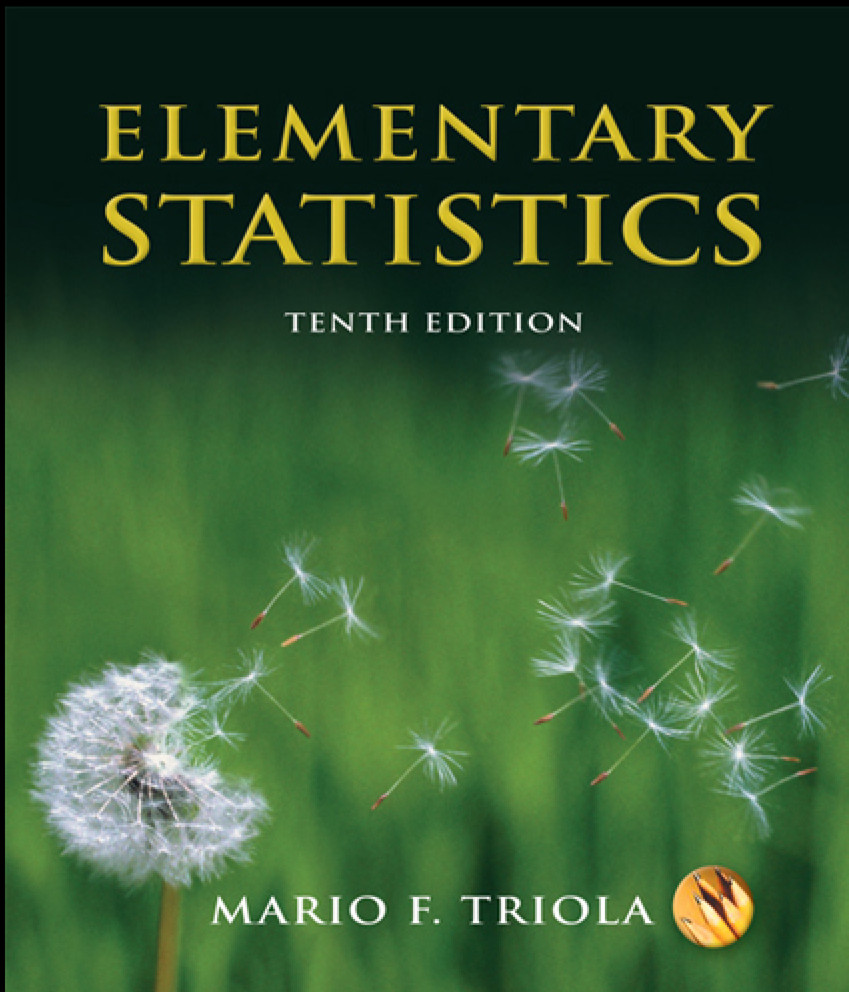
Lecture Slides



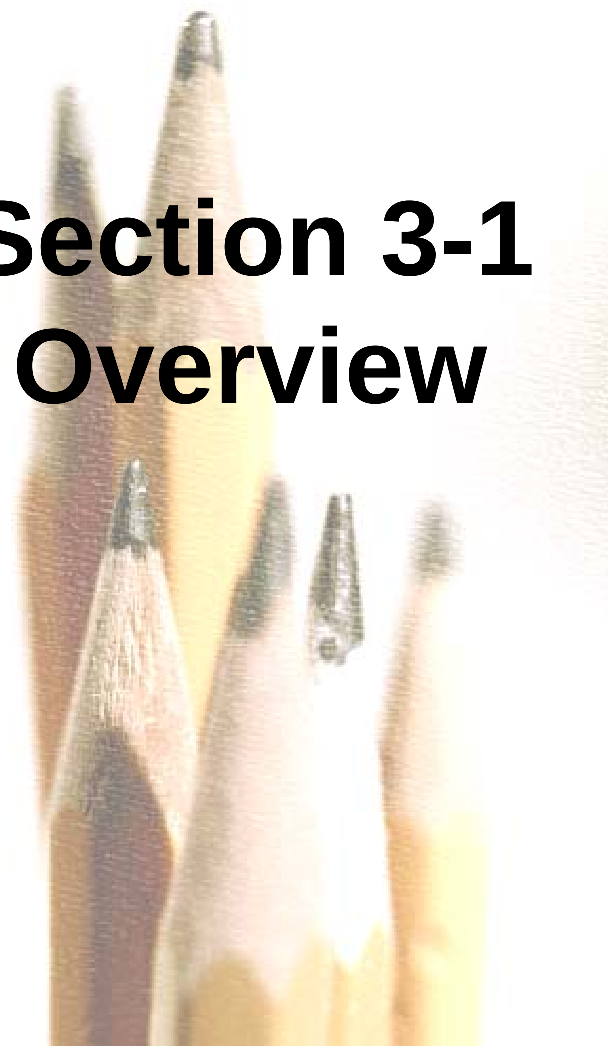
Elementary Statistics
Tenth Edition

and the Triola Statistics Series

by Mario F. Triola

Chapter 3
Statistics for Describing, Exploring, and Comparing Data

Section 3-1
Overview



Created by Tom Wegleitner, Centreville, Virginia

Overview

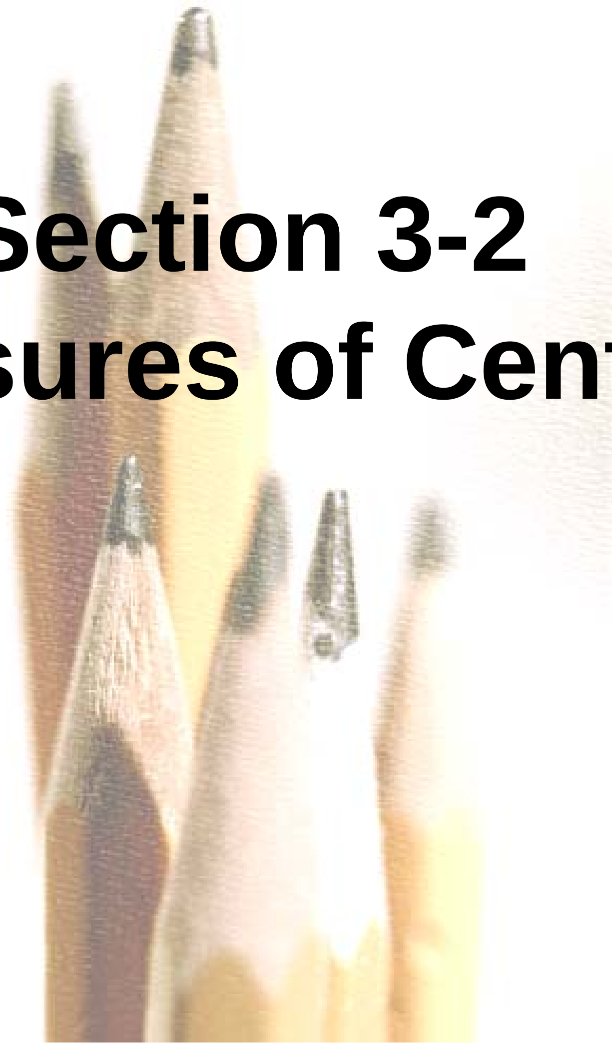
- **Descriptive Statistics**

  summarize or describe the important characteristics of a known set of data

- **Inferential Statistics**

  use sample data to make inferences (or generalizations) about a population

Section 3-2
Measures of Center



Created by Tom Wegleitner, Centreville, Virginia

Key Concept

When describing, exploring, and comparing data sets, these characteristics are usually extremely important: center, variation, distribution, outliers, and changes over time.

Definition

- **Measure of Center**

  the value at the center or middle of a data set

Definition

Arithmetic Mean

(Mean)

the measure of center obtained by adding the values and dividing the total by the number of values

Notation

Σ denotes the **sum** of a set of values.

x is the **variable** usually used to represent the individual data values.

n represents the **number of values** in a **sample**.

N represents the **number of values** in a **population**.

Notation

$\bar{x}$ is pronounced 'x-bar' and denotes the mean of a set of **sample** values

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

μ is pronounced 'mu' and denotes the mean of all values in a **population**

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

Definitions

- **Median**

  the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

- **often denoted by $\tilde{x}$ (pronounced 'x-tilde')**

- **is not affected by an extreme value**

Finding the Median

- If the number of values is odd, the median is the number located in the exact middle of the list.

- If the number of values is even, the median is found by computing the mean of the two middle numbers.

