2. SUMMARIZING OF DATA

2.1 Measures of Central 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. 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.

Objectives:

- ☞ 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 $X_1+X_2+X_3+...+X_N$

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

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. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

The sum of the product of the two variables could be written:

$$\sum_{i=1}^{N} (X_i * Y_i) = (X_1 * Y_1) + (X_2 * Y_2) + \ldots + (X_N * Y_N)$$

2.2. 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
- Quantiles (Quartiles, Deciles and Percentiles)

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

2.2.1. Mean

The Arithmetic Mean

- Is defined as the sum of the magnitude of the items divided by the number of items.
- The mean of $X_1, X_2, X_3 \ldots X_n$ 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

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

Exercises:

1. Marks of 75 students are summarized in the following frequency distribution:

Marks	No. of students
40-44	7
45-49	10
50-54	22
55-59	f_4
60-64	f_5
65-69	6
70-74	3

If 20% of the students have marks between 55 and 59

i. Find the missing frequencies f_4 and f_5.
ii. Find the mean.

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

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

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 with out 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 kg. Calculate the correct average weight.

Solutions:

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

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

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, …, Xn 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, …X_n be the value of items of a series and W_1, W_2, …W_n their corresponding weights , then the weighted mean denoted $\overline{X}_w$ 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:

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

The Geometric Mean

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

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

- Taking the logarithms of both sides

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

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

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

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

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

Example:

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

Solutions:

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

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

The Harmonic Mean

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

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

In a case of frequency distribution:

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

If observations $X_1, X_2, \ldots X_n$ have weights $W_1, W_2, \ldots W_n$ respectively, then their harmonic mean is given by

$$\boxed{\mathbf{H.M} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{W_i}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{W_i}/\mathbf{X_i}}}$$, 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

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

2.2.1. 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 modal 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:

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

Where:

$\hat{X} = the \bmod e\ of\ the\ distribution$

$w = the\ size\ of\ the \bmod al\ class$

$\Delta_1 = f_{mo} - f_1$

$\Delta_2 = f_{mo} - f_2$

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

$f_1 = frequency\ of\ the\ class\ preceeding\ the \bmod al\ class$

$f_2 = frequency\ of\ the\ class\ following\ the \bmod al\ class$

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* mod *al class,* sin *ce it is a class with the highest frequency.*

$L_{mo} = 45$

$w = 10$

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

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

$f_{mo} = 31$

$f_1 = 29$

$f_2 = 5$

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

$$= 45.71$$

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.

2.2.3. 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 $\hat{X}$.

Median for ungrouped data

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

Example: Find the median of the following numbers.

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

$$L_{med}=49.5, \quad w=5$$
$$n=75, \quad c=17, \quad f_{med}=22$$

$$\Rightarrow \tilde{X}=L_{med}+\frac{w}{f_{med}}\left(\frac{n}{2}-c\right)$$
$$=49.5+\frac{5}{22}(37.5-17)$$
$$=54.16$$

2.3. Measure of locations (Quantiles)

When a distribution is arranged in order of magnitude of items, the median is the value of the middle term. Their measures that depend up on their positions in distribution quartiles, deciles, and percentiles are collectively called quintiles.

Quartiles:

- Quartiles are measures that divide the frequency distribution in to four equal parts.

- The value of the variables corresponding to these divisions are denoted Q_1, Q_2, and Q_3 often called the first, the second and the third quartile respectively.
- Q_1 is a value which has 25% items which are less than or equal to it. Similarly Q_2 has 50%items with value less than or equal to it and Q_3 has 75% items whose values are less than or equal to it.
- To find Q_i (i=1, 2, 3) we count $\frac{iN}{4}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$Q_i = L_{Q_i} + \frac{w}{f_{Q_i}}\left(\frac{iN}{4} - c\right) \quad , i = 1,2,3$$

Where:

L_{Q_i} = *lower class boundary of the quartile class.*

w = *the size of the quartile class*

N = *total number of observations.*

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

f_{Q_i} = *the frequency of the quartile class.*

Remark:

The quartile class (class containing Q_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{iN}{4}$.

