Statistics &Data Analysis Concepts for Data Science and ML 3

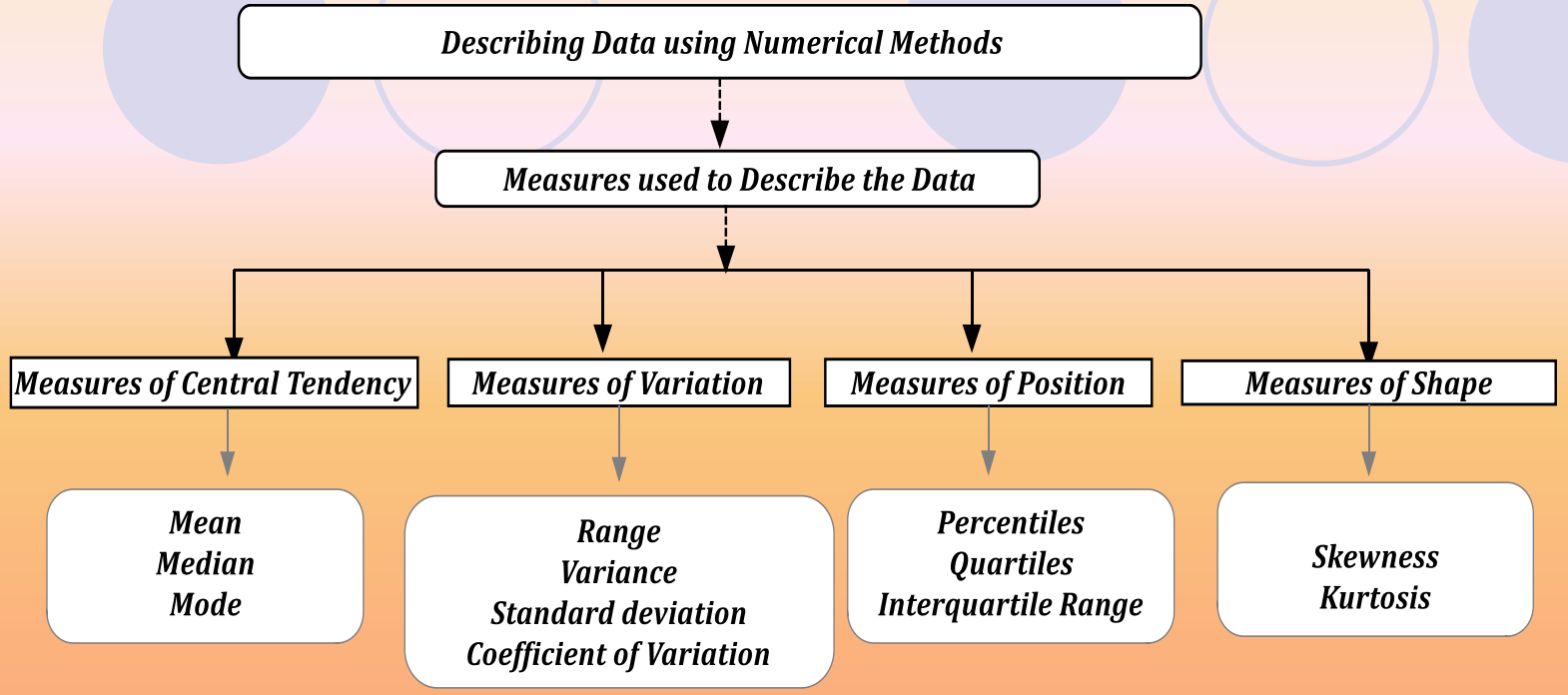
Descriptive Statistics: Numerical Methods

Learning Objectives

- **Master the techniques of describing data using numerical methods, and use these methods to compare and draw meaningful conclusions from data**
- **Calculate and apply the measures of central tendency including the mean, median, mode for both ungrouped and grouped data**
- **Calculate the measures of position: percentiles and quartiles and interpret them**
- **Calculate and apply various measures of variation— range, interquartile range, variance, and standard deviation for both grouped and ungrouped data**
- **Compare the mean, median, mode, and standard deviation to draw meaningful conclusions from the data**
- **Relate the mean and standard deviation using the Chebyshev's and Empirical rules and understand the importance of Empirical rule in statistics and data analysis**
- **Describe the relationship between two variables — covariance and coefficient of correlation**
- **Use computer packages to compute the above measures and interpret the results**

Some Key Concepts

- Qualitative and quantitative data
- Discrete and continuous data
- Ungrouped and grouped data (frequency distribution)
- Difference between a population and a sample
- Population parameters and sample statistics, and
- Symbols used to describe the population parameters and sample statistics.



Population Parameters

μ = population mean

σ^2 = population variance

σ = population standard deviation

p= population proportion

N = size of the population

* μ is read as "mu" and σ is read as "sigma."

Sample statistics

$\bar{x}$ = sample mean

s^2 = sample variance

s = sample standard deviation

$\bar{p}$ = sample proportion

n = sample size

*($\bar{x}$ is read as "x-bar")

Chapter 3 : Different Measures of Describing Data - Flow Chart (1)

Population Parameters & Sample Statistics

A Population is described by its ***parameters***

The population parameters are:

μ = population mean σ^2 = population variance

σ = population standard deviation

p= population proportion N = population size

A ***sample*** is described by its ***statistics***. These are:

$\bar{x}$ = sample mean s^2 = sample variance

s = sample standard deviation $\bar{p}$ = sample proportion

n = sample size

($\bar{x}$ is read as “x -bar”)

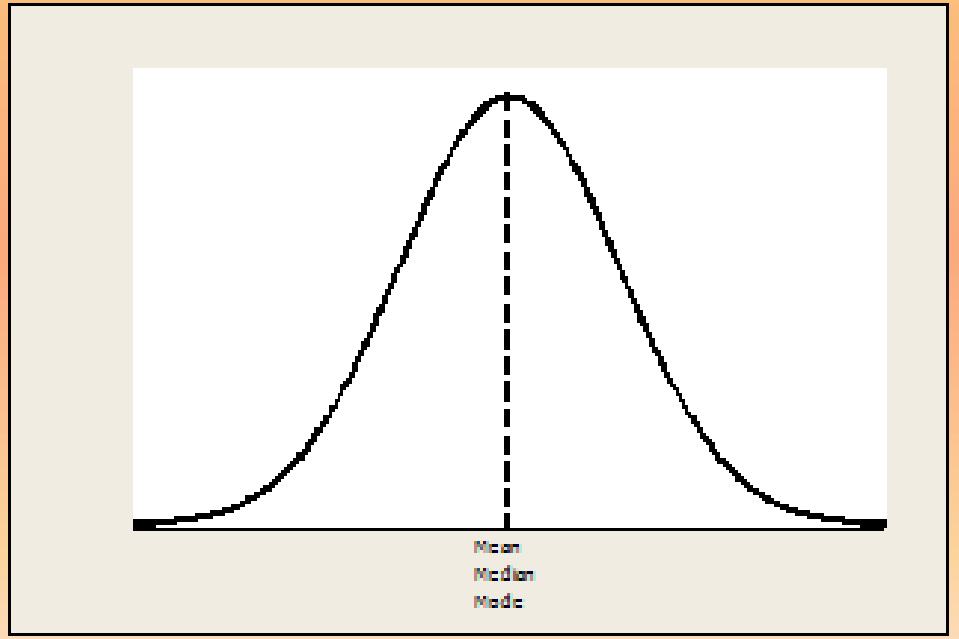
Measures of Central Tendency or Measures of Location

- Central tendency is described by the “average” and is also called the measure of location.
- The central tendency of a data set is the tendency of the data to cluster about certain numerical values such as, the mean or median.
- ***Measures of Central Tendency:***

(1) Mean

(2) Median

(3) Mode



Mean or the Average

- The mean or the average is commonly used to obtain a typical representation of a group as a whole
- The mean of a data set is sum of the values divided by the number of observations. The mean of *n* observations $x_1, x_2, x_3, .., x_n$ is given by

$$Mean = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$$

$$Mean = \frac{\sum allvalues}{n} = \frac{\sum x}{n}$$

The stock price for a technology company for the past five days are $12, $13, $10, $15, and $10. The mean stock price is

$$Mean = \frac{12+13+10+15+10}{5} = \frac{60}{5} = \$12$$

Sample and Population Means

Sample Mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

Population Mean

$$\mu = \frac{\sum x_i}{N}$$

The price of certain product (in dollars) is given below. What is the mean or the average price?

5 8 10 7 10 14

The sample mean can be calculated as

$$\bar{x} = \frac{\sum x}{n} = \frac{5+8+10+7+10+14}{6} = 9$$

Interpretation of Mean

The mean can be interpreted in the following ways:

- it provides a single number presenting the whole data set
- It gives us the significance of the whole
- It is unique because a given data set has only one mean
- It is useful for comparing different data sets in terms of average

The mean can be affected by extreme values (extreme values are very high or very low values in a data set that may be difficult to detect in a large data set).

Median

- The ***median*** is the middle value of a data set when the data are arranged in increasing (or decreasing) order.

- The median is the value that divides the data into two equal parts, such that half of the values lie above the median and the other half below it. It measures the central item in the data.

- For the ungrouped data (data not grouped into a frequency distribution) the median is calculated based on whether the number of observations is odd or even.

CALCULATING MEDIAN WHEN THE NUMBER OF OBSERVATIONS IS ODD

If the number of observations is odd, the median can be calculated by

- Arranging the data in increasing order
- Locating the middle value after the values have been arranged in ascending order of magnitude
- There is a distinct median when the number of observations is odd
- Unlike the mean, the median is not affected by extreme values

MEDIAN WHEN THE NUMBER OF OBSERVATIONS IS EVEN

When the number of observations is even, there are two middle values, and the median is obtained by taking the arithmetic mean of the middle terms.

Example:

Suppose we have the following observations arranged in increasing order.

1	2	3	4	5	6	7
8.2	8.3	8.9	9.6	9.8	10.2	12.0

The number of observations is seven (n=7) which is odd therefore, the middle value or the median is 9.6.

Example:

The data below are the annual incomes in thousands of dollars for a sample of eight employees of a manufacturing company for the past year. Find the median.

EXAMPLES ON MEDIAN…CONT.

1	2	3	4	5	6	7	8
70	62	60	45	40	56	38	35

The number of observations is: n=8 (even). To find the median, arrange the data in increasing order

1	2	3	4	5	6	7	8
35	38	40	45	56	60	62	70

Next, find the location of the median. The location of the median for the data set with even number of observations is given by

$$\frac{n+1}{2} = \frac{8+1}{2} = 4.5$$

Therefore, the median is the average of 4th and 5th values

$$Median = \frac{45+56}{2} = 50.5$$

MODE

The mode is the value that occurs most frequently in a set or, it is the value that is repeated most often in a data set. Sometimes chance causes a single often- repeated value to be the most frequent value in the data set.

Example:

The following data represent the number of hours of use of personal computer per day by a sample 20 employees at work:

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

The mode for this data is 3 hours because this value is repeated the maximum number of times. Therefore, ***Mode = 3 hours***

ADVANTAGES OF MODE

- It is easy to calculate; in some cases it can be located by inspection.
- It is not affected by extreme values.

DISADVANTAGES OF MODE

- **It is not always possible to find a unique mode in a data set. Some data may have more than one mode. If the data have a unique mode, then we have a *unimodal* data. If a data set has two modes, then it is *bi-modal*. In case of more than two modes, we say that we have *multimodal* data.**

- **Sometimes there may not be any mode in the data as none of the values repeat, or there may be cases where several values repeat the same number of times. In cases where there are several modes in the data, the mode becomes a useless measure. If the data set contains two, three or many modes, interpretation becomes difficult.**

COMPARING MEAN, MEDIAN, AND MODE

- If the values of the mean, median and mode are equal or approximately equal, the shape or the distribution of the data is ***symmetrical***.

