STATISTICS IS THE GRAMMAR OF SCIENCE

PROBABILITY AND STATISTICS

LECTURE – 4

DESCRIPTIVE MEASURES

(MEASURES OF LOCATION AND DISPERSION)

PREPARED BY
HAZBER SAMSON
FAST NUCES ISLAMABAD

MEASURES OF LOCATION

AVERAGE A single value which represents the whole data set is called an ***average*** or ***measure of location*** or ***measure of central tendency***.

There is not just one measure of location in fact, there are many. At present we will consider the following measures of location

- Arithmetic Mean
- Median
- Mode
- Combined Mean
- Weighted Mean
- Geometric Mean
- Harmonic Mean

ARITHMETIC MEAN

It is one of the types of average. It is denoted by $\bar{x}$. It is a single value which represents the whole data. It can be obtained with the help of following formulas

$Case-1: When\ Raw\ Data\ is\ Given$

$$\bar{x} = \frac{\Sigma x}{n}$$

where

$\bar{x}: sample\ mean$

$\Sigma x: sum\ of\ values$

$n: number\ of\ values$

EXAMPLE-1 Find mean of the following values $1,2,3,4,5$

SOLUTION $Now\ \bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

$Case-2: When\ Frequency\ is\ Given$

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

$where\ \bar{x}: sample\ mean$

$\Sigma fx: sum\ of\ values$

$\Sigma f: sum\ of\ frequencies$

EXAMPLE-2 Find mean of the following values

x	3	8	13	18	23
f	1	2	3	4	5

SOLUTION

x	*f*	*x.f*
3	1	3
8	2	16
13	3	39
18	4	72
23	5	115
65	15	245

$$Now\ \bar{x} = \frac{\sum fx}{\sum f} = \frac{245}{15} = 16.33$$

PROPERTIES OF ARITHMETIC MEAN

1. The mean is unique ie a set of data has only one mean.
2. All values are included in computing the mean.
3. Mean of constant values is constant.
4. Mean is effected by change of origin ie $if\ y = x \pm b\ then\ \bar{y} = \bar{x} \pm b$
5. Mean is effected by change of Scale ie $if\ y = a\,x\ then\ \bar{y} = a\bar{x}$
6. Mean is is effected by change of origin and scale ie
   $if\ y = ax + b\ then\ \bar{y} = a\bar{x} + b$
7. Sum of deviations of observations from mean is zero. ie $\Sigma\left(x - \bar{x}\right) = 0$
8. To compute a mean, the data must be measured at the interval or ratio level.

MEDIAN

It is one of the types of average. It is denoted by $\tilde{x}$. It is a single value which represents the whole data. Median is the value of the middle term in the ranked data set. It can be obtained with the help of following formulas.

MEDIAN The midpoint of the values after they have been ordered from the minimum to the maximum values.

To compute the median for a set of data, you first rank the values from smallest to largest and then use Equation below to compute the rank of the value that is the median

MEDIAN

$$\text{Median} = \frac{n+1}{2} \text{ ranked value}$$

$Case-1: When\ Raw\ Data\ is\ Given$

$$\tilde{x} = \left(\frac{n+1}{2}\right)^{th} value\ in\ the\ ranked\ data\ set$$

$where\ \tilde{x} : sample\ median$

$n : number\ of\ values$

We compute the median by following one of two rules:

- ***RULE 1*** If the data set contains an odd number of values, the median is the measurement associated with the middle-ranked value.

- ***RULE 2*** If the data set contains an even number of values, the median is the measurement associated with the average of the two middle-ranked values.

EXAMPLE- 3 The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

SOLUTION

Step 1 Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

Step 2 Select the middle value.

292, 300, 311, 401, 595, 618, 713

↑

Median

Hence, the median is 401 rooms.

USING FORMULA

Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

$$\tilde{x} = \left(\frac{n+1}{2}\right)^{th} \text{ value in the ranked data set}$$

$$\tilde{x} = \left(\frac{7+1}{2}\right)^{th} \text{ value}$$

$$\tilde{x} = 4^{th} \text{ value}$$

$$\tilde{x} = 401$$

EXAMPLE- 4 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 Arranging the data in order

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

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

USING FORMULA

Arranging the data in order

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

$$\tilde{x} = \left(\frac{n+1}{2}\right)^{th} \text{ value in the ranked data set}$$

$$\tilde{x} = \left(\frac{8+1}{2}\right)^{th} \text{ value}$$

$$\tilde{x} = 4.5^{th} \text{ value}$$

$$\tilde{x} = 4th \text{ value} + 0.5(5th - 4th) \text{ value}$$

$$\tilde{x} = 764 + 0.5\,(856 - 764)$$

$$\tilde{x} = 810$$

Case − 2 : *When Frequency is Given*

In this case first find $\frac{\Sigma f}{2} = m\,(say)$,

Now search Corresponding number in x for m.

