Descriptive Statistics

The following data represent marks of three students' groups and each group with different teacher, find the mean of each group:

A: 59, 61, 62, 58, 60

B: 50, 60, 66, 54, 70

C: 19, 65, 46, 78, 72

$\overline{A} = 60$

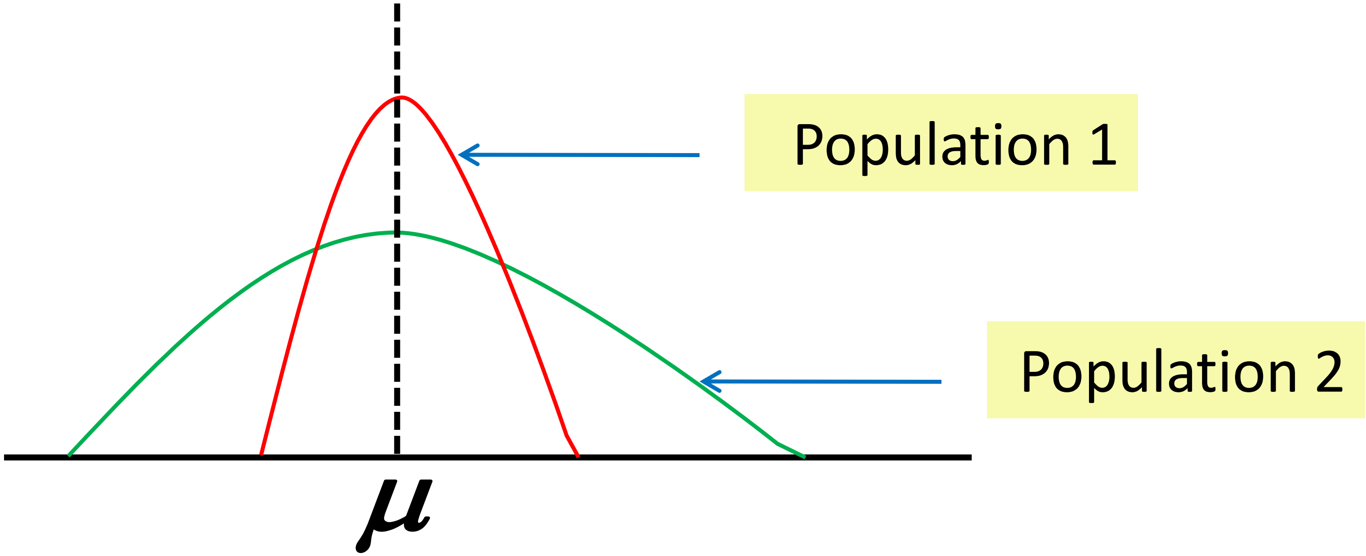
$\overline{B} = 60$

$\overline{C} = 60$

Definition of Measures of variation

The **measure of variation** in a set of observations refers to how spread out the observations are from each other.

When the measure of variation is small, this means that the values are close together (but not the same)



Population 1

Population 2

μ

fourth lecture

We will learn in this lecture:

1- Range

2- Deviation

3- Variance and Standard Deviation

4-Coefficient of variance

5- The standard score

6-Coefficient of skewness

Measures of variation

First:
The Range

Definition of range:

The **range** of a data set is the difference between the maximum and the minimum data entries in the set.

R = maximum data entry – minimum data entry

Example (1):