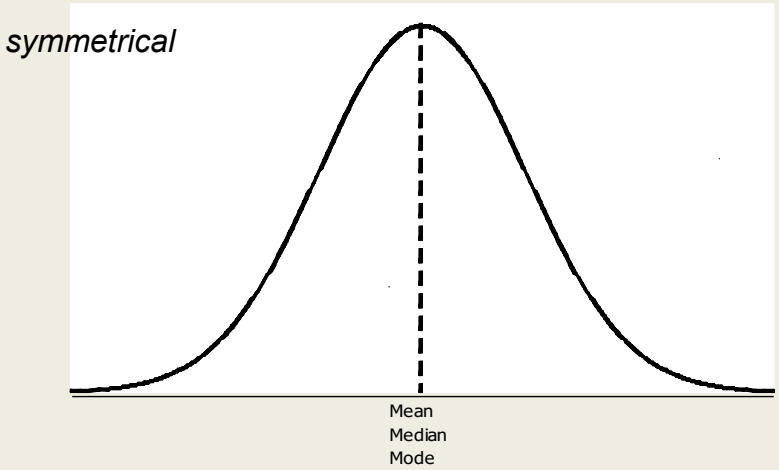
- If Mean > Median, and
  Mean > Mode

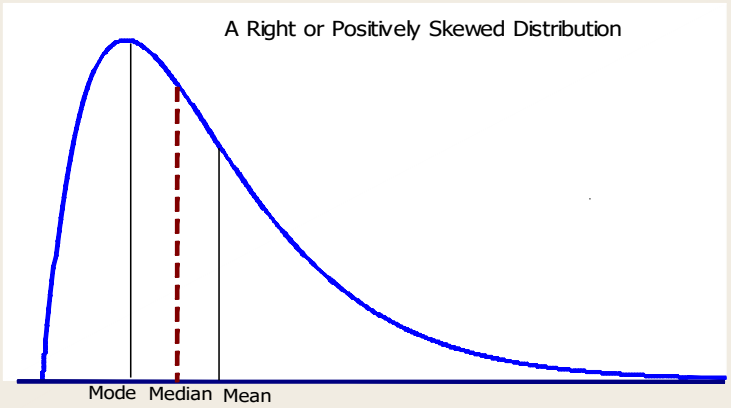
Then the shape of the data is ***right skewed*** or ***positively skewed***

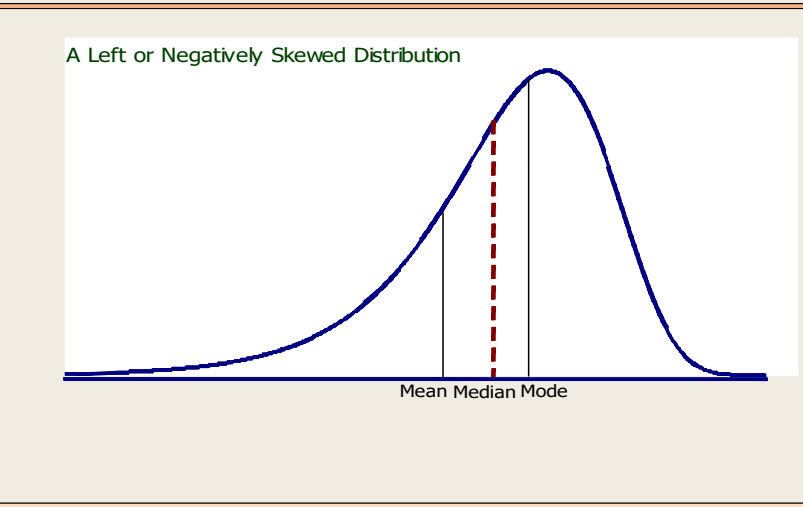
- If Mean < Median, and
  Mean < Mode

Then the shape is ***left skewed*** or we have a ***negatively skewed*** data

COMPARING MEAN, MEDIAN, AND MODE







Note that when the distribution is skewed — negatively or positively — the *median* is always midway between the mean and the mode. Therefore, median is often used to describe a skewed data and is the best measure for central tendency whenever data are skewed.

MEASURES OF POSITION: PERCENTILES AND QUARTILES

Percentile and quartiles are measures to describing the data. These are known as ***measures of position*** and are described below.

Percentile: A *percentile* is a point below which a stated percentage or proportion of observations lie.

A percentile tells us how the data values are spread out over the interval from the smallest value to the largest value.

The p^{th} percentile of a data set is a value, such that at least p percent of the values are less than or equal to this value, and at least (100-p) percent of the values are greater than or equal to this value.

Quartiles

Quartiles: are special percentiles which divide the observations into groups of successive size, each containing 25% of the data points. The quartiles are denoted by

Q_1: the first quartile or 25th percentile;

Q_2: the second quartile or 50th percentile (which is also the median); and

Q_3: the third quartile or 75th percentile.

Interquartile range: It is the difference between the third quartile and the first quartile and encompasses the middle 50% of the values.

IQ Range = $Q_3 - Q_1$

The five measure summary provides a visual display of the data in form of a plot known as the ***box plot***. This is a plot of the minimum, maximum and three quartiles, Q_1, Q_2, and Q_3. The box plot shows the data extremes, the range, the median, the quartiles, and the interquartile range.

CALCULATING PERCENTILES AND QUARTILES

To find a percentile or a quartile

- **Arrange the data in increasing order**
- **Find the location of the percentile using the following formula**

$$L_p = (n+1)\frac{P}{100}$$

where, L_p = location of the percentile
n = total number of observations
P = desired percentile

Example

Find the median, the first quartile, and the third quartile for the data below.

2038	1758	1721	1637	2097	2047	2205	1787	2287	1940	2311
2054	2406	1471	1460							

CALCULATING PERCENTILES AND QUARTILES

Solution: The number of observations is 15 (n=15).

- Arrange the data in increasing order. The sorted values are shown below.

Sorted data									
1460	1471	1637	1721	1758	1787	1940	2038	2047	2054
2097	2205	2287	2311	2406					

(a) Calculate the median or Q2 (50th percentile)

First, calculate the position of the median or Q2 using.

The median or Q2 (50th percentile) is located at

$$L_p = (n+1)\frac{P}{100} = 16\left(\frac{50}{100}\right) = 8$$

Therefore, the 8th value in the sorted data (Table 3.2) is the median or Q2. This value is 2038. Therefore,

Median = 2038

CALCULATING PERCENTILES AND QUARTILES

(b) Calculate the first quartile or Q1 (25th percentile)

Calculate the position of Q1. The first Quartile (Q1) or 25th percentile is located at

$$L_p = (n+1)\frac{P}{100} = 16(\frac{25}{100}) = 4$$

The 4th value in the sorted data (Table on the previous page) is Q1. This value is 1721. Therefore,

Q1 = 1721

(c) Calculate the third quartile or Q3 (75th percentile)

$$L_p = (n+1)\frac{P}{100} = 16(\frac{75}{100}) = 12$$

The 12th value in the sorted data is Q3. Therefore,

Q3 = 2205

PERCENTILES AND QUARTILES USING MINITAB

The table below shows the values of Q1, Q2 (Median) and Q3 calculated using MINITAB. These values match the values calculated manually in the previous examples

Descriptive Statistics Calculations using MINITAB

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Data	15	1947.9	2038.0	1950.2	298.8	77.1

Variable	Minimum	Maximum	Q1	Q3
Data	1460.0	2406.0	1721.0	2205.0

EXAMPLE ON PERCENTILES

Example:

The test scores for 25 students in a class are shown below.

Test Scores

65 84 95 64 86 96 67 87 98 73 89 99 75

90 100 76 92 96 78 93 100 80 93 93 85

(a) Find the 80th and the 90th percentile of the test score.

Solution: The first step is to arrange the data in increasing order or in an ordered array. The sorted values are shown below.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
64	65	67	73	75	76	78	80	84	85	86	87	89	90	92	93	93	93

19	20	21	22	23	24	25
95	96	96	98	99	100	100

Solution…cont.

(a) ***80th Percentile – the 80th percentile is located at***

$$L_p = (n+1)\frac{P}{100} = (25+1)\frac{80}{100} = 20.8$$

Lp is the location of the percentile, n is the number of observations (n=25 in this case), and P is the required percentile.

To find the value corresponding to the 80th percentile, locate the 20th and 21st value (from the sorted data on the previous page) and determine the distance between these two values. Then multiply this difference by 0.8 and add the result to the smaller value. This will give us the 80th percentile. The value corresponding to the 80th percentile is

$$96 + 0.8(96 - 96) = 96$$

90th Percentile

$$L_p = (n+1)\frac{P}{100} = (25+1)\frac{90}{100} = 23.4$$

The 23rd value in the sorted data is 99 and the 24th value is 100. Therefore, the value corresponding to 90th percentile is

$$99 + 0.4(100 - 99) = 99.4$$

(b) If you are at 80th percentile, what does it mean?

If you are at the 80th percentile, it means that 80% of those who took the test are below you and the rest are above you.

(c) Find Q1, Q2, and Q3

Using the approach in the previous example:

Q1 = 77
Q2 = 89
Q3 = 95.5

You should verify these results.

(d) Find the interquartile range.

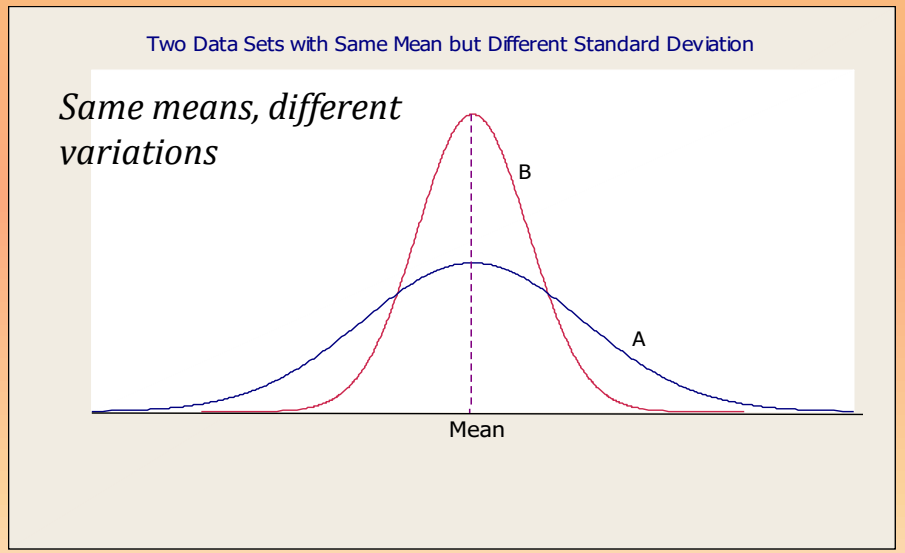
The interquartile range is

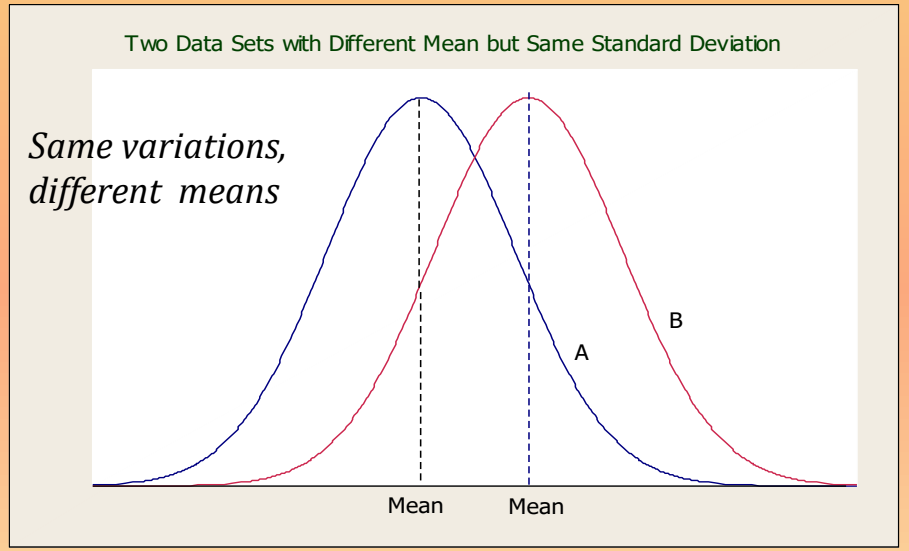
IQ Range = Q3 - Q1 = 95.5 - 77 = 18.5

MEASURES OF VARIATION

The measures of central tendency (mean, median, and mode) are not sufficient to give us a complete description of the data. They must be supported by other measures. These measures are the *measures of variation* or *measures of dispersion*. They tell us about the variation or dispersion of the data values around the average.

