Disclaimer

- This presentation is purely for academic purpose and does not carry any commercial value.
- All images and photos used in this presentation are property of respective image holder(s) and due credit is provided to them. Images are used only for indicative purpose and does not carry any other meaning.
- All information and data in this slide are collected from open domain.

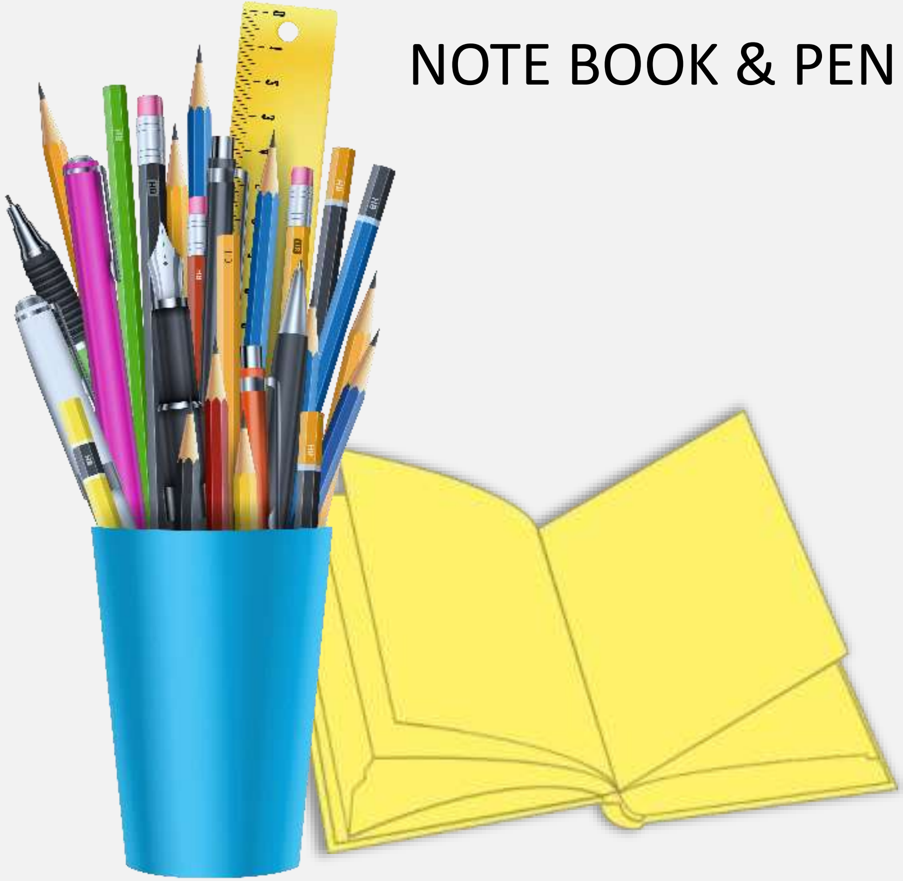
MANISH GODSE, Ph.D.(IIT Bombay)

Request & Instructions

PLEASE OPEN

CALCULATOR

NOTE BOOK & PEN



LAPTOP OR DESKTOP,
IF YOU HAVE.

PLEASE FOLLOW THIS

SILENCE

MUTE
MIC

RAISE
HAND

NO
CHAT

SILENT
MODE

DESCRIPTIVE STATISTICS

BOOKS & REFERENCES

TEXT BOOK

- Albright, Winston. Business Analytics – Data Analysis and Decision Making. Cengage Learning.
  Part 1 – Exploring Data, Chapter 2 and 3

PROBLEMS & PYTHON CODE

- Class notes

- Anderson, Sweeney, Williams, Camm, Cochran (2014). Business Statistics, Cengage Learning (12th Edition)

<a href="https://www.freepik.com/free-photos-vectors/background">Background photo created by freepik - www.freepik.com</a>

TABLE OF CONTENTS

1. Learnings In This Module
2. Introduction
3. Measure of Location
4. Measure of Variability
5. Measure of Shape of a Distribution
6. Exploratory Data Analysis
7. Measures of Association Between Two Variables
8. Weighted Mean and Group Data

LEARNINGS IN THIS MODULE

1

LEARNINGS IN THIS CHAPTER

- You will learn **mean**, **median**, and **mode**.
- You will also study descriptive statistics such as **range**, **variance**, **standard deviation**, **percentiles**, and **correlation**.
- These numerical measures can be computed separately for **each (single) variable**. However, in the **two-variable** case, you will also develop measures of the **relationship** between the variables.
- These numerical measures will assist in the **understanding** and **interpretation** of data.
- If the measures are computed for data from a population, they are called **population parameters**.
- In statistical inference, a sample statistic is referred to as the **point estimator** of the corresponding population parameter.

Numerical Summary Measures

SINGLE VARIABLE

Location
(Ungrouped Data)

- Mean
- Weighted Mean
- Median
- Mode
- Percentile
- Quartile

Dispersion

- Range
- Variance
- Standard deviation
- Coefficient of Variation

Shape

- Skewness
- Relative location (z-score)
- Outlier

Exploratory Data Analysis

- Five Number Summary
- Box Plot

Location
(Grouped/ Frequency Data)

- Mean
- Variance
- Standard Deviation

TWO VARIABLES

- Covariance
- Correlation

INTRODUCTION

2

WHAT IS DESCRIPTIVE STATISTICS?

Reference: *http://rcompanion.org/handbook/C_02.html*

DESCRIPTIVE STATISTICS is used to summarize data in more meaningful way so that it allows for simpler interpretation of data and helps generate insight into the information contained in the data.

- Choosing which summary statistics are appropriate depend on the type of variable being examined.
- Different statistics should be used for ratio, ordinal, and nominal data.
- Raw data is used in descriptive statistics.
- Descriptive statistics is important to establish validity that sample is representing correct population.

WHAT IS EXAMINED IN DESCRIPTIVE STATISTICS?

Reference: *http://rcompanion.org/handbook/C_02.html*

LOCATION is called central tendency. It is a measure of the values of the data. Measures of location include mean, median etc.

VARIATION is also called dispersion. It is a measure of how far the data points lie from one another. Common statistics include standard deviation and coefficient of variation. For data that aren't normally-distributed, percentiles or the interquartile range might be used.

SHAPE refers to the distribution of values. The best tools to evaluate the shape of data are histograms and related plots. Statistics include skewness and kurtosis, though they are less useful than visual inspection. We can describe data shape as normally-distributed, log-normal, uniform, skewed, bi-modal, and others.

DESCRIPTIVE STATISTICS

MEASURE	STATISTICS METHOD
Size of Data	Number of rows and columns
Data Types	Check data type of each variable or attribute
Measure of Location (or Center)	Mean, Median, Mode, Geometric Mean, Harmonic Mean
Measure of Dispersion (or Dispersion)	Minimum Value, Maximum Value, Range of Data, Variance, Standard Deviation, Coefficient of variance
Measure of Distribution (or Shape)	Percentile, Quantile, Quartile, Inter-Quartile Range, z-score, skewness & kurtosis, Box plot
Measure of Association	Scatter chart, Covariance, Correlation coefficient

PLOTTING DISTRIBUTION

Number of data	Data type	Graph
Univariate data	Discrete data	Histogram
	Continuous data	Polygon
Bivariate data		Scatter plot

MEASURE OF LOCATION

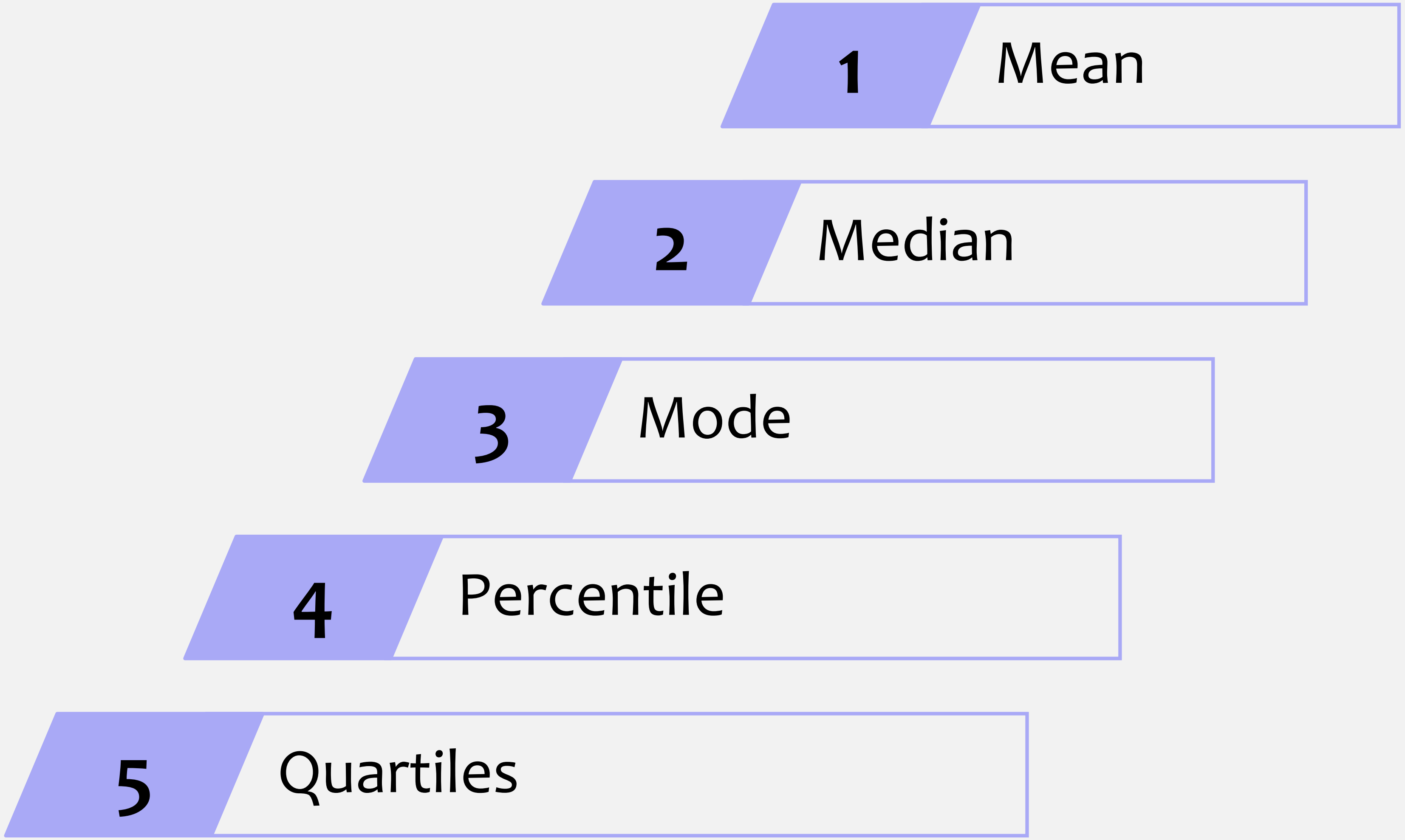
3

MEASURE OF CENTRAL LOCATION

A measure of central Location is a **single value** that attempts to **describe a set of data** by identifying the central position within that set of data. As such, measures of central Location are sometimes called **measures of central tendency**.

Central tendency is numerical value around which most numerical values in dataset tends to cluster.

SUMMARIZING QUALITATIVE DATA



1. Mean
2. Median
3. Mode
4. Percentile
5. Quartiles

MEAN

- The MEAN provides a measure of central location for the data.
- If the data are for a **sample**, the mean is denoted by $\bar{x}$.
- If the data are for a **population**, the mean is denoted by the greek letter μ.

- A sample with n observations, the formula for the **sample mean** is as follows.

$$\textit{Sample Mean}, \bar{x} = \frac{\sum x_i}{n}$$

→

$$\textit{Sample Mean}, \bar{x} = \frac{x_1 + x_2 + x_3 + \text{......} + x_n}{n}$$

$$\textit{Population Mean}, \mu = \frac{\sum x_i}{N}$$

MEAN - EXAMPLE

Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries.

Calculate mean monthly starting salary?

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$

$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$

$$= \frac{42{,}480}{12} = 3540$$

Graduate	Salary ($)
1	3450
2	3550
3	3650
4	3480
5	3355
6	3310
7	3490
8	3730
9	3540
10	3925
11	3520
12	3480

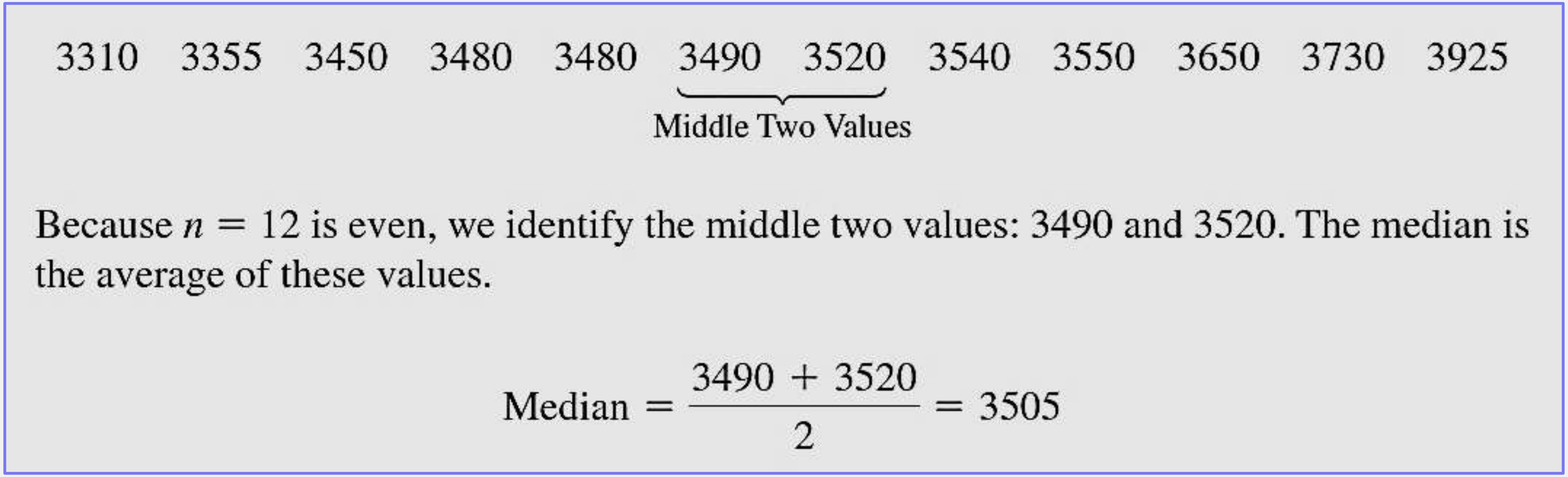
MEDIAN

- The median is the value in the **middle** when the **data** are **arranged** in **ascending** order (smallest value to largest value).
- With an **odd number** of observations, the median is the **middle value**.
- An **even number** of observations has **no single middle value.** In this case, we follow convention and define the median as the **average of the values** for the middle two observations.

MEDIAN - EXAMPLE

Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries. **Calculate median monthly starting salary?**

First arrange the data in ascending order



$$3310 \quad 3355 \quad 3450 \quad 3480 \quad 3480 \quad \underbrace{3490 \quad 3520}_{\text{Middle Two Values}} \quad 3540 \quad 3550 \quad 3650 \quad 3730 \quad 3925$$

Because $n = 12$ is even, we identify the middle two values: 3490 and 3520. The median is the average of these values.

$$\text{Median} = \frac{3490 + 3520}{2} = 3505$$

MEAN OR MEDIAN, WHICH IS BETTER?

- Although the mean is the more commonly used measure of central location, in some situations the median is preferred.
- The mean is influenced by **extremely small** and **large data values**.
- For instance, suppose that one of the graduates had a starting salary of $10,000 per month. If we recompute the mean, the sample mean changes from $3540 to $4046. The median of $3505, however, is unchanged, because $3490 and $3520 are still the middle two values.
- With the extremely high starting salary included, the median provides a better measure of central location than the mean.
- We can generalize to say that whenever a data set contains **extreme values**, the **median** is often the **preferred measure** of central location.

TRIMMED MEAN, WHY IS BETTER?

- It is better to use the median than the mean as a measure of central location when a data set contains extreme values.
- Another measure, sometimes used when extreme values are present, is the trimmed mean. It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values. For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values.

MODE

- The **mode** is the value that occurs with **greatest frequency**.
- Situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.
  - If the data contain exactly two modes, we say that the data are **bimodal**.
  - If data contain more than two modes, we say that the data are **multimodal**.
  - In **multimodal** cases the mode is almost **never reported** because listing three or more modes would **not** be particularly helpful in **describing a location for the data**.