Deciles:

- Deciles are measures that divide the frequency distribution in to ten equal parts.
- The values of the variables corresponding to these divisions are denoted D_1, D_2,.. D_9 often called the first, the second,…, the ninth deciles respectively.
- To find D_i (i=1, 2,..9) we count $\frac{iN}{10}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$D_i = L_{Di} + \frac{w}{f_{D_i}}(\frac{iN}{10} - c) \quad , i = 1,2,...,9$$

Where:

L_{D_i} = *lower class boundary of the decile class.*

w = *the size of the decile class*

N = *total number of observations.*

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

f_{D_i} = *the frequency of the decile class.*

Remark:

The deciles class (class containing D_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{iN}{10}$.

Percentiles:

- Percentiles are measures that divide the frequency distribution in to hundred equal parts.
- The values of the variables corresponding to these divisions are denoted P_1, P_2,.. P_{99} often called the first, the second,…, the ninety-ninth percentile respectively.
- To find P_i (i=1, 2,..99) we count $\frac{iN}{100}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

$$P_i = L_{P_i} + \frac{w}{f_{P_i}}(\frac{iN}{100} - c) \quad , i = 1,2,...,99$$

Where:

L_{P_i} = *lower class boundary of the percentile class.*

w = *the size of the percentile class*

N = *total number of observations.*

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

f_{P_i} = *the frequency of the percentile class.*

Remark:

The percentile class (class containing P_i) is the class with the small cumulative frequency (less than type) greater than or equal to $\frac{iN}{100}$.

Example: Considering the following distribution

Calculate:

a) All quartiles.
b) The 7^{th} decile.
c) The 90^{th} percentile.

Values	Frequency
140- 150	17
150- 160	29
160- 170	42
170- 180	72
180- 190	84
190- 200	107
200- 210	49
210- 220	34
220- 230	31
230- 240	16
240- 250	12

Solutions:

- First find the less than cumulative frequency.
- Use the formula to calculate the required quantile.

Values	Frequency	Cum.Freq(less than type)
140- 150	17	17
150- 160	29	46
160- 170	42	88
170- 180	72	160
180- 190	84	244
190- 200	107	351
200- 210	49	400
210- 220	34	434
220- 230	31	465
230- 240	16	481
240- 250	12	493

a) Quartiles:

i. Q_1

- determine the class containing the first quartile.

$\frac{\mathbf{N}}{\mathbf{4}} = \mathbf{123.25}$

$\Rightarrow \mathbf{170-180}$ **is the class containing the first quartile.**

$L_{Q_1} = 170$, $w = 10$

$N = 493$, $c = 88$, $f_{Q_1} = 72$

$$\Rightarrow Q_1 = L_{Q_1} + \frac{w}{f_{Q_1}}\left(\frac{N}{4} - c\right)$$

$$= 170 + \frac{10}{72}(123.25 - 88)$$

$$= \underline{\underline{174.90}}$$

ii. Q_2

- determine the class containing the second quartile.

$$\frac{2*N}{4} = 246.5$$

$$\Rightarrow 190 - 200 \quad \textit{is the class containing the second quartile.}$$

$L_{Q_2} = 190$, $w = 10$

$N = 493$, $c = 244$, $f_{Q_2} = 107$

$$\Rightarrow Q_2 = L_{Q_2} + \frac{w}{f_{Q_2}}\left(\frac{2*N}{4} - c\right)$$

$$= 170 + \frac{10}{72}(246.5 - 244)$$

$$= \underline{\underline{190.23}}$$

iii. Q_3

- determine the class containing the third quartile.

$$\frac{3*N}{4} = 369.75$$

$$\Rightarrow 200 - 210 \quad \textit{is the class containing the third quartile.}$$

$L_{Q_3} = 200$, $w = 10$

$N = 493$, $c = 351$, $f_{Q_3} = 49$

$$\Rightarrow Q_3 = L_{Q_3} + \frac{w}{f_{Q_3}}\left(\frac{3*N}{4} - c\right)$$

$$= 200 + \frac{10}{49}(369.75 - 351)$$

$$= \underline{\underline{203.83}}$$

b) D_7

- determine the class containing the 7^{th} decile.

$$\frac{7*N}{10} = 345.1$$

$$\Rightarrow 190 - 200 \quad \textit{is the class containing the seventh decile.}$$

$L_{D_7} = 190$, $w = 10$

$N = 493$, $c = 244$, $f_{D_7} = 107$

$$\Rightarrow D_7 = L_{D_7} + \frac{w}{f_{D_7}}(\frac{7*N}{10} - c)$$
$$= 190 + \frac{10}{107}(345.1 - 244)$$
$$= \underline{\underline{199.45}}$$

c) P_{90}

- determine the class containing the 90^{th} percentile.

$$\frac{90*N}{100} = 443.7$$

$$\Rightarrow 220 - 230 \ \ is\ the\ class\ containing\ the\ 90^{th}\ \ percentile.$$

$$L_{P_{90}} = 220, \qquad w = 10$$

$$N = 493, \qquad c = 434, \qquad f_{P_{90}} = 3107$$

$$\Rightarrow P_{90} = L_{P_{90}} + \frac{w}{f_{P_{90}}}(\frac{90*N}{100} - c)$$
$$= 220 + \frac{10}{31}(443.7 - 434)$$
$$= \underline{\underline{223.13}}$$

2.4 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.