Why study variation?





Different Measures of Variation or Dispersion

Measures of variation are used to measure the variability in the data. These are:

(1) **Range**
(2) Variance
(3) Standard Deviation
(4) Coefficient of Variation
(5) Interquartile Range

(1) RANGE

- Range is the simplest measure of variability. It is the difference between the maximum (largest) and minimum (smallest) value in the data set.

$$Range = Maximum\ Value - Minimum\ Value$$

- Range is easy to calculate and understand but its use is limited.
- The range takes into account only the largest and the smallest value and ignores the other observations.
- It is affected by extreme values.
- It takes into account only two values that can change drastically from sample to sample.
- This measure of variation is not suggested for large data set

Example:

The data shows the monthly salaries (in dollars) for 15 employees of a company. Calculate the range.

Monthly Salary ($)

2038	1758	1721	1637	2097	2047	2205
1787	2287	1940	2311	2054	2406	1471
1460						

Solution: It is easier to calculate the range if the data are sorted. The sorted data from the above Table are shown below.

1460 1471 1637 1721 1758 1787 1940 2038 2047 2054
2097 2205 2287 2311 2406

The largest value in the data is 2406 and the smallest value is 1460. Therefore,

Range = 2406 - 1460 = 946 dollars

Note that larger the value of range, larger is the variation.

(2) VARIANCE

- The variance measures the average of the squared deviation of the values from a fixed point, which is mean and can be calculated using both the sample and population data
- The variance calculated from a sample data is called the sample variance (s^2)
- The variance calculated from the population data is known as the population variance(σ^2)
- It is important to note the distinction between the sample and the population variance. They are denoted by different symbols, but have the same concept. They slightly differ in values.

CALCULATING THE SAMPLE VARIANCE

Sample variance is the sum of the squared differences between each of the observations and the mean. It is the average of squared distances.

Suppose we have *n* number of observations $x_1, x_2, \ldots, x_n$

then the variance, s^2 is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots + (x_n - \bar{x})^2}{n-1}$$

or,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (1)$$

Where,

$\bar{x}$ = sample mean

x_i = ith value of the data point x

(note that x_1 is the first value of the data point, x_2 is the second value of the data point and so on)

$\sum (x_i - \bar{x})^2$ is the sum of all squared differences between each of values and the mean.

Another formula to calculate the sample variance

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

$\sum x$ is the sum of the values of the variable,

$\sum x^2$ is the sum of the squared values of the variable, and n is the number of observations

CALCULATING THE SAMPLE VARIANCE, s^2

To calculate the variance using equation (1),

- Calculate the mean of the data
- Obtain the difference between each observation and the mean
- Square each difference
- Add the squared differences
- Divide the sum by (n-1)

Example (1)

The following data represents the price of certain item in dollars

5, 8, 10, 7, 10, 14

Calculate the variance using equation (1).

Solution: First, calculate the sample mean using the formula

$$\bar{x} = \frac{\sum x}{n} = \frac{5+8+10+7+10+14}{6} = 9$$

x_i	$(x_i - \bar{x})^2$
5	$(5 - 9)^2 = 16$
8	$(8 - 9)^2 = 1$
10	$(10 - 9)^2 = 1$
7	$(7 - 9)^2 = 4$
10	$(10 - 9)^2 = 1$
14	$(14 - 9)^2 = 25$
	$\sum (x_i - \bar{x})^2 = 48$

Therefore, the ***sample variance***

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{48}{5} = 9.6(dollars)^2$$

The variance is in squared units which is a difficult configuration to interpret. This is the reason we take the square root of the variance which is the *standard deviation*.

Calculating the variance using equation (2)

The calculation of variance using the equation (2) is explained below.
First, set up a table as shown:

x_i	x_i^2
5	25
8	64
10	100
7	49
10	100
14	196
$\sum x_i = 54$	$\sum x_i^2 = 534$

Using equation (2), the variance can be calculated as

$$s^2 = \frac{\sum x^2 - \left[\frac{\left(\sum x\right)^2}{n}\right]}{n-1} = \frac{534 - \left[\frac{(54)^2}{6}\right]}{5} = \frac{48}{5} = 9.6$$

The variance obtained by this method is the same as using equation (1)

- variance can never be negative
- if all the values in the data set are the same, the variance and standard deviation are zero, indicating no variability
- usually, no random phenomena will ever have the same measured values therefore, it is important to know the variation in the data.

(3) STANDARD DEVIATION

The sample standard deviation (denoted by *s)* is calculated by taking the square root of the variance. The standard deviation can be calculated using the formulas below.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad \text{or,} \qquad s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

Example (2)

Calculate the standard deviation of the data in Example (1).

Solution: To calculate the standard deviation, we first calculate the variance as demonstrated in the examples. Using the variance, the standard deviation can be calculated as

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{48}{5}} = 3.1$$

The sample standard deviation, s = 3.1 dollars

INTERPRETING THE VARIANCE (s^2) & STANDARD DEVIATION (s)

- **The variance and standard deviation measure the average deviation (or the scatter) around the mean.**
- **The variance is the average of squared distances from the mean. In calculating the variance, the computation results in squared units, such as dollar squared, inch squared, etc. This makes the interpretation difficult. Therefore, for practical purposes we calculate the standard deviation by taking the square root of the variance.**
- **Taking the square root of the variance results in the original unit of data (it is no more dollars squared or inch squared but, dollars or inches).**
- **In other words, the variance is the measure of variation affected by the units of measurement, whereas, the standard deviation is measured in the same unit as the data.**
- **In the previous example, s = 3.1 dollars tells us that the average deviation of the price is 3.1 dollars.**

(4) Coefficient of Variation (CV)

The coefficient of variation is a relative measure of dispersion expressed as percentage. It tells us how large the standard deviation is in relation to the mean. It is calculated using the following formula:

Sample coefficient of variation (CV)

$$CV = \frac{S\tan dardDeviation}{Mean} * 100\%$$

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

Population coefficient of variation (CV)

$$CV = \frac{\sigma}{\mu} * 100\%$$

EXAMPLE : COEFFICIENT OF VARIATION (CV)

Data below are the price of certain item in dollars: 5, 8, 10, 7, 10, 14
Where, n = 6 (the number of observations). Calculate the coefficient of variation (CV).

Solution:

The mean and the standard deviation of the data were calculated in the example above. Recall that the mean for the data was

$$\bar{x} = \frac{\sum x}{n} = \frac{54}{6} = 9$$

and the standard deviation, s = 3.1 dollars (see the previous example for the calculation of standard deviation). Therefore, the coefficient of variation (CV)

$$CV = \frac{s}{\bar{x}} * 100\% = \frac{3.1}{9.0} * 100\% = 34.44\%$$

This tells us that the standard deviation is 34.44% of the sample mean. Note that the coefficient of variation is expressed as a percent, which means it has no unit.

(5) Interquartile Range

- Interquartile range is another measure of variation that is calculated by taking the difference between the third quartile (Q_3) and the first quartile (Q_1).
- The third quartile is the 75th percentile and the first quartile is the 25th percentile. The interquartile range or IQR is given by

$$IQR = Q_3 - Q_1$$

The interquartile range is the range of the middle 50% of the values in the data set. This is a better measure of variability than the simple range because it avoids the extreme values.

EXAMPLE

Find the first quartile, the third quartile, and the interquartile range (IQR) for the monthly salary data of 20 employees in Table 3.8.

2038	1758	1721	1637	2097	2047	2205
1787	2287	1940	2311	2054	2406	1471
1460	1500	2250	1650	2100	1850	

(5) Interquartile Range…cont.

Solution: *First, sort the data*

1460	1471	1500	1637	1650	1721	1758	1787
1850	1940	2038	2047	2054	2097	2100	2205
2250	2287	2311	2406				

The number of observations, n=20. To calculate the first quartile (Q1) or 25th percentile, first find the location of Q1 first using

$$L_p = (n+1)\frac{P}{100} = 21(\frac{25}{100}) = 5.25$$

Q1 is the 5th value in the sorted data + 0.25, times the distance between the 5th and the 6th value, which is

1650 + (0.25) (1721-1650) =1667.75

Therefore,

Q1 = 1667.75 or 1667.8

(5) Interquartile Range…cont.

Similarly, to calculate the third quartile (Q3) or 75th percentile, find the location of Q3 using

$$L_p = (n+1)\frac{P}{100} = 21(\frac{75}{100}) = 15.75$$

Thus, Q3 is the 15th value in the sorted data + 0.75 times the distance between the 15th and the 16th value, which is

2100 + (0.75) (2205-2100) = 2178.75

Therefore,

$$Q_3 = 2178.75 \text{ or } 2178.8$$

The interquartile range,

$$IQR = Q_3 - Q_1 = 2178.8 - 1667.8 = 511$$

Summary of Formulas:

Sample mean : $\bar{x} = \dfrac{\sum x_i}{n}$

Population mean : $\mu = \dfrac{\sum x_i}{N}$

Population variance : $\sigma^2 = \dfrac{\sum (x_i - \mu)^2}{N}$

Sample variance: $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

Sample variance can also be calculated using:

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

Summary of Formulas

Sample standard deviation : $s = \sqrt{s^2}$
Population standard deviation: $\sigma = \sqrt{\sigma^2}$
Coefficient of variation (sample): $c.v = \frac{s}{\bar{x}} * 100$ ***Coefficient of variation (population):*** $CV = \frac{\sigma}{\mu} * 100$ ***Interquartile range :*** $IQR = Q_3 - Q_1$

Relationship between the Mean and Standard Deviation

There are two important rules that describe the relationship between the mean and standard deviation. These are

- **Chebyshev's Theorem, and**
- **Empirical Rule**

CHEBYSHEV'S THEOREM

This theorem states that ***no matter what the shape of the distribution*** (symmetrical or skewed),

- at least 75% of all observations will fall within ± 2 standard deviations of the mean
- at least 89% of the observations will fall within ± 3 standard deviations of the mean
- at least 94% of the observations will fall within ±4 standard deviations of the mean

Example : Chebyshev's Theorem

Suppose a sample of 100 students (n=100) were given a statistics test. The average or the mean of the test score was 80 with a standard deviation 5. Then according to Chebyshev's rule, *at least* 75 students will have a score between 80±2(5) or between 70 and 90; *at least* 89 of those who took the test will have scores between 80±3(5) or 65 and 95; and *at least* 94 of the students will have scores between 80±4(5) or 60 and 100. These percentages are irrespective of the shape of the test score data. The term "***at least***" in the theorem statement makes it very general.

