CHAPTER -3
MEASURES OF CENTERAL TENDENCY

MEASURES OF CENTERAL TENDENCY

Introduction

- When we want to make comparison between groups of numbers it is good to have a single value that is considered to be a good representative of each group. This single value is called the **average** of the group. Averages are also called measures of central tendency.
- An average which is representative is called typical average and an average which is not representative and has only a theoretical value is called a descriptive average

Importance:

- To comprehend the data easily.
- To facilitate comparison.
- To make further statistical analysis.

The Summation Notation:

- Let $X_{1,}$ $X_{2,}$ X_3 ...X_N be a number of measurements where N is the total number of observation and X_i is i^{th} observation.
- Very often in statistics an algebraic expression of the form $X_1+X_2+X_3+...+X_N$ is used in a formula to compute a statistic. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the summation notation.
- The symbol $\sum_{i=1}^{N} X_i$ is a mathematical shorthand for $\sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$

The expression is read, "the sum of X sub i from i equals 1 to N." It means "add up all the numbers."

Example: Suppose the following were scores made on the first homework assignment for five students in the class: 5, 7, 7, 6, and 8. In this example set of five numbers, where N=5, the summation could be written:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 5 + 7 + 7 + 6 + 8 = 33$$

The "i=1" in the bottom of the summation notation tells where to begin the sequence of summation. If the expression were written with "i=3", the summation would start with the third number in the set. For example:

$$\sum_{i=3}^{N} X_i = X_3 + X_4 + \ldots + X_N$$

In the example set of numbers, this would give the following result:

$$\sum_{i=3}^{N} X_i = X_3 + X_4 + X_5 = 7 + 6 + 8 = 21$$

The "N" in the upper part of the summation notation tells where to end the sequence of summation. If there were only three scores then the summation and example would be:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 5 + 7 + 7 = 21$$

Sometimes if the summation notation is used in an expression and the expression must be written a number of times, as in a proof, then a shorthand notation for the shorthand notation is employed. When the summation sign "" is used without additional notation, then "i=1" and "N" are assumed.

For example:

$$\sum X = \sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$$

PROPERTIES OF SUMMATION

1. $\sum_{i=1}^{n} k = nk$ where k is any constant
2. $\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$ where k is any constant
3. $\sum_{i=1}^{n} (a + bX_i) = na + b\sum_{i=1}^{n} X_i$ where a and b are any constant
4. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$
5. $\sum_{i=1}^{N} (X_i * Y_i) = (X_1 * Y_1) + (X_2 * Y_2) + \ldots + (X_N * Y_N)$

Example: considering the following data determine

X	Y
5	6
7	7
7	8
6	7
8	8

a) $\sum_{i=1}^{5} X_i$

b) $\sum_{i=1}^{5} Y_i$

c) $\sum_{i=1}^{5} 10$

d) $\sum_{i=1}^{5} (X_i + Y_i)$

e) $\sum_{i=1}^{5} (X_i - Y_i)$

f) $\sum_{i=1}^{5} X_i Y_i$

g) $\sum_{i=1}^{5} X_i^{\ 2}$

h) $(\sum_{i=1}^{5} X_i)(\sum_{i=1}^{5} Y_i)$

Solutions:

a) $\sum_{i=1}^{5} X_i = 5+7+7+6+8 = 33$

b) $\sum_{i=1}^{5} Y_i = 6+7+8+7+8 = 36$

c) $\sum_{i=1}^{5} 10 = 5*10 = 50$

d) $\sum_{i=1}^{5} (X_i + Y_i) = (5+6)+(7+7)+(7+8)+(6+7)+(8+8) = 69 = 33+36$

e) $\sum_{i=1}^{5} (X_i - Y_i) = (5-6)+(7-7)+(7-8)+(6-7)+(8-8) = -3 = 33-36$

f) $\sum_{i=1}^{5} X_i Y_i = 5*6+7*7+7*8+6*7+8*8 = 241$

g) $\sum_{i=1}^{5} X_i^{\ 2} = 5^2 + 7^2 + 7^2 + 6^2 + 8^2 = 223$

h) $(\sum_{i=1}^{5} X_i)(\sum_{i=1}^{5} Y_i) = 33 * 36 = 1188$

- **Properties of measures of central tendency (a typical average should posses the following)**
  - It should be rigidly defined.
  - It should be based on all observation under investigation.
  - It should be as little as affected by extreme observations.
  - It should be capable of further algebraic treatment.
  - It should be as little as affected by fluctuations of sampling.
  - It should be ease to calculate and simple to understand.