EXAMPLE - MODE

These data show the time in days required to complete year-end audits for a sample of 20 clients of Sanderson and Clifford, a small public accounting firm.

Year-End Audit Times *(In Days)*		
12	14	19
18	15	15
18	17	20
27	22	23
22	21	33
28	14	18
16	13	

Audit Time	Frequency
12	1
13	1
14	**2**
15	**2**
16	1
17	1
18	**3**

Audit Time	Frequency
19	1
20	1
21	1
22	**2**
27	1
28	1
33	1

MODE = 18

PERCENTILE

A **percentile** provides information about how the data are spread over the interval from the smallest value to the largest value.

The ***p*^*th*^** percentile is a value such that at least ***p*** percent of the observations are less than or equal to this value and at least ***(100 - p)*** percent of the observations are greater than or equal to this value.

PERCENTILE

Colleges and universities frequently report admission test scores in terms of percentiles. For instance, suppose an applicant obtains a raw score of 54 on the verbal portion of an admission test. How this student performed in relation to other students taking the same test may not be readily apparent. However, if the raw score of 54 corresponds to the 70th percentile, we know that approximately 70% of the students scored lower than this individual and approximately 30% of the students scored higher than this individual.

CALCULATE PERCENTILE

STEP-1 : Arrange the data in ascending order (smallest value to largest value).

STEP-2 : Compute an index i

$$\boldsymbol{i = \left(\frac{p}{100}\right)n}$$

where *p* is the percentile of interest and *n* is the number of observations.

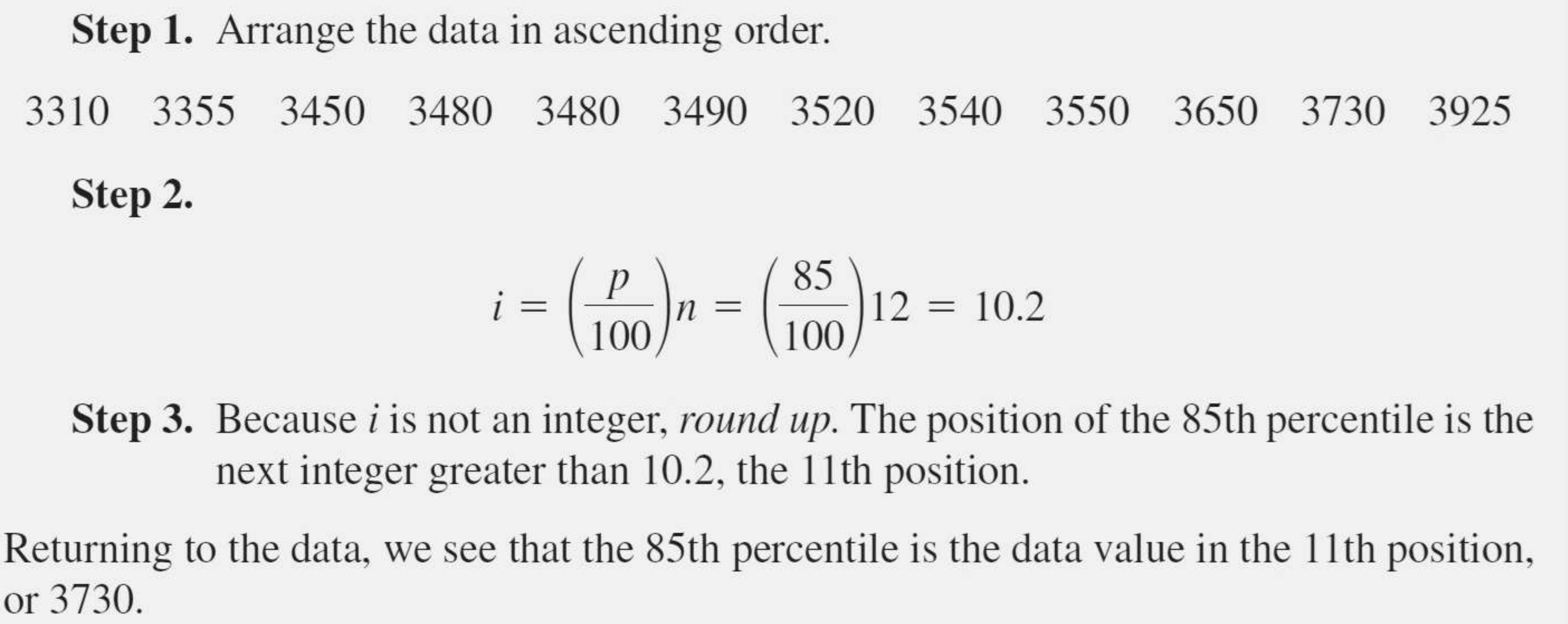
STEP-3 :

a) If *i is not an integer, round up.* The next integer *greater* than *i* denotes the position of the *p*th percentile.

b) If *i is an integer*, the *p*th percentile is the average of the values in positions *i* and *i* + 1

PERCENTILE - EXAMPLE

Calculate 85th percentile for salary data as shown in earlier example.



Step 1. Arrange the data in ascending order.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Step 2.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

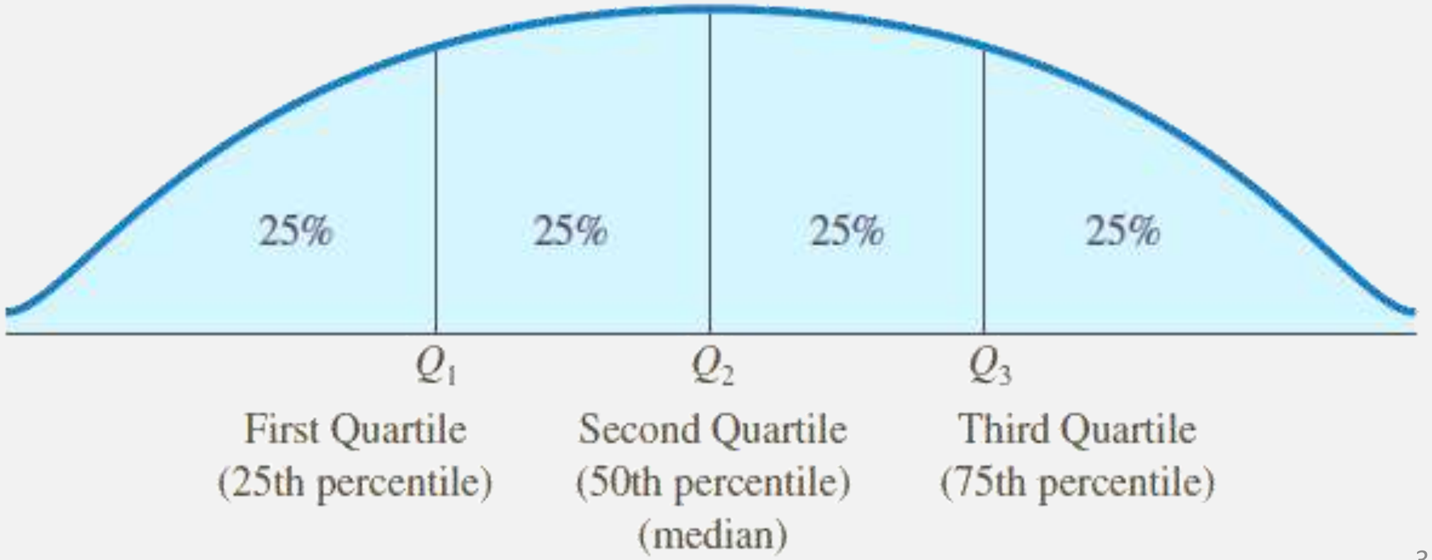
Step 3. Because i is not an integer, *round up*. The position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730.

QUARTILE

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. Figure shows a data distribution divided into four parts. The division points are referred to as the **quartiles** and are defined as

- $Q1$ First quartile, or 25th percentile
- $Q2$ Second quartile, or 50th percentile (also the median)
- $Q3$ Third quartile, or 75th percentile



QUARTILE – EXAMPLE (SLIDE-1/2)

Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries.

Calculate quartiles

First arrange the data in ascending order

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Quartile	Calculate Quartile index
Q1	(25/100) × 12 = 3
Q2	(50/100) × 12 = 6
Q3	(75/100) × 12 = 9
Q4	(100/100) × 12 = 12

$$i = \left(\frac{p}{100}\right)n$$

i = Index

n = Number of Elements
= 12

QUARTILE – EXAMPLE *(SLIDE-2/2)*

1	2	3	4	5	6	7	8	9	10	11	12
3310	**3355**	**3450**	**3480**	**3480**	**3490**	**3520**	**3540**	**3550**	**3650**	**3730**	**3925**
8%	17%	25%	33%	42%	50%	58%	67%	75%	83%	92%	100%

Q1 **Q2** **Q3**

Average of 3450 and 3480

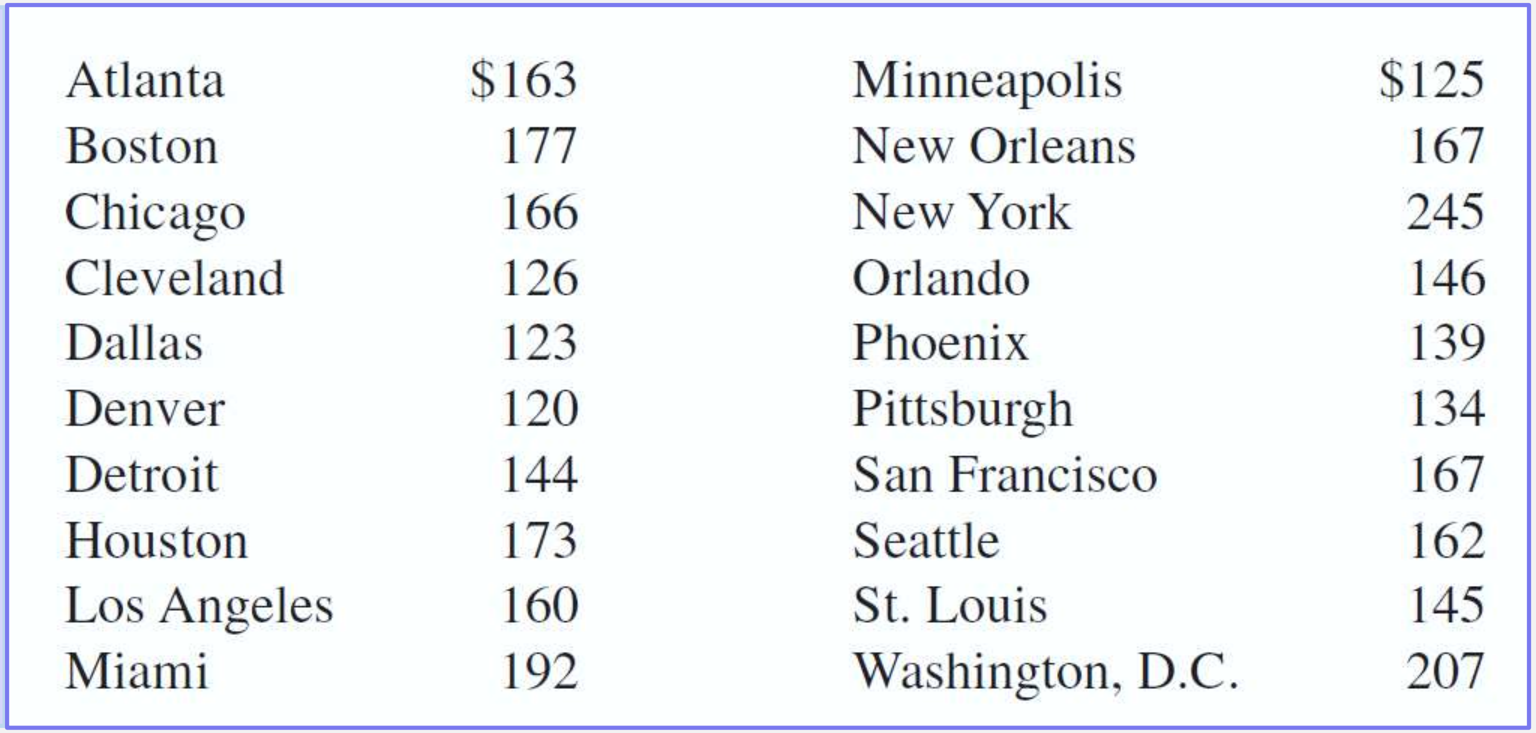
Average of 3490 and 3520

Average of 3550 and 3650

Quartile	Calculate Quartile	Value of Quartile
Q1	(25/100)×12 = 3	(3450 + 3480)/2 = 3465
Q2	(50/100)×12 = 6	(3490 + 3520)/2 = 3505
Q3	(75/100)×12 = 9	(3550 + 3650)/2 = 3600

SOLVE THIS PROBLEM

The Dow Jones Travel Index reported what business travelers pay for hotel rooms per night in major U.S. cities (The Wall Street Journal, January 16, 2004). The average hotel room rates for 20 cities are as given:



Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

1. What is the **mean** hotel room rate?
2. What is the **median** hotel room rate?
3. What is the **mode**?
4. What is the **first quartile**?
5. What is the **third quartile**?

MEASURE OF VARIABILITY

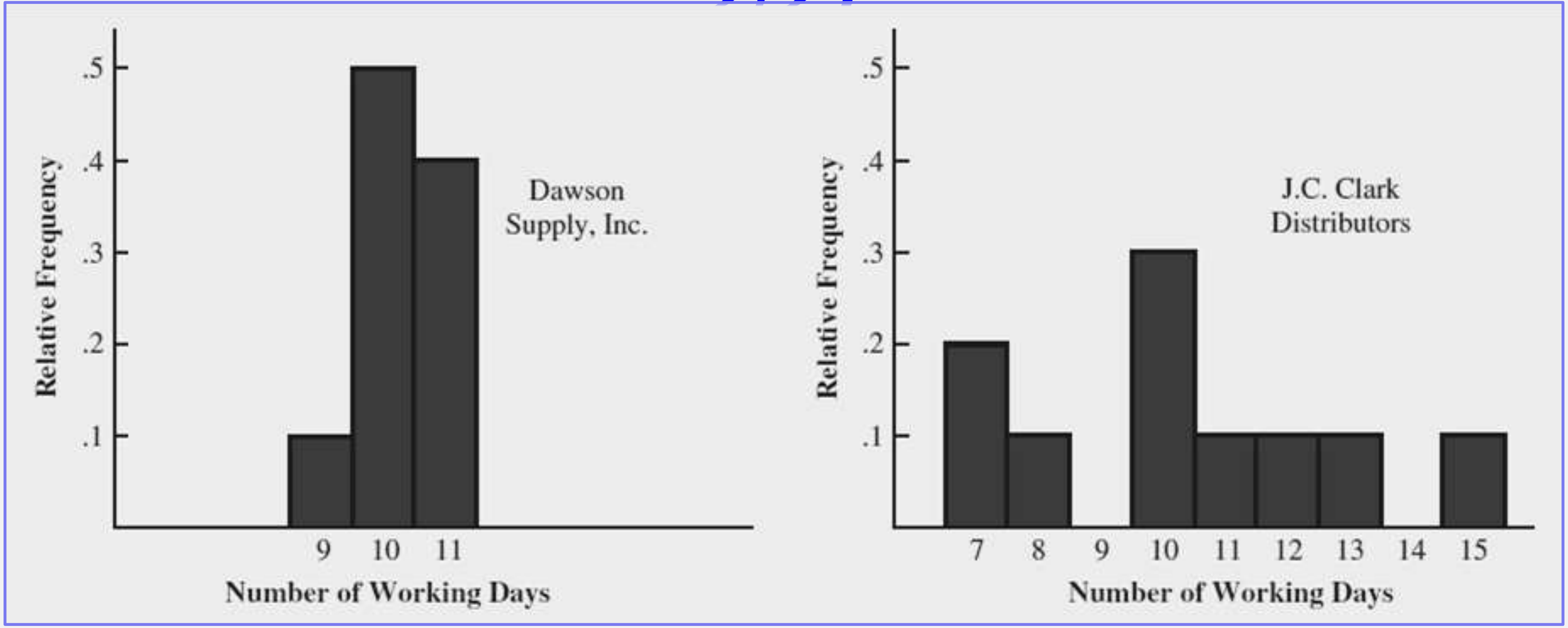
4

CASE STUDY ON VARIABILITY *(SLIDE-1/2)*

Suppose that you are a purchasing agent for a large manufacturing firm and that you regularly place orders with two different suppliers.