GENERAL FORM OF CHEBYSHEV'S THEOREM

Within k standard deviation of the mean, at least $(1-\frac{1}{k^2})$ *percent of the values occur.*

where, k is given by

$$k = \frac{x-\bar{x}}{s} \; or, k = \frac{x-\mu}{\sigma}$$

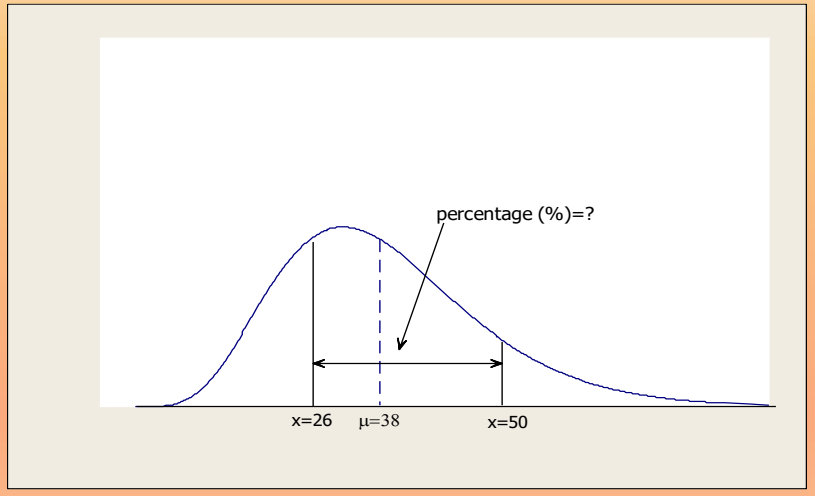
k determines how far the data value is from the mean, in terms of standard deviation units. In this equation: s= sample standard deviation/ σ = population standard deviation

x_i = data values, $\bar{x}$ = sample mean , μ= population mean

Example : Chebyshev's Theorem

Suppose that the distribution of a data is skewed with mean, μ = 38 and the standard deviation, σ = 6.What proportion of the values would fall between 26 and 50?

Solution: The situation is explained below



The percent of values between 26 and 50 can be found using the Chebyshev's theorem. This theorem states that the percentage of the observations within k standard deviations of the mean is

$$(1-\frac{1}{k^2})$$

k is determined using

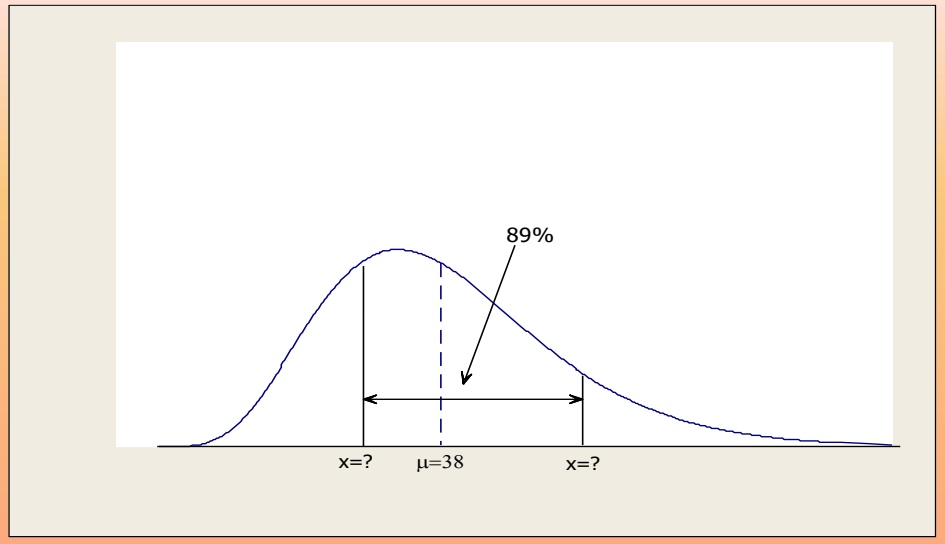
$$k=\frac{x-\mu}{\sigma}=\frac{26-38}{6}=-2$$

$$k=\frac{x-\mu}{\sigma}=\frac{50-38}{6}=+2$$

Therefore, the value of k = ±2 and the required percentage is $\left(1-\frac{1}{k^2}\right)=1-\frac{1}{(2)^2}0.75$

Example : Chebyshev's Theorem…cont.

Between what two values at least 89% of the observations fall?
The situation is explained below



Given

$$(1-\frac{1}{k^2})=0.89$$

Solving this equation for k yields,
k =±3.015.

We know that the mean μ = 38 and the standard deviation σ = 6. The two values can be calculated using

$$\mu \pm k\sigma$$
$$38 \pm 3.015(6)$$
$$38 \pm 18.09$$
$$19.91 and 56.09$$

Therefore, 89% of the observations will fall between 19.91 and 56.09.

Empirical rule

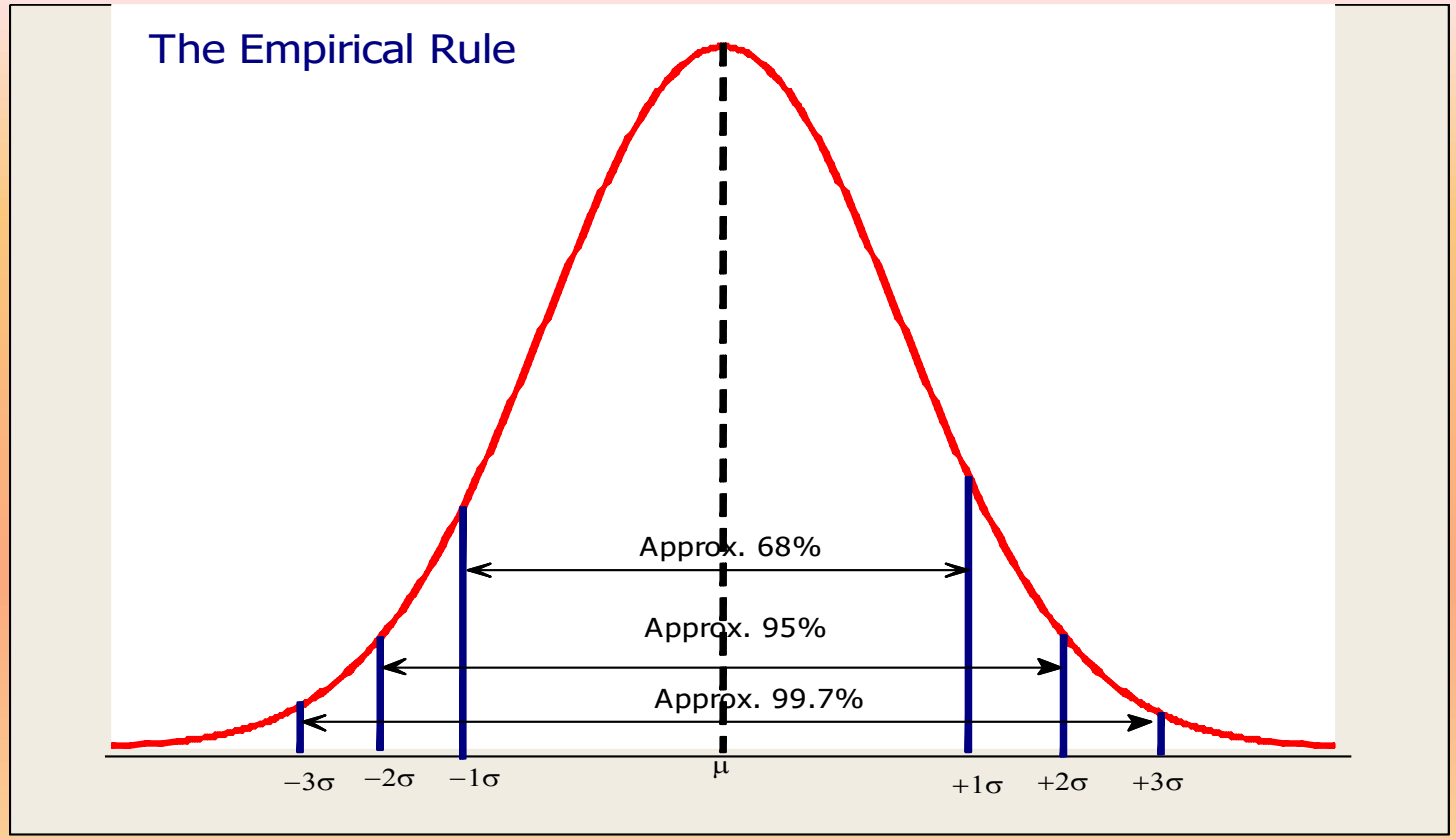
The empirical rule applies to ***symmetrical or bell shaped distribution*** also known as the ***normal distribution. Empirical rule combines the mean and standard deviation and provides one of the most useful statistical results.***

The empirical rule states that if the data are symmetrical or bell shaped:

- ***approximately 68% of the observations will lie between the mean and ± 1 standard deviation***
- ***approximately 95% of the observations will lie between the mean and ± 2 standard deviations***
- ***approximately 99.7% of the observations will lie between the mean and ± 3 standard deviations.***

$\mu \pm 1\sigma$	will contain approximately 68% of the observations
$\mu \pm 2\sigma$	will contain approximately 95% of the observations
$\mu \pm 3\sigma$	will contain approximately 99.7% of the observations

Empirical Rule – Graphically



Empirical Rule – Example

Consider a data set with bell-shaped or symmetrical distribution with mean μ = 80 and standard deviation σ=5. Determine the proportion of observations within each of the following ranges: (a)75 and 85 (b) 70 and 90 (c) 65 and 95.

Solution: Since the data follow a bell-shaped or symmetrical distribution, we can apply the empirical rule to find the percent of observation within each of the range.

[a] Note that the range 75 to 85 is within one standard deviation of the mean. That is,

$$\mu \pm 1\sigma = 80 \pm 5 = 75, 85$$

From the empirical rule we know that the mean ± 1 standard deviation contains approximately 68% of the observations. ***Therefore, approximately 68% of the observations will be contained within 75 and 85.***

(b)The range 70 to 90 is within two standard deviation of the mean. That is,

$$\mu \pm 2\sigma = 80 \pm (2)5 = 70, 90$$

Empirical Rule – Example..cont.

Again, from the empirical rule we know that mean ± 2 standard deviations contain approximately 95% of the observations. Therefore, approximately 95% of the observations will fall between 70 and 90.

(c) The range 65 to 95 is within three standard deviation of the mean. That is,

$$\mu \pm 3\sigma = 80 \pm (3)5 = 65, 95$$

From the empirical rule we know that mean ± 3 standard deviation contains approximately 99.7% of the observations. Therefore, approximately 99.7% of the observations will fall between 65 and 95.

Using the empirical rule, we can easily determine the percent of observations as long as the values are within one, two, or three standard deviations from the mean.

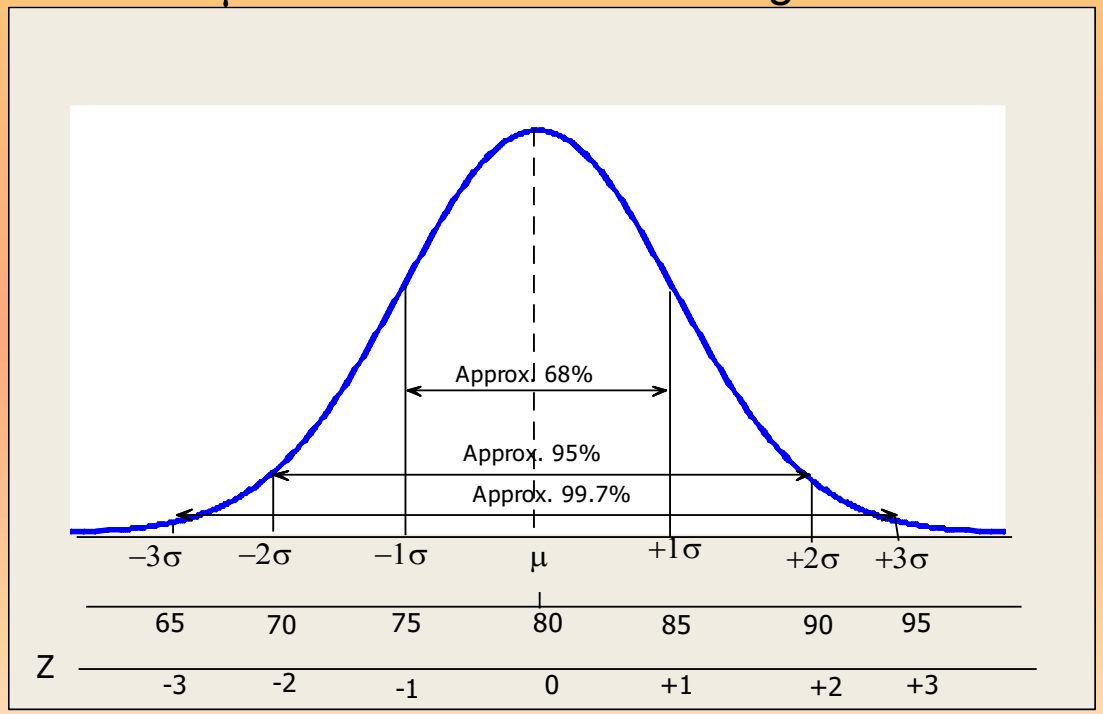
How do we find the percent when the values or the points of interest are other than one, two, or three standard deviations from the mean? This can be determined by using a formula known as the z-score formula.