Types of measures of central tendency

There are several different measures of central tendency; each has its advantage and disadvantage.

- The Mean (Arithmetic, Geometric and Harmonic)
- The Mode
- The Median
- Quintiles (Quartiles, Deciles and Percentiles)

The choice of these averages depends up on which best fit the property under discussion.

The Arithmetic Mean

- Is defined as the sum of the magnitude of the items divided by the number of items.
- The mean of X_1, X_2 ,X_3 …X_n is denoted by A.M ,m or $\overline{X}$ and is given by:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\Rightarrow \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- If X_1 occurs f_1 times
- If X_2occurs f_2 times
- If X_n occurs f_n times

Then the mean will be $\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$, where k is the number of classes and $\sum_{i=1}^{k} f_i = n$

Example: Obtain the mean of the following number

2, 7, 8, 2, 7, 3, 7

Solution:

X_i	f_i	$X_i f_i$
2	2	4
3	1	3
7	3	21
8	1	8
Total	7	36

$$\overline{X} = \frac{\sum_{i=1}^{4} f_i X_i}{\sum_{i=1}^{4} f_i} = \frac{36}{7} = 5.15$$

Arithmetic Mean for Grouped Data

If data are given in the shape of a continuous frequency distribution, then the mean is obtained as follows:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}, Where$$

X_i =the class mark of the i^{th} class and f_i = the frequency of the i^{th} class

Example: calculate the mean for the following age distribution.

Class	frequency
6- 10	35
11- 15	23
16- 20	15
21- 25	12
26- 30	9
31- 35	6

Solutions:

- First find the class marks
- Find the product of frequency and class marks
- Find mean using the formula.

Class	f_i	X_i	$X_i f_i$
6- 10	35	8	280
11- 15	23	13	299
16- 20	15	18	270
21- 25	12	23	276
26- 30	9	28	252
31- 35	6	33	198
Total	100		1575

$$\overline{X} = \frac{\sum_{i=1}^{6} f_i X_i}{\sum_{i=1}^{6} f_i} = \frac{1575}{100} = 15.75$$

If the values in a series or mid values of a class are large enough, coding of values is a good device to simplify the calculations**.**

Special properties of Arithmetic mean

1. The sum of the deviations of a set of items from their mean is always zero. i.e. $\sum_{i=1}^{n} (X_i - \overline{X}) = 0.$
2. The sum of the squared deviations of a set of items from their mean is the minimum. i.e. $\sum_{i=1}^{n} (Xi - \overline{X})^2 < \sum_{i=1}^{n} (X_i - A)^2, A \neq \overline{X}$
3. If $\overline{X}_1$ is the mean of n_1 observations

   If $\overline{X}_2$ is the mean of n_2 observations

   .

   .

   If $\overline{X}_k$ is the mean of n_k observations

   Then the mean of all the observation in all groups often called the combined mean is given by:

$$\mathbf{\overline{X}_c = \frac{\overline{X}_1 n_1 + \overline{X}_2 n_2 + \ldots + \overline{X}_k n_k}{n_1 + n_2 + \ldots n_k} = \frac{\sum_{i=1}^{k} \overline{X}_i n_i}{\sum_{i=1}^{k} n_i}}$$

Example: In a class there are 30 females and 70 males. If females averaged 60 in an examination and boys averaged 72, find the mean for the entire class.

Solutions:

Females	*Males*
$\overline{X}_1 = 60$	$\overline{X}_2 = 72$
$n_1 = 30$	$n_2 = 70$

$$\overline{X}_c = \frac{\overline{X}_1 n_1 + \overline{X}_2 n_2}{n_1 + n_2} = \frac{\sum_{i=1}^{2} \overline{X}_i n_i}{\sum_{i=1}^{2} n_i}$$

$$\Rightarrow \overline{X}_c = \frac{30(60) + 70(72)}{30 + 70} = \frac{6840}{100} = 68.40$$

4. If a wrong figure has been used when calculating the mean the correct mean can be obtained without repeating the whole process using:

$$\mathbf{CorrectMean = WrongMean + \frac{(CorrectValue - WrongValue)}{n}}$$

Where n is total number of observations.

Example: An average weight of 10 students was calculated to be 65.Latter it was discovered that one weight was misread as 40 instead of 80 k.g. Calculate the correct average weight.

