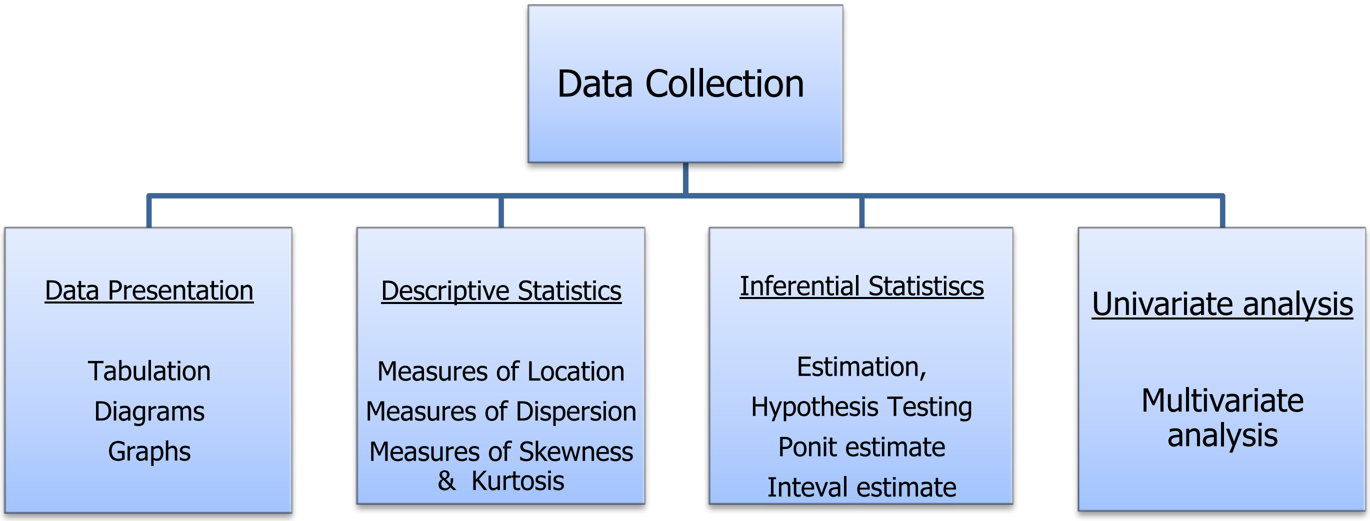
INVESTIGATION



EXAMPLE:

7,8,9,10,11	n=5,	x=45,	$\bar{x}$ =45/5=9
3,4,9,12,15	n=5,	x=45,	$\bar{x}$ =45/5=9
1,5,9,13,17	n=5,	x=45,	$\bar{x}$ =45/5=9

S.D. : (1) 1.58 (2) 4.74 (3) 6.32

So we use

Measures of Dispersion
Or
Measures of variability

Measures of Dispersion

Measures of dispersion summarize differences in the data, how the numbers differ from one another.

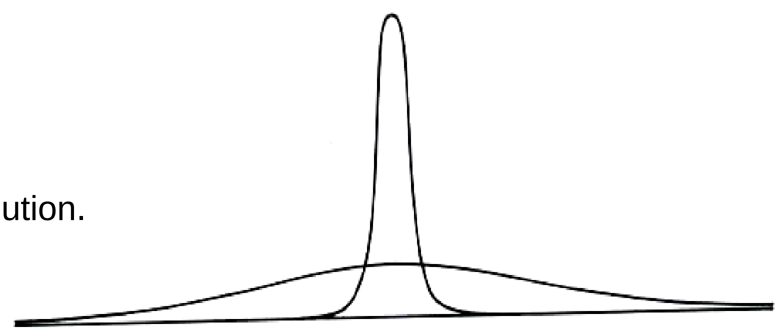
series I: 70 70 70 70 70 70 70 70 70 70

Series II: 66 67 68 69 70 70 71 72 73 74

Series III: 1 19 50 60 70 80 90 100 110 120

Measures of Variability

A single summary figure that describes
the spread of observations within a distribution.



MEASURES OF DESPERSION

1- **Range**
Difference between the smallest and largest observations.

2- **Interquartile Range**
Range of the middle half of scores.

3- **Variance**
Mean of all squared deviations from the mean.

4- **Standard Deviation**
Rough measure of the average amount by which observations deviate from the mean. The square root of the variance.