After several months of operation, you find that the mean number of days required to fill orders is 10 days for both of the suppliers. The histograms summarizing the number of working days required to fill orders from the suppliers are shown in Figure. Although the mean number of days is 10 for both suppliers, do the two suppliers demonstrate the same degree of reliability in terms of making deliveries on schedule? Note the dispersion, or variability, in delivery times indicated by the histograms. Which supplier would you prefer?

CASE STUDY ON VARIABILITY *(SLIDE-2/2)*



For most purchasing agents, the lower variability shown for Dawson Supply, Inc., would make Dawson the preferred supplier.

RANGE

The range of a dataset is the difference between the largest and smallest values in that dataset.

Range = Largest value - Smallest value

The reason is that the range is based on only two of the observations and thus is highly influenced by extreme values.

Suppose one of the graduates received a starting salary of $10,000 per month. In this case, the range would be 10,000 - 3310 = 6690 rather than 615. This large value for the range would not be especially descriptive of the variability in the data because 11 of the 12 starting salaries are closely grouped between 3310 and 3730.

INTERQUARTILE RANGE

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR). This measure of variability is the difference between the third quartile, Q3, and the first quartile, Q1. In other words, the interquartile range is the range for the middle 50% of the data.

For the data on monthly starting salaries, the quartiles are Q3 3600 and Q1 3465.
Thus, the interquartile range is
3600 - 3465 = 135

VARIANCE

lack of consistency or fixed pattern; liability to vary or change

- The **variance** is a measure of **variability** that utilizes all the data.
- The variance is based on the difference between the value of each observation (x_i) and the mean. The difference between each x_i and the mean ($\bar{x}$ for a sample, μ for a population) is called a *deviation about the mean.* For a sample, a deviation about the mean is written $(x_i - \bar{x})$; for a population, it is written $(x_i - \mu)$. In the computation of the variance, the deviations about the mean are *squared.*
- The variance is useful in **comparing the variability of two** or **more** variables. In a comparison of the variables, the one with the **largest variance** shows the most **variability**.

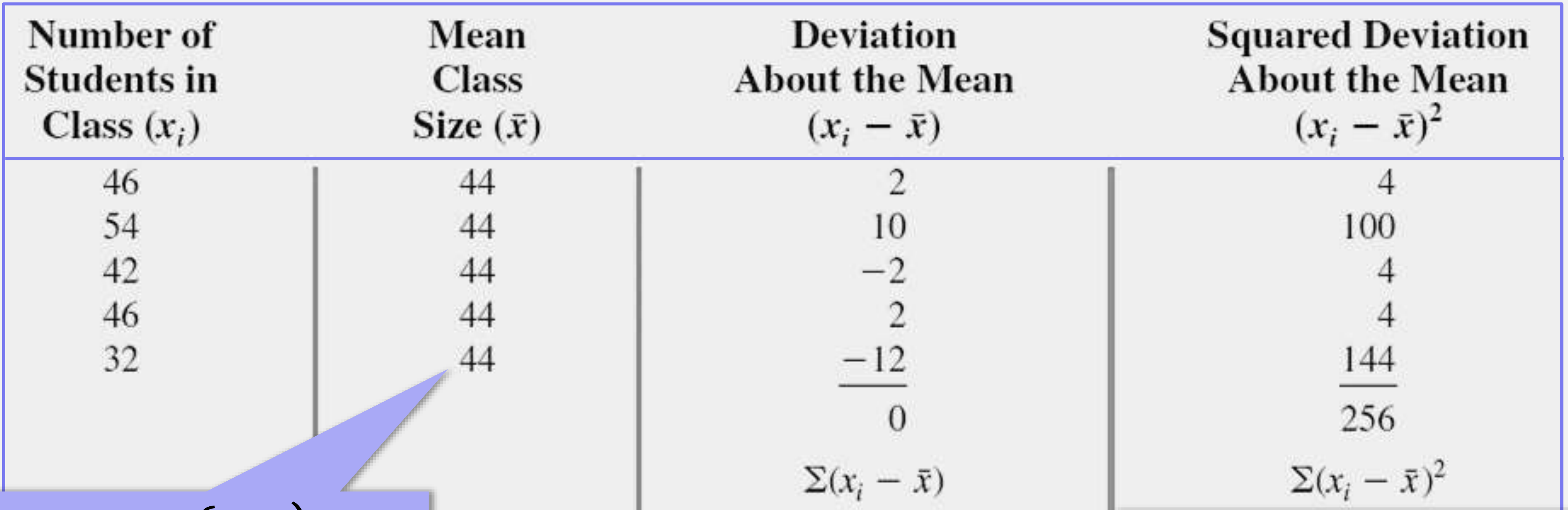
VARIANCE – POPULATION AND SAMPLE

If the data are for a **population**, the average of the squared deviations is called the **population variance**, and for **sample** it is called as **sample variance**.

Population Variance $$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

Sample Variance $$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

COMPUTE VARIANCE - EXAMPLE



Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

(46+54+42+46+32)÷5 = 44

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64$$

For any data set, the sum of the deviations about the mean will *always equal zero.* Note that in Tables 3.3 and 3.4, (*xi*) 0. The positive deviations and negative deviations cancel each other, causing the sum of the deviations about the mean to equal zero.

STANDARD DEVIATION

The standard deviation is defined to be the positive square root of the variance.

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

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

From earlier example

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

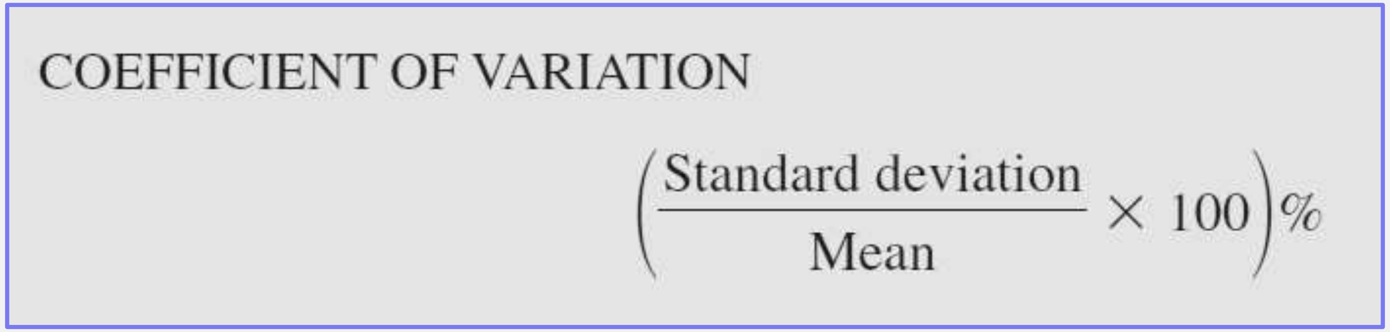
$$s = \sqrt{64} = 8$$

The standard deviation is measured in the **same units** as the **original data**. For this reason the standard deviation is more easily **compared** to the **mean** and **other statistics** that are measured in the same units as the original data.

- Standard deviation is impacted by outliers and extreme values.
- Bigger the standard deviation bigger is the volatility.

COEFFICIENT OF VARIANCE

- The coefficient of variation is a relative measure of variability; it measures the standard deviation relative to the mean.
- The coefficient of variation is a useful statistic for comparing the variability of variables that have different standard deviations and different means.



COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

COEFFICIENT OF VARIANCE

- For the class size data, we found a sample **mean** of 44 and a sample **standard deviation** of 8.

  The **coefficient of variation** is [(8/44) × 100]% = 18.2%. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean.

- For the starting salary data with a sample **mean** of 3540 and a sample **standard deviation** of 165.65, the **coefficient of variation**, [(165.65/3540) × 100]% = 4.7%, tells us the sample standard deviation is only 4.7% of the value of the sample mean.

STANDARD DEVIATION

	Dataset-1	Dataset-2
Mean	60	45
Standard Deviation	10	12
Coefficient of Variance	17	29

Standard deviation looks very similar however Coefficient of Variance are very different, and explains variability.

EXAMPLE – CAR RENTAL RATES

Car rental rates per day for a sample of seven Eastern U.S. cities are as follows (*The Wall Street Journal*, January 16, 2004).

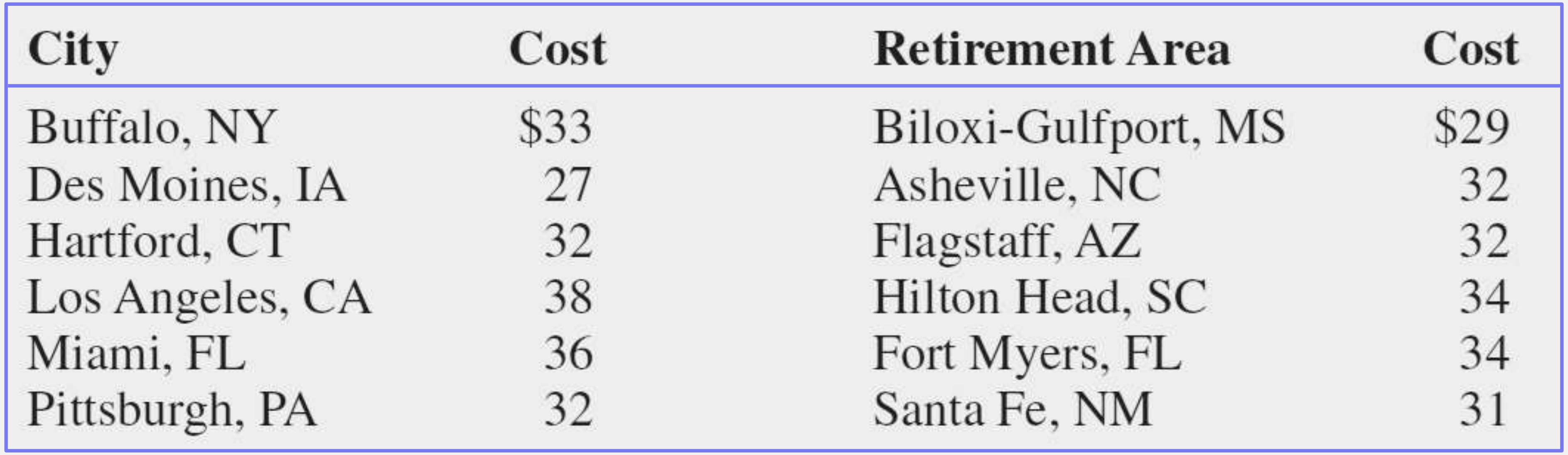
1. Compute the mean, variance, and standard deviation for the car rental rates.
2. A similar sample of seven Western U.S. cities showed a sample mean car rental rate of \$38 per day. The variance and standard deviation were 12.3 and 3.5, respectively. Discuss any difference between the car rental rates in Eastern and Western U.S. cities.

City	Daily Rate ($)
Boston	43
Atlanta	35
Miami	34
New York	58
Orlando	30
Pittsburgh	30
Washington D.C.	36

EXAMPLE - GROCERY COSTS ACROSS THE COUNTRY

How do grocery costs compare across the country? Using a market basket of 10 items including meat, milk, bread, eggs, coffee, potatoes, cereal, and orange juice, *Where to Retire* magazine calculated the cost of the market basket in six cities and in six retirement areas across the country (*Where to Retire,* November/December 2003).

1. Compute the mean, variance, and standard deviation for the sample of cities and the sample of retirement areas.
2. What observations can be made based on the two samples?



City	Cost	Retirement Area	Cost
Buffalo, NY	$33	Biloxi-Gulfport, MS	$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Angeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

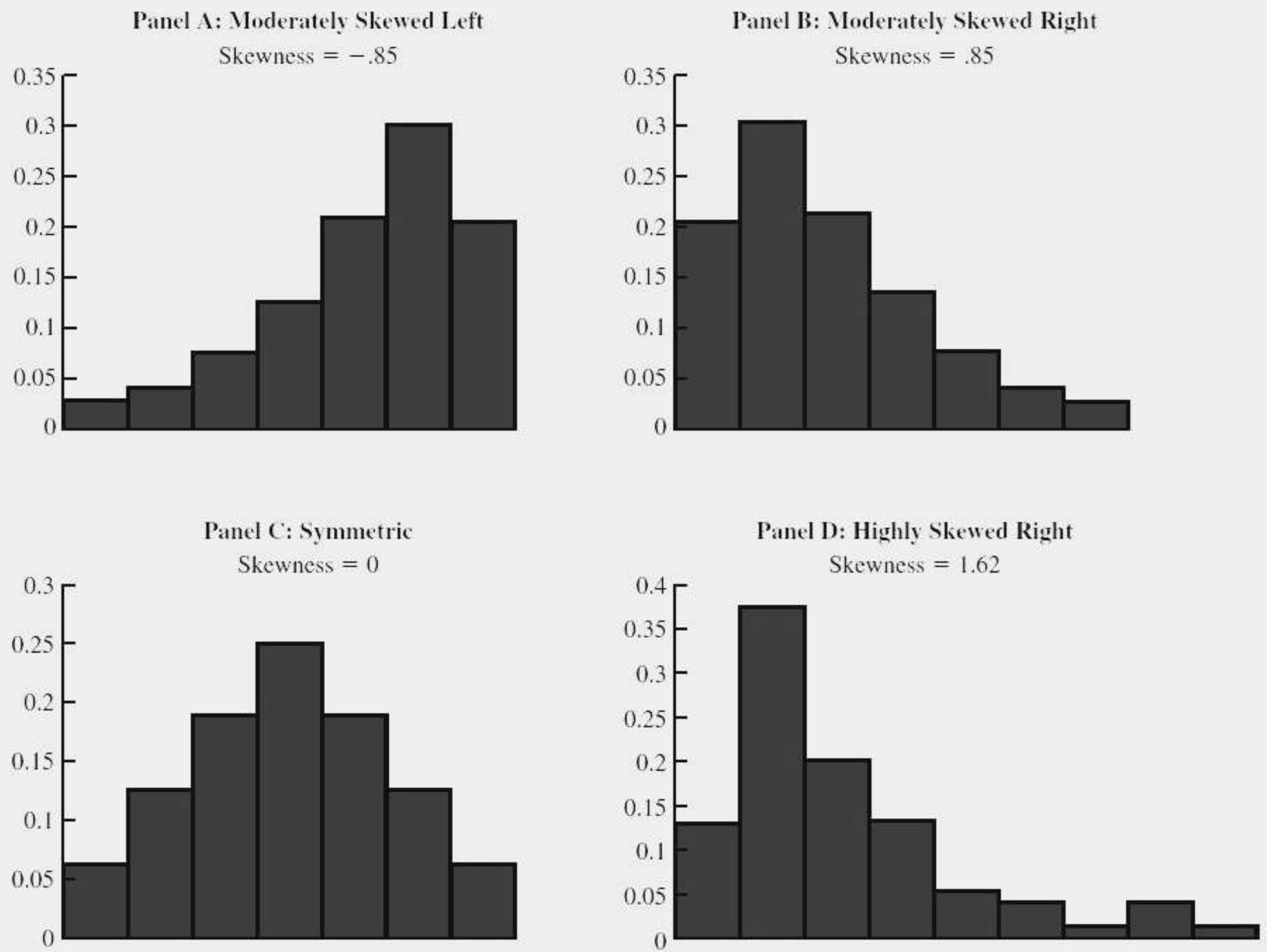
MEASURE OF SHAPE OF A DISTRIBUTION

5

An important numerical measure of the shape of a distribution is called **skewness.**

- For data skewed to the left, the skewness is **negative**.
- For data skewed to the right, the skewness is **positive**.
- If the data are symmetric, the skewness is **zero**.