Solutions:

$$\mathbf{CorrectMean = WrongMean + \frac{(CorrectValue - WrongValue)}{n}}$$

$$\mathbf{CorrectMean = 65 + \frac{(80 - 40)}{10} = 65 + 4 = 69k.g.}$$

5. The effect of transforming original series on the mean.
   a) If a constant *k* is added/ subtracted to/from every observation then the new mean will be *the old mean± k* respectively.
   b) If every observations are multiplied by a constant *k* then the new mean will be *k*old mean*

Example:

1. The mean of n Tetracycline Capsules $X_1, X_2, \ldots, X_n$ are known to be 12 gm. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ (i = 1, 2, …, n) then what will be the mean of the new set of capsules

Solutions:

$$\mathbf{NewMean = 2*OldMean - 0.5 = 2*12 - 0.5 = 23.5}$$

2. The mean of a set of numbers is 500.
   a) If 10 is added to each of the numbers in the set, then what will be the mean of the new set?
   b) If each of the numbers in the set are multiplied by -5, then what will be the mean of the new set?

Solutions:

$$\mathbf{a).NewMean = OldMean + 10 = 500 + 10 = 510}$$

$$\mathbf{b).NewMean = -5 * OldMean = -5 * 500 = -2500}$$

Weighted Mean

- When a proper importance is desired to be given to different data a weighted mean is appropriate.
- Weights are assigned to each item in proportion to its relative importance.
- Let $X_1, X_2, \ldots X_n$ be the value of items of a series and $W_1, W_2, \ldots W_n$ their corresponding weights , then the weighted mean denoted $\bar{X}_w$ is defined as:

$$\overline{X}_w = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i-1}^{n} W_i}$$

Example:

A student obtained the following percentage in an examination:
English 60, Biology 75, Mathematics 63, Physics 59, and chemistry 55.Find the students weighted arithmetic mean if weights 1, 2, 1, 3, 3 respectively are allotted to the subjects.

Solutions:

$$\overline{X}_w = \frac{\sum_{i=1}^{5} X_i W_i}{\sum_{i-1}^{5} W_i} = \frac{60*1+75*2+63*1+59*3+55*3}{1+2+1+3+3} = \frac{615}{10} = 61.5$$

Merits and Demerits of Arithmetic Mean

Merits:

- It is rigidly defined.
- It is based on all observation.
- It is suitable for further mathematical treatment.
- It is stable average, i.e. it is not affected by fluctuations of sampling to some extent.
- It is easy to calculate and simple to understand.

Demerits:

- It is affected by extreme observations.
- It cannot be used in the case of open end classes.
- It cannot be determined by the method of inspection.
- It cannot be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.
- It can be a number which does not exist in a serious.
- Sometimes it leads to wrong conclusion if the details of the data from which it is obtained are not available.
- It gives high weight to high extreme values and less weight to low extreme values.

The Geometric Mean

- The geometric mean of a set of n observation is the n^{th} root of their product.
- The geometric mean of X_1, X_2 ,X_3 …X_n is denoted by G.M and given by:

$$G.M = \sqrt[n]{X_1 * X_2 * \ldots * X_n}$$

- Taking the logarithms of both sides

$$\log(G.M) = \log(\sqrt[n]{X_1 * X_2 * \ldots * X_n}) = \log(X_1 * X_2 * \ldots * X_n)^{\frac{1}{n}}$$

$$\Rightarrow \log(G.M) = \frac{1}{n}\log(X_1 * X_2 * \ldots . * X_n) = \frac{1}{n}(\log X_1 + \log X_2 + \ldots + \log X_n)$$

$$\Rightarrow \log(G.M) = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

$\Rightarrow$The logarithm of the G.M of a set of observation is the arithmetic mean of their logarithm.

$$\Rightarrow G.M = \text{Anti}\log\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)$$

<u>Example:</u>

Find the G.M of the numbers 2, 4, 8.

Solutions:

$$G.M = \sqrt[n]{X_1 * X_2 * \ldots * X_n} = \sqrt[3]{2*4*8} = \sqrt[3]{64} = 4$$

Remark: The Geometric Mean is useful and appropriate for finding averages of ratios.

The Harmonic Mean

The harmonic mean of X_1, X_2 , X_3 …X_n is denoted by H.M and given by:

$$\boxed{H.M = \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}}$$, This is called simple harmonic mean.

In a case of frequency distribution:

$$\boxed{H.M = \frac{n}{\sum_{i=1}^{k}\frac{f_i}{X_i}}} \ , \ n = \sum_{i=1}^{k} f_i$$

If observations X_1, X_2… Xn have weights W_1, W_2 …Wn respectively, then their harmonic mean is given by

$$\boxed{H.M = \frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} W_i / X_i}}$$, This is called Weighted Harmonic Mean.