Calculate the range of marks of the students:
82, 40, 62, 70, 30, 80

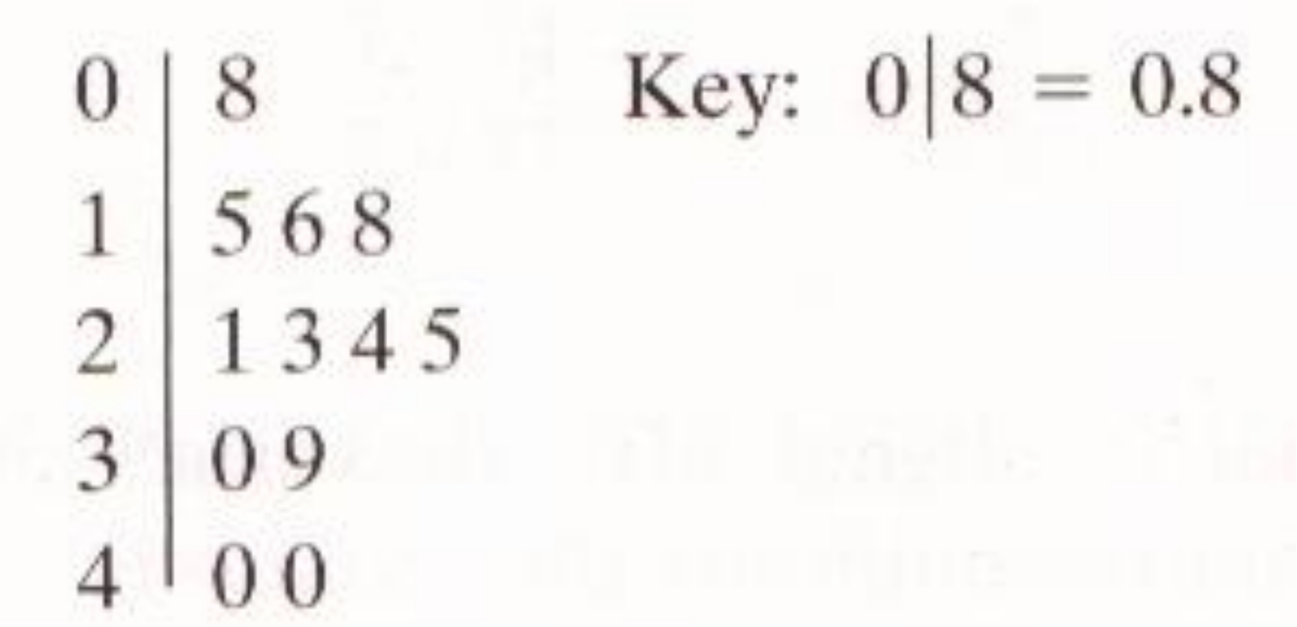
Solution:

$$R = \text{maximum data entry} - \text{minimum data entry}$$

$$R = 82 - 30 = 52$$

Example (2):

Find the range of the data set represented by the steam-and-leaf plot:



0	8
1	5 6 8
2	1 3 4 5
3	0 9
4	0 0

Key: 0|8 = 0.8

Solution: R = maximum data entry – minimum data entry

$$R = 4 - 0.8 = \boxed{3.2}$$

Advantages of the range :

- It’s easy to calculate
- It gives a quick idea about the nature of data, often used in quality control and describe the weather.

Disadvantages of the range :

- It uses only two entries from the data set.
- Affected by extreme values .therefore it’ approximate measurement.

Second:

Deviation

Definition:

The deviation of an entry x in a population data set is the difference between the entry and the mean μ of
the data set

$$\text{Deviation of } x = x - \mu$$

Example(3):

Find the deviation of each starting salary for corporation A:

Starting salaries for corporation A(1000s of dollars)

salary	**41**	**38**	**39**	**45**	**47**	**41**	**44**	**41**	**37**	**42**

Solution:

Deviation of $x = x - \mu$

$$\mu = \sum X/n$$

Salary(x)	**41**	**38**	**39**	**45**	**47**	**41**	**44**	**41**	**37**	**42**	$\sum X = 415$

$$\mu = \sum X/n$$

$$= 415/10 = 41.5$$

Salary(x)	**Deviation** **X-μ**
41	-.5
38	-3.2
39	-2.5
45	3.5
47	5.5
41	-.5
44	2.5
41	-.5
37	-4.5
42	0.5
∑X=415	∑ (X-μ) =0

Third : Variance

Definition:

The population variance of a population data set of N entries is

population variance= $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Definition:

The population standard deviation of a population data set of N entries is the square root of the population variance

population standard deviation =

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Example(4):

Find the population variance and standard deviation of the starting salary for corporation A:

Starting salaries for corporation A(1000s of dollars)

salary	**41**	**38**	**39**	**45**	**47**	**41**	**44**	**41**	**37**	**42**

Solution:

Salary(x)	Deviation X-µ	Squares $(X-\mu)^2$
41	-0.5	0.25
38	-3.2	12.25
39	-2.5	6.25
45	3.5	12.25
47	5.5	30.25
41	-0.5	0.25
44	2.5	6.25
41	-0.5	0.25
37	-4.5	20.25
42	0.5	0.25
∑X=415	∑ (X-µ) =0	SSx =88.5

$$SSx = 88.5$$

$$N = 10$$

$$\sigma^2 = \frac{88.5}{10} \approx 8.9$$

$$\sigma = \sqrt{8.85} \approx 3.0$$

Definition:

The sample variance and sample standard deviation of a sample data set of n entries are listed below.