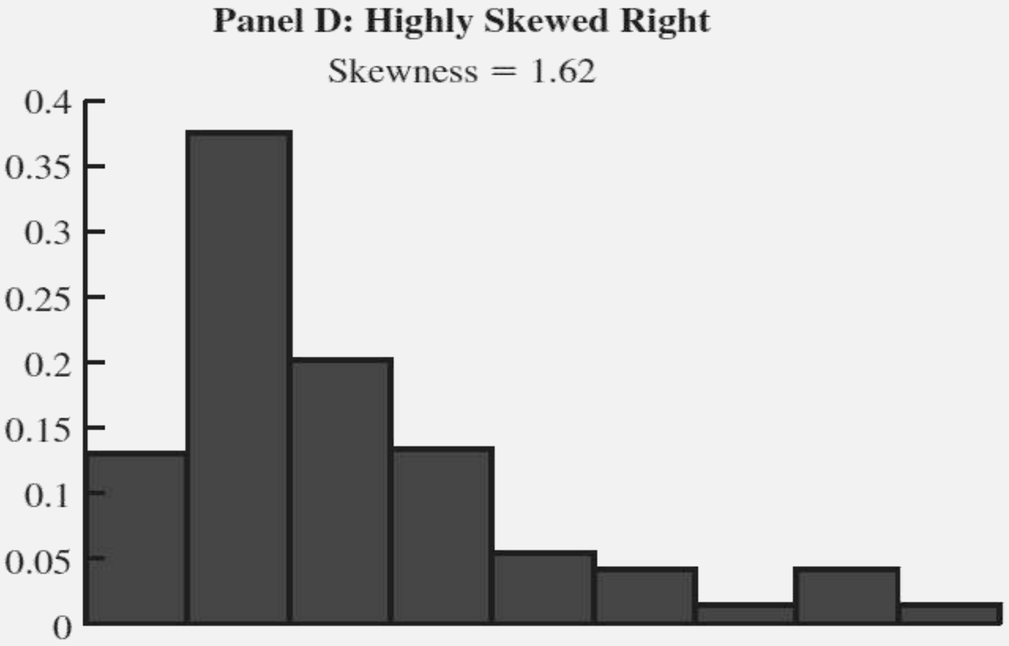
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$



SKEWNESS

- For a symmetric distribution, the mean and the median are equal.
- When the data are positively skewed, the mean will usually be greater than the median; when the data are negatively skewed, the mean will usually be less than the median.
- The median provides the preferred measure of location when the data are highly skewed.

Histogram in Panel D are customer purchases at a women's apparel store. The mean purchase amount is $77.60 and the median purchase amount is $59.70. The relatively few large purchase amounts tend to increase the mean, while the median remains unaffected by the large purchase amounts.



z - SCORES

- In addition to measures of location, variability, and shape, we are also interested in the **relative location of values within a data set**.
- Measures of relative location help us determine how far a particular value is from the mean. By using both the mean and standard deviation, we can determine the relative location of any observation
- The *z*-score for any observation can be interpreted as a measure of the relative location of the observation in a data set.

$$z_i = \frac{x_i - \bar{x}}{s}$$

z_i	= z-score of x_i
$\bar{x}$	= Sample mean
x_i	= Value of each item under consideration
s	= Sample standard deviation

z – SCORES INTERPRETATION

- The z-score is often called the **standardized value**.
- The z-score, $\mathrm{z_i}$, can be interpreted as the number of standard deviations $\mathrm{x_i}$ is from the **mean** . For example, $\mathrm{z_1}$ = 1.2 would indicate that $\mathrm{x_1}$ is 1.2 standard deviations greater than the sample mean. Similarly, $\mathrm{z_2}$ = 0.5 would indicate that $\mathrm{x_2}$ is .5, or 1/2, standard deviation less than the sample mean.
- A **z-score greater than zero** occurs for observations with a value greater than the mean, and a **z-score less than zero** occurs for observations with a value less than the mean. A **z-score of zero** indicates that the value of the observation is equal to the mean.
- Observations in two different data sets with the same *z*-score can be said to have the same relative location in terms of being the same number of standard deviations from the mean.

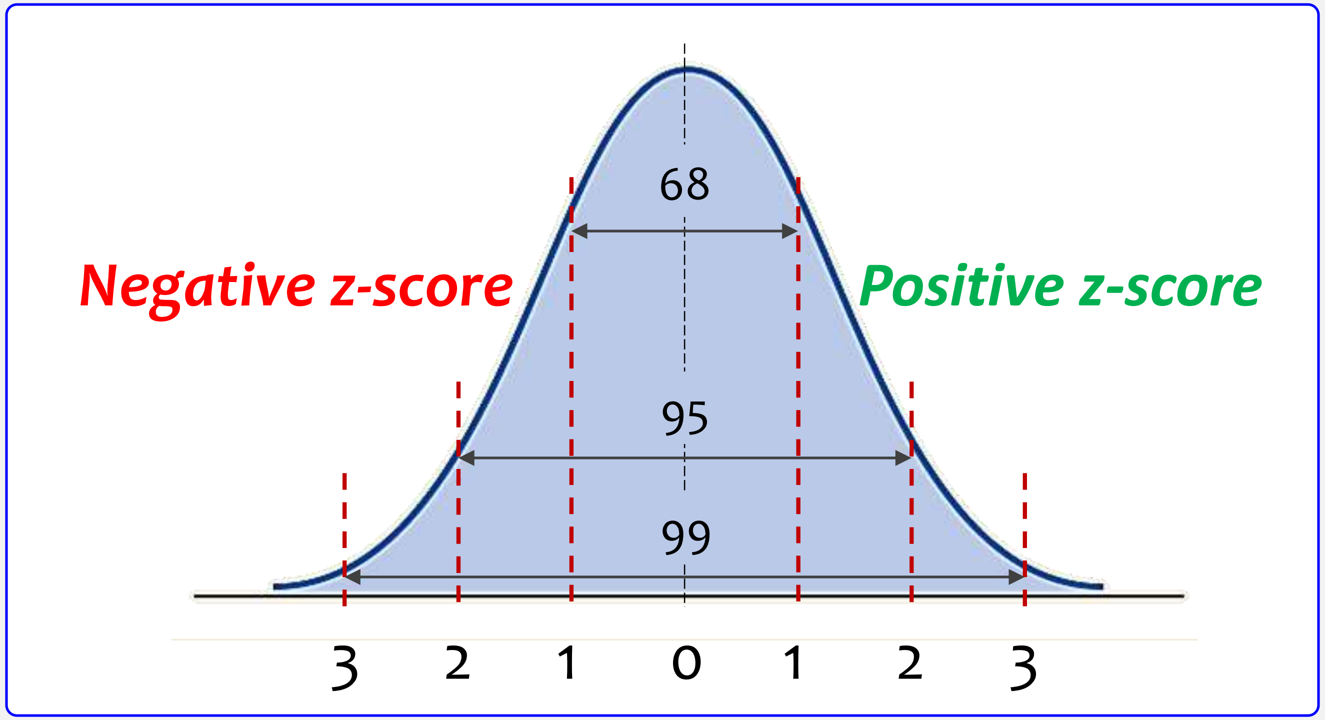
Z-SCORE & OUTLIER

In most of the data sets, 99% of values have a Z-score between -3 and 3. It means, they lie within three standard deviations above or below the mean.

- Z-score equal to 1 indicates that the data are one standard deviation above the mean.
- Z-score equal to -1 indicates that the data are one standard deviation below the mean.

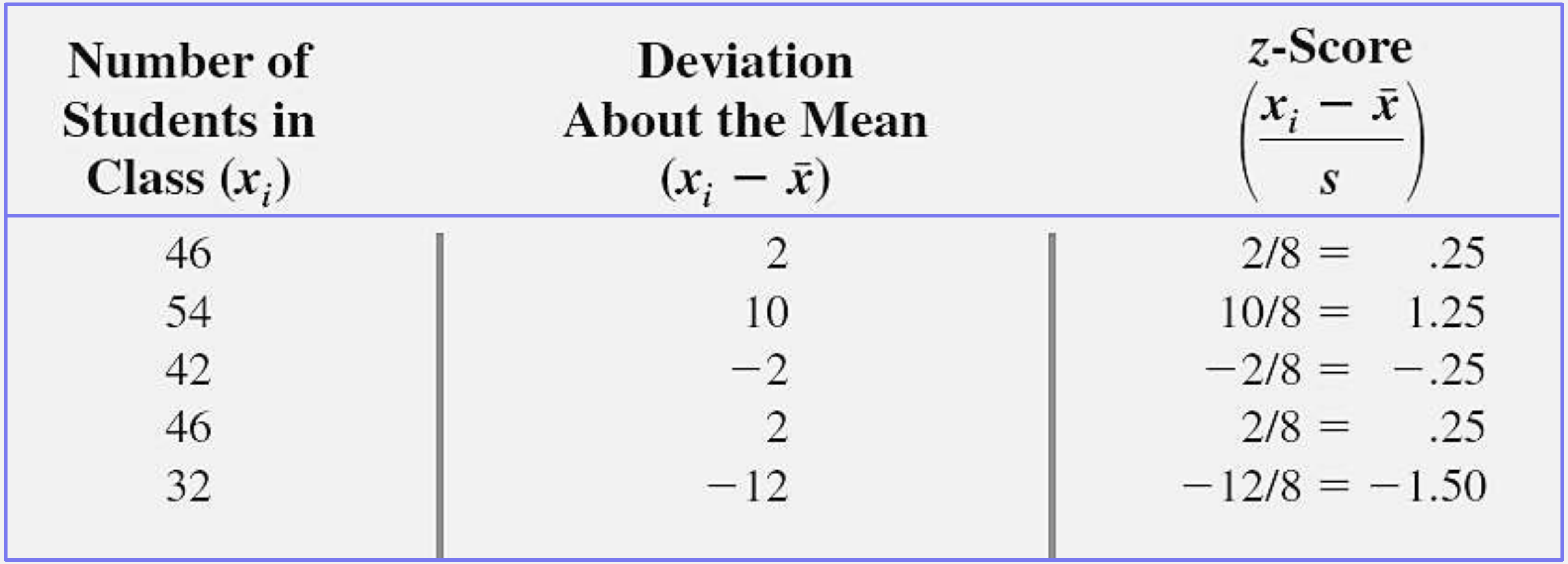
Typically, Z-score values greater than or less than + 3 or – 3, respectively, are considered **outliers**.

However, z-score value for outlier can be changed depending upon the nature of problem and domain.



z – SCORES - EXAMPLE

mean = 44

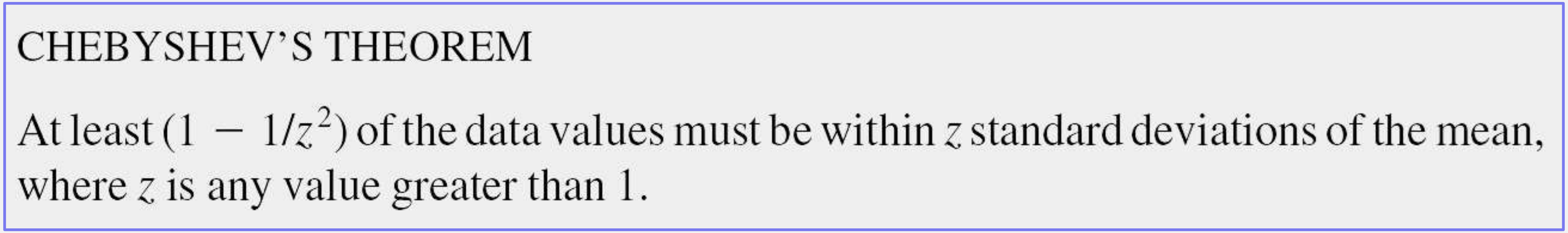


Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\dfrac{x_i - \bar{x}}{s}\right)$
46	2	$2/8 = .25$
54	10	$10/8 = 1.25$
42	−2	$-2/8 = -.25$
46	2	$2/8 = .25$
32	−12	$-12/8 = -1.50$

The *z*-score of 1.50 for the fifth observation shows it is farthest from the mean; it is 1.50 standard deviations below the mean.

CHEBYSHEV’S THEOREM

Chebyshev’s theorem enables to decide the proportion of data values that must be within a specified number of standard deviations of the mean.



CHEBYSHEV’S THEOREM

At least $(1 - 1/z^2)$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.

One of the advantages of Chebyshev’s theorem is that it applies to any data set regardless of the shape of the distribution of the data.

IMPLICATIONS OF CHEBYSHEV'S THEOREM

Some of the implications of this theorem, with z = 2, 3, and 4 standard deviations, follow.

- At least 0.75, or 75%, of the data values must be within z =2 standard deviations of the mean.
- At least 0.89, or 89%, of the data values must be within z =3 standard deviations of the mean.
- At least 0.94, or 94%, of the data values must be within z =4 standard deviations of the mean.

CHEBYSHEV’S THEOREM - EXAMPLE

Suppose that the midterm test scores for 100 students in a college business statistics course had a **mean** of 70 and a **standard deviation** of 5. How many students had test scores between 60 and 80? How many students had test scores between 58 and 82?

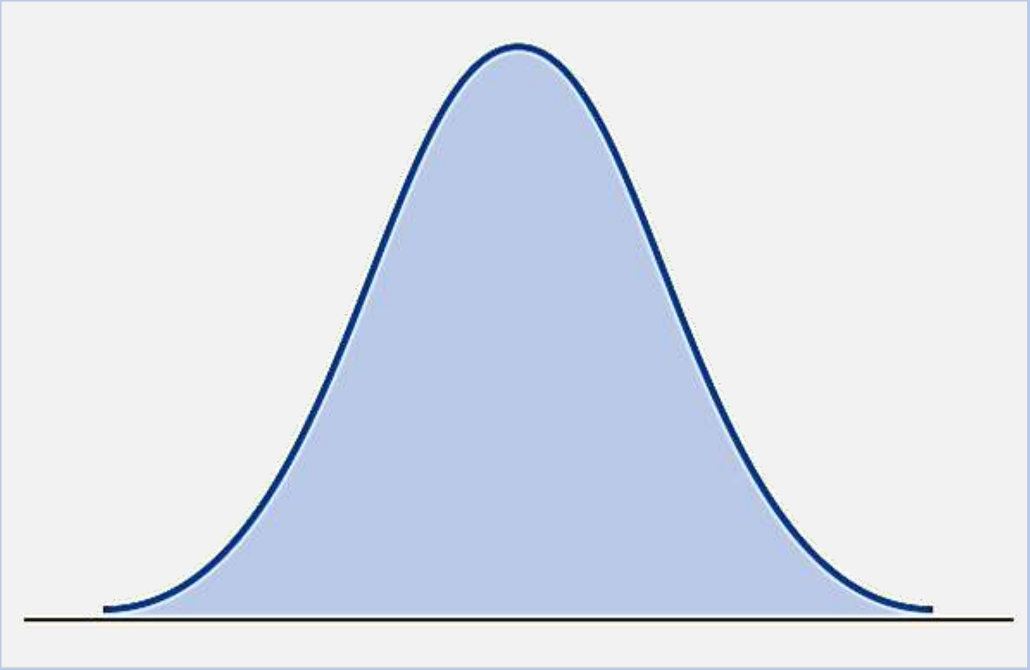
- For the test scores between 60 and 80, we see that (60 - 70)/5=2 indicates 60 is 2 standard deviations below the mean and that (80 - 70)/5=2 indicates 80 is 2 standard deviations above the mean. Applying Chebyshev’s theorem with z = 2, at least 75% of the students must have test scores between 60 and 80.
- For the test scores between 58 and 82, we see that (58 - 70)/5=2.4 indicates 58 is 2.4 standard deviations below the mean and that (82 - 70)/5=2.4 indicates 82 is 2.4 standard deviations above the mean. Applying Chebyshev’s theorem with z = 2.4, at least 82.6% of the students must have test scores between 58 and 82.

$$\boxed{\left(1-\frac{1}{z^2}\right)=\left(1-\frac{1}{(2.4)^2}\right)=.826}$$

EMPIRICAL RULE

- The empirical rule is based on the **normal probability distribution**.
- In many practical applications, data sets exhibit a **symmetric mound-shaped** or **bell-shaped distribution** like the one shown in Figure.
- When the data are believed to approximate this distribution, the **empirical rule** can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

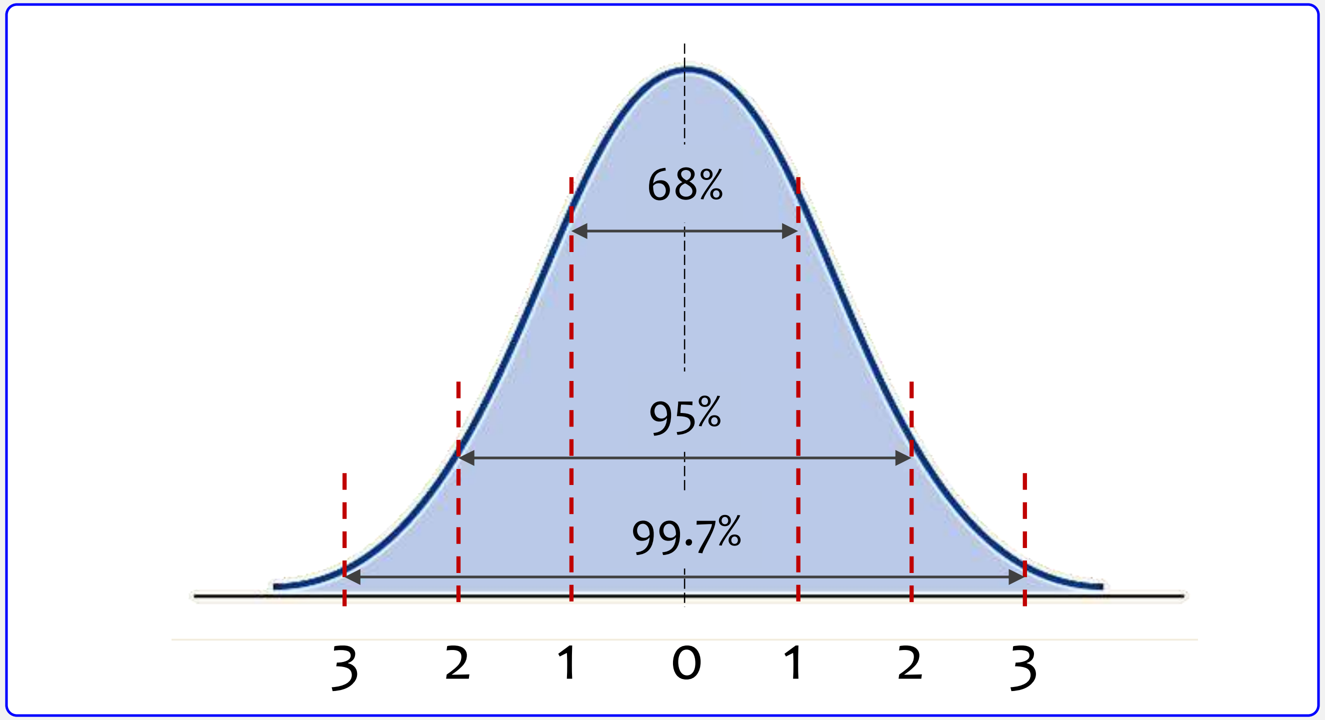
A Symmetric Mound-shaped or Bell-shaped Distribution



EMPIRICAL RULE & BELL-SHAPE

For data having a bell-shaped distribution:

- Approximately **68%** of the data values will be within **ONE** standard deviation of the mean.
- Approximately **95%** of the data values will be within **TWO** standard deviations of the mean.
- Approximately **99.7%** of the data values will be within **THREE** standard deviations of the mean.