Remark: The Harmonic Mean is useful and appropriate in finding average speeds and average rates.

Example: A cyclist pedals from his house to his college at speed of 10 km/hr and back from the college to his house at 15 km/hr. Find the average speed.

Solution: Here the distance is constant

➔The simple H.M is appropriate for this problem.

X_1= 10km/hr X_2=15km/hr

$$H.M = \frac{2}{\frac{1}{10} + \frac{1}{15}} = 12 km/hr$$

The Mode

- Mode is a value which occurs most frequently in a set of values
- The mode may not exist and even if it does exist, it may not be unique.
- In case of discrete distribution the value having the maximum frequency is the model value.

Examples:

1. Find the mode of 5, 3, 5, 8, 9

Mode =5

2. Find the mode of 8, 9, 9, 7, 8, 2, and 5.
   It is a bimodal Data: 8 and 9
3. Find the mode of 4, 12, 3, 6, and 7.
   No mode for this data.

- The mode of a set of numbers $X_1, X_2, \ldots, X_n$ is usually denoted by $\hat{X}$.

Mode for Grouped data

If data are given in the shape of continuous frequency distribution, the mode is defined as:

$$\hat{X} = L_{mo} + w\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$$

Where:

$\hat{X} = the\ \text{mode}\ of\ the\ distribution$

$w = the\ size\ of\ the\ \text{mod}al\ class$

$\Delta_1 = f_{mo} - f_1$

$\Delta_2 = f_{mo} - f_2$

$f_{mo} = frequency\ of\ the\ \text{mod}al\ class$

$f_1 = frequency\ of\ the\ class\ preceeding\ the\ \text{mod}al\ class$

$f_2 = frequency\ of\ the\ class\ following\ the\ \text{mod}al\ class$

Note: The modal class is a class with the highest frequency.

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farms	No. of farms
5-15	8
15-25	12
25-35	17
35-45	29
45-55	31
55-65	5
65-75	3

Solutions:

$45 - 55$ **is the modal class, since it is a class with the highest frequency.**

$L_{mo} = 45$

$w = 10$

$\Delta_1 = f_{mo} - f_1 = 2$

$\Delta_2 = f_{mo} - f_2 = 26$

$f_{mo} = 31$

$f_1 = 29$

$f_2 = 5$

$$\Rightarrow \hat{X} = 45 + 10\left(\frac{2}{2+26}\right)$$
$$= 45.71$$

Merits and Demerits of Mode

Merits:

- It is not affected by extreme observations.
- Easy to calculate and simple to understand.
- It can be calculated for distribution with open end class

Demerits:

- It is not rigidly defined.
- It is not based on all observations
- It is not suitable for further mathematical treatment.
- It is not stable average, i.e. it is affected by fluctuations of sampling to some extent.
- Often its value is not unique.

Note: being the point of maximum density, mode is especially useful in finding the most popular size in studies relating to marketing, trade, business, and industry. It is the appropriate average to be used to find the ideal size.

The Median

- In a distribution, median is the value of the variable which divides it in to two equal halves.

- In an ordered series of data median is an observation lying exactly in the middle of the series. It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

-If $X_1, X_2 \ldots X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]} \ldots X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value.

$\Rightarrow X_{[1]} < X_{[2]} < \ldots < X_{[n]}$

-Median is denoted by $\tilde{X}$.

Median for ungrouped data

$$\tilde{X} = \begin{cases} X_{[(n+1)/2]} & \text{, If n is odd.} \\ \frac{1}{2}(X_{[n/2]} + X_{[(n/2)+1]}), & \text{If n is even} \end{cases}$$

Example: Find the median of the following numbers.

a) 6, 5, 2, 8, 9, 4.
b) 2, 1, 3, 5, 8.

Solutions:

a) First order the data: 2, 4, 5, 6, 8, 9

Here n=6

$$\tilde{X} = \frac{1}{2}(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]})$$
$$= \frac{1}{2}(X_{[3]} + X_{[4]})$$
$$= \frac{1}{2}(5+6) = 5.5$$

b) Order the data :1, 2, 3, 5, 8

Here n=5

$$\tilde{X} = X_{[\frac{n+1}{2}]}$$
$$= X_{[3]}$$
$$= 3$$

Median for grouped data If data are given in the shape of continuous frequency distribution, the median is defined as:

$$\tilde{X} = L_{med} + \frac{w}{f_{med}}(\frac{n}{2} - c)$$

Where :

L_{med} = lower class boundary of the median class.

w = the size of the median class

n = total number of observations.

c = the cumulative frequency (less than type) preceeding the median class.

f_{med} = the frequency of the median class.