Sample variance= $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

Remark

If the sample size (n) is large (greater than 30) then σ^2, s^2 are equal approximately.

Example (5):

Calculate the standard deviation of the following sample: (8,9,7,6,5)?

Solution:

1-Calculating the average: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \frac{35}{5} \quad = 7$

2- Calculating the variance: $\boldsymbol{s}^2 = \frac{1}{\boldsymbol{n}-1}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}(\boldsymbol{x}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
8	1	1
9	2	4
7	0	0
6	-1	1
5	-2	4
Total	0	10

$$\boldsymbol{s}^2 = \frac{1}{\boldsymbol{n}-1}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}(\boldsymbol{x}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2 = \frac{10}{5-1} = 2.5$$

3- Calculating the standard deviation:

$$s = \sqrt{s^2} = \sqrt{2.5} = 1.591$$

We can use this formula to calculate the sample variance:

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

Example (6):

Calculate the standard deviation of the following sample: (8,9,7,6,5)?

Solution:

x	x^2
8	64
9	81
7	49
6	36
5	25
35	225

The variance is:

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$s^2 = \frac{1}{5-1}\left[255 - \frac{(35)^2}{5}\right]$$

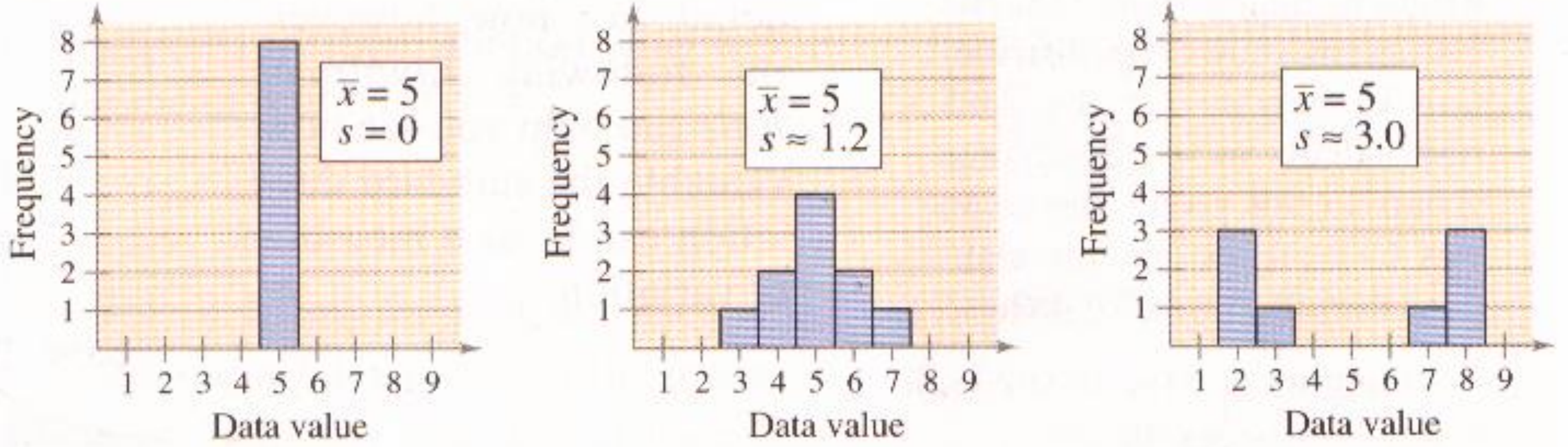
$$= \frac{1}{4}(255-245) \quad = \frac{10}{4} = 2.5$$

The standard deviation is:

$$s = \sqrt{2.5} = 1.581$$

Interpreting Standard Deviation:

when interpreting the standard deviation , remember that it is a measure of the typical amount an entry deviates from the mean. The more the entries are spread out, the greater the standard deviation



Standard Deviation for Grouped Data:

You learned that large data sets are usually best represented by a frequency distribution . The formula for the sample standard deviation for a frequency distribution is :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i$$

Where n=∑f is the number of entries in the data set.