EMPIRICAL RULE - EXAMPLE

The liquid detergent cartons are filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is .25 ounces, we can use the empirical rule to draw the following conclusions.

- Approximately 68% of the filled cartons will have weights between 15.75 and 16.25 ounces (within one standard deviation of the mean).
- Approximately 95% of the filled cartons will have weights between 15.50 and 16.50 ounces (within two standard deviations of the mean).
- Almost all filled cartons will have weights between 15.25 and 16.75 ounces (within three standard deviations of the mean).

OUTLIER

Sometimes a data set will have one or more observations with **unusually large** or **unusually small** values. These extreme values are called **outliers**.

- An outlier may be a data value that has been **incorrectly recorded**. If so, it can be **corrected** before further analysis.
- An outlier may also be from an observation that was **incorrectly included** in the data set; if so, it can be **removed**.
- Finally, an outlier may be an unusual data value that has been **recorded correctly** and belongs in the data set. In such cases it should **remain**.

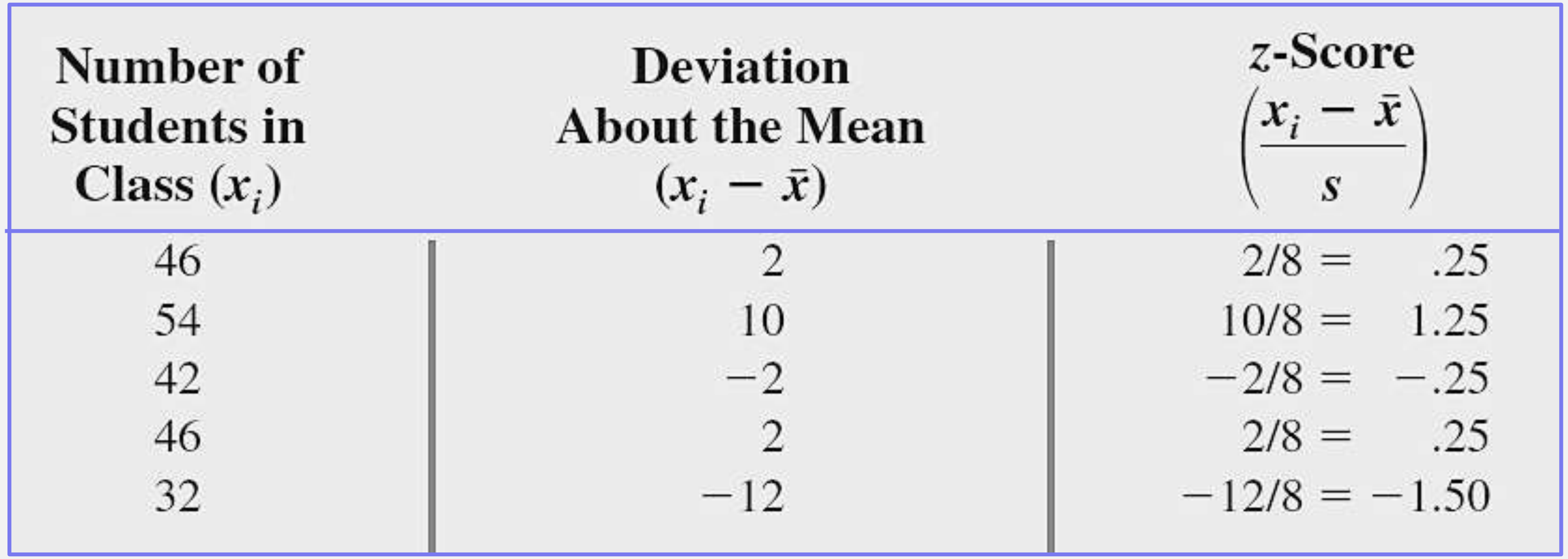
Standardized values (z-scores) can be used to identify outliers.

OUTLIER AND Z-SCORE

- It is a good idea to check for outliers before making decisions based on data analysis.
- Outliers should not necessarily be deleted, but their accuracy and appropriateness should be verified.
- The empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within **three standard deviations of the mean**. Hence, in using z-scores to identify **outliers**, we recommend treating any data value with a z-score **less than 3 or greater than 3 as an outlier**. Such data values can then be reviewed for accuracy and to determine whether they belong in the data set.

OUTLIER - EXAMPLE

The z-score of -1.50 shows the fifth class size is farthest from the mean. However, this standardized value is well within the -3 to +3 guideline for outliers. Thus, the z-scores do not indicate that outliers are present in the class size data.



Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	$2/8 = .25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -.25$
46	2	$2/8 = .25$
32	-12	$-12/8 = -1.50$

Consumer Review posts reviews and ratings of a variety of products on the Internet. The following is a sample of 20 speaker systems and their ratings (www.audioreview.com). The ratings are on a scale of 1 to 5, with 5 being best.

Speaker	**Rating**	**Speaker**	**Rating**
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

a. Compute the mean and the median.

b. Compute the first and third quartiles.

c. Compute the standard deviation.

d. The skewness of this data is -1.67. Comment on the shape of the distribution.

e. What are the z-scores associated with Allison One and Omni Audio?

f. Do the data contain any outliers? Explain.

EXPLORATORY DATA ANALYSIS

6

FIVE NUMBER SUMMARY

In a **five-number summary**, the following five numbers are used to summarize the data.

1. Smallest value
2. First quartile (Q1)
3. Median (Q2)
4. Third quartile (Q3)
5. Largest value

STEPS TO DEVELOP FIVE NUMBER SUMMARY

The easiest way to develop a five-number summary is to first place the data in ascending order. Then it is easy to identify the smallest value, the three quartiles, and the largest value.

FIVE NUMBER SUMMARY – EXAMPLE

Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries.

Calculate 5 number summary.

First arrange the data in ascending order

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Quartile	Calculate Quartile index
Q1	(25/100) × 12 = 3
Q2	(50/100) × 12 = 6
Q3	(75/100) × 12 = 9
Q4	(100/100) × 12 = 12

$$i = \left(\frac{p}{100}\right)n$$

i = Index

n = Number of Elements

= 12

FIVE NUMBER SUMMARY – EXAMPLE

1	2	3	4	5	6	7	8	9	10	11	12
3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
8%	17%	25%	33%	42%	50%	58%	67%	75%	83%	92%	100%
		Q1			Q2			Q3			

Average of 3450 and 3480 Average of 3490 and 3520 Average of 3550 and 3650

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925

Q1 = 3450

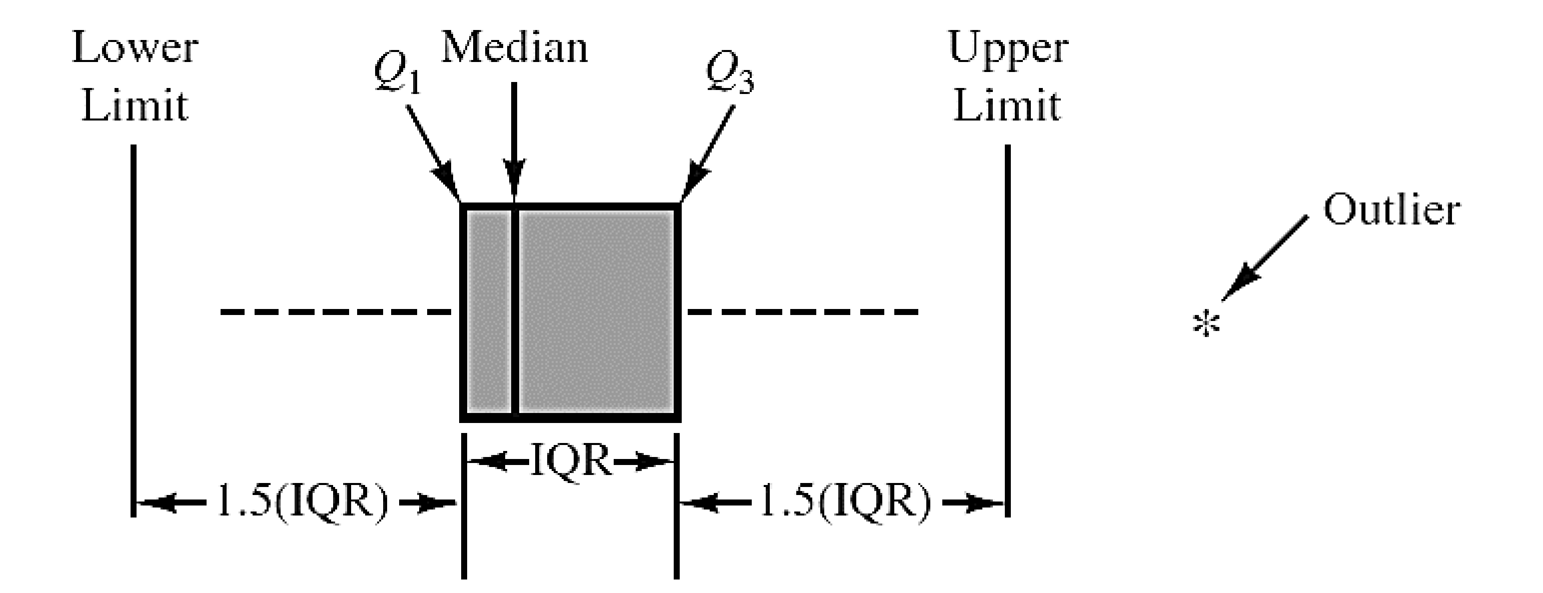
Q2 = 3505
(Median)

Q3 = 3600

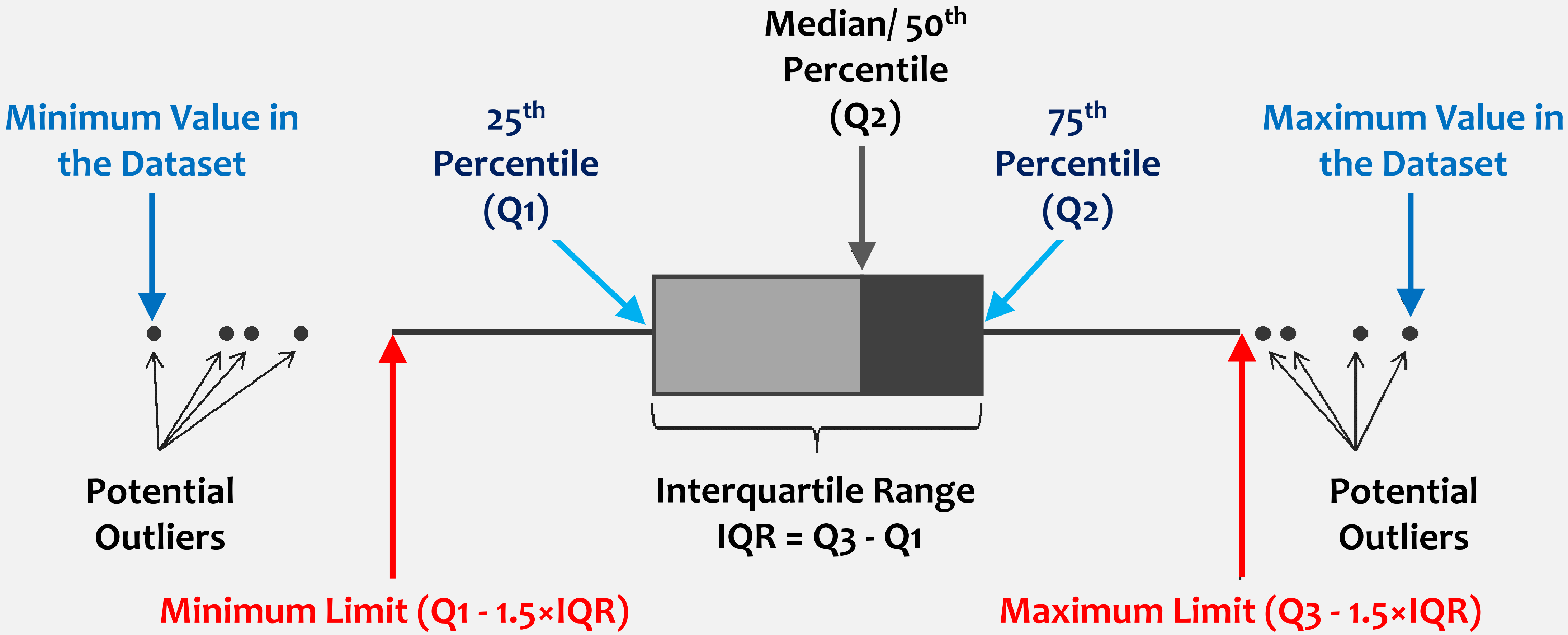
BOX PLOT

- A box plot is a graphical summary of data that is based on a **five-number summary**.
- A key to the development of a **box** plot is the computation of the median, the quartiles, Q1 and Q3, and lowest and highest limit. This box contains the **middle 50%** of the data.
- In a box plot, draw is drawn from the **first quartile** to the **third quartile**. A vertical line is drawn at the median of box. The **whiskers** go from each quartile to the minimum and maximum limit.

By using the ***interquartile range***, **IQR = Q3 - Q1,** limits are located. The **limits** for the box plot are **1.5×(IQR) below Q1** and **1.5×(IQR) above Q3**. Data outside these limits are considered **outliers**.