5.40	1.10	0.42	0.73	0.48	1.10
0.42	0.48	0.73	1.10	1.10	5.40

(in order - even number of values – no exact middle shared by two numbers)

$$\frac{0.73 + 1.10}{2}$$

MEDIAN is 0.915

5.40	1.10	0.42	0.73	0.48	1.10	0.66
0.42	0.48	0.66	0.73	1.10	1.10	5.40

(in order - odd number of values)

exact middle **MEDIAN is 0.73**

Definitions

- **Mode**

  the value that occurs most frequently

- **Mode is not always unique**

- **A data set may be:**

  Bimodal

  Multimodal

  No Mode

Mode is the only measure of central tendency that can be used with nominal data

Mode - Examples

a. 5.40 1.10 0.42 0.73 0.48 1.10	⇦ Mode is 1.10
b. 27 27 27 55 55 55 88 88 99	⇦ Bimodal - 27 & 55
c. 1 2 3 6 7 8 9 10	⇦ No Mode

Definition

- **Midrange**

the value midway between the maximum and minimum values in the original data set

$$\textbf{Midrange} = \frac{\text{maximum value + minimum value}}{2}$$

Round-off Rule for Measures of Center

Carry one more decimal place than is present in the original set of values.

Mean from a Frequency Distribution

Assume that in each class, all sample values are equal to the class midpoint.

Mean from a Frequency Distribution

use class midpoint of classes for variable *x*

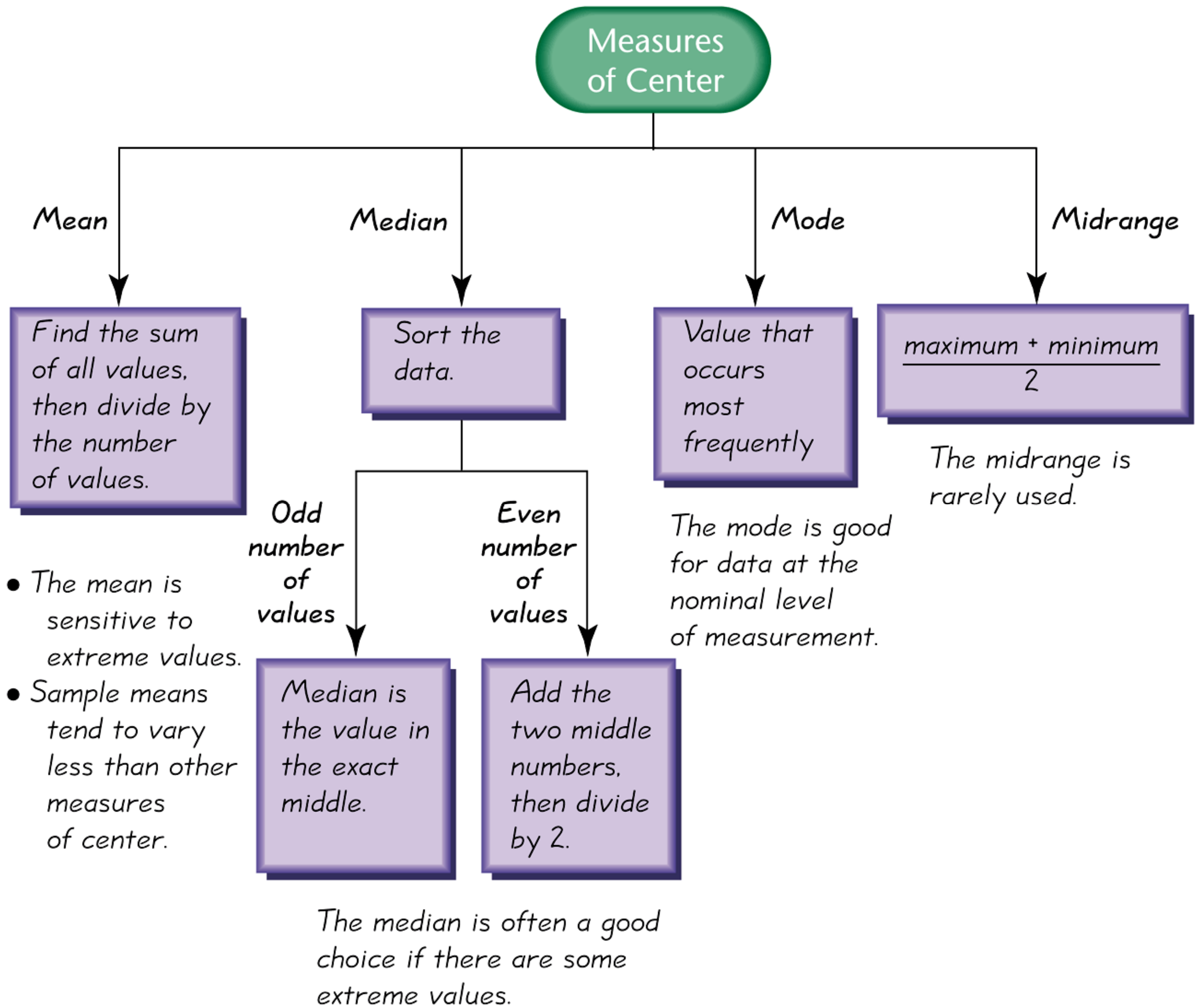
$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f}$$