2.4.1. 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) Variance, Standard deviation and coefficient of variation.

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.

$$R = L - S \qquad , L = l\arg est\ observation$$
$$S = smallest\ observation$$

Range for grouped data:

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

$$R = UCL_k - LCL_1, \qquad UCL_k \ \ is\ upperclass\lim it\ of\ the\ last\ class.$$
$$UCL_1 \ \ is\ lower\ class\lim it\ of\ the\ first\ class.$$

This is some times expressed as:

$$R = X_k - X_1, \qquad X_k \ \ is\ class\ mark\ of\ the\ last\ class.$$
$$X_1 \ \ is\ classmark\ of\ the\ first\ class.$$

Relative Range (RR)

It is also sometimes 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

Solution: (2)

$R = 4 \Rightarrow L - S = 4$ _____(1)

$RR = 0.25 \Rightarrow L + S = 16$ _____(2)

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

$L = 10 \ and \ S = 6$

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.

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

For the case of frequency distribution it is expressed as:

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

We usually use the following short cut formula.

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

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

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.

$$\textit{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

$$\textit{Sample standard deviation} = s = \sqrt{S^2}$$

The following steps are used to calculate the sample variance:

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).

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. $\overline{X} = 11$

X_i	5	10	12	17	Total
$(X_i - \overline{X})^2$	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

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}{n-1} = \frac{4400}{74} = 59.46.$$
$$\Rightarrow S = \sqrt{S^2} = \sqrt{59.46} = 7.71.$$

Special properties of Standard deviations

1. $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} < \sqrt{\frac{\sum(X_i - A)^2}{n-1}}$, $A \neq \bar{X}$
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. **Chebyshev's Theorem**

For any data set ,no matter what the pattern of variation, the proportion of the values that fall with in k standard deviations of the mean or $(\bar{X} - kS,\ \bar{X} + kS)$ will be at least $1 - \frac{1}{k^2}$, where k is a number greater than 1. i.e. the proportion of items falling beyond k standard deviations of the mean is at most $\frac{1}{k^2}$

Example: Suppose a distribution has mean 50 and standard deviation 6. What percent of the numbers are:

a) Between 38 and 62
b) Between 32 and 68
c) Less than 38 or more than 62.
d) Less than 32 or more than 68.

Solutions:

a) 38 and 62 are at equal distance from the mean,50 and this distance is 12

$\Rightarrow ks = 12$

$\Rightarrow k = \frac{12}{S} = \frac{12}{6} = 2$

➔ Applying the above theorem, at least $(1-\frac{1}{k^2})*100\% = 75\%$ of the numbers lie between 38 and 62.

b) Similarly done.

c) It is just the complement of a) i.e. at most $\frac{1}{k^2}*100\% = 25\%$ of the numbers lie less than 32 or more than 62.

d) Similarly done.

Exercise: The average score of a special test of knowledge of wood refinishing has a mean of 53 and standard deviation of 6. Find the range of values in which at least 75% the scores will lie.

4. If the standard deviation of $X_1, X_2, ...X_n$ *is* S, then the standard deviation of

a) $X_1 + k,\ X_2 + k, ...X_n + k$ *will alsobe* S

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

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

Exercise: Verify each of the above relationship, considering k and a as constants.

Examples:

1. The mean and standard deviation of n cement $X_1, X_2, ...X_n$ are known to be 12 gm and 3 gm respectively. New set of cement 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 cements.

2. The mean and the standard deviation of a set of numbers are respectively 500 and 10.
   a) If 10 are 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 $= |k|S = 5*10 = 50$

Coefficient of Variation (C.V)

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

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

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

Example: 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.

Exercise: 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?

2.5. 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

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

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

- 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.

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

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

➔ Student A performed better relative to his section because the score of student A is two standard deviations 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 deviations 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.