Z-SCORE

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

where, z = distance from the mean to the point of interest in terms of standard deviation unit, x = point of interest, μ = mean, σ = standard deviation

Determine z-score within each of the following ranges: (a)75 and 85 (b) 70 and 90 (c) 65 and 95 when μ= and σ = 5. See the figure below.



Z-scores

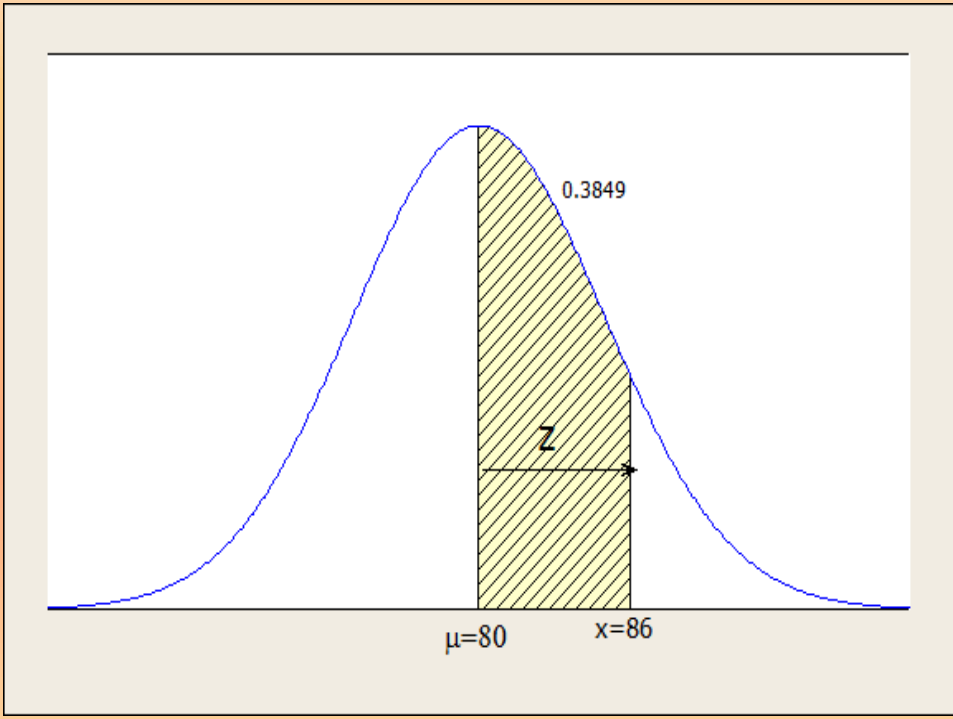
Application of z-score

Suppose that the test scores in a statistics class have bell-shaped or symmetrical distribution with mean μ = 80 and standard deviation σ=5. determine the percentage of students who have a score between 80 and 86?

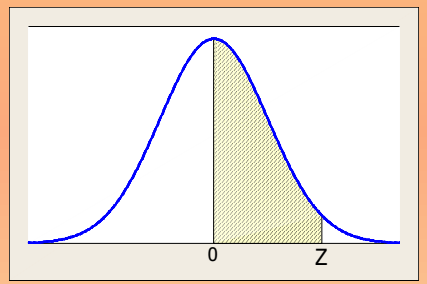
SOLUTION

Note that the value 85 is +1 standard deviation from the mean [see Figure on the previous slide], therefore, 86 will be slightly higher than 1 standard deviation. To know how far the value 86 is from the mean, we use ***the z-score formula***. This can be found using the z-score formula:

$$z = \frac{x - \mu}{\sigma} = \frac{86 - 80}{5} = 1.2$$



Z=1.2 means that the value 86 is 1.2 standard deviations from the mean. According to the empirical rule approximately 68% of the observations falls within ± 1 standard deviation of the mean. This means that + 1 standard deviation will contain approximately half of 68% or approximately 34%. So what percentage corresponds to z= +1.2 standard deviations? It must be higher than 34%. To know the exact percentage, we need to look into the ***Standard Normal Table or z-table*** shown below. This percentage is 0.3849 or 38.49%



Part of a Standard Normal Table

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441

Using the z-score formula, we can determine how far any value is from the mean in terms of standard deviations. Figure on the previous page shows the z-scores for various values from the mean (for example, how far is the value 85 in standard deviations). Note that z=1.0 means that the value is one standard deviation from the mean. The other values of z-score can be interpreted in the same way.

If the z-score is ±1, ±2, or ±3, we know the percent for these from the empirical rule. If the z-score value is any value other than one, two, or three - the percent of observations corresponding to those values can be found using a standard normal table.

Exploratory Data Analysis: Box plot

A box plot uses a ***five-number summary*** as a graphical representation of data. These five numbers are

- The smallest or the minimum data value
- Q1: the first quartile, or 25th percentile
- Q2: the second quartile, or the median or 50th percentile
- Q3: the third quartile, or 75th percentile
- The largest or the maximum data value

EXAMPLE OF A BOX PLOT

The utility bill for 50 customers (n=50) rounded to the nearest dollar was collected. The data were sorted using a computer software. Table below shows the sorted data. Construct a box plot of the utility bill data

Sorted data

82	90	95	96	102	108	109	111	114	116
119	123	127	128	129	130	130	135	137	139
141	143	144	147	148	149	149	150	151	153
154	157	158	163	165	166	167	168	171	172
175	178	183	185	187	191	197	202	206	213

Descriptive Statistics for utility bill

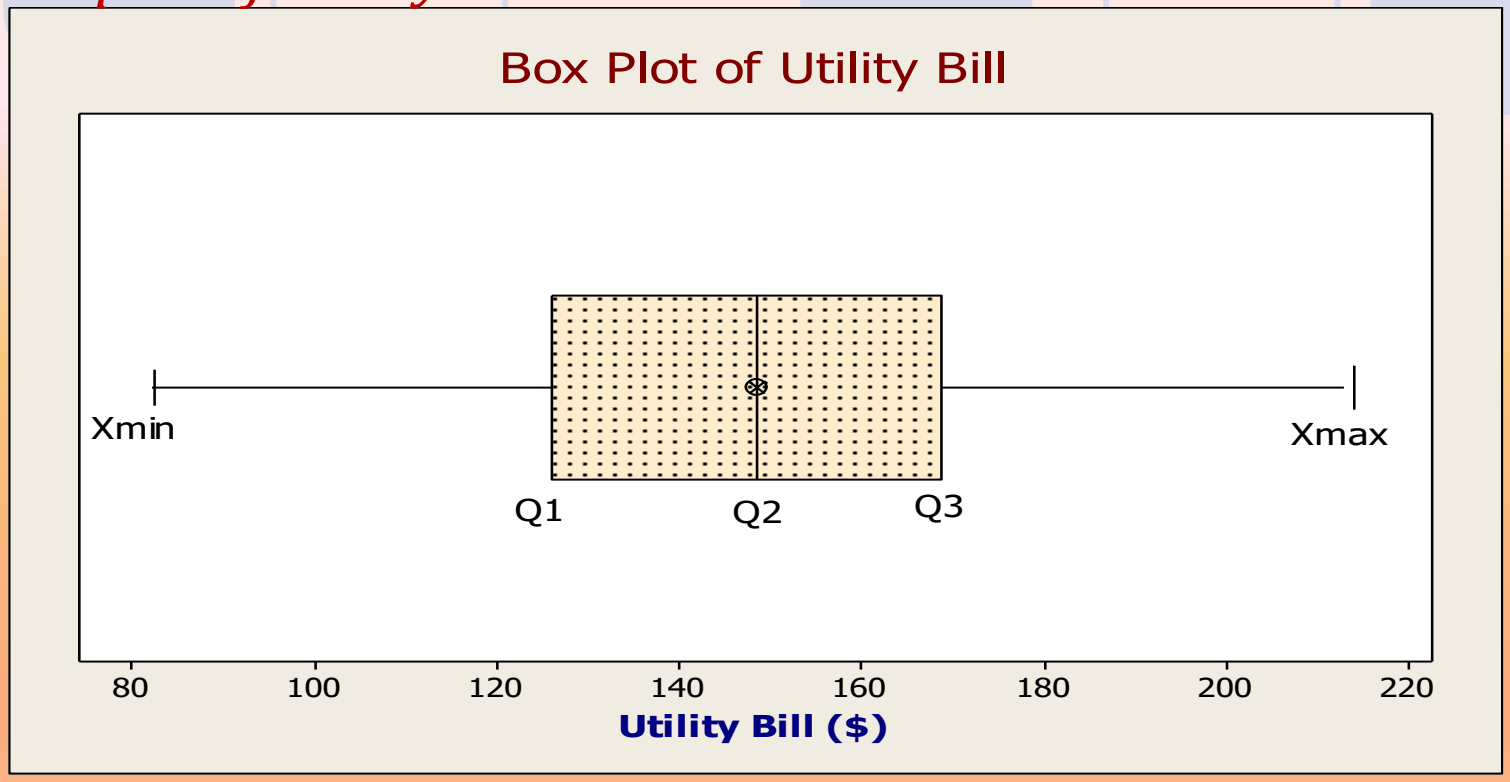
Descriptive Statistics: C1

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Utility Bill	50	147.06	148.50	146.93	31.69	4.48

Variable	Minimum	Maximum	Q1	Q3
Utility Bill	82.00	213.00	126.00	168.75

The highlighted values in the above table are plotted as a box plot

Box plot of Utility Data



From the box plot, the shape of the data can be determined. In this plot, Q1, Q2, and Q3 are enclosed in a box. Q2 is the median. If Q2 or the median divides the box in approximately two halves, and if the distance from the X_{min} to Q1 and Q3 to X_{max} are equal or approximately equal, then the data are symmetrical.

Descriptive Statistics using Excel

Table below shows the utility bill for 50 customers rounded to the nearest dollars.