EXAMPLE-5

(a) Find median of the following values

x	3	8	13	18	23
f	1	2	3	4	5

SOLUTION Here $\frac{\Sigma f}{2} = \frac{15}{2} = 7.5$ so it shows that central value is 8th value.

Now the value corresponding to the 8th position is 18 so here $median = 18$

(b) Find median of the following values

x	2	4	6	8	12
f	1	2	3	3	9

SOLUTION Here $\frac{\Sigma f}{2} = \frac{18}{2} = 9$ so it shows that central values are 9th and 10th values.

Now the value corresponding to the 9th and 10th positions are 8 and 12 so here

$$median = \frac{8+12}{2} = 10$$

PROPERTIES OF MEDIAN

1. Median of constant values is constant.
2. Median is unique ie A set of data has only one median.
3. Median is not affected by extreme values
4. It can be computed for Ratio Level, Interval Level and Ordinal Level data.

MODE

It is the most repeated value in the data set. It is one of the types of average. It is denoted by $\widehat{x}$. Mode is the value of the observation that appears most frequently in the data set. It can be obtained with the help of following formulas

NOTE: If a data set has no mode, 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.

$$Case-1: When\ Raw\ Data\ is\ Given$$

$$\hat{x} = Most\ repeated\ value\ in\ the\ data\ set$$

EXAMPLE-6(a) Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are 18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10'

SOLUTION It is helpful to arrange the data in order although it is not necessary.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Since $10 million occurred 3 times—a frequency larger than any other number—the mode is $10 million.

(b) 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.

(c) 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	109	109	110
109	111	112	111	109

SOLUTION Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

$$Case-2: When\ Frequency\ is\ Given$$

$$\hat{x} = Value\ of\ x\ corresponding\ to\ highest\ \ frequency$$

EXAMPLE-7 Find mode of the following values

x	3	8	13	18	23
f	1	2	3	4	5

SOLUTION Here highest frequency in the data set is 5 and the value corresponding to that frequency is 23 so $Mode = 23$.

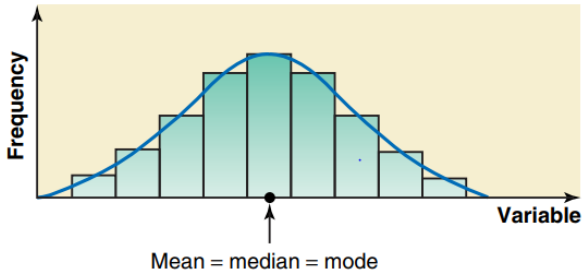
PROPERTIES OF MODE

1. Mode is not unique ie A set of data my have more then one values of mode.
2. It is not affected by extreme values.
3. It can be computed for Ratio Level, Interval Level and Nominal Level data.

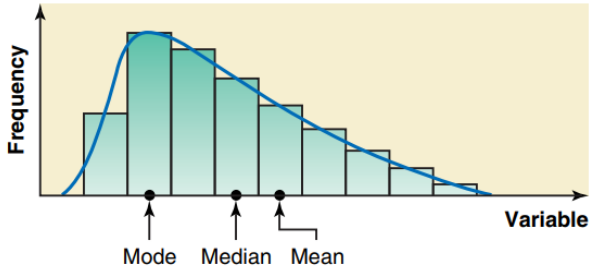
RELATIONSHIPS AMONG THE MEAN, MEDIAN AND MODE

Histogram or a frequency distribution curve may be symmetric or skewed. This section describes the relationships among the mean, median, and mode for histograms and frequency distribution curves. Knowing the values of the mean, median, and mode can give us some idea about the shape of a frequency distribution curve.

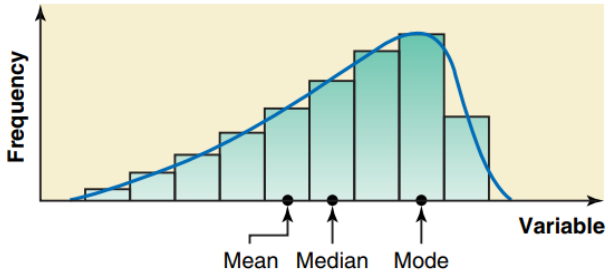
1. For a **symmetric histogram** and frequency distribution curve with one peak, the values of the mean, median, and mode are identical, and they lie at the center of the distribution.



