CHAPTER 2:
Basic Summary Statistics

- ***Measures of Central Tendency (or location)***

§ Mean – mode – median

- ***Measures of Dispersion (or Variation)***

§ Variance – standard deviation – coefficient of variation

2.1. Introduction:

For the population of interest, there is a population of values of the variable of interest.

Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown) of the variable of interest. The population size = N

Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known) The sample size = n

(i) A **parameter** is a measure (or number) obtained from the population values $X_1, X_2, \ldots, X_N$

(parameters are unknown in general)

A **statistic** is a measure (or number) obtained from the sample values $x_1, x_2, \ldots, x_n$

(statistics are known in general)

2.2. Measures of Central Tendency: (Location)

- The values of a variable often tend to be concentrated around the center of the data.

Some of these measures are: the **mean, mode, median**

- These measures are considered as representatives (or typical values) of data.

:

Mean:

(1) Population mean μ :

If $X_1, X_2, \ldots, X_N$ are the population values of the variable of interest , then the population mean is:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{(unit)}$$

- The population mean μ is a parameter (it is usually unknown)

(2) Sample mean $\bar{x}$:

If $x_1, x_2, \ldots, x_n$ are the sample values, then the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(unit)

The sample mean $\bar{x}$ is a statistic (it is known)

The sample mean $\bar{x}$ is used to approximate (estimate) the population mean μ.

Example:

Consider the following population values:

$$X_1 = 30,\ X_2 = 22,\ X_3 = 35,\ X_4 = 27,\ X_5 = 41.$$

Suppose that the sample values obtained are:

$$x_1 = 30,\ x_2 = 35,\ x_3 = 27.$$

Then:

$$\mu = \frac{30+22+35+27+41}{5} = \frac{155}{5} = 31 \quad \text{(unit)}$$

$$\bar{x} = \frac{30 + 35 + 27}{3} = \frac{92}{3} = 30.67 \quad \text{(unit)}$$

Notes:

- The mean is simple to calculate.
- There is only one mean for a given sample data.
- The mean can be distorted by extreme values.
- The mean can only be found for quantitative variables

Median:

The median of a finite set of numbers is that value which divides the **ordered** set into two equal parts.

Let $x_1, x_2, \ldots, x_n$ be the sample values . We have **two cases**:

(1) If the sample size, *n*, is **odd**:

- The median is the **middle value** of the **ordered** observations.

The middle observation is the ordered $\frac{n+1}{2}$ observation

The median = The $\frac{n+1}{2}$ order observation.

Ordered set → (smallest to largest)

Rank (or order) →

*	*	…	Middle value= MEDIAN	…	*
1	2	…	$\frac{n+1}{2}$	…	*n*

Example:

Find the median for the sample values: 10, 54, 21, 38, 53.

Solution:

$n = 5$ (odd number)

The rank of the middle value (median) = $\frac{n+1}{2}$ = (5+1)/2 = 3

Ordered set →	10	21	**38**	53	54
Rank (or order) →	1	2	$\frac{n+1}{2}$ = 3	4	5

The median =38 (unit)

(2) If the sample size, n, is **even**:

The median is the mean (average) of the two middle values of the **ordered** observations.

The middle two values are the ordered $\frac{n}{2}$ and $\frac{n}{2}+1$ observations.

- The median =

$$\frac{1}{2}\left\{\left(\frac{n}{2}\right)th \text{ ordered observations} + \left(\frac{n}{2}+1\right)th \text{ ordered observations}\right\}$$

Ordered set →

Rank (or order) →

*	*	…	Middle value	Middle value	…	*
1	2	…	$\frac{n}{2}$	$\frac{n}{2}+1$	…	n

Example:

Find the median for the sample values: 10, 35, 41, 16, 20, 32

Solution:

.$n = 6$ (even number)

The rank of the middle values are

$\frac{n}{2} = 6 / 2 = 3$

$\frac{n}{2} + 1 = (6 / 2) + 1 = 4$

Ordered set →	10	16	**20**	**32**	35	41
Rank (or order) →	1	2	**3**	**4**	5	6

The median $= \frac{20+32}{2} = \frac{52}{2} = 26$ (unit)

Note:

The median is simple to calculate.

There is only one median for given data.

The median is not affected too much by extreme values.

The median can only be found for quantitative variables

Mode:

The mode of a set of values is that value which occurs with the highest frequency.

- ·If all values are different or have the same frequency, there is no mode.
- · A set of data may have more than one mode.

