Chapter Three

Data Description

1. Measures of Central Tendency
2. Measures of Variation
3. Measures of Position
4. Exploratory Data Analysis

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Three
Data Description

1. *Measures of Central Tendency*

- Measures of average are called *measures of central tendency* and include several measurements such as: *mean, median, mode, midrange, etc. Two concepts must be defined:*

1. **Statistic** is a characteristic or measure obtained by using the data values from a sample.
2. **Parameter** is a characteristic or measure obtained by using all the data values from a specific population.

1.1. The Mean

The *mean,* also is known as the *arithmetic average*

The **mean** is the sum of the values, divided by the total number of values. The symbol $\bar{X}$ represents the sample mean.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \ldots\ldots\ldots + X_n}{n}$$ **Raw data**

$$\bar{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{X_1 f_1 + X_2 f_2 + \ldots\ldots\ldots + f_n X_n}{\sum f n}$$ **Tabulated data**

- where n represents the total number of values in the sample, X_i is the statistic and f_i is the frequency.
- For a population, the Greek letter (μ) is used for the mean and N the represents the total number of values in the population.

Example 1: The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean. 110 76 29 38 105 31

Solution:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{110+76+29+38+105+31}{6} = 64.8$$

Example 2: The data represent the number of miles run during one week for a sample of 20 runners. Find the mean.

Classes	5.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5	35.5-40.5
Frequency	1	2	3	5	4	3	2

Solution

1. **Make a table as shown.**
2. **Find the midpoints of each class and enter them in column C.**

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13$$

3. **For each class, multiply the frequency by the midpoint:**
   $f_1 \times X_1 = 1 \times 8 = 8$, $f_2 \times X_2 = 2 \times 13 = 26$ etc.

Classes	Frequency	Midpoint X_m	$f_i \times X_m$
5.5-10.5	1	8	8
10.5-15.5	2	13	26
15.5-20.5	3	18	54
20.5-25.5	5	23	115
25.5-30.5	4	28	112
30.5-35.5	3	33	99
35.5-40.5	2	38	76

4. **Find the sum of $\sum X_i f_i$**

$$\sum f_i = 20 \qquad \sum X_i f_i = 490$$

5. **Divide the sum by $\sum f_i$ to get the mean.**

$$\overline{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{490}{20} = 24.5\ mile.$$

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

1.2. The Median

- The ***median*** is the halfway point in a data set. Before you can find this point, the data must be arranged in order. When the data set is ordered, it is called a **data array.**

- **Steps in computing the median of a data array**

(1) Arrange the data in order. (2) Select the middle point.

Note: (Raw data)

- **For odd number of values in the data set; the median was an actual data value.** $MD = \frac{n}{2}$
- **When there are an even number of values in the data set, the median will fall between two given values (average of the two values)** $MD = \frac{\frac{n}{2}+(\frac{n}{2}+1)}{2}$

Example 3: The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median. 684, 764, 656, 702, 856, 1133, 1132, 1303

Solution: (Odd values) (Arrange values)

656, 684, 702, 764, 856, 1132, 1133, 1303

n/2 ↑ (n/2)+1

Median

MD = (764 + 856)/2 = 810

Example 4: The number of children with asthma during a specific year in seven local districts is shown. Find the median. 253, 125, 328, 417, 201, 70, 90

Solution: (Even values)

70, 90, 125, 201, 253, 328, 417

↑

Median (n/2) = 201

Tabulated Data

1. Determine ascending C.F.
2. Determine the order of median ($\frac{\sum f_i}{2} = \frac{20}{2} = 10$).
3. Determine the class of median (between 20.5 to 30.5). (where **10** is located between **6 and 11**.
4. Determine MD using the next Eq.

$$MD = L_1 + \left\{\frac{\left[\left(\frac{\sum f_i}{2}\right) - F_i\right]}{f_M}\right\} \times W$$

W = Class width = 25.5-20.5 = 5

L_1 = Lower boundary limits of MD = 20.5

F_i = C.F. before the MD class = 3

f_M = median class frequency = 11-6= 5

Classes	Frequency	Midpoint X_m	*Ascending C.F.*
5.5-10.5	1	8	0
10.5-15.5	2	13	1
15.5-20.5	3	18	3
20.5-25.5	5	23	6
25.5-30.5	4	28	11
30.5-35.5	3	33	15
35.5-40.5	2	38	18
			20