Example (7):
the following data represent the number of children in 50 households:

1	3	1	1	1	1	2	2	1	0
1	1	0	0	0	1	5	0	3	6
3	0	3	1	1	1	1	6	0	1
3	6	6	1	2	2	3	0	1	1
4	1	1	2	2	0	3	0	2	4

x	f	$x\ f$	$x - \bar{x}$	$\left(x - \bar{x}\right)^2$	$\left(x - \bar{x}\right)^2 f$
0	10	0	-1.8	3.24	32.40
1	19	19	-0.8	0.64	12.16
2	7	14	0.2	0.04	0.28
3	7	21	1.2	1.44	10.08
4	2	8	2.2	4.84	9.68
5	1	5	3.2	10.24	10.24
6	4	24	4.2	17.64	70.56
Total	**∑f=50**	**∑fx=91**	**------**	---------	145.40

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

$$= \frac{91}{50} = 1.8$$

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}$$

$$s^2 = \frac{145.4}{49} = 2.97$$

$$s = \sqrt{s^2} \qquad = \sqrt{\frac{145.4}{49}} \approx 1.7$$

Examlpe (8) :

Calculate the standard deviation of the scores of students following

class	f_i	t.limit	m.point x_i
40-49	2	39.5-49.5	44.5
50-59	9	49.5-59.5	54.5
60-69	15	59.5-69.5	64.5
70-79	11	69.5-79.5	74.5
80-89	2	79.5-89.5	84.5
90-99	1	89.5-99.5	94.5

$$s^2 = \frac{1}{n-1}\left[\sum x^2 f - \frac{\left(\sum xf\right)^2}{n}\right]$$

class	f_i	t.limit	x_i	$x_i f_i$	$x_i^2 f_i$
40-49	2	39.5-49.5	44.5	89	3960.5
50-59	9	49.5-59.5	54.5	490.5	26732.25
60-69	15	59.5-69.5	64.5	960.5	62403.75
70-79	11	69.5-79.5	74.5	819.5	61052.75
80-89	2	79.5-89.5	84.5	169	12280.5
90-99	1	89.5-99.5	94.5	94.5	8930.25
Total	40	--------	-----	2630	177360

$$s^2 = \frac{1}{n-1}\left[\sum x^2 f - \frac{\left(\sum xf\right)^2}{n}\right]$$

$$s^2 = \frac{1}{40-1}\left(177360 - \frac{2630^2}{40}\right) = 113.78$$

$$s = \sqrt{s^2} = \sqrt{113.78} = 10.67$$

Remark:

The standard deviation is always **positive**.

Remark:

Example (9): Calculate the standard deviation of the following sample: 8, 8, 8, 8, 8

x	x^2
8	64
8	64
8	64
8	64
8	64
40	320

The variance is:

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$s^2 = \frac{1}{5-1}\left[320 - \frac{(40)^2}{5}\right] = 0$$

The standard deviation is:

$$s = 0$$

If data are equal then the standard deviation is 0.

Example (10):

Find the date such that:

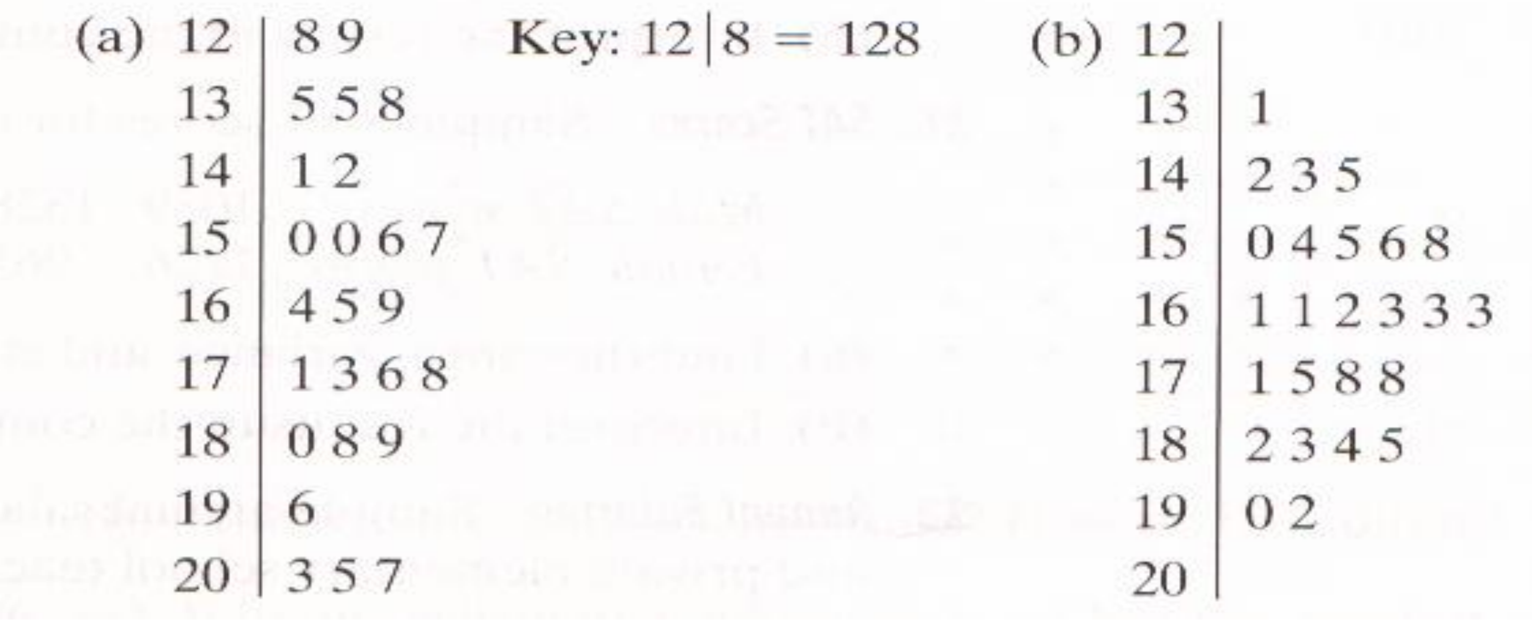
$$\bar{x} = 7 \qquad s = 0 \qquad n = 5$$

Solution:

7, 7, 7, 7, 7

Example (11):

Both data sets have a mean of 165.One has a standard deviation of 16,and the other has a standard deviation of 24.Which is which ? Explain your reasoning?



Key: 12|8 = 128

(a)

12	8 9
13	5 5 8
14	1 2
15	0 0 6 7
16	4 5 9
17	1 3 6 8
18	0 8 9
19	6
20	3 5 7

(b)

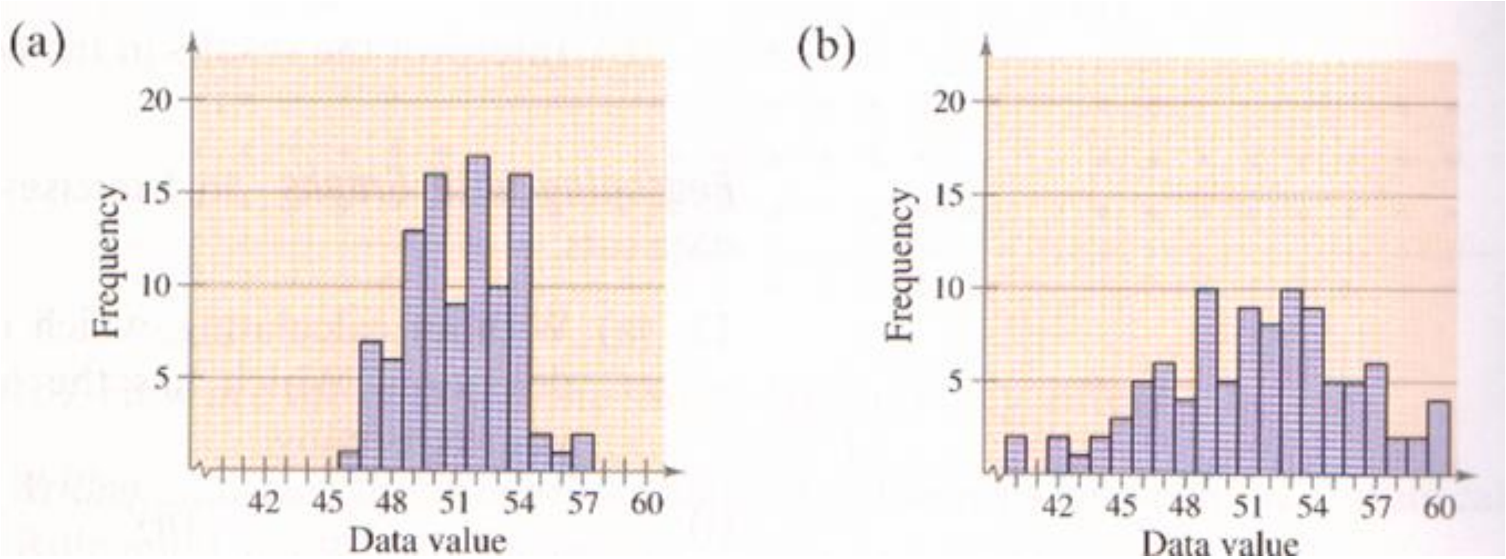
12	
13	1
14	2 3 5
15	0 4 5 6 8
16	1 1 2 3 3 3
17	1 5 8 8
18	2 3 4 5
19	0 2
20	