Example:

Data set	Mode(s)
26, 25, 25, 34	25 (unit)
3, 7, 12, 6, 19	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	No mode
3, 3, 12, 6, 8, 8	3 and 8 (unit)

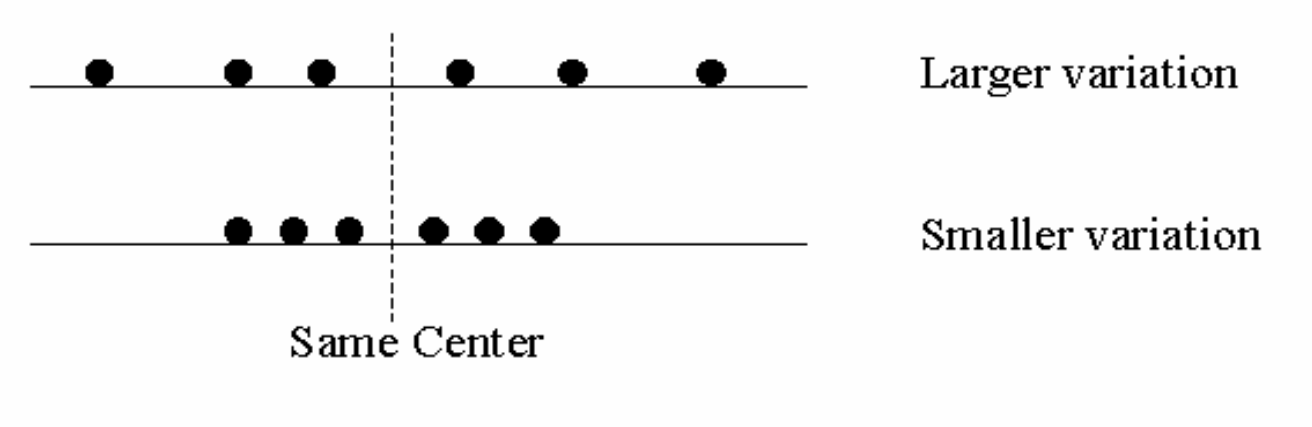
Note:

- The mode is simple to calculate but it is not “good”.
- The mode is not affected too much by extreme values.
- The mode may be found for both quantitative and qualitative variables.

2.3. Measures of Dispersion (Variation):

The variation or dispersion in a set of values refers to how spread out the values are from each other.

- The variation is small when the values are close together.
- There is no variation if the values are the same.



Smaller variation

Larger variation

Some measures of dispersion:

Range – Variance – Standard deviation

Coefficient of variation

Range:

Range is the difference between the largest (Max) and smallest (Min) values.

Range = Max – Min

Example:

Find the range for the sample values: 26, 25, 35, 27, 29, 29.

Solution:

Range = 35 – 25 = 10 (unit)

Note:

The range is **not useful** as a measure of the variation since it only takes into account two of the values. (it is not good)

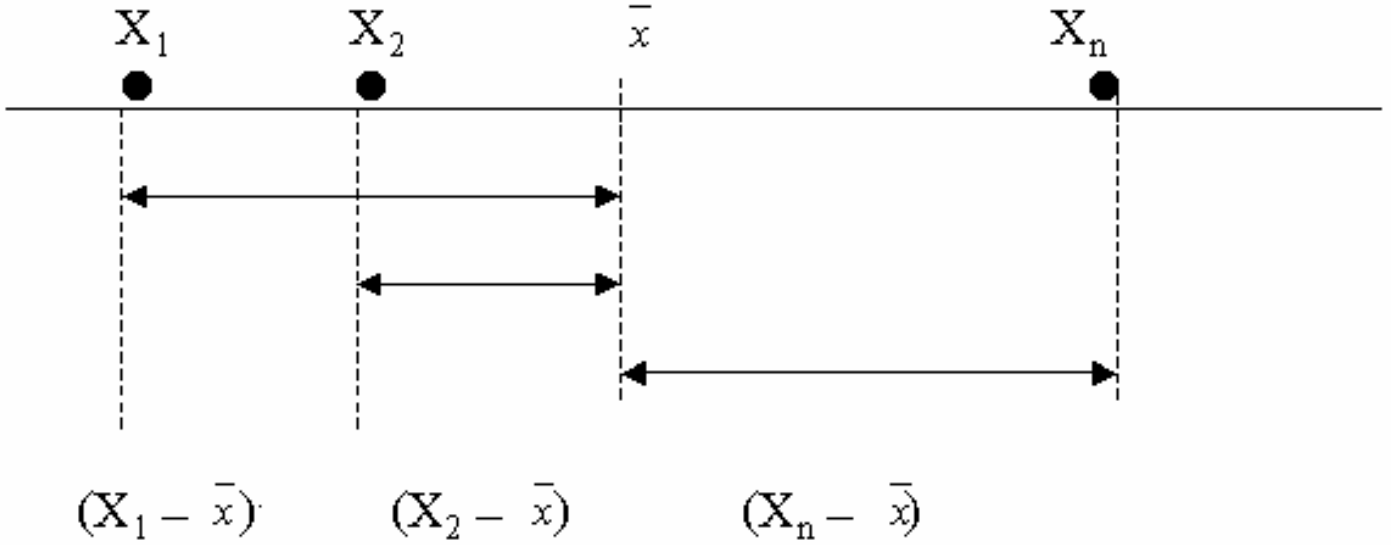
Variance:

The variance is a measure that uses the mean as a point of reference.