2. For a histogram and a frequency distribution curve **skewed to the right**, the value of the mean is the largest that of the mode is the smallest, and the value of the median lies between these two.



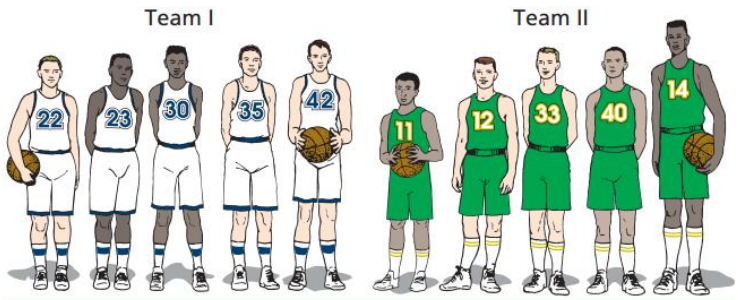
3. If a histogram and a frequency distribution curve are **skewed to the left**, the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two.



MEASURES OF DISPERSION

Up to this point, we have discussed only descriptive measures of center, specifically, the mean, median, and mode. However, two data sets can have the same mean, median, or mode and still differ in other respects. For example, consider the heights of the five starting players on each of two men's college basketball teams, as shown in Fig. below

Five starting players on two basketball teams



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

The two teams have the same mean height, 75 inches; the same median height, 76 inches; and the same mode, 76 inches. Nonetheless, the two data sets clearly differ. In particular, the heights of the players on Team II vary much more than those on Team I. To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set. Such descriptive measures are referred to as **measures of variation** or **measures of spread or measures of dispersion.**

Dispersions is defined as the degree of spreadness or scatterness of observations from some average value (mean, median or mode) of the data.

There are two types of measures of dispersion

1. Absolute Measure of Dispersion
2. Relative Measure of Dispersion

ABSOLUTE MEASURE OF DISPERSION

It is used to measure degree of scatterness in the data about some central point. These measures give result in same unit as the unit of data. Following are the types of absolute measure of dispersion Range, Variance, Standard Deviation, Mean Deviation, Quartile Deviation

RELATIVE MEASURE OF DISPERSION

A relative measure of dispersion is one that is expressed on the form of ratio, coefficient or percentage and is independent of units of measurement. It is useful for comparison of data of different nature. Following are the types of relative measure of dispersion
Co-efficient of Range, Co-efficient of Variation, Co-efficient of Standard Deviation, Co-efficient of Mean Deviation, Co-efficient of Quartile Deviation

ABSOLUTE MEASURES OF DISPERSION

RANGE

The simplest measure of dispersion is the range. Range is defined as
Range is the difference between the maximum and minimum values in a data set

It is denoted by R. Mathematically

$$R = x_m - x_0$$

where R : *Range*

x_m : *Largest value in data set*

x_0 : *Smallest value in data set*

EXAMPLE-1 For the given data $1,\ 2,\ 3,\ 4,\ 5$. Find the range for this data.

SOLUTION Given Data is $X = 1,\ 2,\ 3,\ 4,\ 5$

Range is given by $R = x_m - x_0 = 5 - 1 = 4$

EXAMPLE-2 A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the Range of each group.

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

SOLUTION

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

VARIANCE AND STANDARD DEVIATION

Karl Pearson introduced the statistical concepts of the variance and standard deviation. It measures the mean amount by which the values in a sample, vary from their mean.

VARIANCE The arithmetic mean of the squared deviations from the mean.

We shall study variance and standard deviation for sample data.

SAMPLE VARIANCE AND STANDARD DEVIATION

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

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

where

$\bar{X}$ = sample mean

n = sample size

Formula for the Sample Standard Deviation

The standard deviation of a sample (denoted by s) is

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

where

X = individual value

$\bar{X}$ = sample mean

n = sample size

SHORT CUT FORMULAS FOR VARIANCE AND STANDARD DEVIATION

- $Sample\ Variance = S^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)$
- $Sample\ S.D = S = \sqrt{\frac{1}{n-1}\left(\sum x^2 - \frac{(\Sigma x)^2}{n}\right)}$

Shortcut or Computational Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

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

PROPERTIES OF VARIANCE

1. Variance of constant is zero. ie $V(c)=0$
2. Variance is invariant to change of the origin. ie $V(X\pm a)=V(X)$
3. Variance is affected by change of scale. ie $V(aX)=a^2.V(X)$