Weighted Mean

In some cases, values vary in their degree of importance, so they are weighted accordingly.

$$\bar{x} = \frac{\Sigma\,(w \cdot x)}{\Sigma\, w}$$

Best Measure of Center



Definitions

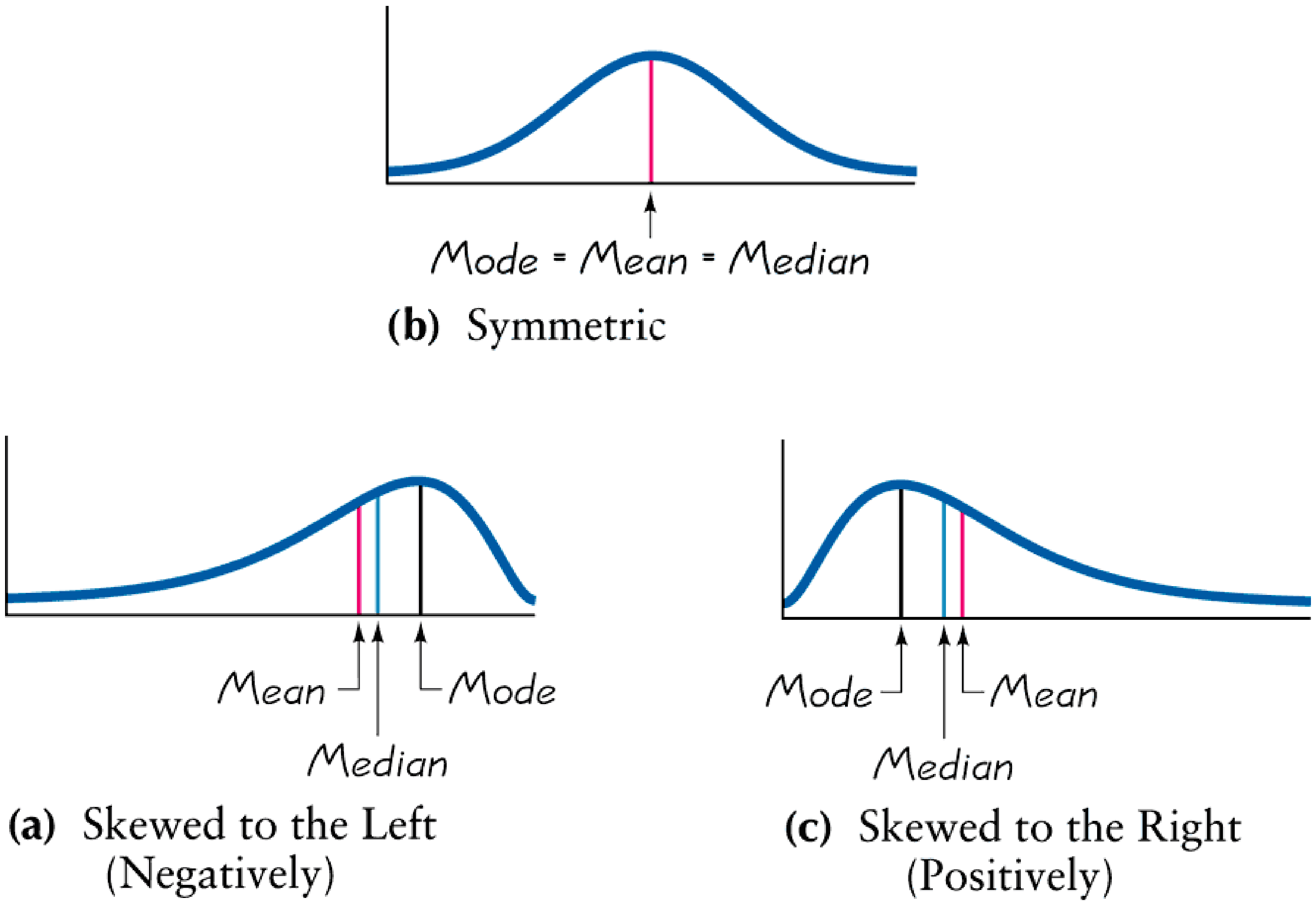
- **Symmetric**

  distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half

- **Skewed**

  distribution of data is skewed if it is not symmetric and if it extends more to one side than the other

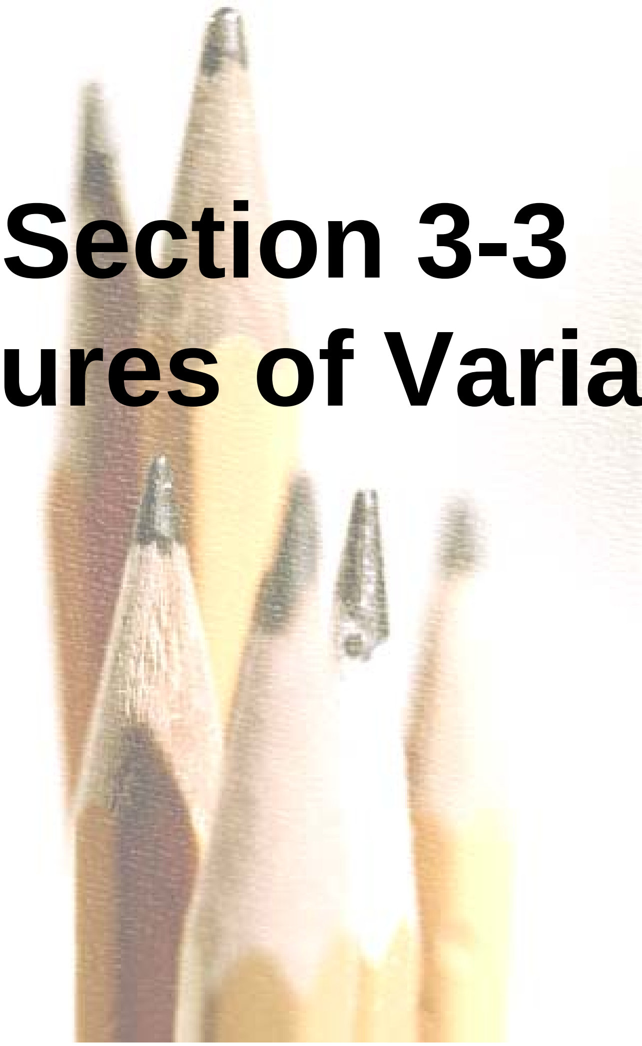
Skewness


Mode = Mean = Median
(b) Symmetric
Mean
Mode
Median
(a) Skewed to the Left
(Negatively)
Mode
Mean
Median
(c) Skewed to the Right
(Positively)

Recap

In this section we have discussed:

- **Types of measures of center**
  - **Mean**
  - **Median**
  - **Mode**
- **Mean from a frequency distribution**
- **Weighted means**
- **Best measures of center**
- **Skewness**



Section 3-3
Measures of Variation

Created by Tom Wegleitner, Centreville, Virginia

Key Concept

Because this section introduces the concept of variation, which is something so important in statistics, this is one of the most important sections in the entire book.

Place a high priority on how to interpret values of standard deviation.

Definition

The range of a set of data is the difference between the maximum value and the minimum value.

Range = (maximum value) – (minimum value)

Definition

The standard deviation of a set of sample values is a measure of variation of values about the mean.

Sample Standard Deviation Formula

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

Sample Standard Deviation (Shortcut Formula)

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

Standard Deviation - Important Properties

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation s is usually positive.
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation s are the same as the units of the original data values.

Population Standard Deviation

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

This formula is similar to the previous formula, but instead, the population mean and population size are used.

Definition

- The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.

- Sample variance: Square of the sample standard deviation *s*

- Population variance: Square of the population standard deviation σ

Variance - Notation

standard deviation squared

Notation $\begin{cases} s^2 & \text{Sample variance} \\ \sigma^2 & \text{Population variance} \end{cases}$

Round-off Rule for Measures of Variation

Carry one more decimal place than is present in the original set of data.