Median class

$$MD = 20.5 + \left\{\frac{\left[\left(\frac{20}{2}\right) - 3\right]}{5}\right\} \times 5$$

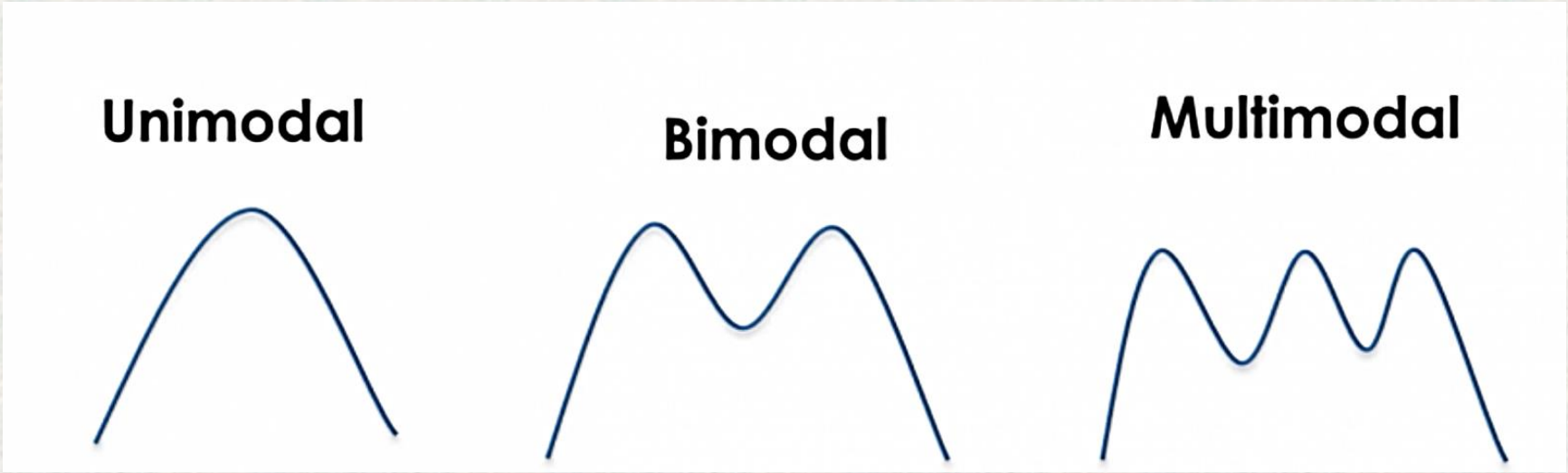
MD= 27.5

1.3. The Mode

The value that occurs most often in a data set is called the **mode.** For example, the set of data (3, 5, 8, 9, 5, 2, 5, 7, 5) **the mode is (5).**

A data set that has only :

- one value that occurs with the greatest frequency is called **unimodal.**
- two values with the same greatest frequency **bimodal.**
- more than two values with the same greatest frequency **multimodal.**



Example 4: The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

104 104 104 104 104
107 109 104 109 110
103 111 112 111 109

Solution

Since the values 104 occurred 6 times, the modes is 104. The data set is said to be unimodal.

Example:5 Find the mode for the number of branches that six banks have.

401, 344, 209, 201, 227, 353

Solution

Since each value occurs only once, there is no mode.

Note: Do not say that the mode is zero. That would be incorrect, because in some data, such as temperature, zero can be an actual value.

Tabulated Data

For tabulated data, mode can be calculated using the following formula:

$$M_o = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times W$$

Where:

$\mathbf{L_1}$ is the lower boundary limits For mode's class.

$\mathbf{d_1}$ = the deference between mode's class and the previous class.

$\mathbf{d_2}$ = the deference between mode's class and the next class.

W is the class width.

Example 6: The data represent the spot speed of a passenger cars in (km/hr) passing with a section of road. Find the mode.

Solution:

- Determine the mode's class (50.5-550.5) where it is the highest frequency.
- $\mathbf{L_1}$ = is the lower boundary limits For mode's class (50.5).
- $\mathbf{d_1}$ = the deference between mode's class and the previous class (20-12 = 8).
- $\mathbf{d_2}$ = the deference between mode's class and the next class (20 -17 = 3).

W is the class width (5).