Sorted data

82	90	95	96	102
108	109	111	114	116
119	123	127	128	129
130	130	135	137	139
141	143	144	147	148
149	149	150	151	153
154	157	158	163	165
166	167	168	171	172
175	178	183	185	187
191	197	202	206	213

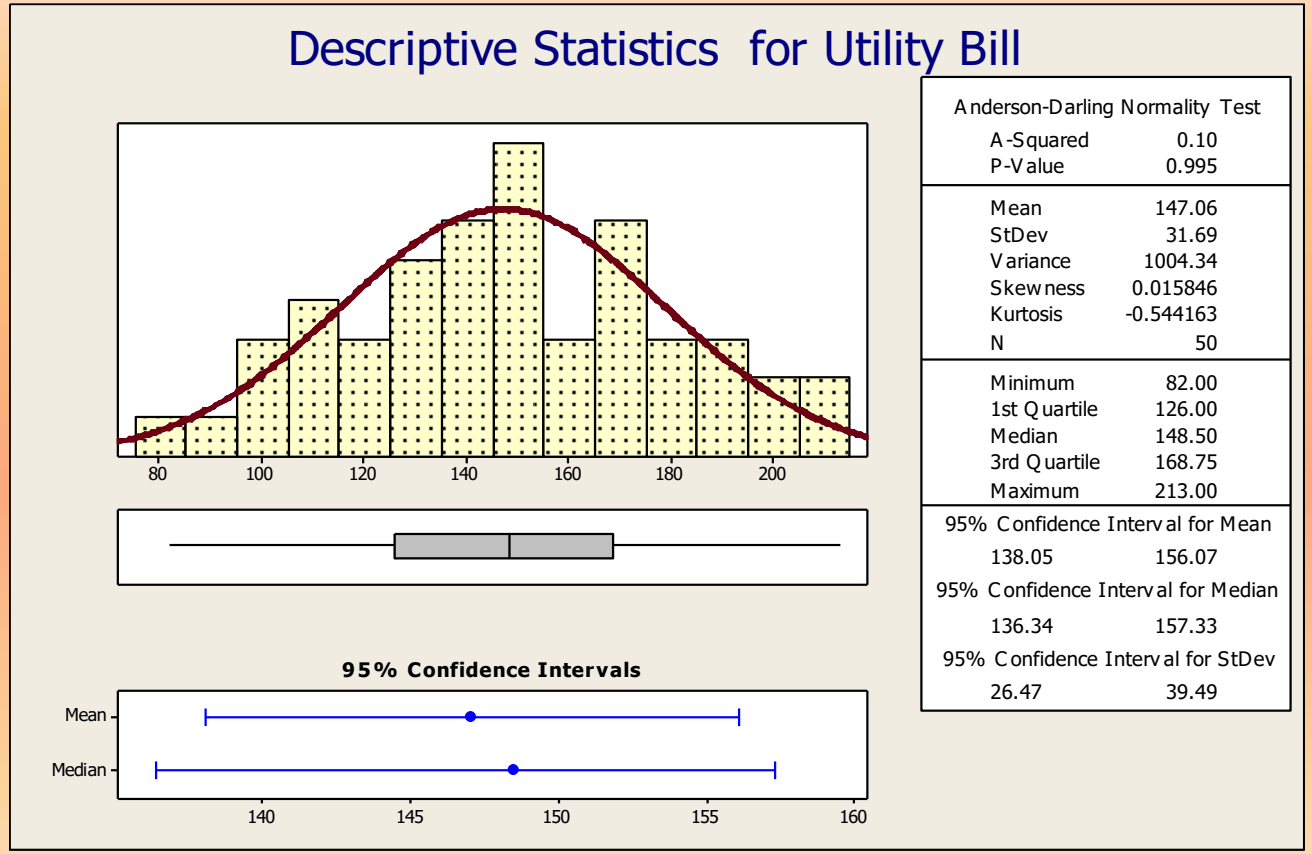
The steps to calculate these statistics using EXCEL are described in appendix of Chapter 3.

The descriptive of the data using EXCEL

Utility Bill	
Mean	147.06
Standard Error	4.481837269
Median	148.5
Mode	130
Standard Deviation	31.69137525
Sample Variance	1004.343265
Kurtosis	-0.544163238
Skewness	0.015845641
Range	131
Minimum	82
Maximum	213
Sum	7353
Count	50

DESCRIPTIVE STATISTICS USING MINITAB

Figure shows the descriptive statistics along with the histogram and the box plot of the utility bill data on the previous page. This summary was created using MINITAB. Such a summary statistics with the graphs are very useful in obtaining the detailed description of data.



Measures of Association between two Quantitative Variables: the Covariance and the Coefficient of Correlation

THE COVARIANCE

The covariance is a measure of strength of linear relationship between two quantitative variables x and y. For n observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ the sample covariance is defined using the following relationship:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

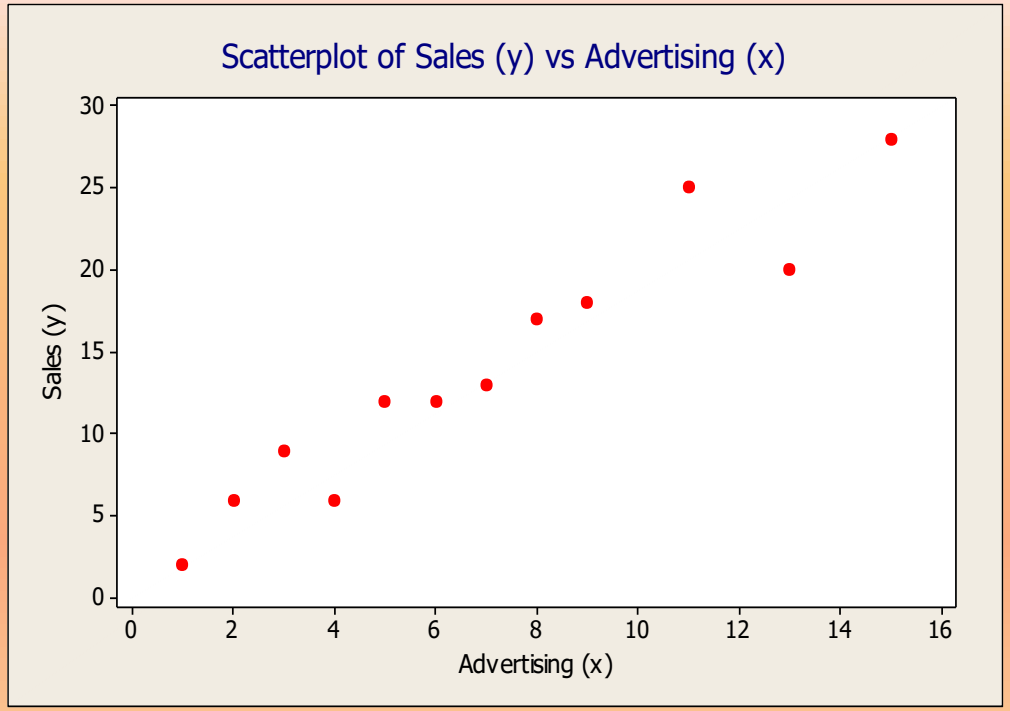
EXAMPLE

Table 3.25 shows the advertising expenditures and the corresponding sales for 12 companies. Both the sales and advertising are in millions of dollars.

Advertising (x)	1	2	11	9	7	6	15	3	13	5	4	8
Sales (y)	2	6	25	18	13	12	28	9	20	12	6	17

Solution:

The scatterplot of the data shows an increase in advertising expenditure with an increase in sales. This indicates a positive relationship between sales and advertising.



The scatter plot shows a positive relationship between x and y ; that is, as the advertising expenditure (x) increases, the value of sales (y) also increases. This shows a positive covariance which is confirmed by the calculated value of Sxy= 33.18 (on the next slide)

CALCULATIONS FOR COVARIANCE

$$\bar{x} = \frac{\sum x}{n} = \frac{84}{12} = 7$$

$$\bar{y} = \frac{\sum y}{n} = \frac{168}{12} = 14$$

Calculations for Covariance

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	1	2	(1-7)= -6	(2-14)= -12	72
	2	6	(2-7)= -5	(6-14)= -8	40
	11	25	(11-7)= 4	(25-14)= 11	44
	9	18	(9-7)= 2	(18-14)= 4	8
	7	13	(7-7)= 0	(13-14)= -1	0
	6	12	(6-7)= -1	(12-14)= -2	2
	15	28	(15-7)= 8	(28-14)= 14	112
	3	9	(3-7)= -4	(9-14)= -5	20
	13	20	(13-7)= 6	(20-14)= 6	36
	5	12	(5-7)= -2	(12-14)= -2	4
	4	6	(4-7)= -3	(6-14)= -8	24
	8	17	(8-7)= 1	(17-14)= 3	3
Totals	84	168	0	0	365

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{365}{11} = 33.18$$

INTERPRETATION OF COVARIANCE:

- The positive value of Sxy indicates a positive linear relationship between x and y . This means that as the value of x increases, the value of y also increases.
- A negative value of Sxy is an indication of a negative linear relationship between x and y . If the covariance is negative, the value of y decreases as the value of x increases.
- A value of Sxy close to zero indicates no or very weak relationship between x and y .

LIMITATION OF COVARIANCE

A large positive value of the covariance does not mean a strong positive linear relationship between x and y . Similarly, a large negative value of the covariance does not necessarily mean a strong negative linear relationship.

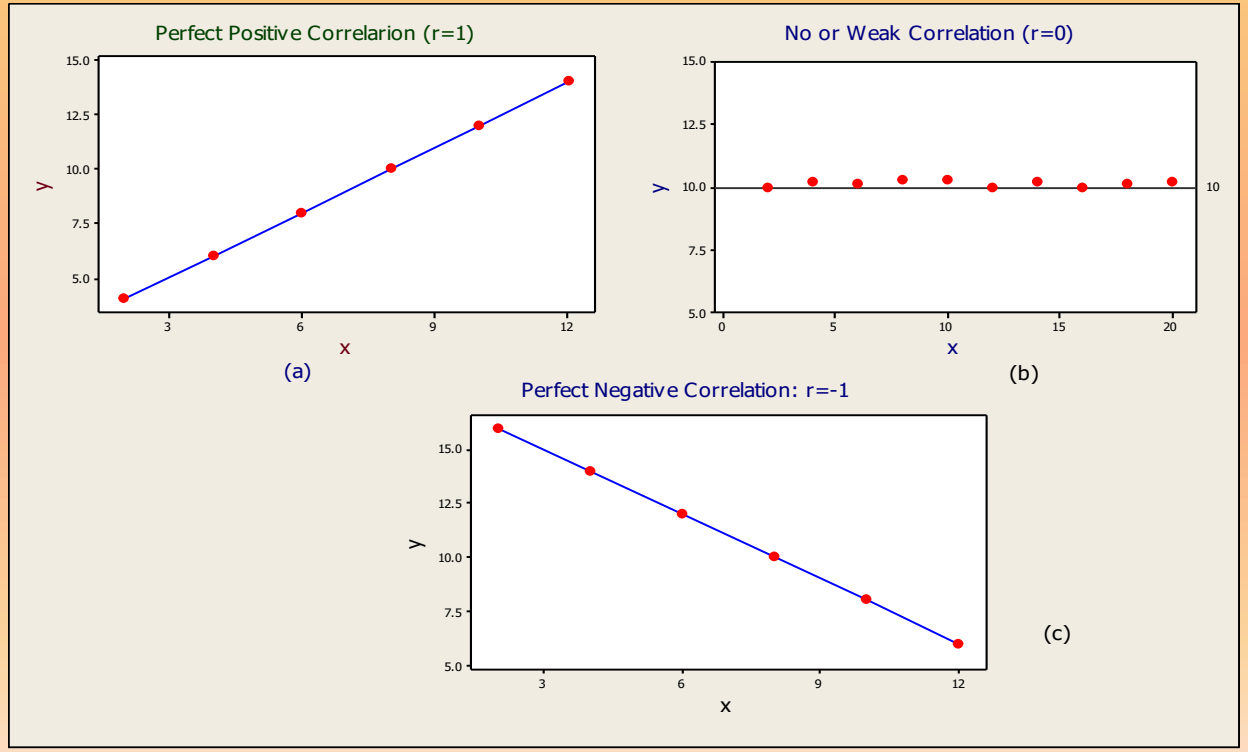
The value of the covariance is a quantity that depends on the units of measurement for x and y .

A better measure relationship between two quantitative variable is correlation coefficient or coefficient of correlation.

The Coefficient of Correlation

The sample coefficient of correlation (r_{xy}) is a measure of relative strength of a linear relationship between two quantitative variables. This is a unit less quantity.

The coefficient of correlation has a value between -1 and +1 where a value of -1 indicates a perfect negative correlation and a value of +1 indicates a perfect positive correlation.