The variance is small when all values are close to the mean. The variance is large when all values are spread out from the mean.

deviations from the mean:

Deviations from the mean:



(1) Population variance:

Let $X_1, X_2, \ldots, X_N$ be the population values.

The population variance is defined by

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_N - \mu)^2}{N} \qquad (\text{unit})^2$$

where $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ is the population mean

Notes:

- σ^2 is a parameter because it is obtained from the population values (it is unknown in general).
- $\sigma^2 \geq 0$

(2) Sample Variance:

Let $x_1, x_2, \ldots, x_n$ be the sample values.
The sample variance is defined by:

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \cdots + \left(x_N - \bar{x}\right)^2}{n-1} \qquad (\text{unit})^2$$

where $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean.

Notes:

- S^2 is a statistic because it is obtained from the sample values (it is known).
- S^2 is used to approximate (estimate) σ^2 .
- $S^2 \geq 0$

Example:

We want to compute the sample variance of the following sample values: 10, 21, 33, 53, 54.

Solution: n=5

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2 \quad \text{(unit)}$$

$$\therefore S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{5} (x_i - 34.2)^2}{5-1}$$

$$S^2 = \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad \text{(unit)}^2$$

Another method:

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$
10	-24.2	585.64
21	-13.2	174.24
33	-1.2	1.44
53	18.8	353.44
54	19.8	392.04
$\sum_{i=1}^{5} x_i = 171$	$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 1506.8$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5}$$

$$= \frac{171}{5} = 34.2$$

$$S^2 = \frac{1506.8}{4}$$

$$= 376.7$$

Calculating Formula for S^2:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

* Simple

* More accurate

Note:

To calculate S^2 we need:

- n = sample size

$\sum x_i$ = The sum of the values

$\sum x_i^2$ = The sum of the squared values

For the above example:

x_i	10	21	33	53	54	$\sum x_i = 171$
x_i^2	100	441	1089	2809	2916	$\sum x_i^2 = 7355$

$$S^2 = \frac{7355 - (5)(34.2)^2}{5-1} = \frac{1506.8}{4} = 376.7$$

Standard Deviation:

- The standard deviation is another measure of variation.
- It is the **square root** of the variance.

(1) Population standard deviation is: $\sigma = \sqrt{\sigma^2}$ (unit)

(2) Sample standard deviation is: $S = \sqrt{S^2}$ (unit)

Coefficient of Variation (C.V.):

- The variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population (or sample).
- If we want to compare the variation of two variables we cannot use the variance or the standard deviation because:

1. **The variables might have different units**.
2. **The variables might have different means.**

· We need a measure of the **relative variation** that will not depend on either the units or on how large the values are. This measure is the coefficient of variation (C.V.) which is defined by:

$$C.V. = \frac{S}{\bar{x}} * 100\%$$ (free of unit or unit less)

	Mean	St.dev.	C.V.
1st data set	$\bar{x}_1$	S_1	$C.V_1 = \frac{S_1}{\bar{x}_1} 100\ \%$
2nd data set	$\bar{x}_2$	S_2	$C.V_2 = \frac{S_2}{\bar{x}_2} 100\%$

· The relative variability in the 1st data set is larger than the relative variability in the 2nd data set if $C.V_1 > C.V_2$ (and vice versa).

Example:

1st data set: $\bar{x}_1 = 66$ kg, $S_2 = 4.5$ kg

$$\Rightarrow C.V_1 = \frac{4.5}{66} * 100\% = 6.8\%$$

2nd data set: $\bar{x}_2 = 36$ kg, $S_2 = 4.5$ kg

$$\Rightarrow C.V_2 = \frac{4.5}{36} * 100\% = 12.5\%$$

Since $C.V_1 < C.V_2$, the relative variability in the 2nd data set is larger than the relative variability in the 1st data set.

Notes: (Some properties of $\bar{x}$, S, and S^2:

Sample values are : $x_1, x_2, \ldots, x_n$

a and *b* are constants

Sample Data	Sample mean	Sample st.dev	Sample Variance
$x_1, x_2, \ldots, x_n$	$\bar{x}$	S	S^2
$ax_1, ax_2, \ldots, ax_n$	$a\bar{x}$	$\lvert a \rvert S$	a^2S^2
$x_1 + b, \ldots, x_n + b$	$\bar{x} + b$	S	S^2
$ax_1 + b, \ldots, ax_n + b$	$a\bar{x} + b$	$\lvert a \rvert S$	a^2S^2

Absolute value:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0 \end{cases}$$

Example:

	Sample	Sample mean	Sample St..dev.	Sample Variance
	1,3,5	3	2	4
(1) (2) (3)	-2, -6, -10 11, 13, 15 8, 4, 0	-6 13 4	4 2 4	16 4 16

Data (1) $-2x_1, -2x_2, -2x_3$ (a = –2)

(2) $x_1 + 10, x_2 + 10, x_3 + 10$ (b = 10)

(3) $-2x_1 + 10, -2x_2 + 10, -2x_3 + 10$ (a = –2, b = 10)