Range

- The difference between the lowest and highest values in the data set.
- The range can be misleading with outliers

data: 2,4,5,2,5,6,1,6,8,25,2
Sorted data: 1,2,2,2,3,4,5,6,6,8,25

Range = maximum – minimum
= 25 – 1
= 24

Hotel Rates
52, 76, 100, 136, 186, 196, 205, 150, 257, 264, 264, 280, 282, 283, 303, 313, 317, 317, 325, 373, 384, 384, 400, 402, 417, 422, 472, 480, 643, 693, 732, 749, 750, 791, 891

Range = 891-52 = 839

Measures of Position

Quartiles, Deciles, Percentiles

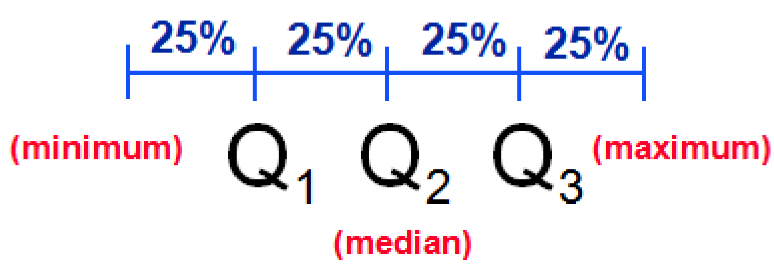
- **Quartiles**

Q_1, Q_2, Q_3 Divides ranked scores into four equal parts

$Q_1 = \frac{n+1}{4}$ th

$Q_2 = \frac{2\,(n+1)}{4} = \frac{n+1}{2}$ th

$Q_3 = \frac{3(n+1)}{4}$ th



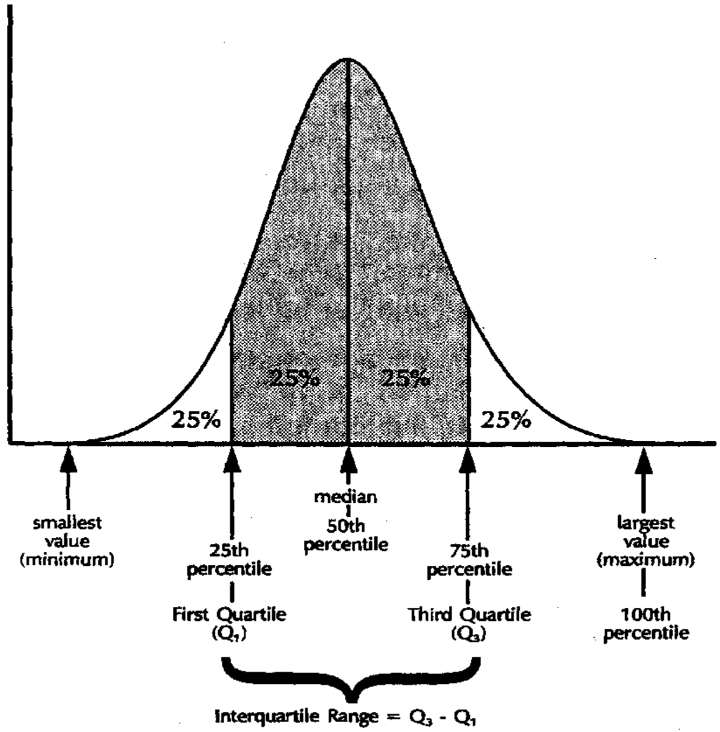
- **Inter quartile Range**

The inter quartile range is Q_3-Q_1

$$IQR = Q_3 - Q_1$$

50% of the observations in the distribution are in the inter quartile range.

The following figure shows the interaction between the quartiles, the median and the inter quartile range



Sort The Values First

Sample Number	Unsorted Values	Ranked Values	
1	25	14	Minimum
2	27	16	
3	20	16	
4	23	18	
5	26	19	LQ or Q_1
6	24	20	
7	19	20	
8	16	21	
9	25	23	
10	18	24	Md or Q_2
11	30	24	
12	29	25	
13	32	25	
14	26	26	
15	24	26	UQ or Q_3
16	21	27	
17	28	27	
18	27	28	
19	20	29	
20	16	30	Maximum

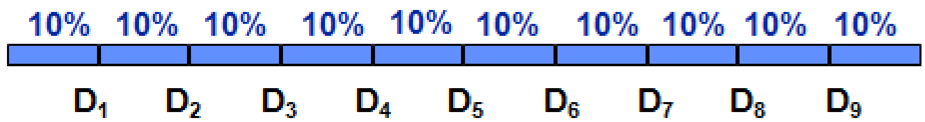
- Quartiles (Q_x) & Precentile (P_x)

$Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$

- Deciles:

D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, $\mathbf{D_9}$

divides ranked data into **ten** equal parts



Deciles & Percentiles:

$D_1 = P_{10}$, $D_2 = P_{20}$, $D_3 = P_{30}$, …… , $D_9 = P_{90}$

- Fractiles (Quantiles):

  partitions data into approximately equal parts

  examples are the **Quartiles, Deciles, Percentiles**

- Percentiles:

  Maximum is 100th percentile: 100% of values lie at or below the maximum
  Median is 50th percentile: 50% of values lie at or below the median

  Any percentile can be calculated, But the most common are 25th (1st Quartile) and 75th (3rd Quartile)

  - Locating Percentiles in a Frequency Distribution

A percentile is a score below which a specific percentage of the distribution falls(the median is the 50th percentile.

The 75th percentile is a score below which 75% of the cases fall.

The median is the 50th percentile: 50% of the cases fall below it

Another type of percentile :The quartile lower quartile is 25th percentile and the upper quartile is the 75th percentile

25th percentile

50th percentile

80th percentile

NUMBER OF CHILDREN

25% included here

50% included here

80% included here

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	260	26.6	26.6	26.6
	1	161	16.4	16.5	43.1
	2	260	26.6	26.6	69.7
	3	155	15.8	15.9	85.6
	4	70	7.2	7.2	92.7
	5	31	3.2	3.2	95.9
	6	21	2.1	2.1	98.1
	7	11	1.1	1.1	99.2
	EIGHT OR MORE	8	.8	.8	100.0
	Total	977	99.8	100.0	
Missing	NA	2	.2		
Total		979	100.0		

Notice the **Cumulative Percent**

26.6% have 0 children → so, the P_{25} located there

69.7% have 2 children → so, the P_{50} located there

85.6% have 3 children → so, the P_{80} located there

Illustration

Table 7.1 Haemoglobin Values (g%) of 26 Normal Children

11.8	12.9	12.4	13.3	13.8
11.4	12.3	11.7	12.9	12.2
10.4	10.8	12.7	13.2	
11.6	12.0	12.2	14.2	
10.8	10.5	11.6	13.5	
12.2	11.2	12.6	13.0	

10.4	11.2	11.7	12.2	12.6	13.0	13.8
10.5	11.4	11.8	12.2	12.7	13.2	14.2
10.8	11.6	12.0	12.3	12.9	13.3	
10.8	11.6	12.2	12.4	12.9	13.5	

The lower quartile Q_1 is 11.6 i.e. about 25% of the number of observations fall below the value 11.6. The upper quartile Q_3 is 12.9 i.e. nearly 25% of the number of observations are above the value 12.9. Therefore, the interquartile range is 11.6 to 12.9

Table 7.2 Protein Intake of 400 Families

Protein intake/consumption unit/day (gram)	No. of families
15–25	30
25–35	40
35–45	100
45–55	110
55–65	80
65–75	30
75–85	10
Total	400

Using the data given in Table 6.2, a few of the centiles are computed as follows:

First quartile or 25th percentile

$$P_{25} = L + \frac{(25N/100 - cf)}{f} \times C$$
$$= L + \frac{(N/4 - cf)}{f} \times C$$
$$= 35 + \frac{100 - 70}{100} \times 10 = 35 + \frac{30}{10}$$
$$= 35 + 3 = 38$$

Third quartile or 75th percentile

$$P_{75} = L + \frac{(75N/100 - cf)}{f} \times C$$
$$= L + \frac{(3N/4 - cf)}{f} \times C$$
$$= 55 + \frac{300 - 280}{80} \times 10 = 55 + \frac{200}{80}$$
$$= 57.5$$

Third percentile

$$P_3 = L + \frac{(3N/100 - cf)}{f} \times C$$
$$= 15 + \frac{12 - 0}{30} \times 10 = 15 + \frac{120}{30} = 15 + 4$$
$$= 19$$