BOX PLOT - OUTLIER



Median/ 50th Percentile (Q2)

Minimum Value in the Dataset

25th Percentile (Q1)

75th Percentile (Q2)

Maximum Value in the Dataset

Potential Outliers

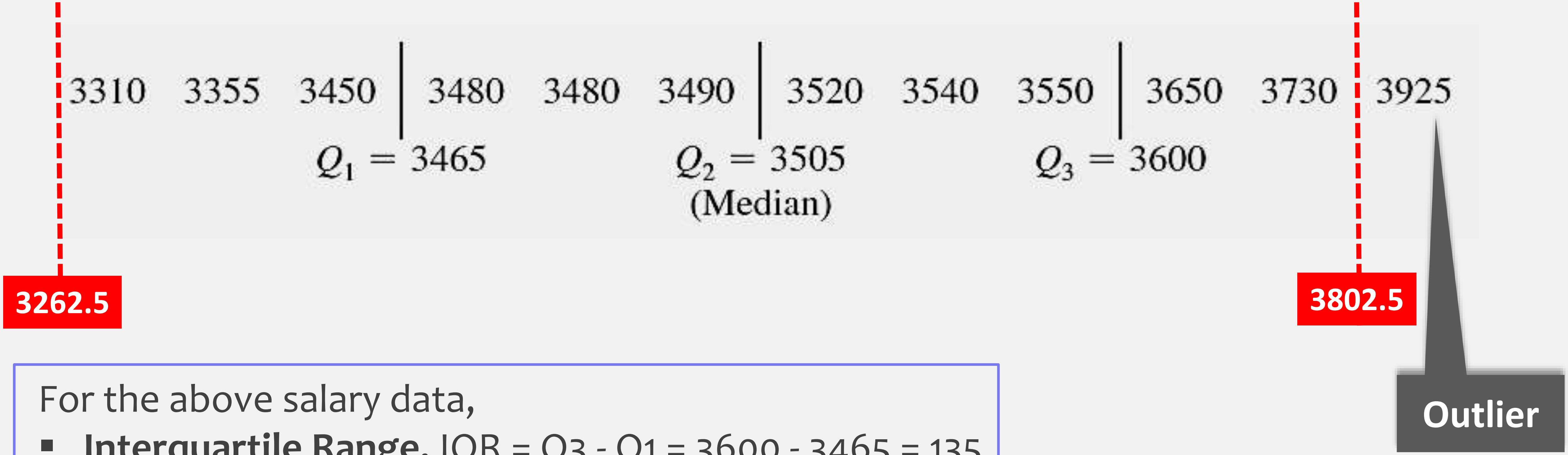
Interquartile Range IQR = Q3 - Q1

Potential Outliers

Minimum Limit (Q1 - 1.5×IQR)

Maximum Limit (Q3 - 1.5×IQR)

BOX PLOT – EXAMPLE



3310 3355 3450 | 3480 3480 3490 | 3520 3540 3550 | 3650 3730 3925

$Q_1 = 3465$ $Q_2 = 3505$ (Median) $Q_3 = 3600$

For the above salary data,

- **Interquartile Range,** IQR = Q3 - Q1 = 3600 - 3465 = 135

The limits are

- **Minimum Limit** = 3465 - 1.5×(135) = 3262.5

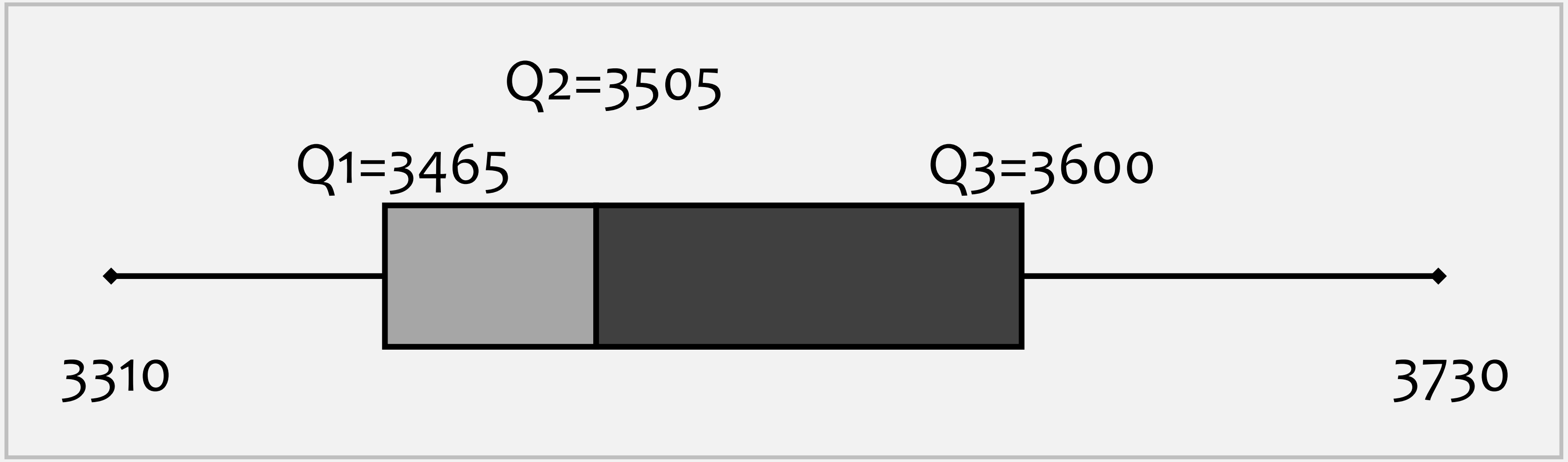
and

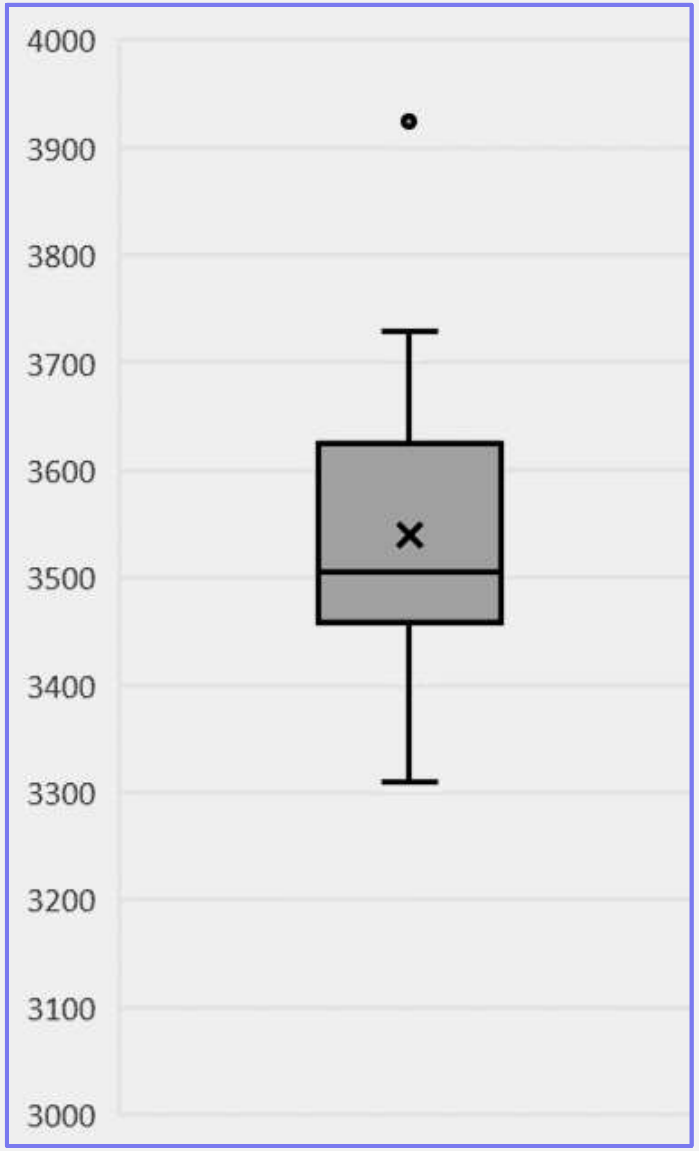
- **Maximum Limit** = 3600 - 1.5×(135) = 3802.5

Any number below this limit is outlier.

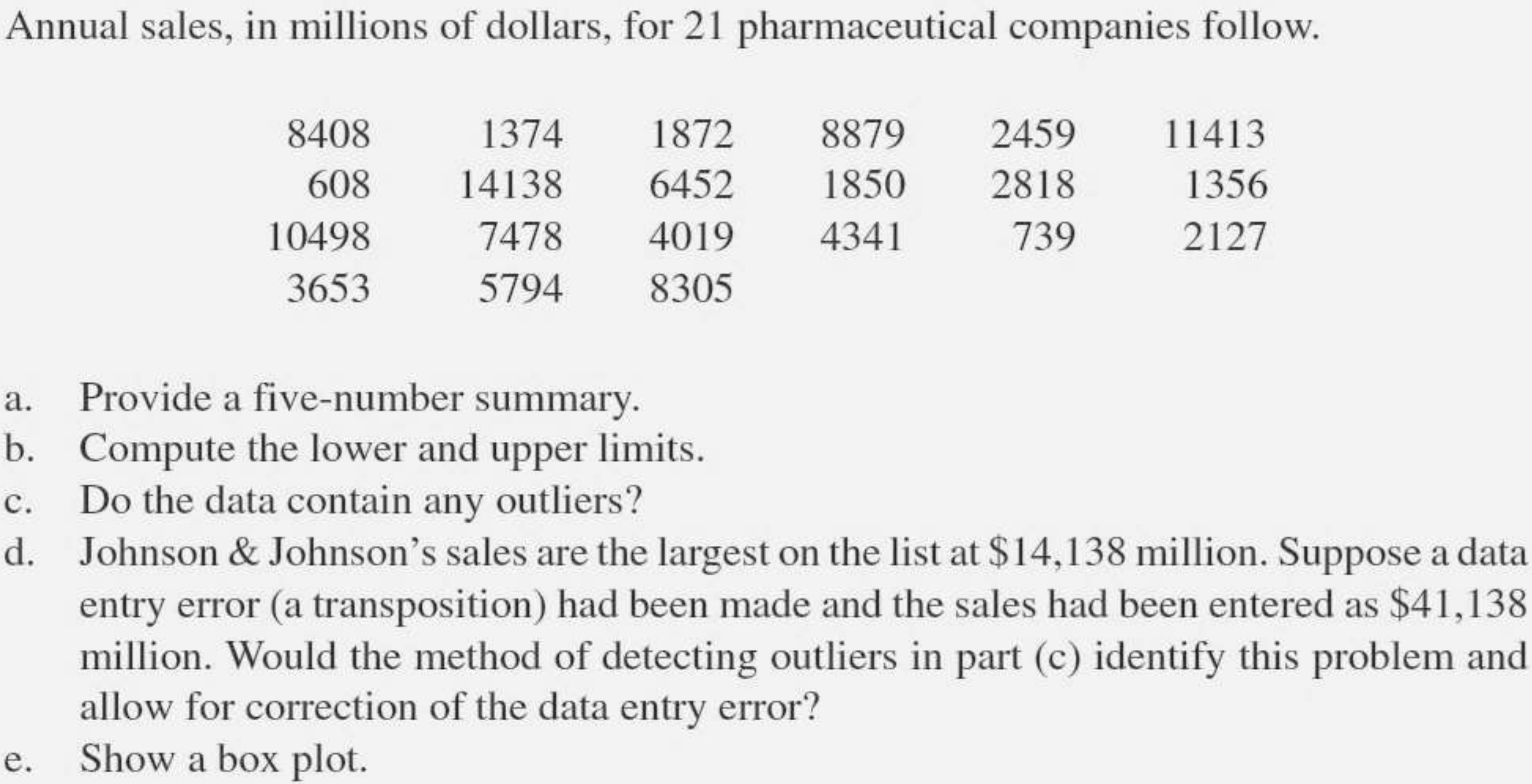
Any number above this limit is outlier.

BOX PLOT - PRESENTATION





EXAMPLE - ANNUAL SALES



Annual sales, in millions of dollars, for 21 pharmaceutical companies follow.

8408	1374	1872	8879	2459	11413
608	14138	6452	1850	2818	1356
10498	7478	4019	4341	739	2127
3653	5794	8305			

a. Provide a five-number summary.

b. Compute the lower and upper limits.

c. Do the data contain any outliers?

d. Johnson & Johnson's sales are the largest on the list at $14,138 million. Suppose a data entry error (a transposition) had been made and the sales had been entered as $41,138 million. Would the method of detecting outliers in part (c) identify this problem and allow for correction of the data entry error?

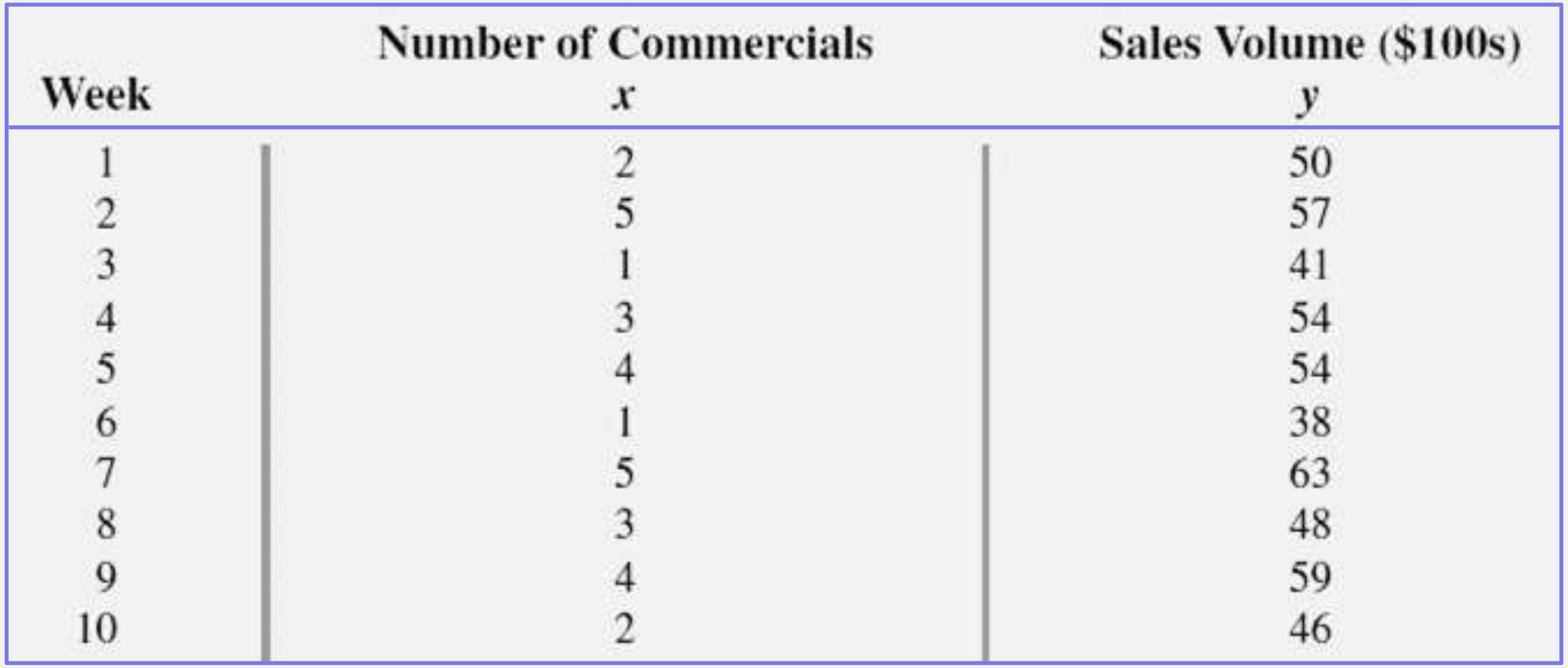
e. Show a box plot.