PROPERTIES OF STANDARD DEVIATION

1. Standard Deviation of constant is zero. ie $S.D(c)=0$
2. Standard Deviation is invariant to change of the origin. ie $S.D(X\pm a)=S.D(X)$
3. Standard Deviation is affected by change of scale. ie $S.D(aX)=|a|.S.D(X)$

EXAMPLE-3 Find the variance and standard deviation for the following data

35, 45, 30, 35, 40, 25

SOLUTION First of all we shall find mean

X	$X-\overline{X}$	$(X-\overline{X})^2$
35	0	0
45	10	100
30	-5	25
35	0	0
40	5	25
25	-10	100
210	0	250

Now mean is given by

$$\overline{X}=\frac{\sum X}{n}=\frac{210}{6}=35$$

Variance is given by

$$S^2=\frac{\sum\left(X-\overline{X}\right)^2}{n-1}=\frac{250}{6-1}=\frac{250}{5}=10$$

Standard Deviation is given by

$$S=\sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n-1}}=\sqrt{10}=3.16$$

EXAMPLE-4 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

Step 1 Find the sum of the values.

$$\Sigma X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

Step 2 Square each value and find the sum.

$$\Sigma X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

Step 3 Substitute in the formulas and solve.

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$$
$$= \frac{6(958.94) - 75.6^2}{6(6-1)}$$
$$= \frac{5753.64 - 5715.36}{6(5)}$$
$$= \frac{38.28}{30}$$
$$= 1.276$$

The variance is 1.28 rounded.

$$s = \sqrt{1.28} = 1.13$$

Hence, the sample standard deviation is 1.13.

NOTE

Note that ΣX^2 is not the same as $(\Sigma X)^2$. The notation ΣX^2 means to square the values first, then sum; $(\Sigma X)^2$ means to sum the values first, then square the sum.

Table 3–2 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

VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

Following are the basic formulas that are used to calculate the sample variances and standard deviations for grouped data:

$$Sample\ Variance = S^2 = \frac{\Sigma f\left(x-\bar{x}\right)^2}{n-1}$$

$$Sample\ S.D = S = \sqrt{\frac{\Sigma f\left(x-\bar{x}\right)^2}{n-1}}$$

SHORT CUT FORMULAS FOR VARIANCE AND STANDARD DEVIATION

Short-cut formulas are more efficient for calculating the variance and standard deviation.

$$Sample\ Variance = S^2 = \frac{1}{n-1}\left(\sum x^2 f - \frac{\left(\sum xf\right)^2}{n}\right)$$

$$Sample\ S.D = \sqrt{\frac{1}{n-1}\left(\sum x^2 f - \frac{(\Sigma xf)^2}{n}\right)}$$

EXAMPLE-5 *CALCULATING THE SAMPLE VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA*

The following data, give the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail order company.

Number of Orders	*f*
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the variance and standard deviation.

SOLUTION All the information required for the calculation of the variance and standard deviation appears in Table below

Number of Orders	x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$f\,(x-\bar{x})^2$
10-12	11	4	-5.64	31.8096	127.2384
13-15	14	12	-2.64	6.9696	83.6352
16-18	17	20	0.36	0.1296	2.592
19-21	20	14	3.36	11.2896	158.0544
	62	50	-4.56	50.1984	371.52

Now mean is given by

$$\overline{X}=\frac{\sum f\,x}{\sum f}=\frac{832}{50}=16.64$$

Variance is given by

$$S^2=\frac{\sum f\left(x-\bar{x}\right)^2}{n-1}=\frac{371.52}{50-1}=\frac{371.52}{49}=7.58$$

Standard Deviation is given by

$$S=\sqrt{\frac{\sum f\left(x-\bar{x}\right)^2}{n-1}}=\sqrt{7.58}=2.75$$

VARIANCE AND STANDARD DEVIATION USING SHORT CUT FORMULAS

Number of Orders	x	f	$x\,f$	$x^2\,f$
10-12	11	4	44	484
13-15	14	12	168	2352
16-18	17	20	340	5780
19-21	20	14	280	5600
	62	50	832	14216

Variance is given by

$$S^2=\frac{1}{n-1}\left(\sum x^2 f-\frac{\left(\sum xf\right)^2}{n}\right)=\frac{1}{49}\left(14216-\frac{(832)^2}{50}\right)=7.58$$

Standard Deviation is given by

$$S=\sqrt{\frac{1}{n-1}\left(\sum x^2 f-\frac{(\Sigma xf)^2}{n}\right)}=\sqrt{7.58}=2.75$$

RELATIVE MEASURES OF DISPERSION

A relative measure of dispersion is one that is expressed on the form of ratio, coefficient or percentage and is independent of units of measurement. It is useful for comparison of data of different nature. Following are the types of relative measure of dispersion