First decile or 10th percentile

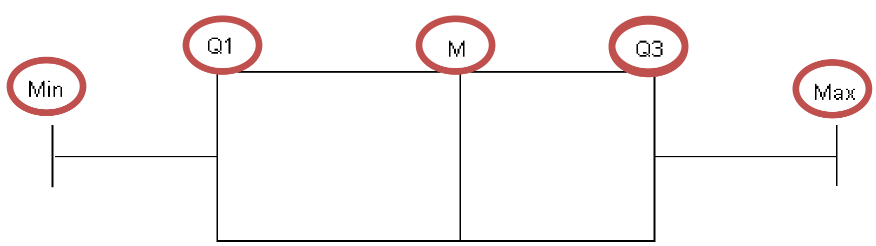
$$P_{10} = L + \frac{(10N/100 - cf)}{f} \times C$$
$$= 25 + \frac{40 - 30}{40} \times 10 = 25 + \frac{100}{40} = 25 + 2.5$$
$$= 27.5$$

97th percentile

$$P_{97} = L + \frac{(97N/100 - cf)}{f} \times C$$
$$= 65 + \frac{388 - 360}{30} \times 10$$
$$= 65 + 9.3 = 74.3$$

- Five Number Summary

Représentation d'un Box-Plot



- ✓ Minimum Value
- ✓ 1st Quartile
- ✓ Median
- ✓ 3rd Quartile
- ✓ Maximum Value

- VARIANCE:

Deviations of each observation from the mean, then averaging the sum of **squares** of these deviations.

- STANDARD DEVIATION:

4 words:

" ROOT – MEANS – SQUARE – DEVIATIONS "

Variance

The average amount that a score deviates from the typical score.

Example: 1, 2, 3, 4, 5 → Mean= 3

Score – Mean = Difference Score

1-3= -2, 2-3= -1, 3-3= 0, 4-3= 1, 5-3=2

Average of Difference Scores = $\frac{(-2)+(-1)+\ 0+\ 1+\ 2}{5} = 0$

(This happens with such values where variance is not helpful, But not always) So

In order to make this number not 0, square the difference scores (no negatives to cancel out the positives)

And Variance = $\frac{4+\ 1\ +\ 0+\ 1+\ 4}{5} = 2$

Computational Formula Of Variance:

Population: $\sigma^2 = \frac{N\sum X^2 - (\sum X)^2}{N^2}$

Sample: $S^2 = \frac{n\sum X^2 - (\sum X)^2}{n^2}$

- **Standard Deviation**

- ✓ To “undo” the squaring of difference scores, take the square root of the variance.
- ✓ Return to original units rather than squared units.
- ✓ Quantifying Uncertainty

Standard deviation

measures the variation of a variable in the sample.

Technically,

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Standard Deviation

Population: $\sigma = \sqrt{\sigma^2}$

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

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

Sample: $s = \sqrt{s^2}$

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

$$s = \sqrt{\frac{n\sum X^2 - (\sum X)^2}{n^2}}$$

Example:

Data: X = {6, 10, 5, 4, 9, 8}; N = 6

X	$X-\bar{X}$	$(X-\bar{X})^2$
6	-1	1
10	3	9
5	-2	4
4	-3	9
9	2	4
8	1	1
Total= 42		Total= 28

Mean:

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

Variance:

$$s^2 = \frac{\sum (\bar{X}-X)^2}{N} = \frac{28}{6} = 4.67$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

Calculating a Mean and a Standard Deviation

	Data x	Deviation x - Mean	Absolute Deviation |x - Mean|	Squared Deviation $(x\text{-Mean})^2$
	10	-20	20	400
	20	-10	10	100
	30	0	0	0
	40	10	10	100
	50	20	20	400
Sums	150	0	60	1000
Means	30	0	12	200
				Variance

Standard deviation = $\sqrt{\text{Variance}}$	14.1

Example of SD with discrete data

Marks achieved by 7 students: 3, 4, 6, 2, 8, 8, 5

Mean of these marks = $^{36}/_{7}$ = 5.14

Deviations from mean…

X	X - X	$(x - x)^2$
3	3 - 5.14 = -2.14	4.59
4	4 - 5.14 = -1.14	1.31
6	6 - 5.14 = 0.86	0.73
2	2 - 5.14 = -3.14	9.88
8	8 - 5.14 = 2.86	8.16
8	2.86	8.16
5	5 - 5.14 = -0.14	0.02
	Total = 0	**Total = 32.85**

Variance = $^{32.85} / {}_{7}$ = 4.69

SD = $\sqrt{4.69}$ = 2.17

Variability Example: Standard Deviation

X	X²
3	9
4	16
4	16
4	16
6	36
7	49
7	49
8	64
8	64
9	81
Sum: 60	Sum: 400

Mean = 6, SD = 2

Two ways to calculate the SD:

1- $$S=\sqrt{\frac{\sum(X-\overline{X})^2}{n}}$$

$$S=\sqrt{\frac{(3-6)^2+(4-6)^2+(4-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(7-6)^2+(8-6)^2+(8-6)^2+(9-6)^2}{10}}$$

$$\boxed{S=\sqrt{\frac{40}{10}}=2.0}$$

2- $$S=\sqrt{\frac{n\sum X^2-(\sum X)^2}{n^2}}$$

$$S=\sqrt{\frac{10(400)-(60)^2}{10^2}} \rightarrow S=\sqrt{\frac{4000-3600}{100}} \rightarrow S=\sqrt{4.0} \rightarrow \boxed{S=2.0}$$

able 7.5 Calculation of Standard Deviation for the Data of Table 7.1

Serial No.	Haemoglobin values	Deviation from arithmetic mean 12.2	Square of deviation
1	11.8	– 0.4	0.16
2	11.4	– 0.8	0.64
3	10.4	– 1.8	3.24
4	11.6	– 0.6	0.36
5	10.8	– 1.4	1.96
6	12.2	0	0
7	12.9	0.7	0.49
8	12.3	0.1	0.01
9	10.8	– 1.4	1.96
10	12.0	– 0.2	0.04
11	10.5	– 1.7	2.89
12	11.2	– 1.0	1.00
13	12.4	– 0.2	0.04
14	11.7	– 0.5	0.25
15	12.7	– 0.5	0.25
16	12.2	0	0
17	11.6	– 0.6	0.36
18	12.6	0.4	0.16
19	13.3	1.1	1.21
20	12.9	0.7	0.49
21	13.2	1.0	1.00
22	14.2	2.0	4.00
23	13.5	1.3	1.69
24	13.0	0.8	0.64
25	13.8	1.6	2.56
26	12.2	0.0	0
tal	317.2	0	25.40

Arithmetic mean is 12.2

$$\text{Standard deviation} = s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} = \sqrt{\frac{25.40}{25}}$$

$$s = \sqrt{1.016} = 1.01 \text{ g\%}$$

Table 7.6 Calculation of Standard Deviation for Data of Table 7.2

Protein intake/ consumption unit/day (g)	No. of families	Midpoint of class interval	Deviation of midpoint from arithmetic mean*	Squared deviation	Frequency × sq. deviation
Class interval	f	x	$(x-\bar{x})$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
15–25	30	20	– 27.5	756.25	22687.5
25–35	40	30	– 17.5	306.25	12250.0
35–45	100	40	– 7.5	56.25	5625.0
45–55	110	50	2.5	6.25	687.5
55–65	80	60	12.5	156.25	12500.0
65–75	30	70	22.5	506.25	15187.5
75–85	10	80	32.5	1056.25	10562.5
Total	400				79500.0

*Arithmetic mean = 47.5

From this table we get

$$\Sigma f(x-\bar{x})^2 = 79500.0$$

$$\Sigma f = 400$$

Therefore,

$$\text{Standard deviation} = S = \sqrt{\frac{79500.0}{400}} = 14.10 \text{ g}$$

$$\text{Variance} = S^2 = 198.75.$$

Illustration (ii)

Table 7.8 Calculation of Standard Deviation for Data of Table 7.2

Protein intake/ consumption unit/day (g)	No. of families	Midpoint of class interval		Square of midpoint of class interval	Frequency × square
Class interval	f	x	$f \cdot x$	x^2	$f \cdot x^2$
15–25	30	20	600	400	12000
25–35	40	30	1200	900	36000
35–45	100	40	4000	1600	160000
45–55	110	50	5500	2500	275000
55–65	80	60	4800	3600	288000
65–75	30	70	2100	4900	147000
75–85	10	80	800	6400	64000
Total	400		19000		982000

$$\text{Standard deviation} = \sqrt{\frac{1}{\Sigma f}\left[\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}\right]}$$

$$= \sqrt{1/400\left(982000 - \frac{(19000)^2}{400}\right)}$$

$$= \sqrt{1/400\ (982000 - 902500)}$$

$$= \sqrt{198.75}$$

$$= 14.10 \text{ g}$$

7.6 ALTERNATIVE METHOD OF CALCULATING STANDARD DEVIATION

This method can be computationally easier using values and the alternative formula given.