Boundary limits	fi	Class midpoint
30.5-35.5	1	33
35.5-40.5	5	38
40.5-45.5	7	43
45.5-50.5	12	48
50.5-55.5	20	53
55.5-60.5	17	58
60.5-65.5	14	63
65.5-70.5	10	68
70.5-75.5	7	73
75.5-80.5	4	78
80.5-85.5	2	83
85.5-90.5	1	88

$$M_o = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times W$$

$$M_o = 50.5 + \left(\frac{8}{8+3}\right) \times 5 = 54.14 \text{ km/hr}$$

2. Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean is affected by extremely high or low values,

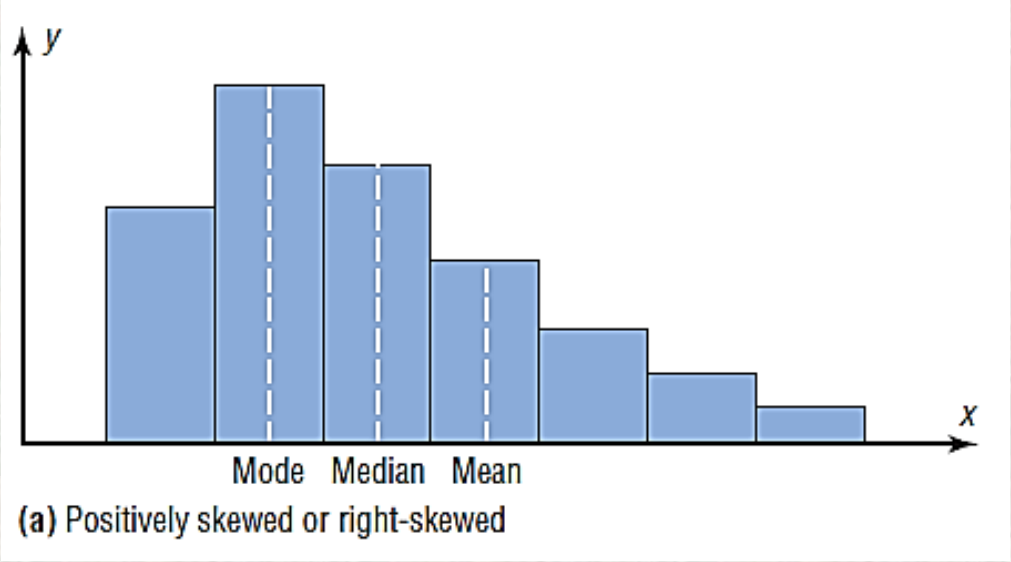
The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is affected less than the mean by extremely high or extremely low values.

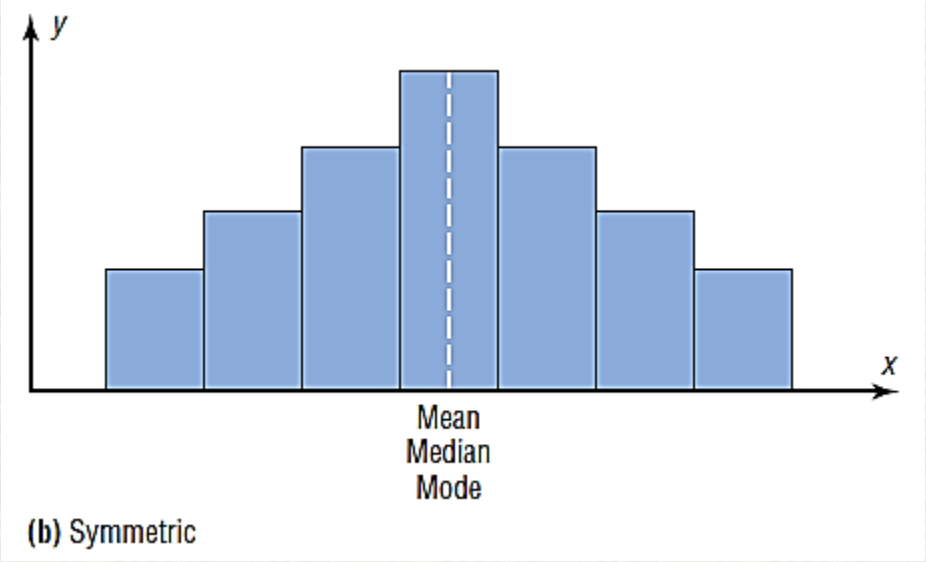
The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