Remark:

The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{n}{2}$.

Example: Find the median of the following distribution.

Class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Solutions:

- First find the less than cumulative frequency.
- Identify the median class.
- Find median using formula.

Class	Frequency	Cumu.Freq(less than type)
40-44	7	7
45-49	10	17
50-54	22	39
55-59	15	54
60-64	12	66
65-69	6	72
70-74	3	75

$$\frac{n}{2} = \frac{75}{2} = 37.5$$

39 is the first cumulative frequency to be greater than or equal to 37.5

$\Rightarrow 50-54$ is the median class.

$L_{med} = 49.5, \quad w = 5$

$n = 75, \quad c = 17, \quad f_{med} = 22$

$$\Rightarrow \tilde{X} = L_{med} + \frac{w}{f_{med}}\left(\frac{n}{2} - c\right)$$

$$= 49.5 + \frac{5}{22}(37.5 - 17)$$

$$= 54.16$$

<u>Merits and Demerits of Median</u>

Merits:

- Median is a positional average and hence not influenced by extreme observations.
- Can be calculated in the case of open end intervals.
- Median can be located even if the data are incomplete.

Demerits:

- It is not a good representative of data if the number of items is small.
- It is not amenable to further algebraic treatment.
- It is susceptible to sampling fluctuations.

CHAPTER -4

Measures of Dispersion (Variation)

Introduction and objectives of measuring Variation

-The scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.

-Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

Objectives of measuring Variation:

- To judge the reliability of measures of central tendency
- To control variability itself.
- To compare two or more groups of numbers in terms of their variability.
- To make further statistical analysis.

Absolute and Relative Measures of Dispersion

The measures of dispersion which are expressed in terms of the original unit of a series are termed as *absolute measures*. Such measures are not suitable for comparing the variability of two distributions which are expressed in different *units of measurement* and different average size. Relative measures of dispersions are a ratio or percentage of a measure of absolute dispersion to an appropriate measure of central tendency and are thus pure numbers independent of the *units of measurement*. For comparing the variability of two distributions (even if they are measured in the same unit), we compute the relative measure of dispersion instead of absolute measures of dispersion.

Types of Measures of Dispersion

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

1) Range and relative range
2) Standard deviation ,coefficient of variation and standard scores

The Range (R)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1:	32	35	36	36	37	38	40	42	42	43	43	45
Distribution 2:	32	32	33	33	33	34	34	34	34	34	35	45

For this reason, among others, the range is not the most important measure of variability.

$$R = L - S \qquad , L = \text{largest observation}$$
$$S = \text{smallest observation}$$

Range for grouped data:

If data are given in the shape of continuous frequency distribution, the range is computed as:

$$R = UCL_k - UCL_1, \quad UCL_k \text{ is upper class limit of the last class.}$$
$$UCL_1 \text{ is lower class limit of the first class.}$$

This is some times expressed as:

$$R = X_k - X_1, \quad X_k \text{ is class mark of the last class.}$$
$$X_1 \text{ is class mark of the first class.}$$

Merits and Demerits of range

Merits:

- It is rigidly defined.
- It is easy to calculate and simple to understand.

Demerits:

- It is not based on all observation.
- It is highly affected by extreme observations.
- It is affected by fluctuation in sampling.
- It is not liable to further algebraic treatment.
- It can not be computed in the case of open end distribution.
- It is very sensitive to the size of the sample.

Relative Range (RR)

-it is also some times called coefficient of range and given by:

$$RR = \frac{L - S}{L + S} = \frac{R}{L + S}$$

Example:

1. Find the relative range of the above two distribution.(exercise!)
2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of:
   a) Smallest observation
   b) Largest observation

Solutions :(2)

$$R = 4 \Rightarrow L - S = 4 \underline{\hspace{4cm}} (1)$$

$$RR = 0.25 \Rightarrow L + S = 16 \underline{\hspace{3cm}} (2)$$

$$Solving\,(1)\,and\,(2)\,at\ the\ same\,time, one\,can\,obtain\,the\ following\,value$$

$$L = 10 \quad and \quad S = 6$$

The Variance

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2, \quad i = 1,2,....N$$

For the case of frequency distribution it is expressed as:

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum f_i (X_i - \mu)^2, \quad i = 1,2,....k$$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample\ Varince = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad i = 1,2,.....,n$$

For the case of frequency distribution it is expressed as:

$$Sample\ Varince = S^2 = \frac{1}{n-1}\sum f_i (X_i - \overline{X})^2, \quad i = 1,2,....k$$

We usually use the following short cut formula.

$$S^2 = \frac{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2}{n-1}, \ for\ raw\,data.$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i X_i^{\,2} - n\overline{X}^2}{n-1}, \; for \; frequency\, distribution.$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Population\, s\tan dard\ deviation = \sigma = \sqrt{\sigma^2}$$

$$Sample\, s\tan dard\ deviation = s = \sqrt{S^2}$$

The following steps are used to calculate the sample standard deviation

1. Find the arithmetic mean.
2. Find the difference between each observation and the mean.
3. Square these differences.
4. Sum the squared differences.
5. Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e., n-1 (where n is equal to the number of observations in the data set).
6. Square root the result obtained from step 5

Examples: Find the variance and standard deviation of the following sample data

1. 5, 17, 12, 10.
2. The data is given in the form of frequency distribution.

Class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Solutions:

1. $\bar{X} = 11$

X_i	5	10	12	17	Total
$(X_i - \bar{X})^2$	36	1	1	36	74

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{74}{3} = 24.67.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{24.67} = 4.97.$$

2. $\bar{X} = 55$

X_i(C.M)	42	47	52	57	62	67	72	Total
$f_i(X_i - \bar{X})^2$	1183	640	198	60	588	864	867	4400

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}{n-1} = \frac{4400}{74} = 59.46.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{59.46} = 7.71.$$

<u>Special properties of Standard deviations</u>

1. $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} < \sqrt{\frac{\sum(X_i - A)^2}{n-1}} \quad , A \neq \bar{X}$

2. For normal (symmetric distribution the following holds.

- Approximately 68.27% of the data values fall within one standard deviation of the mean. i.e. with in $(\bar{X} - S, \ \bar{X} + S)$
- Approximately 95.45% of the data values fall within two standard deviations of the mean. i.e. with in $(\bar{X} - 2S, \ \bar{X} + 2S)$
- Approximately 99.73% of the data values fall within three standard deviations of the mean. i.e. with in $(\bar{X} - 3S, \ \bar{X} + 3S)$

3. If the standard deviation of $X_1, X_2, \ldots. X_n \ is \ S$, then the standard deviation of

a) $X_1 + k,\ X_2 + k, \ldots.X_n + k\ \ will\, also\, be\ S$

b) $kX_1,\ kX_2, \ldots.kX_n\ would\ be|k|S$

c) $a + kX_1,\ a + kX_2, \ldots.a + kX_n\ would\ \ be|k|\ S$

Exercise: Verify each of the above relation ship, considering *k* and *a* as constants.

Examples:

1. The mean and standard deviation of n Tetracycline Capsules $X_1, X_2, \ldots.X_n$ are known to be 12 gm and 3 gm respectively. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i – 0.5$ (i = 1, 2, …, n) then what will be the standard deviation of the new set of capsules
2. The mean and the standard deviation of a set of numbers are respectively 500 and 10.
   a. If 10 is added to each of the numbers in the set, then what will be the variance and standard deviation of the new set?
   b. If each of the numbers in the set are multiplied by -5, then what will be the variance and standard deviation of the new set?

Solutions:

1. Using c) above the new standard deviation $= |k|S = 2*3 = 6$
2. a. They will remain the same.
   b. New standard deviation$= = |k|S = 5*10 = 50$

<u>Coefficient of Variation (C.V)</u>

- Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{S}{\overline{X}}*100$$

- The distribution having less C.V is said to be less variable or more consistent.

Examples:

1. An analysis of the monthly wages paid (in Birr) to workers in two firms A and B belonging to the same industry gives the following results

Value	Firm A	Firm B
Mean wage	52.5	47.5
Median wage	50.5	45.5
Variance	100	121

In which firm A or B is there greater variability in individual wages?

Solutions:

Calculate coefficient of variation for both firms.

$$C.V_A = \frac{S_A}{\overline{X}_A} * 100 = \frac{10}{52.5} * 100 = 19.05\%$$

$$C.V_B = \frac{S_B}{\overline{X}_B} * 100 = \frac{11}{47.5} * 100 = 23.16\%$$

Since $C.V_A < C.V_B$, in firm B there is greater variability in individual wages.