Calculating the Coefficient of Correlation

The sample coefficient of correlation can be calculated using the following equation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where: r_{xy} = sample coefficient of correlation

s_{xy} = sample covariance

s_x = sample standard deviation of x

s_y = sample standard deviation of y

Note:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

EXAMPLE: COEFFICIENT OF CORRELATION

Calculate the sample coefficient of correlation for the data given below (x and y values are the given data where x is the advertising dollars and y is the sales both in millions of dollars).

	x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	1	2	$(1-7)^2 = 36$	$(2-14)^2 = 144$
	2	6	$(2-7)^2 = 25$	$(6-14)^2 = 64$
	11	25	$(11-7)^2 = 16$	$(25-14)^2 = 121$
	9	18	$(9-7)^2 = 4$	$(18-14)^2 = 16$
	7	13	$(7-7)^2 = 0$	$(13-14)^2 = 1$
	6	12	$(6-7)^2 = 1$	$(12-14)^2 = 4$
	15	28	$(15-7)^2 = 64$	$(28-14)^2 = 196$
	3	9	$(3-7)^2 = 16$	$(9-14)^2 = 25$
	13	20	$(13-7)^2 = 36$	$(20-14)^2 = 36$
	5	12	$(5-7)^2 = 4$	$(12-14)^2 = 4$
	4	6	$(4-7)^2 = 9$	$(6-14)^2 = 64$
	8	17	$(8-7)^2 = 1$	$(17-14)^2 = 9$
Totals			$\sum (x_i - \bar{x})^2 = 212$	$\sum (y_i - \bar{y})^2 = 684$

EXAMPLE: COEFFICIENT OF CORRELATION…CONT.

Note: The following values can be calculated from the table on the previous page

$$\bar{x} = \frac{\sum x}{n} = \frac{84}{12} = 7 \qquad \bar{y} = \frac{\sum y}{n} = \frac{168}{12} = 14$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{212}{11}} = 4.39 \qquad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{684}{11}} = 7.89$$

From the previous example, $s_{xy} = 33.18$, the sample coefficient of correlation is

The Sxy is the covariance. The calculation of Sxy is shown in the previous example. Using all the above values, the coefficient of correlation:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{33.18}{(4.39)(7.89)} = +0.96$$

Alternate way of Calculating coefficient of Correlation…cont.

We will use the data of previous example to calculate the coefficient of correlation using an alternate formula. First set up the table and perform the necessary calculations from the given values of x and y.

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	1	2	2	1	4
	2	6	12	4	36
	11	25	275	121	625
	9	18	162	81	324
	7	13	91	49	169
	6	12	72	36	144
	15	28	420	225	784
	3	9	27	9	81
	13	20	260	169	400
	5	12	60	25	144
	4	6	24	16	36
	8	17	136	64	289
Totals	$\sum x_i = 84$	$\sum y_i = 168$	$\sum x_i y_i = 1541$	$\sum x_i^2 = 800$	$\sum y_i^2 = 3036$

Using the values on the previous page and the formula shown below, the coefficient of correlation:

$$r_{xy} = \frac{\sum x_i y_i - \left(\frac{\sum x_i \sum y_i}{n} \right)}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} * \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}} = \frac{1541 - \left(\frac{(84)(168)}{12} \right)}{\sqrt{800 - \frac{(84)^2}{12}} * \sqrt{3036 - \frac{(168)^2}{12}}} = \frac{365}{(14.56)(26.15)} = 0.96$$

This value is the same as the one obtained using the formula in the previous example.

EXAMPLES OF COEFFICIENT OF CORRELATION

Figures (a) through (d) show several scatterplots with the correlation coefficient.

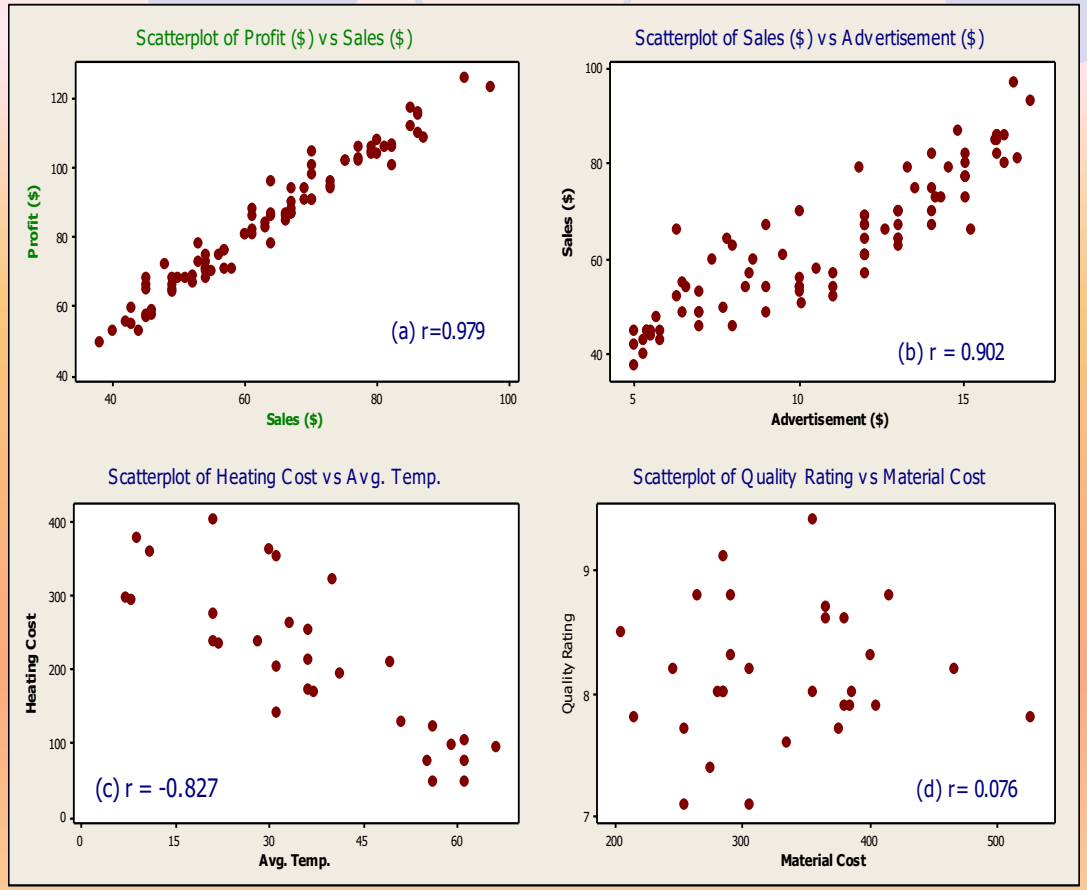


Figure (a): positive correlation between sales and profit with a correlation coefficient value r= + 0.979.

Figure(b): a positive relationship between the sales and advertisement expenditures with a calculated correlation coefficient, r= +0.902.

Figure(c): shows a negative relationship between the heating cost and the average temperature with coefficient of correlation (r=-0.827)

Figure(d): a weak relationship between the quality rating and the material cost with r=0.076.

Measures of Shape: Skewness and Kurtosis

SKEWNESS (s_k)

Skewness is lack in symmetry. It is a measure of departure from symmetry. If the skewness, is zero, the data are symmetrical; if it is greater than zero (positive), the data are positively or right skewed and if the skewness is less than zero (negative), the data are negatively or left skewed. The value of skewness has a range from – 3 to 3.

KURTOSIS

Kurtosis measures the peakedness of the data. It is used to measure the height of the peak in the distribution. A ***laptokurtic*** distribution is more peaked than ***platykurtic*** (flatter distribution). Between ***laptokurtic*** and ***platykurtic*** is ***mesokurtic,*** which is the normal distribution. The kurtosis, K_r does not provide any information by itself. Instead, it must be compared to other distribution.

The formulas and calculations of skewness is explained in Chapter 3.

Grouped Data – MEAN FOR THE GROUPED DATA

Given the following Frequency Distribution created from lifetime of 200 television components. Calculate the mean for this grouped data.

Class- interval	Frequency (f)
200 - 230	4
230 - 260	11
260 - 290	30
290 - 320	46
320 - 350	49
350 - 380	40
380 - 410	14
410 - 440	6
	$\sum f = 200$

The sample mean for the grouped data is calculated using the following formula:

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

f_i = the frequency of class i (i=1 is class interval 1, i=2 is class interval 2 and so

M_i =midpoint of class i, and n= number of observations which is the same as $\sum f$

The midpoint is calculated by: (Lower class limit +Upper class limit)/2

MEAN…CONT.

The calculations for the mean of grouped data are explained in the table below.

Class Interval	Midpoint (M)	Frequency ((f_i)	fM
200 - 230	215	4	860
230 - 260	245	11	2695
260 - 290	275	30	8250
290 - 320	305	46	14030
320 - 350	335	49	16415
350 - 380	365	40	14600
380 - 410	395	14	5530
410 - 440	425	6	2550
		$\sum f = 200 = n$	$\sum fM = 64930$

The sample mean $$\bar{x} = \frac{\sum fM}{n} = \frac{64930}{200} = 324.65$$

MEDIAN FOR THE GROUPED DATA

The sample median for the grouped data is:

$$M_d = L + \left[\frac{(n+1)/2 - F}{f_m}\right] w$$

where, M_d = median

L = lower limit of the ***median class***

n= number of observations

F= sum of the frequencies up to but not including the *median class*

f_m = frequency of the *median class*

w = width of the class

To calculate the median, we first need to calculate the ***median class***. Median class is the class that contains the median. We use the frequency distribution or grouped data shown in the next slide to calculate the median. Before calculating the median, calculate the cumulative frequency as shown and the median class

MEDIAN ...cont.

Class-interval	Frequency, f	Cumulative Freq.
200 - 230	4	4
230 -260	11	15
260 - 290	30	45
290 - 320	46	91
320 - 350	**49**	**140**
350 - 380	40	180
380 - 410	14	194
410 - 440	6	200

$\sum f = 200 = n$

Determine the Median Class

The first cumulative frequency > n/2 contains the median class.

For our example, n=200 and n/2 = 100. The first cumulative frequency greater than 100 is 140, which is in the 320 -350 class (table above). ***Therefore, the median class is 320 -350 -*** the median is contained in this class.

MEDIAN ...cont.

