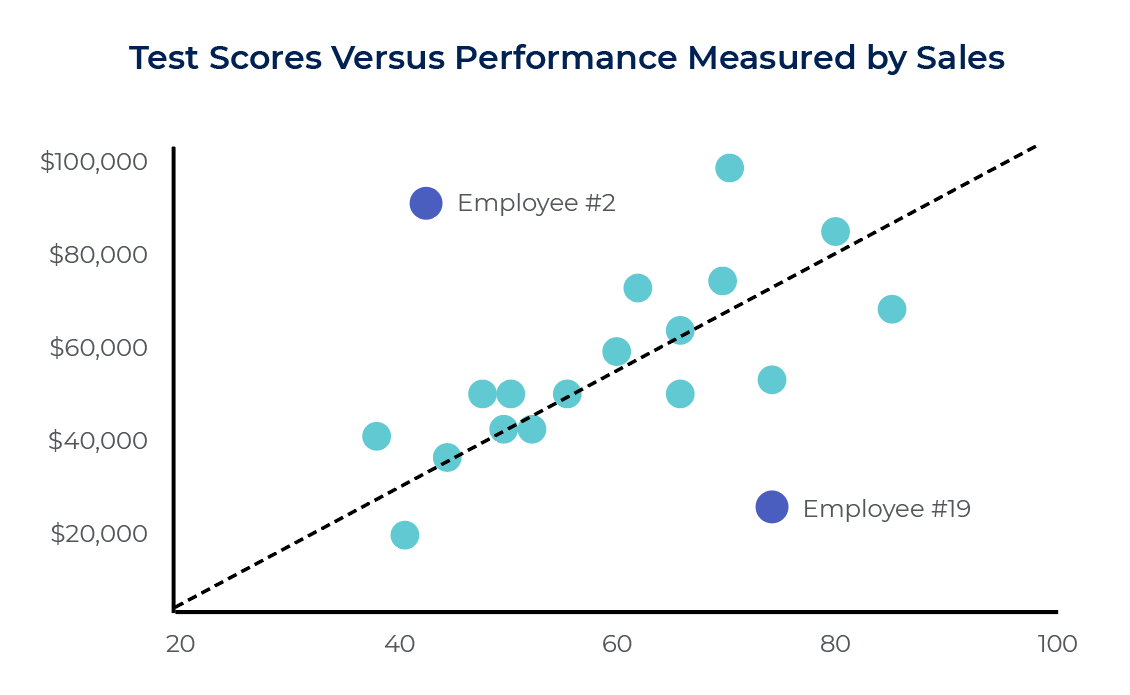
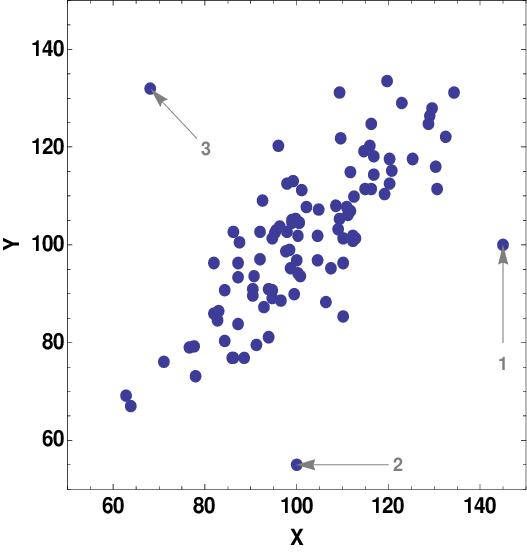
**Questions:**

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

* It gives an overall idea of the data.
* Mean is useful as it takes all of the data in consideration. It is affected by outlier data. It’s not as useful when the data is biased. It is used to see consistency of data
* Median is not affected by extreme or outlier values. It is really useful while dealing with biased data.
* Mode can be used with numerical as well as non numerical data. It is used to find most frequently occurring data.

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?

* An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population.
* Graphical methods like box plot can be used to find outlier data.



The above 2 images show examples of outliers.

Outliers affect calculation of mean or average of the data. Median and mode are not affected to the same degree.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

**Conclusion:** We learned about central tendency and variability of data and how it can be used in statistics and a python program was executed to calculate measures of central tendency and variability.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**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