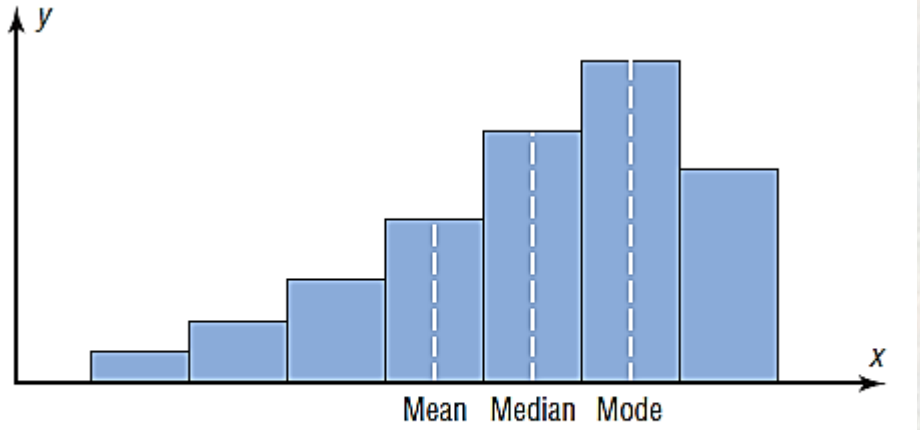
3. Distribution Shapes



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed

2. Measures of Variation

2.1. Population Variance and Standard Deviation

Xi = 80, 85, 90, 98, 104, 115, 122, 130
($\bar{X}$ = 103)
Yi = 101, 90, 113, 102, 103, 104, 106, 105
($\bar{Y}$ = 103)

Both groups have the same mean but the variation of group X seems to be higher than group Y.

- The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is $\boldsymbol{\sigma^2}$ (is the Greek lowercase letter sigma). The formula for the population variance is.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Where: X individual value, μ population mean and N population size

- The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ, The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Example 6: A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading.? The data shows the results of six containers. Find the variance and standard deviation for the data set?

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Solution: The mean for brands A and B are:

For A $\mu = \frac{\sum X}{n}$

$= \frac{10+60+50+30+40+20}{6} = 35$

For B $\mu = \frac{\sum X}{n} = 35$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

	A	B
σ^2	291.67	41.67
σ	17.08	6.45

X_A	X_B	$(X_A-\mu)$	$(X_B-\mu)$	$(X_A-\mu)^2$	$(X_B-\mu)^2$
10	35	-25	0	625	0
60	45	25	10	625	100
50	30	15	-5	225	25
30	35	-5	0	25	0
40	40	5	5	25	25
20	25	-15	-10	225	100
$\mu =$ 35	$\mu =$ 35	$\Sigma =$ 0.0	$\Sigma =$ 0.0	$\Sigma =$ 291.67	$\Sigma =$ 41.67

????

Note: When the means are equal, the larger the variance or standard deviation is, the more variable the data are.

2.2. Sample Variance and Standard Deviation

- The formula for the sample variance, denoted by S^2, is

Where :

n = sample size

$\bar{X}$ = sample mean

X = individual value

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

- The standard deviation of a sample (denoted by S) is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

NOTE: The expression $\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$ does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by *n*, find the variance of the sample by dividing by *n*-1, giving a slightly larger value and an *unbiased* estimate of the population variance.

Example 7: Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars. **11.2, 11.9, 12.0, 12.8, 13.4, 14.3**.

Solution

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

$$S^2 = \frac{6.38}{6-1} = 1.276$$

$$S = \sqrt{S^2} = 1.13$$

X	X-$\bar{X}$	(X-$\bar{X}$)^2
11.2	-1.4	1.96
11.9	-0.7	0.49
12.0	-0.6	0.36
12.8	0.2	0.04
13.4	0.8	0.64
14.3	1.7	2.89
$\bar{X} = 35$		Σ= 6.38

2.3. Variance and Standard Deviation for Tabulated Data

The procedure for finding the variance and standard deviation for tabulated data is similar to that for finding the mean for tabulated data using the midpoints of each class using the following formula:

$$\sigma=\sqrt{\sigma^2}=\sqrt{\frac{f_i(X_m-\bar{X})^2}{\sum f_i-1}}$$

Example 7: Find the variance and the standard deviation for the frequency distribution of the data in **Example 4.**?.