Solution:

(a) has a standard deviation of 24 and (b) has a standard deviation of 16, because the data in (a) have more variability.

Example (12):

Both data sets have represented below have a mean of 50.One has a standard deviation of 2.4,and the other has a standard deviation of 5.Which is which ? Explain your reasoning?

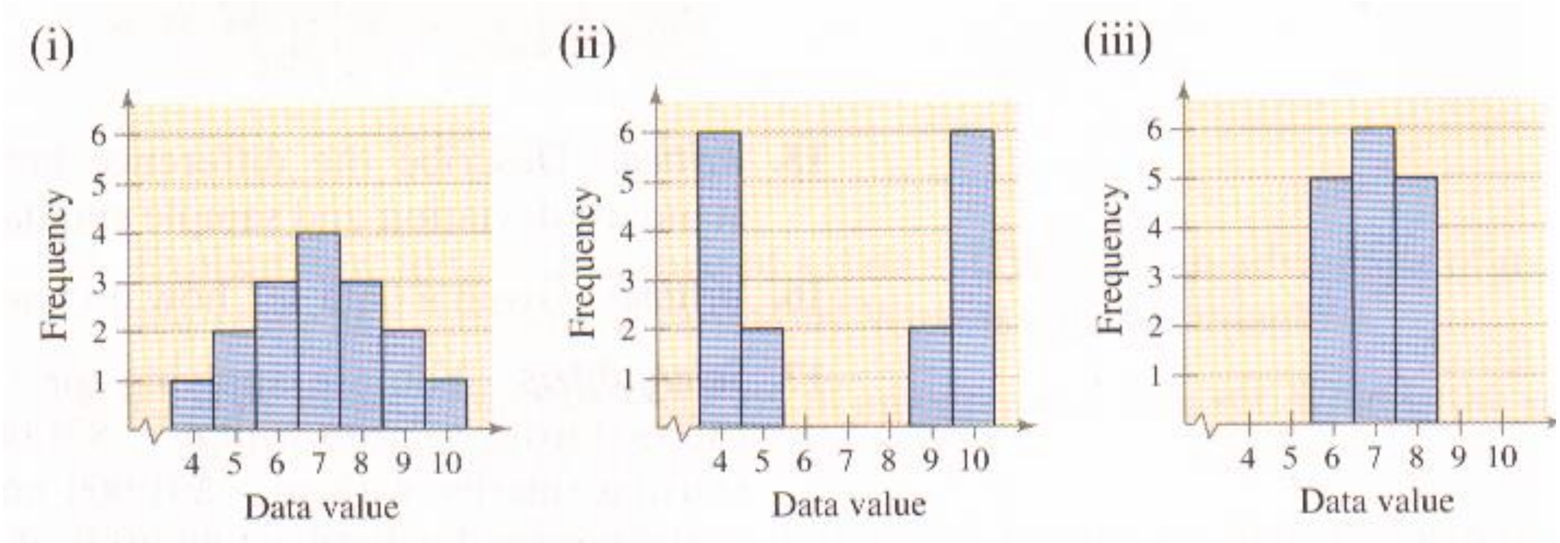


Solution:

(a) has a standard deviation of 2.4 and (b) has a standard deviation of 5, because the data in (a) have less variability.