Illustration (i)

Table 7.7 Calculation of Standard Deviation for Table 7.1

Serial No.	Haemoglobin values (g%)	Square of haemoglobin values
1	11.8	139.24
2	11.4	129.96
3	10.4	108.16
4	11.6	134.56
5	10.8	116.64
6	12.2	148.84
7	12.9	166.41
8	12.3	151.29
9	10.8	116.64
10	12.0	144.00
11	10.5	110.25
12	11.2	125.44
13	12.4	153.76
14	11.7	136.89
15	12.7	161.29
16	12.2	148.84
17	11.6	134.56
18	12.6	158.76
19	13.3	176.89
20	12.9	166.41
21	13.2	174.24
22	14.2	201.64
23	13.5	182.25
24	13.0	169.00
25	13.8	190.44
26	12.2	148.84
Total	317.2	3895.24

$$\text{Standard deviation, } s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}}$$

$$= \sqrt{\frac{3895.24 - \frac{(317.2)^2}{26}}{25}}$$

$$= \sqrt{1.016}$$

$$= 1.01 \text{ g\%}$$

Mean and Standard Deviation

Using the mean and standard deviation together:

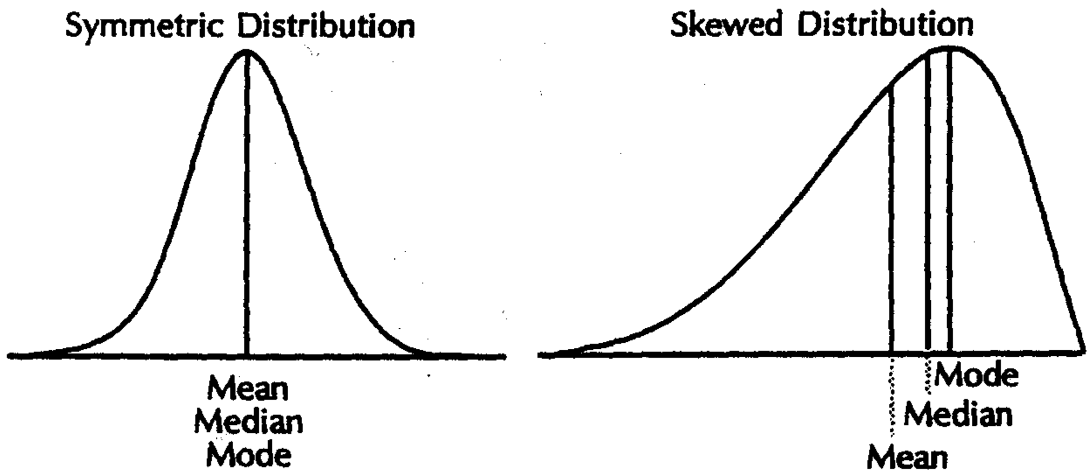
- Is an efficient way to describe a distribution with just two numbers.
- Allows a direct comparison between distributions that are on different scales.

WHICH MEASURE TO USE ?

- ✓ DISTRIBUTION OF DATA IS SYMMETRIC → USE MEAN & S.D.,
- ✓ DISTRIBUTION OF DATA IS SKEWED → USE MEDIAN & QUARTILES

Mean, Median and Mode

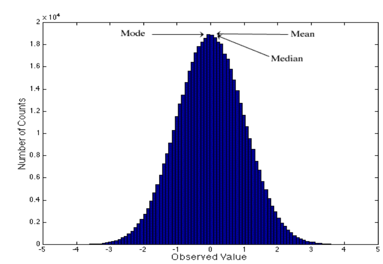
FIGURE 3.11
Effect of skewness on the mean, median, and mode



Distributions

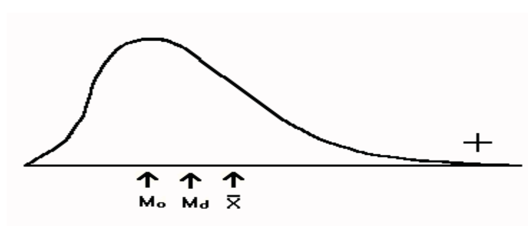
- Bell-Shaped: also known as

  “symmetric” or “normal”



- Skewed:
  - positively (skewed to the right)

    it tails off toward larger values



  - negatively (skewed to the left)

    it tails off toward smaller values