Classes	f_i	X_m	$X_m-\bar{X}$	$(X_m-\bar{X})^2$	$(X_m-\bar{X})^2 \times f_i$
5.5-10.5	1	8	-16.5	272.25	272.25
10.5-15.5	2	13	-11.5	132.25	264.5
15.5-20.5	3	18	-6.5	42.25	126.75
20.5-25.5	5	23	-1.5	2.25	11.25
25.5-30.5	4	28	3.5	12.25	49
30.5-35.5	3	33	8.5	72.25	216.75
35.5-40.5	2	38	13.5	182.25	364.5
Summation	20				1305

Solution

$$\sigma=\sqrt{\sigma^2}=\sqrt{\frac{f_i(X_m-\bar{X})^2}{\sum f_i-1}}$$

$$\bar{X}=\frac{\sum f_i \times X_m}{\sum f_i}=24.5$$

$$\sigma=\sqrt{\frac{f_i(X_m-\bar{X})^2}{\sum f_i-1}}=\sqrt{\frac{1305}{19}}=8.28$$

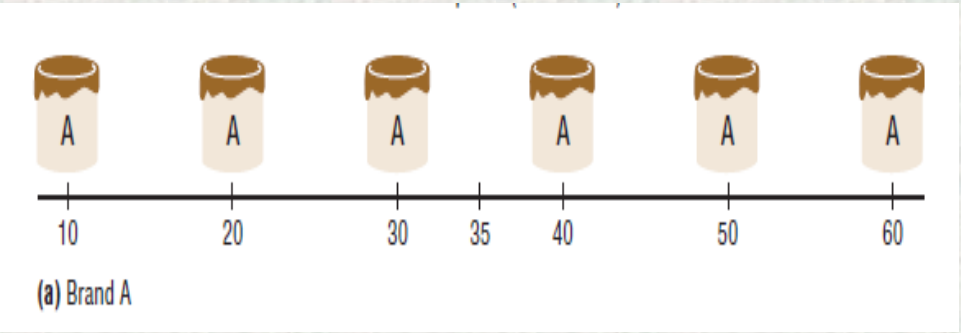
$$\sigma 2=68.68$$

2.4. Range

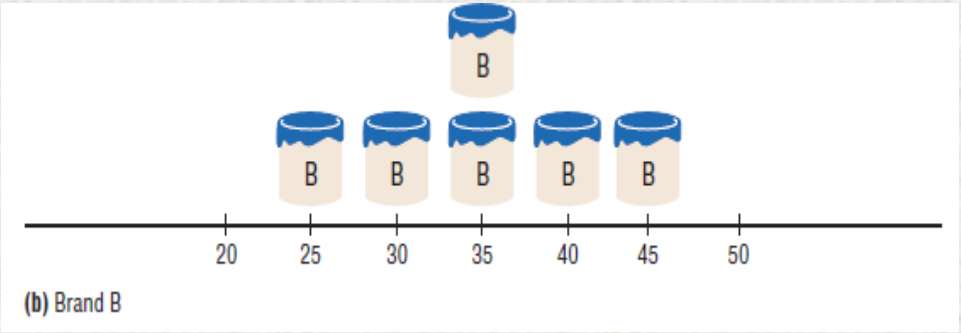
- The range is the simplest measurement of variance.
- The range is the highest value minus the lowest value. The symbol R is used for the range.

R = highest value - lowest value

Example 8: Find the ranges for the paints if last before fading in months?



(a) Brand A



(b) Brand B

Solution

- For brand A, the range is:
  R = 60 – 10 = 50 months
- For brand B, the range is:
  R = 45 – 25 = 20 months

Summary

Summary of Measures of Variation

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Notes:

Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

3. Coefficient of Variation

The coefficient of variation, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples

$$CV_{ar} = \frac{S}{\bar{X}}.100$$

For populations

$$CV_{ar} = \frac{\sigma}{\mu}.100$$

Example 9: The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution

The coefficients of variation are

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{5}{87}.100 = 5.7\%$$ sales

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{773}{5225}.100 = 14.8\%$$ commissions

Note: Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Example 10: The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

Solution

The coefficients of variation are

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{\sqrt{23}}{132}.100 = 3.6\% \quad \text{pages}$$

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{\sqrt{62}}{182}.100 = 4.3\% \quad \text{advertisements}$$

Note: The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.