Once the median class is known, determine the following values

l = lower limit of the ***median class*** = **320**
n= number of observations = **200**
f= sum of the frequencies up to but not including the *median class* = **91**
f_m = frequency of the *median class* = **49**
w = width of the class = **30**
Therefore, the median:

$$M_d = L + \left[\frac{(n+1)/2 - F}{f_m}\right] w = 320 + \left[\frac{201/2 - 91}{49}\right] 30 = 325.81 \quad \text{or,} \quad M_d = 325.81$$

MODE FOR THE GROUPED DATA

In a grouped data, the mode is the average of the ***modal class***. The modal class is the class with maximum frequency. For the grouped data on the previous slide, the mode can be calculated as

$$Mode = \frac{320 + 350}{2} = 335$$

Range, variance, standard deviation, and coefficient of variation for the grouped data

We will calculate the above measures for the following frequency distribution

Class interval (Test Scores)	Frequency (f) Number of students
30 -40	4
40 -50	5
50 -60	6
60- 70	10
70 - 80	12
80 - 90	10
90 - 100	3
	$\sum f = 50$

RANGE

The range is calculated using the following formula:

Range = (Upper limit of the last class – lower limit of the first class)

$$Range = 100 - 30 = 70$$

SAMPLE VARIANCE

The sample variance for the grouped data is calculated using the following formula:

$$s^2 = \frac{\sum fM^2 - n\bar{x}^2}{n-1}$$

f = frequency, M = midpoint, $\bar{x}$ = mean, n= number of observations

(1) Class interval	(2) Midpoint (M)	(3) Frequency (f)	(4) fM	(5) M^2	(6) fM^2
30 - 40	35	4	140	1225	4900
40 - 50	45	5	225	2025	10125
50 - 60	55	6	330	3025	18150
60 - 70	65	10	650	4225	42250
70 - 80	75	12	900	5625	67500
80 - 90	85	10	850	7225	72250
90 - 100	95	3	285	9025	27075
		$\sum f = 50$	$\sum fM = 3300$		$\sum fM^2 = 242250$

$$\bar{x} = \frac{\sum fM}{n} = \frac{3300}{50} = 66 \text{ and } s^2 = \frac{\sum fM^2 - n\bar{x}^2}{n-1} = \frac{242250 - (50)(66)^2}{49} = \frac{242250 - 217800}{49} = 498.99$$

SAMPLE STANDARD DEVIATION

The standard deviation, s is calculated by taking the square root of the variance. Therefore,

$$s = \sqrt{s^2} = \sqrt{498.99} = 22.34$$

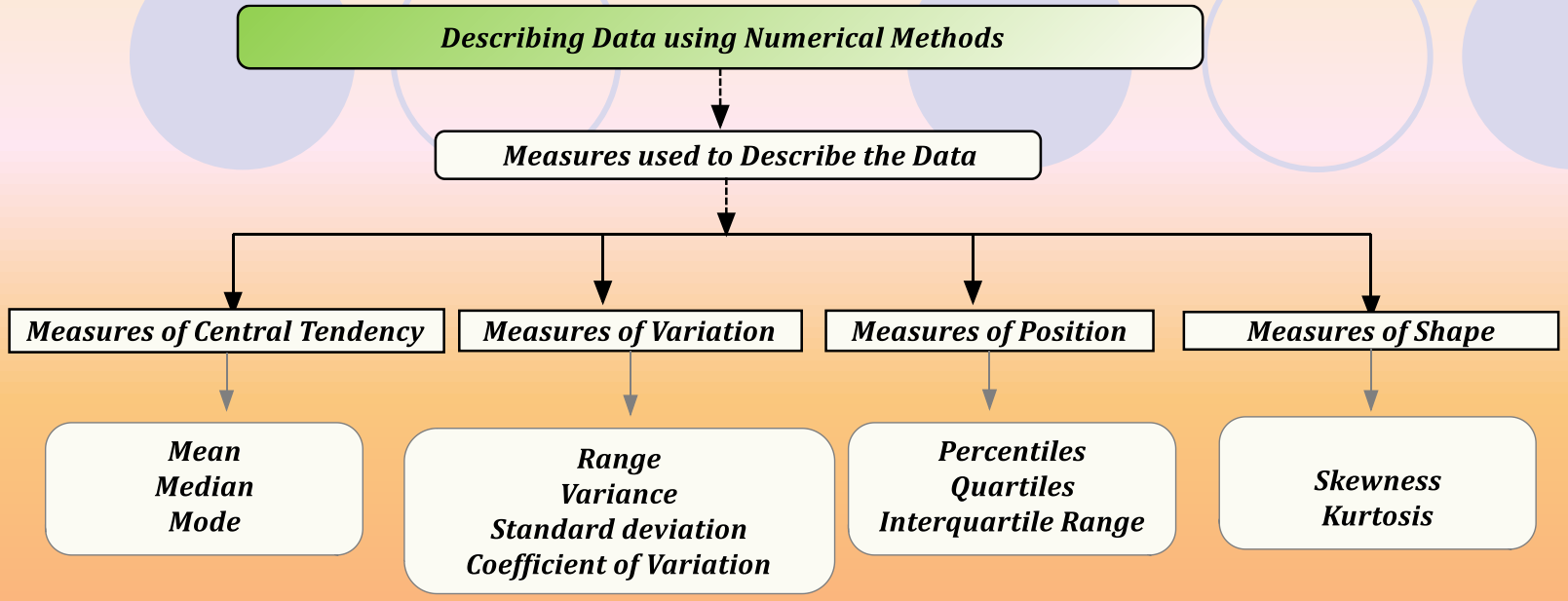
COEFFICIENT OF VARIATION (CV)

The coefficient of variation is calculated by

$$CV = \frac{s}{\bar{x}} * 100\% = \frac{22.34}{66} * 100\% = 33.85\%$$

This means that the standard deviation is 33.85% of the mean.

The above example demonstrates how to calculate the measures of variation for the grouped data.



Population Parameters

μ = population mean

σ^2 = population variance

σ = population standard deviation

p= population proportion

N = size of the population

* μ is read as "mu" and σ is read as "sigma."

Sample statistics

$\bar{x}$ = sample mean

s^2 = sample variance

s = sample standard deviation

$\bar{p}$ = sample proportion

n = sample size

*($\bar{x}$ is read as "x-bar")

Chapter 3 : Different Measures of Describing Data - Flow Chart (1)

Descriptive Statistics : Numerical Methods

Measures of Central Tendency

Sample mean : $\bar{x} = \frac{\sum x_i}{n}$ *Population Mean:* $\mu = \frac{\sum x_i}{N}$

Median:

- Median is the middle value after the values have been arranged in ascending (or descending) order of magnitude.
- There is a distinct median when the number of observations is odd.
- Median is not affected by extreme values

When the number of observations is even

- there are two middle values and the median is obtained by taking the arithmetic mean of the middle terms

Mode:

Mode is the value that occurs most frequently in a set of observations.

Measures of Position

Location of Any Percentile

- Arrange the data in increasing order
- Find the location of the percentile using the following formula:

$$L_p = (n + 1)\frac{P}{100}$$

L_p = location of the percentile

n= total number of observations

P= desired percentile

QUARTILES: The ***quartiles*** divide the data into four parts. For a large data set, it is often desirable to divide the data into four parts. This can be done by calculating the quartiles. The quartiles are defined as

Q_1 = 1st quartile or the 25th percentile

Q_2 = 2nd quartile or the 50th percentile (median)

Q3 = 3rd quartile or the 75th percentile

Chapter 3: Measures of Central Tendency and Measures of Location- Flow Chart (2)

Descriptive Statistics: Numerical Methods...continued

Measures of Variation/Dispersion

Measures of variation or dispersion

(1) Range (2) Interquartile range (3) Variance
(4) Standard deviation (5) Coefficient of variation

Formulas for sample data

Range = *(largest value - smallest value)*

Interquartile range: $IQR = Q_3 - Q_1$

Sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{or,} \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Sample standard deviation: Coefficient of Variation

$$s = \sqrt{s^2} \qquad c.v = \frac{s}{\bar{x}} * 100$$

Formulas for population data

Population variance:

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

Population standard deviation:

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

Coefficient of variation (population):

$$CV = \frac{\sigma}{\mu} * 100$$

Measures of Shape

Measures of Shape: Skewness and Kurtosis

Skewness: Skewness is lack in symmetry. It is a measure of departure from symmetry. If the skewness a_3 is zero, the data are symmetrical; if greater than zero (positive), the data are positively or right skewed, and if the skewness is less than zero (negative), the data are negatively or left skewed.

$$s_k = \frac{n}{(n-1)(n-2)} \sum \left(\frac{(x_i - \bar{x})}{s} \right)^3$$

where:

x_i is the i[th] observation, $\bar{x}$ is mean of the observations

n is the number of non-missing observations, s is the standard deviation

Kurtosis :

Kurtosis is peakedness of the data. It is used to measure the height of the peak in the distribution. A laptokurtic distribution is more peaked than platykurtic (flatter distribution). Between laptokurtic and platykurtic is mesokurtic, which is the normal distribution. The kurtosis a_4 does not provide any information by itself; it must be compared to other distribution.

$$k_r = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{(x_i - \bar{x})}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where: k_r is the kurtosis, x_i is the i[th] observation, $\bar{x}$ is mean of the observations, n is the number of non-missing observations, s is the standard deviation

Measures of Association Between Two Variables

Correlation coefficient : $$r_{xy} = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sqrt{\sum x_i^2 - (\sum x_i)^2/n} * \sqrt{\sum y_i^2 - (\sum y_i)^2/n}}$$

Sample covariance: $$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Chapter 3- Measures of Variation, Measures of Shape- Flow Chart (3)

Descriptive Statistics: Numerical Methods...continued

Relationship between Mean and Standard Deviation

Chebyshev's Theorem

This theorem states that ***no matter what the shape of the distribution*** (symmetrical or skewed),

at least 75% of the observations will fall within ± 2 standard deviation of the mean
at least 89% of the observation will fall within ± 3 standard deviation of the mean
at least 94% of the observations will fall within ±4 standard deviation of the mean

Within k standard deviation of the mean at least $(1-\frac{1}{k^2})$ ***percent of the values occur. Where, k is given by***

$$k=\frac{x-\bar{x}}{s} or, k=\frac{x-\mu}{\sigma}$$

Empirical Rule

The empirical rule applies to ***symmetrical or bell shaped data***. Unlike the Chebyshev's theorem, that applies to any shape (skewed or symmetrical), the empirical rule applies to symmetrical shapes. This rule states that if the data are symmetrical:

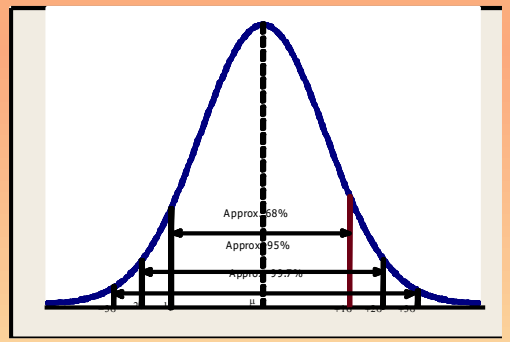
- Approximately 68% of the observations will lie within the mean and ± one standard deviation
- Approximately 95% of the observations will lie within the mean and ± two standard deviation
- Approximately 99.7% of the observations will lie within the mean and ± three standard deviation

That is,

$\mu \pm 1\sigma$ will contain approximately 68% of the observations

$\mu \pm 2\sigma$ will contain approximately 95% of the observations

$\mu \pm 3\sigma$ will contain approximately 99.7% of the observations (See the figure below)



Chapter 3: Chebyshev's and Empirical Rule- Flow Chart (4)

Descriptive Statistics: Numerical Methods...continued

Summary Measures for Grouped Data

Measures of Central Tendency for Grouped Data

Mean for the Grouped Data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

where, f_i = the frequency of class i (i=1 is class interval 1, i=2 is class interval 2 and so on)

M_i = midpoint of class i

n= number of observations which is same as $\sum f$

The midpoint is calculated by : (Lower class limit + Upper class limit)/2

Median for the Grouped Data

The median is the middle value of the data. The sample median for the grouped data is calculated using the following formula

$$M_d = L + \left[\frac{(n+1)/2 - F}{f_m}\right] w$$

Where, M_d = median

L = lower limit of the ***median class***

n=number of observations

F=sum of the frequencies up to but not including the *median class*

f_m = frequency of the *median class*

w = width of the class

Mode for the Grouped Data: In a grouped data, the mode is the average of the ***modal class***. The modal class is the class with maximum frequency.

Measures of Variation for Grouped Data

Range

The range for the above grouped data is calculated using the following formula

Range = Upper limit of the last class - lower limit of the first class

Sample variance

The sample variance of the grouped data is calculated using the following formula

$$s^2 = \frac{\sum fM^2 - n\bar{x}^2}{n-1}$$

where, f = frequency

M = midpoint

n= no. of observations

$\bar{x}$ =mean

Sample Standard Deviation

The standard deviation, s is calculated by taking the square root of the variance. Therefore,

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

Coefficient of Variation (CV)

The coefficient of variation is calculated by

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

Summarizing Data

Statistics Based on Ordered Values	Minimum, First Quartile, Median, Third Quartile, Maximum, Interquartile Range
Statistics Based on Averages	Mean, Standard Deviation, Variance, Skewness, Kurtosis
Describe a symmetrical (bell-shaped) distribution	Mean and Standard Deviation
Relating Continuous Variables	Scatterplots and Correlation

Chapter 3: Measures of Central Tendency and Variation for Grouped Data - Flow Chart (5)