MEASURES OF ASSOCIATION BETWEEN TWO VARIABLES

7

TV COMMERCIALS & SALES – SLIDE (1/2)

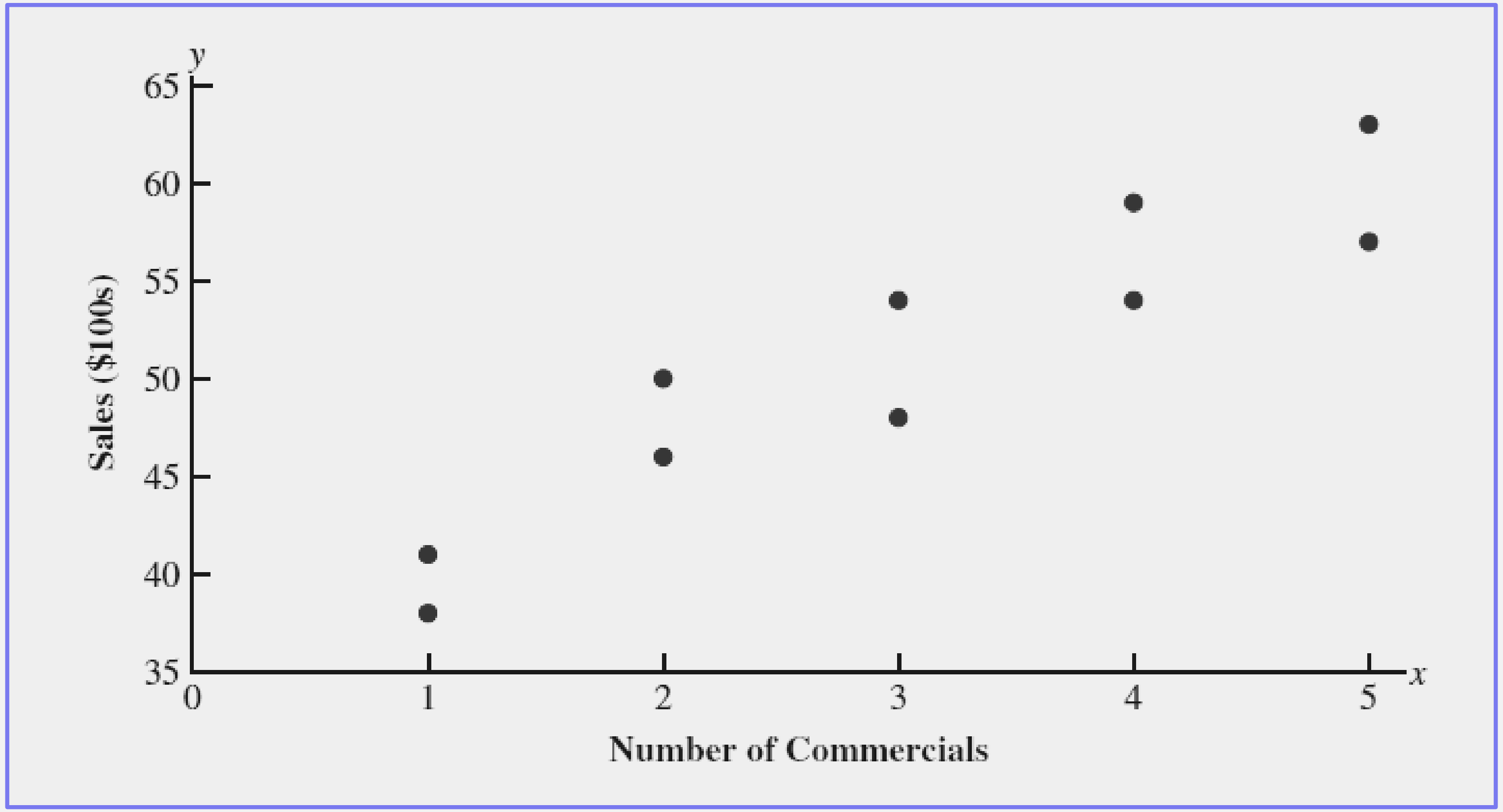
The store’s manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the store during the following week. Sample data with sales expressed in hundreds of dollars are provided in Table. It shows 10 observations (n = 10), one for each week



Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

TV COMMERCIALS & SALES – SLIDE (2/2)

The scatter diagram in Figure shows a positive relationship, with higher sales (y) associated with a greater number of commercials (x). In fact, the scatter diagram suggests that a straight line could be used as an approximation of the relationship.



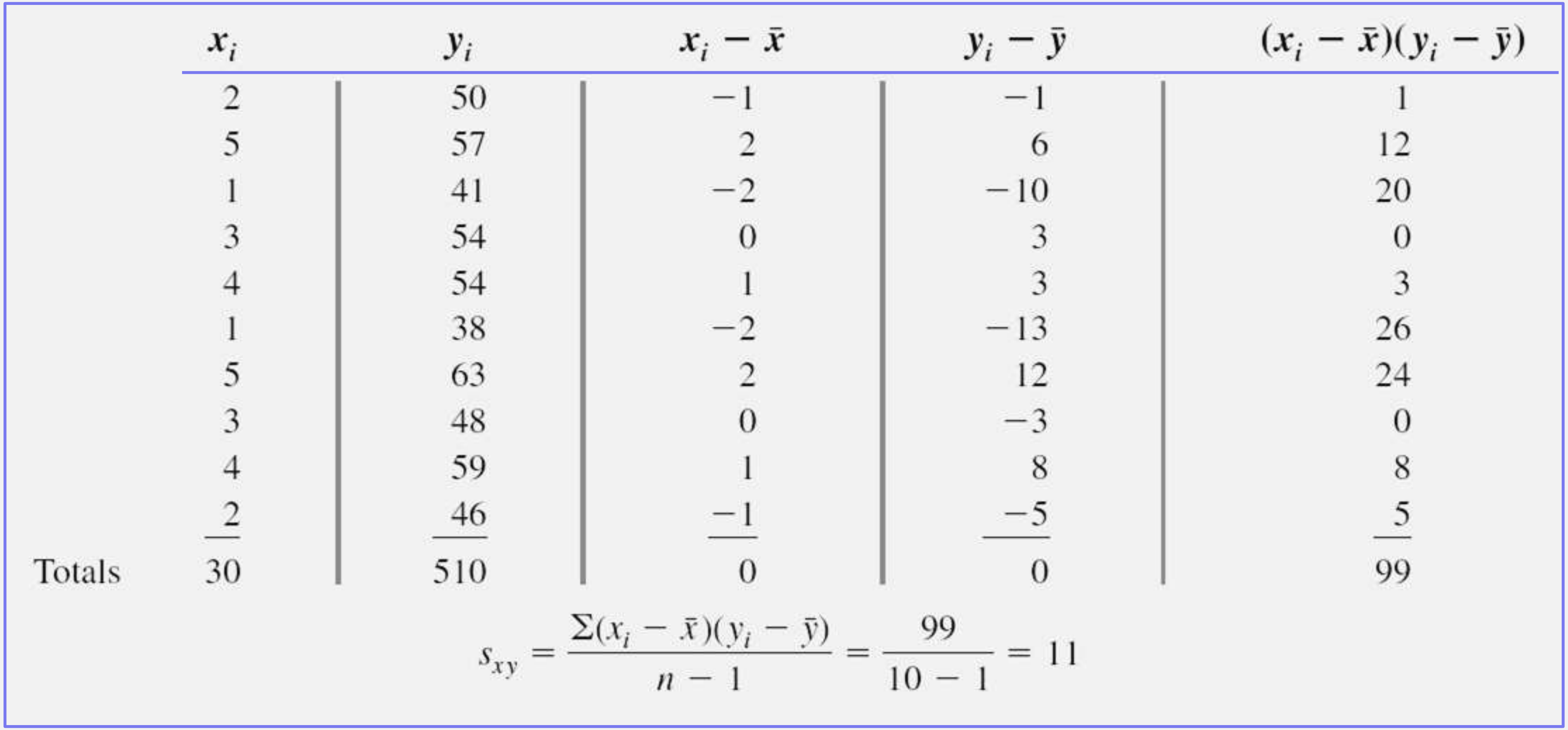
COVARIANCE

- Covariance is a descriptive measure of the linear association between two variables.

- For a sample of size *n* with the observations (*x*1, *y*1), (*x*2, *y*2), and so on, the sample covariance is defined as follows:

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

TV COMMERCIALS & SALES



	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

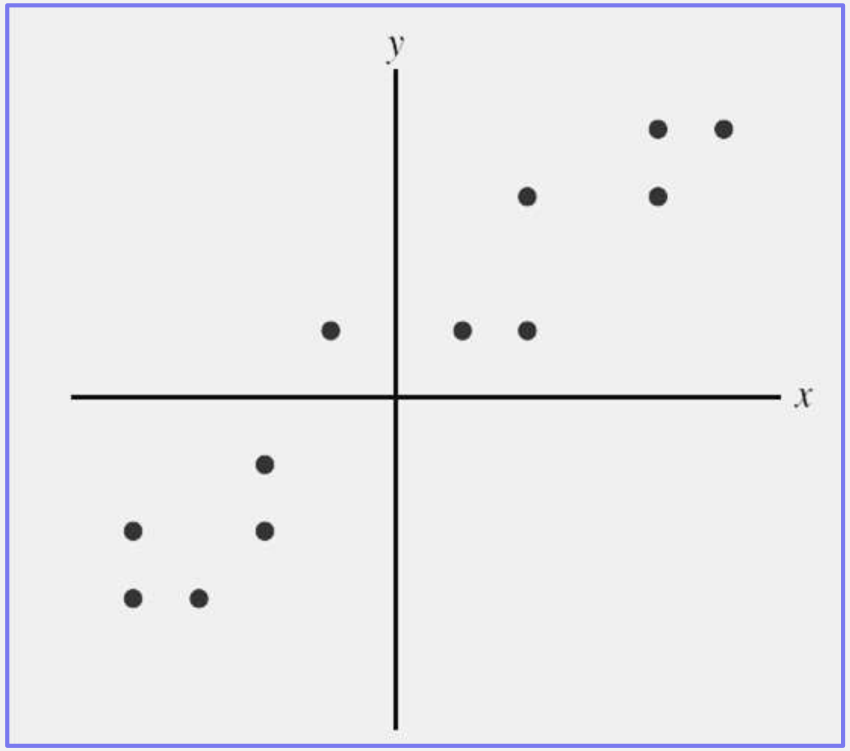
$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

INTERPRETATION OF COVARIANCE

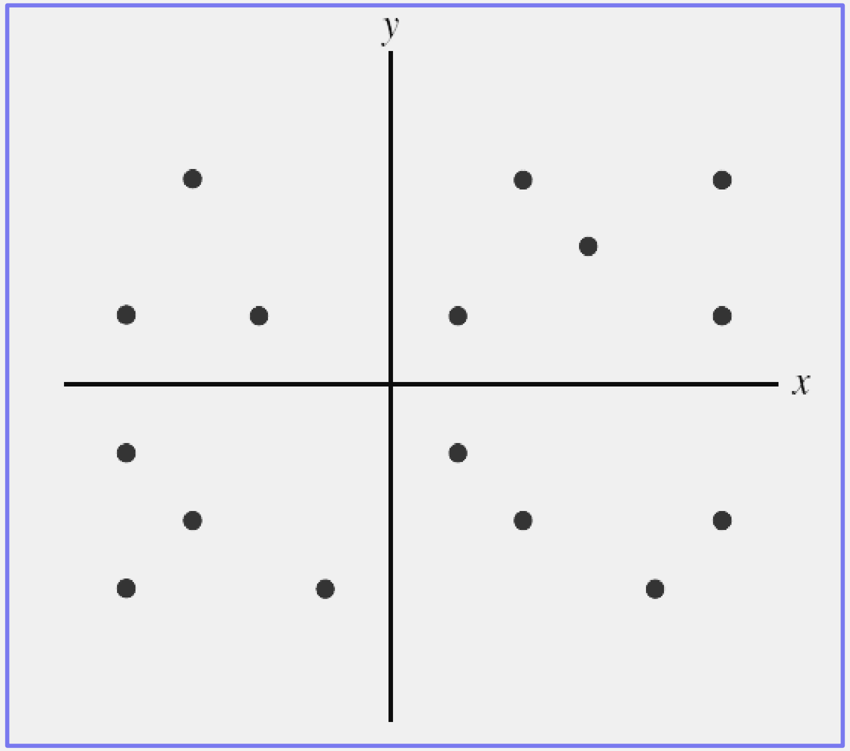
- A **positive value** of covariance indicates a positive linear association between two variables (x and y); i.e. as value of one variable increases value of other variable also increases.
- A **negative value** for covariance indicates a negative linear association between two variables (x and y); that is, as the value of one variable increases, the value of other variable decreases.
- The value of covariance close to **zero**, indicate no linear association between two variables (x and y).

INTERPRETATION OF COVARIANCE USING GRAPH

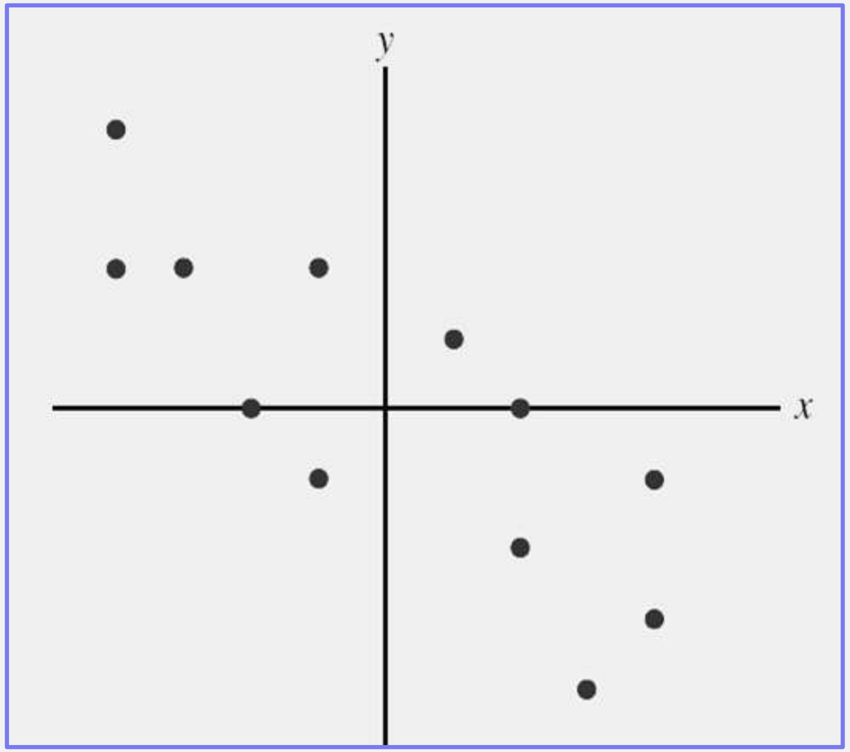
Positive Covariance



Zero Covariance



Negative Covariance



PROBLEM WITH COVARIANCE

- One problem with using covariance as a measure of the strength of the linear relationship is that the value of the covariance depends on the **units of measurement** for x and y.

- Suppose we are interested in the relationship between height *x* and weight *y* for individuals. Clearly the strength of the relationship should be the same whether we measure height in feet or inches. Measuring the height in inches, however, gives us much larger numerical values for $(x_i - \bar{x})$ than when we measure height in feet. Thus, with height measured in inches, we would obtain a **larger value for the numerator** $\sum(x_i - \bar{x})(y_i - \bar{y})$and hence a **larger covariance** — when in fact the relationship does not change.

CORRELATION COEFFICIENT

- Correlation coefficient is used to avoid measure if relationship between two variables x and y having the different units of measurement.
- The correlation coefficient ranges from -1 to +1.
- Values close to -1 or +1 indicate a strong linear relationship. The closer the correlation is to zero, the weaker the relationship.

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

r_{xy} = sample correlation coefficient

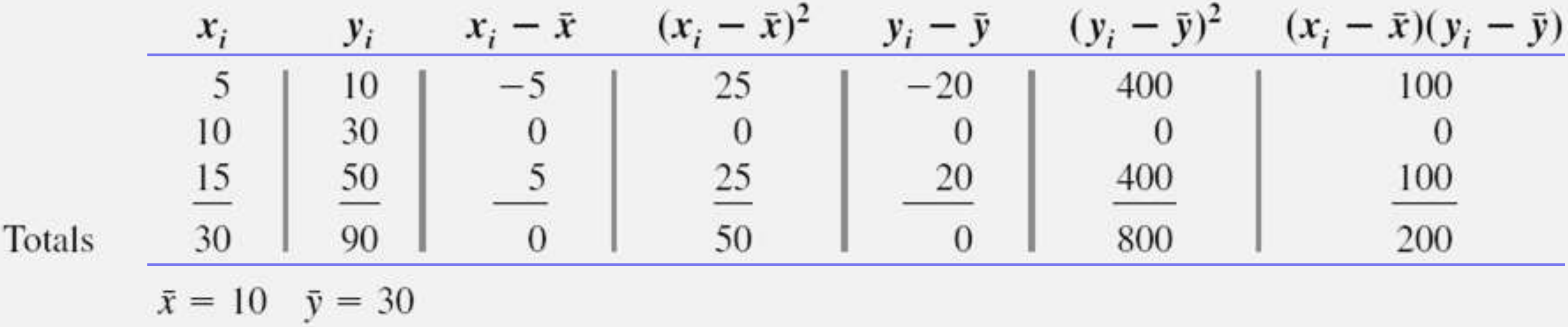
s_{xy} = sample covariance

s_x = sample standard deviation of x

s_y = sample standard deviation of y

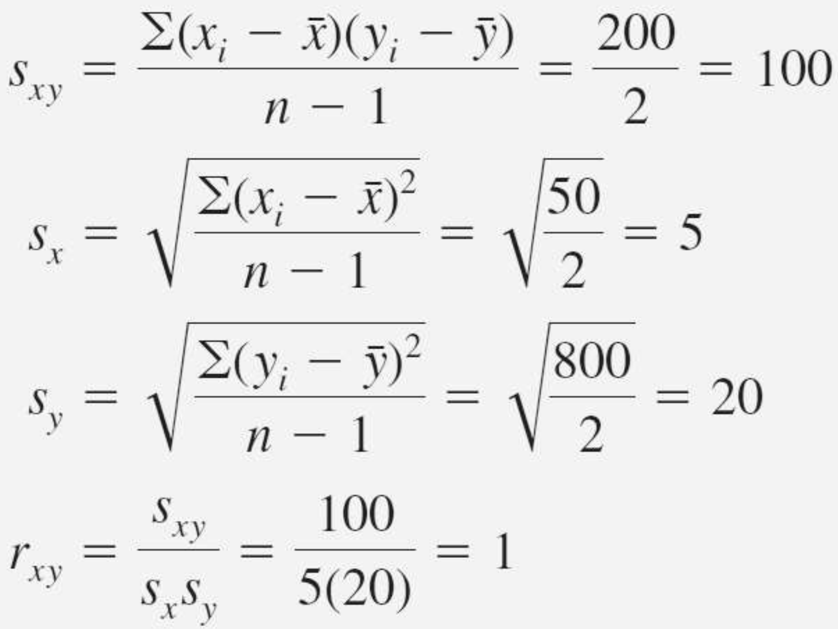
Pearson product moment correlation coefficient simply known as Correlation Coefficient

EXAMPLE - PERFECT +VE LINEAR RELATIONSHIP



	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totals	30	90	0	50	0	800	200

$\bar{x} = 10 \quad \bar{y} = 30$



$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

Thus, we see that the value of the sample correlation coefficient is 1.