2. A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1	25	24	23	26	17
City2	22	21	24	22	20
City3	32	27	35	24	28

Which city have the most consistent temperature, based on these data?

(Exercise)

Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean $\overline{X}$ and standard deviation S, then its value in standard units is

$$Z = \frac{X - \mu}{\sigma}, \quad \textit{for population.}$$

$$Z = \frac{X - \overline{X}}{S}, \quad \textit{for sample}$$

- Z gives the deviations from the mean in units of standard deviation
- Z gives the number of standard deviation a particular observation lie above or below the mean.
- It is used to compare two observations coming from different groups.

Examples:

1. Two sections were given introduction to statistics examinations. The following information was given.

Value	Section 1	Section 2
Mean	78	90
Stan.deviation	6	5

Student A from section 1 scored 90 and student B from section 2 scored 95.Relatively speaking who performed better?

Solutions:

Calculate the standard score of both students.

$$Z_A = \frac{X_A - \overline{X}_1}{S_1} = \frac{90-78}{6} = 2$$

$$Z_B = \frac{X_B - \overline{X}_2}{S_2} = \frac{95-90}{5} = 1$$

➔ Student A performed better relative to his section because the score of student A is two standard deviation above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

2. Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:

Value	Group one	Group two
Mean	10.4 min	11.9 min
Stan.dev.	1.2 min	1.3 min

Relatively speaking:

a) Which group is more consistent in its performance
b) Suppose a person A from group one take 9.2 minutes while person B from Group two take 9.3 minutes, who was faster in performing the task? Why?

Solutions:

a) Use coefficient of variation.

$$C.V_1 = \frac{S_1}{\overline{X}_1} * 100 = \frac{1.2}{10.4} * 100 = 11.54\%$$

$$C.V_2 = \frac{S_2}{\overline{X}_2} * 100 = \frac{1.3}{11.9} * 100 = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

b) Calculate the standard score of A and B

$$Z_A = \frac{X_A - \overline{X}_1}{S_1} = \frac{9.2-10.4}{1.2} = -1$$

$$Z_B = \frac{X_B - \overline{X}_2}{S_2} = \frac{9.3-11.9}{1.3} = -2$$

➔Child B is faster because the time taken by child B is two standard deviation shorter than the average time taken by group 2 while, the time taken by child A is only one standard deviation shorter than the average time taken by group 1.

4.2.3. Moments

- If X is a variable that assume the values X_1, X_2,…..,X_n then

1. The r[th] moment is defined as:

$$\bar{X}^r = \frac{X_1^{\ r} + X_2^{\ r} + \ldots + X_n^{\ r}}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i^{\ r}}{n}$$

- For the case of frequency distribution this is expressed as:

$$\bar{X}^r = \frac{\sum_{i=1}^{k} f_i X_i^{\ r}}{n}$$

- If $r = 1$,it is the simple arithmetic mean, this is called the first moment.

2. The r[th] moment about the mean (the r[th] central moment)

- Denoted by M_r and defined as:

$$M_r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^r}{n} = \frac{(n-1)}{n} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r = \frac{\sum_{i=1}^{k} f_i (X_i - \bar{X})^r}{n}$$

- If $r = 2$, it is population variance, this is called the second central moment. If we assume $n-1 \approx n$,it is also the sample variance.

3. The r[th] moment about any number A is defined as:

- Denoted by $M_r^{\ '}$ and

$$M_r^{\ '} = \frac{\sum_{i=1}^{n} (X_i - A)^r}{n} = \frac{(n-1)}{n} \frac{\sum_{i=1}^{n} (X_i - A)^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r{}' = \frac{\sum_{i=1}^{k} f_i (X_i - A)^r}{n}$$

Example:

1. Find the first two moments for the following set of numbers 2, 3, 7
2. Find the first three central moments of the numbers in problem 1
3. Find the third moment about the number 3 of the numbers in problem 1.