Example (13):(a)Without calculating, which data set has the greatest sample standard deviation? Which has the least sample standard deviation ? Explain your reasoning.

(b)How are the data sets the same ? How do they differ?



Solution:

(a) Data set (ii) has more entries that are farther away from the mean but data set (iii) has more entries that are close to the mean.

(b) The three data sets have the same mean but have different standard deviations.

Example (14):

Let the mean of 4 students’ mark is 5 with standard deviation is 5 and the mean of 6 students’ mark is 5 with standard deviation is 3.5. Find the pooled two-sample variance s_p^2.

Solution:

$$\begin{array}{ll} n_2 = 6 & n_1 = 4 \\ \bar{x}_2 = 5 & \bar{x}_1 = 5 \\ s_2 = 3.5 & s_1 = 3 \end{array}$$

$$\boldsymbol{s_p^2} = \frac{(\boldsymbol{n}_1 - 1)\boldsymbol{s}_1^2 + (\boldsymbol{n}_2 - 1)\boldsymbol{s}_2^2}{\boldsymbol{n}_1 + \boldsymbol{n}_2 - 1}$$

$$\boldsymbol{s_p^2} = \frac{(4-1)(3)^2 + (6-1)(3.5)^2}{4+6-1} = 2.944$$

Advantages and disadvantages of the standard deviation :

It is as like as the mean.

Fourth:

Coefficient of variance

Reasons to use Coefficient of variance (the relative variation) rather than the measure of variation:

- The two variables involved might by measured in different units.
- The means of the two may quit different in size.

Definition of Coefficient of variance:

The Coefficient of variance C.V. describes the standard deviation as a percent of the mean.

$$C.V. = \frac{s}{\bar{x}}$$

You can use the Coefficient of variance to compare data with different units

Example (15) :

Calculate the coefficient of variance for student’s marks such that:

$$s = 10.67 \qquad \bar{x} = 65.75$$

Solution:

$$C.V. = \frac{s}{\bar{x}} = \frac{10.67}{65.75} = 0.167$$

Fifth

The standard score

Definition of The standard score:

The **standard score**, or **z-score**, represent the number or standard deviation a given value x falls from the mean μ . To find the z-score for a given value, use the following formula.

$$z = \frac{value - mean}{standerd\ deviation} = \frac{x - \mu}{\sigma}$$

Remark:

A z-score is used to compare data values within the same data set or to compare data values from different data set.

Example (16):

For the statistics test scores, the mean is 75 and the standard deviation is 10. And for mathematics test scores, the mean is 81 and the standard deviation is 16. If a student gets a 82 on the statistics test and a 89 on the mathematics test, determine on which test the student had a better tests :

Solution:

$$z = \frac{x-\mu}{\sigma}$$

z-score of statistics test

$$z_1 = \frac{82-75}{10} = 0.7$$

z-score of mathematics test

$$z_2 = \frac{89-81}{16} = 0.5$$

The student did better on statistics test.

Remark:

A z-score can be negative, positive, or zero.

If z is negative → $x < \mu$

If z is zero → $x = \mu$

If z is positive → $x > \mu$

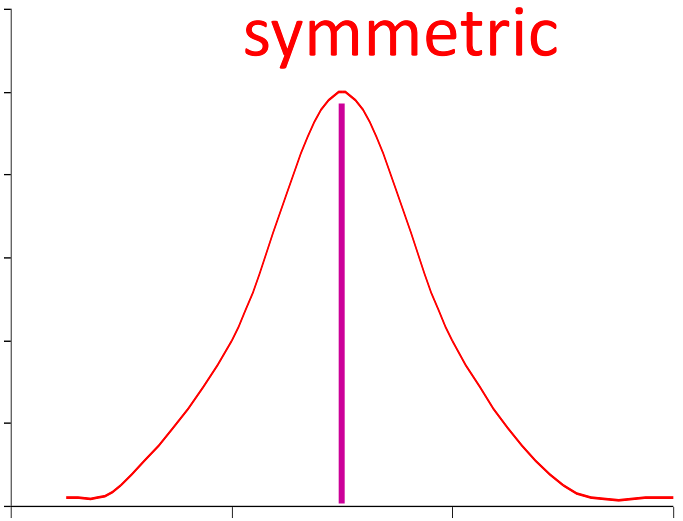
Sixth: Coefficient of skewness

Definition of The Coefficient of skewness:

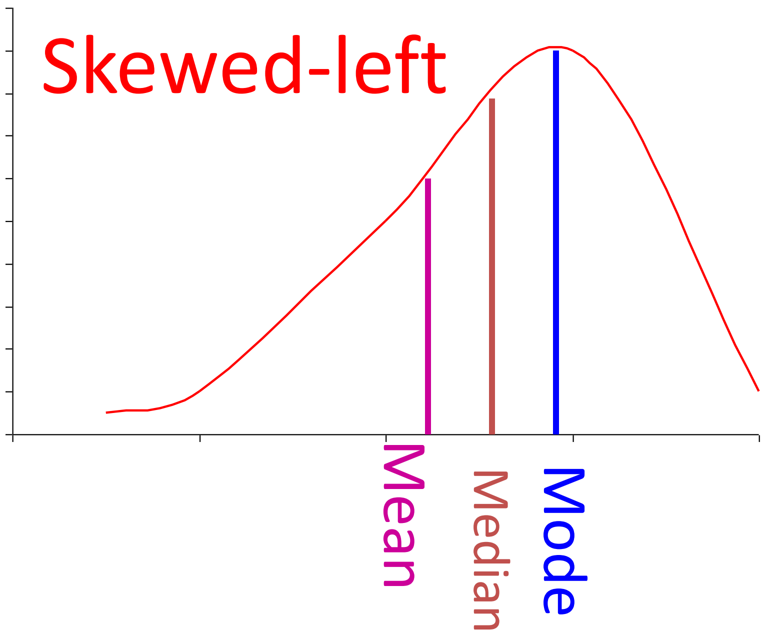
It is the measurement to describes the shape of data.

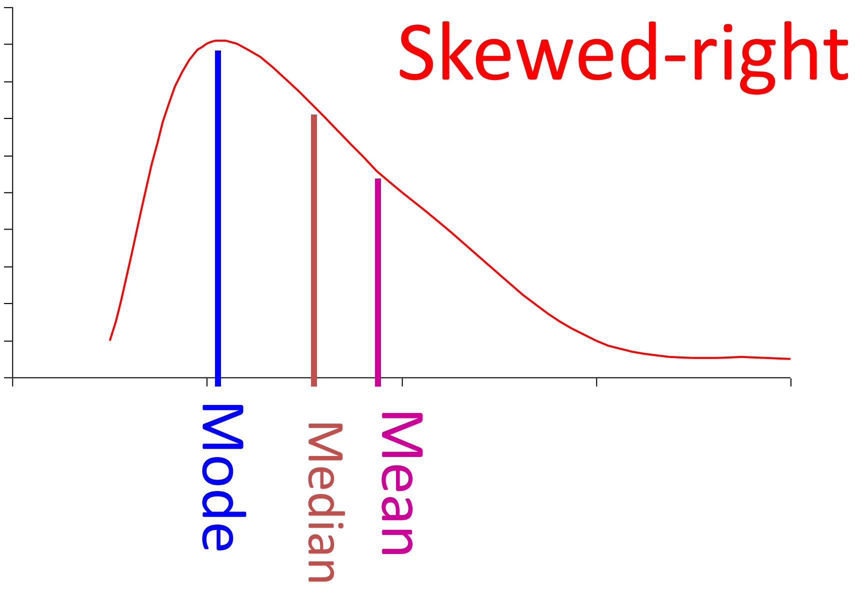
Types of skewness:

- **Skewd left**: a distribution is skewed left (negatively skewed) if its tail extends to the left.
- **Skewd right**: a distribution is skewed right (positively skewed) if its tail extends to the right.
- **Symmetric**: a distribution on one side of the mean is a mirror image of the other side.



Mode =Mean= Median





Remark:

If a distribution is skewed-left → $v < 0$

If a distribution is symmetric → $v = 0$

If a distribution is skewed-right → $v > 0$