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Mean Deviation
5. Co-efficient of Quartile Deviation

CO-EFFICIENT OF DISPERSION Range is an absolute measure of dispersion. Its relative measure is known as co-efficient of dispersion and is defined as

$$Co-efficient of\ \ Dispersion = \frac{x_m - x_0}{x_m + x_0}$$

CO-EFFICIENT OF VARIATION Relative measure of variance is known as co-efficient of variation, it is given by

$$Co-efficient of\ \ Variation = C.V = \frac{S}{\bar{x}} \times 100$$

CO-EFFICIENT OF STANDARD DEVIATION Relative measure of standard deviation is known as co-efficient of variation, it is given by

$$Co-efficient of\ \ S.D = \frac{S.D}{Mean}$$

CO-EFFICIENT OF MEAN DEVIATION Relative measure of mean deviation is known as co-efficient of mean deviation, it is given by

$$Co-efficient of\ \ M.D = \frac{M.D}{Mean}$$

CO-EFFICIENT OF QUARTILE DEVIATION Relative measure of quartile deviation is known as co-efficient of quartile deviation, it is given by

$$Co-efficient of\ \ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is used to compare the variation in two or more sets of data.

USING MEAN AND STANDARD DEVIATION TOGETHER

In the previous sections, we introduced several important descriptive measures that are useful for transforming data into meaningful information. Two of the most important of these measures are the mean and the standard deviation. In this section, we discuss several statistical tools that combine these two. These tools are as follows

- Coefficient of Variation
- The Empirical Rule

COEFFICIENT OF VARIATION

Karl Pearson devised the coefficient of variation to compare the deviations of two different groups such as the heights of men and women. A statistic that allows you to compare standard deviations when the units are different, is called the coefficient of variation. The coefficient of variation (CV) is used to measure the relative variation for distributions with different means.

Population Coefficient of Variation

$$CV = \frac{\sigma}{\mu}(100)\%$$

Sample Coefficient of Variation

$$CV = \frac{s}{\bar{x}}(100)\%$$

EXAMPLE-7 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

$$\text{CVar} = \frac{s}{X} = \frac{5}{87} \cdot 100 = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100 = 14.8\% \qquad \text{commissions}$$

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

THE EMPIRICAL RULE

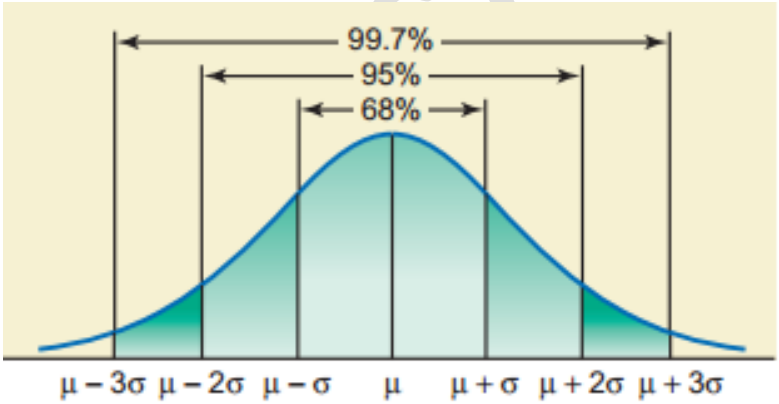
For a symmetrical, bell-shaped distribution, we can be more precise in explaining the dispersion about the mean. These relationships involving the standard deviation and the mean are described by the **Empirical Rule**, sometimes called the **Normal Rule**.

EMPIRICAL RULE For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations will lie within plus and minus one standard deviation of the mean; about 95% of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7%) will lie within plus and minus three standard deviations of the mean.

Empirical Rule For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean.

FIGURE 4.1 Illustration of the empirical rule



EXAMPLE

A sample of the rental rates at University Park Apartments approximates a symmetrical, bell-shaped distribution. The sample mean is \$500; the standard deviation is \$20. Using the Empirical Rule, answer these questions:

1. About 68% of the monthly rentals are between what two amounts?
2. About 95% of the monthly rentals are between what two amounts?
3. Almost all of the monthly rentals are between what two amounts?

SOLUTION

1. About 68% are between \$480 and \$520, found by $\bar{x} \pm 1s = \$500 \pm 1(\$20)$.
2. About 95% are between \$460 and \$540, found by $\bar{x} \pm 2s = \$500 \pm 2(\$20)$.
3. Almost all (99.7%) are between \$440 and \$560, found by $\bar{x} \pm 3s = \$500 \pm 3(\$20)$.