Solutions:

1. Use the rth moment formula.

$$\overline{X}^r = \frac{\sum_{i=1}^{n} X_i^r}{n}$$

$$\Rightarrow \overline{X}^1 = \frac{2+3+7}{3} = 4 = \overline{X}$$

$$\overline{X}^2 = \frac{2^2+3^2+7^2}{3} = 20.67$$

2. Use the rth central moment formula.

$$M_r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^r}{n}$$

$$\Rightarrow M_1 = \frac{(2-4)+(3-4)+(7-4)}{3} = 0$$

$$M_2 = \frac{(2-4)^2+(3-4)^2+(7-4)^2}{3} = 4.67$$

$$M_3 = \frac{(2-4)^3+(3-4)^3+(7-4)^3}{3} = 6$$

3. Use the rth moment about A.

$$M_r = \frac{\sum_{i=1}^{n} (X_i - A)^r}{n}$$

$$\Rightarrow M_3' = \frac{(2-3)^3 + (3-3)^3 + (7-3)^3}{3} = 21$$

4.2.4. Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size.
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$Mean - Mode = 3*(Mean - Median)$$

Measures of Skewness

- Denoted by α_3
- There are various measures of skewness.

1. The Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{S\tan dard\ deviation} = \frac{\bar{X} - \hat{X}}{S}$$

2. The moment coefficient of skewness

$$\alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{M_3}{(\sigma^2)^{3/2}} = \frac{M_3}{\sigma^3}, \ Where\ \sigma\ is\ the\ population\ s\tan dard\ deviation.$$

The shape of the curve is determined by the value of α_3

- $If\ \alpha_3 > 0\ then\ the\ distribution\ is\ positively\ skewed.$
- $If\ \alpha_3 = 0\ then\ the\ distribution\ is\ symmetric.$
- $If\ \alpha_3 < 0\ then\ the\ distribution\ is\ negatively\ skewed.$

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.

Examples:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?

   Solutions:

   Use the Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{Standard\ deviation} = \frac{32 - 30.5}{10} = 0.15$$

$$\alpha_3 > 0 \Rightarrow The\ distribution\ is\ positively\ skewed.$$

2. Some characteristics of annually family income distribution (in Birr) in two regions is as follows:

Region	Mean	Median	Standard Deviation
A	6250	5100	960
B	6980	5500	940

a) Calculate coefficient of skewness for each region
b) For which region is, the income distribution more skewed. Give your interpretation for this Region
c) For which region is the income more consistent?

Solutions: (**exercise)**

3. For a moderately skewed frequency distribution, the mean is 10 and the median is 8.5. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness and the probable mode of the distribution. (**exercise)**
4. The sum of fifteen observations, whose mode is 8, was found to be 150 with coefficient of variation of 20%
   (a) Calculate the Pearsonian coefficient of skewness and give appropriate conclusion.
   (b) Are smaller values more or less frequent than bigger values for this distribution?
   (c) If a constant *k* was added on each observation, what will be the new Pearsonian coefficient of skewness? Show your steps. What do you conclude from this?

(**Exercise**)

4.2.5 Kurtosis

Kurtosis is the degree of peakdness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called *leptokurtic*. If a curve representing a distribution is flat topped, it is called *platykurtic*. The normal distribution which is not very high peaked or flat topped is called *mesokurtic*.

Measures of kurtosis

The moment coefficient of kurtosis:

- Denoted by α_4 and given by

$$\alpha_4 = \frac{M_4}{M_2{}^2} = \frac{M_4}{\sigma^4}$$

$$Where: M_4 \text{ is the fourth moment about the mean.}$$
$$M_2 \text{ is the second moment about the mean.}$$
$$\sigma \text{ is the population standard deviation.}$$

The peakdness depends on the value of α_4.

$$If\ \alpha_4 > 3 \text{ then the curve is leptokurtic.}$$
$$If\ \alpha_4 = 3 \text{ then the curve is mesokurtic.}$$
$$If\ \alpha_4 < 3 \text{ then the curve is platykurtic.}$$

Examples:

1. If the first four central moments of a distribution are:

$$M_1 = 0,\ M_2 = 16,\ M_3 = -60,\ M_4 = 162$$

   a) Compute a measure of skewness
   b) Compute a measure of kurtosis and give your interpretation.

Solutions:

a) $$\alpha_3 = \frac{M_3}{M_2{}^{3/2}} = \frac{-60}{16^{3/2}} = -0.94 < 0$$

$$\Rightarrow \text{The distribution is negatively skewed.}$$

b) $$\alpha_4 = \frac{M_4}{M_2{}^2} = \frac{162}{16^2} = 0.6 < 3$$

$$\Rightarrow \text{The curve is platykurtic.}$$

2. The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4th moment about the mean is 243. Compute the Pearsonian coefficient of skewness and identify the type of skewness. Assume (n-1 = n) (**exercise)**.

3. If the standard deviation of a symmetric distribution is 10, what should be the value of the fourth moment so that the distribution is mesokurtic?

Solutions (**exercise)**.