INTERPRETATION OF CORRELATION COEFFICIENT

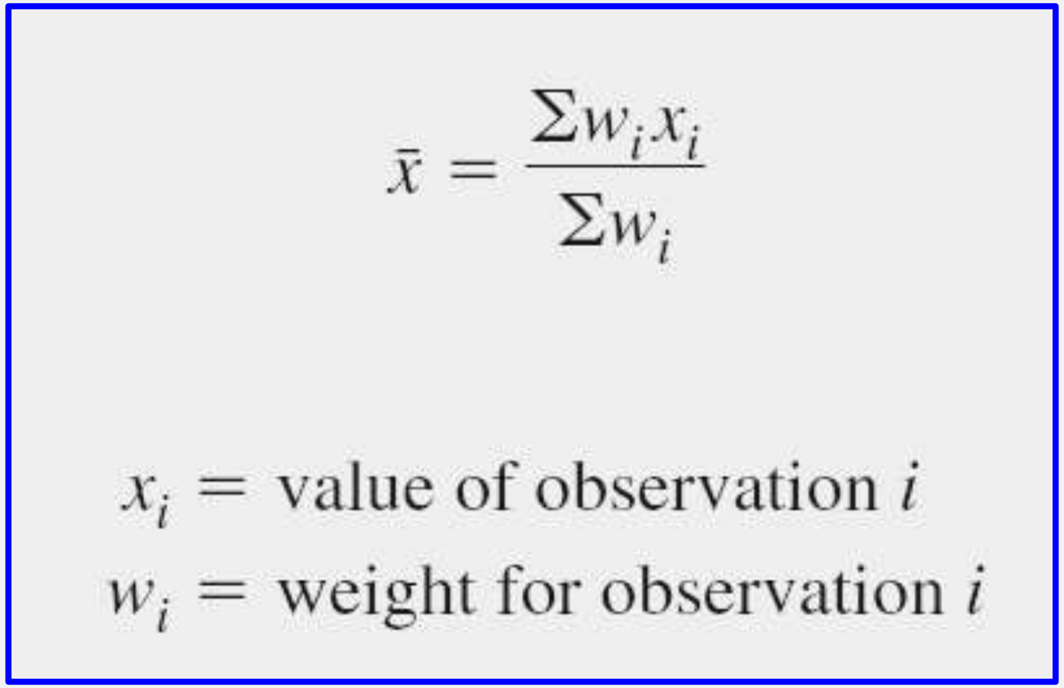
- Let us now suppose that a certain data set indicates a positive linear relationship between *x* and *y* but that the relationship is not perfect.
- The value of r_{xy} will be less than 1, indicating that the points in the scatter diagram are not all on a straight line.
- As the points deviate more and more from a perfect positive linear relationship, the value of r_{xy} becomes smaller and smaller.
- A value of r_{xy} equal to zero indicates no linear relationship between *x* and *y*, and values of r_{xy} near zero indicate a weak linear relationship.

WEIGHTED MEAN
&
GROUP DATA

7

WEIGHTED MEAN

The mean is computed by giving each observation a weight that reflects its importance. A mean computed in this manner is referred to as a weighted mean.



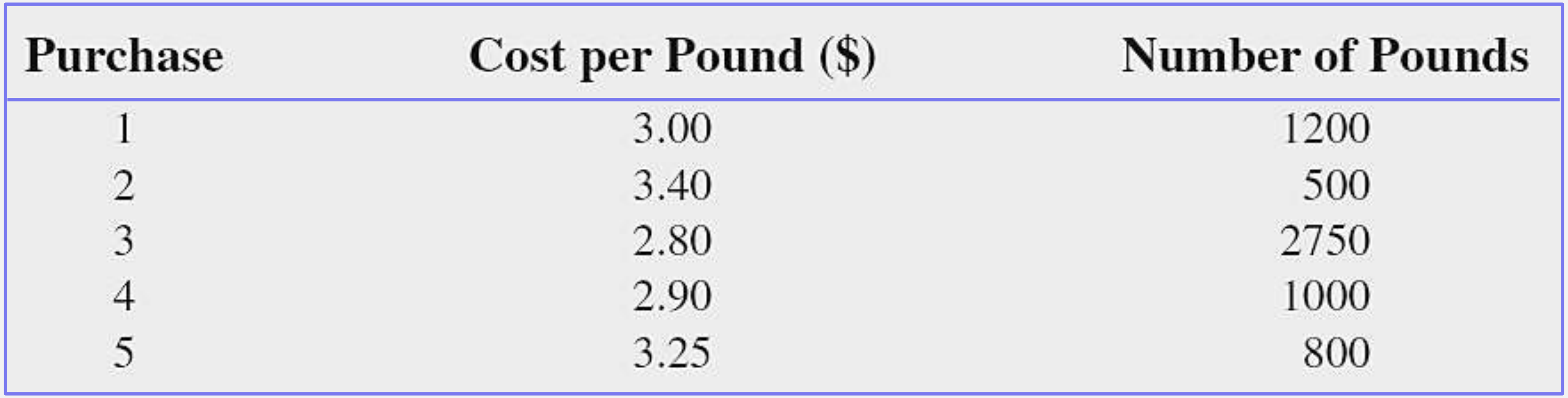
$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

x_i = value of observation i

w_i = weight for observation i

WEIGHTED MEAN - EXAMPLE

Consider the sample of five purchases of a raw material over the past three months. The cost per pound varies from $2.80 to $3.40, and the quantity purchased varies from 500 to 2750 pounds. Suppose that a manager asked for information about the mean cost per pound of the raw material. Because the quantities ordered vary, we must use the formula for a weighted mean.



Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

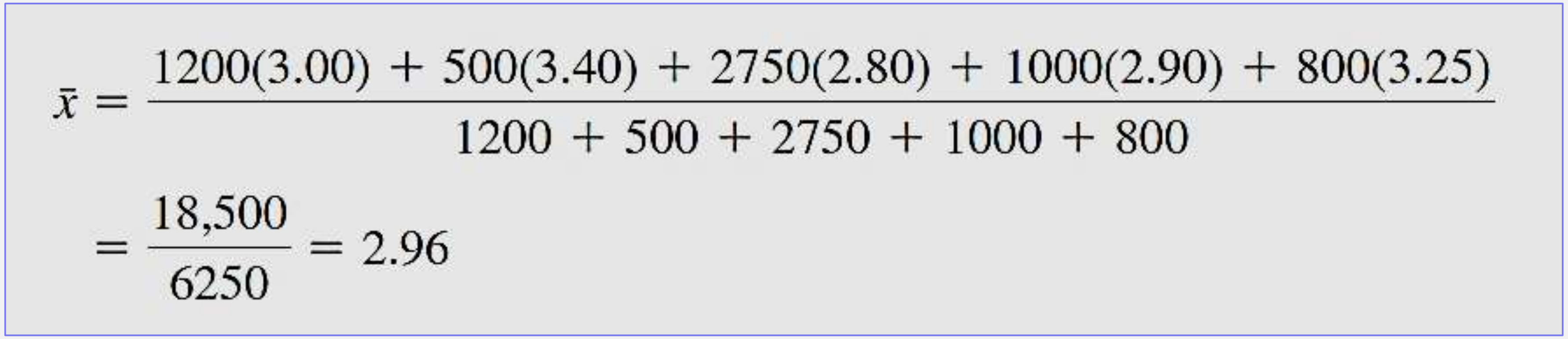
EXAMPLE – COST PER POUND

Calculate Weighted Mean

- The five cost-per-pound data values are

  x1 = 3.00, x2 = 3.40, x3 = 2.80, x4 = 2.90, and x5 = 3.25.

- The weights are w1 = 1200, w2 = 500, w3 = 2750, w4 = 1000, and w5 = 800



$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800}$$

$$= \frac{18{,}500}{6250} = 2.96$$

The weighted mean computation shows that the mean cost per pound for the raw material is $2.96.

GROUPED DATA

- Weighted mean formula can be used to obtain approximations of the mean, variance, and standard deviation for **grouped data**.
- To compute the mean using only the grouped data, we treat the midpoint of each class as being representative of the items in the class.
- Weights are given by the frequencies.

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

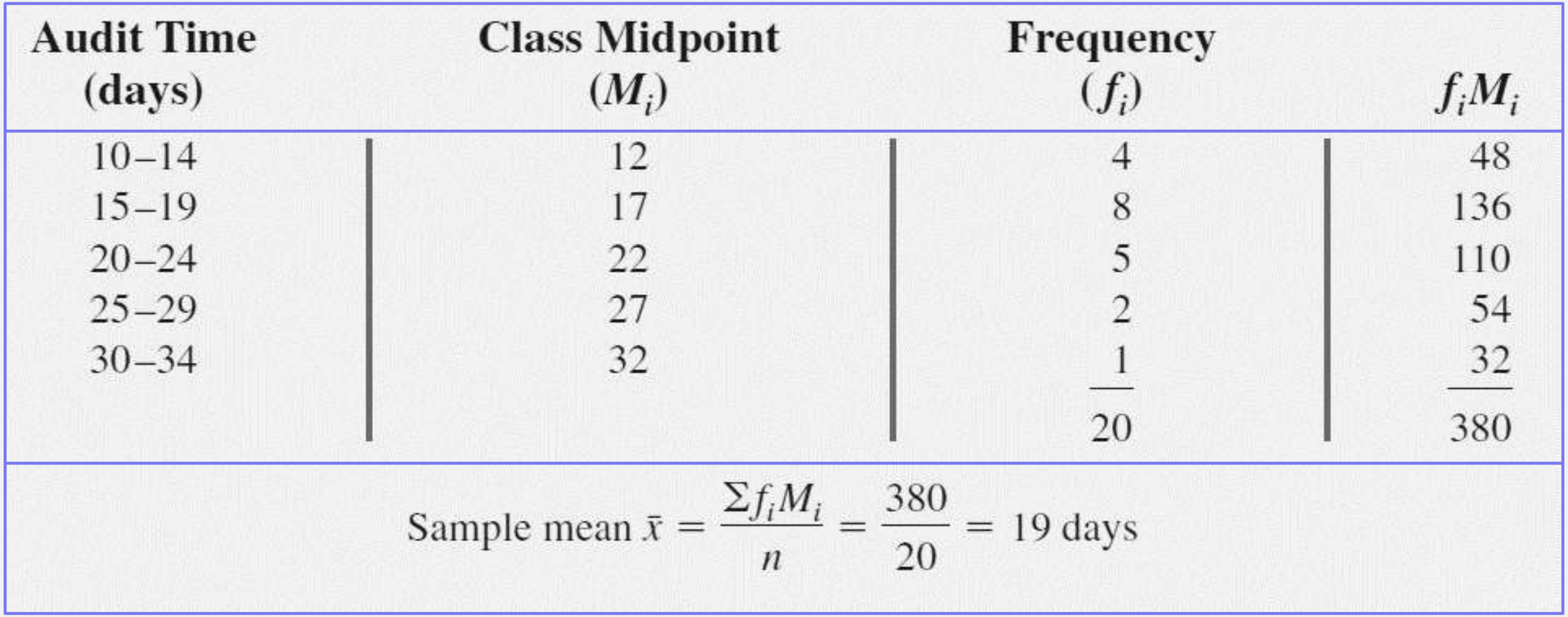
M_i = the midpoint for class i

f_i = the frequency for class i

n = the sample size

EXAMPLE – AUDIT TIME

Calculate Mean

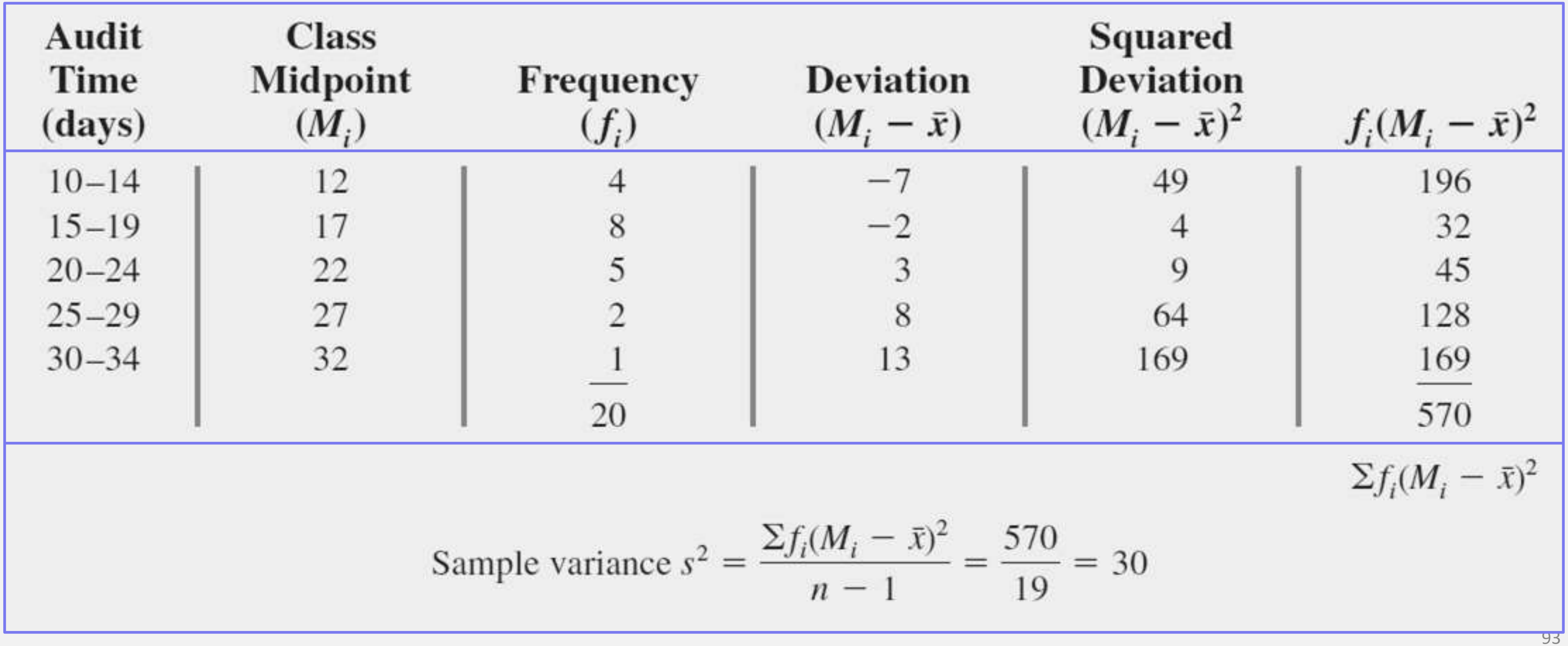


Audit Time (days)	Class Midpoint (M_i)	Frequency (f_i)	f_iM_i
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

$$\text{Sample mean } \bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19 \text{ days}$$

EXAMPLE – VARIANCE

Calculate Variance



Audit Time (days)	Class Midpoint (M_i)	Frequency (f_i)	Deviation ($M_i - \bar{x}$)	Squared Deviation $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
10–14	12	4	−7	49	196
15–19	17	8	−2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Sample variance } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$$

QUESTION AND ANSWERS