Round only the final answer, not values in the middle of a calculation.

Estimation of Standard Deviation Range Rule of Thumb

For estimating a value of the standard deviation *s*,

Use

$$s \approx \frac{\text{Range}}{4}$$

Where range = (maximum value) – (minimum value)

Estimation of Standard Deviation Range Rule of Thumb

For interpreting a known value of the standard deviation *s*, find rough estimates of the minimum and maximum "usual" sample values by using:

Minimum "usual" value = (mean) – 2 X (standard deviation)

Maximum "usual" value = (mean) + 2 X (standard deviation)

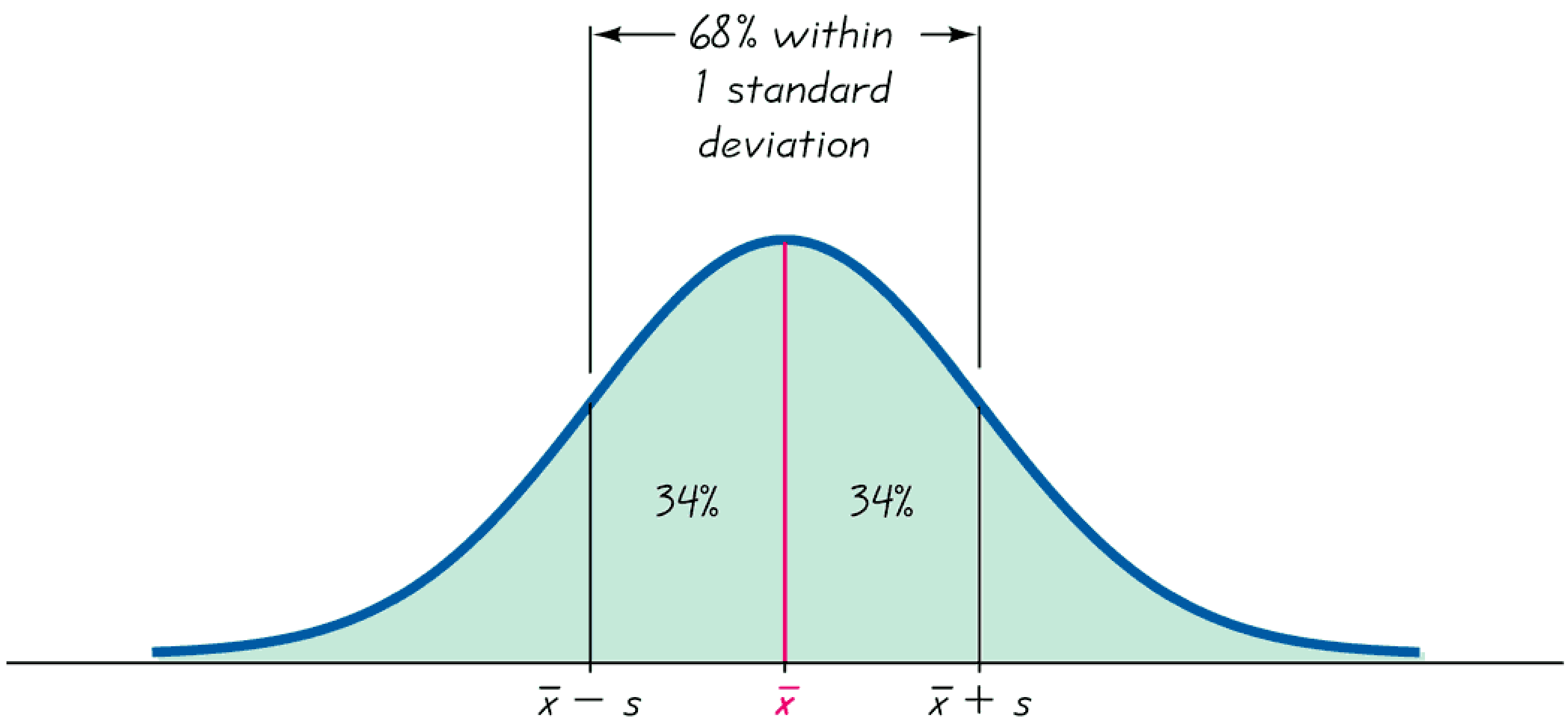
Definition

Empirical (68-95-99.7) Rule

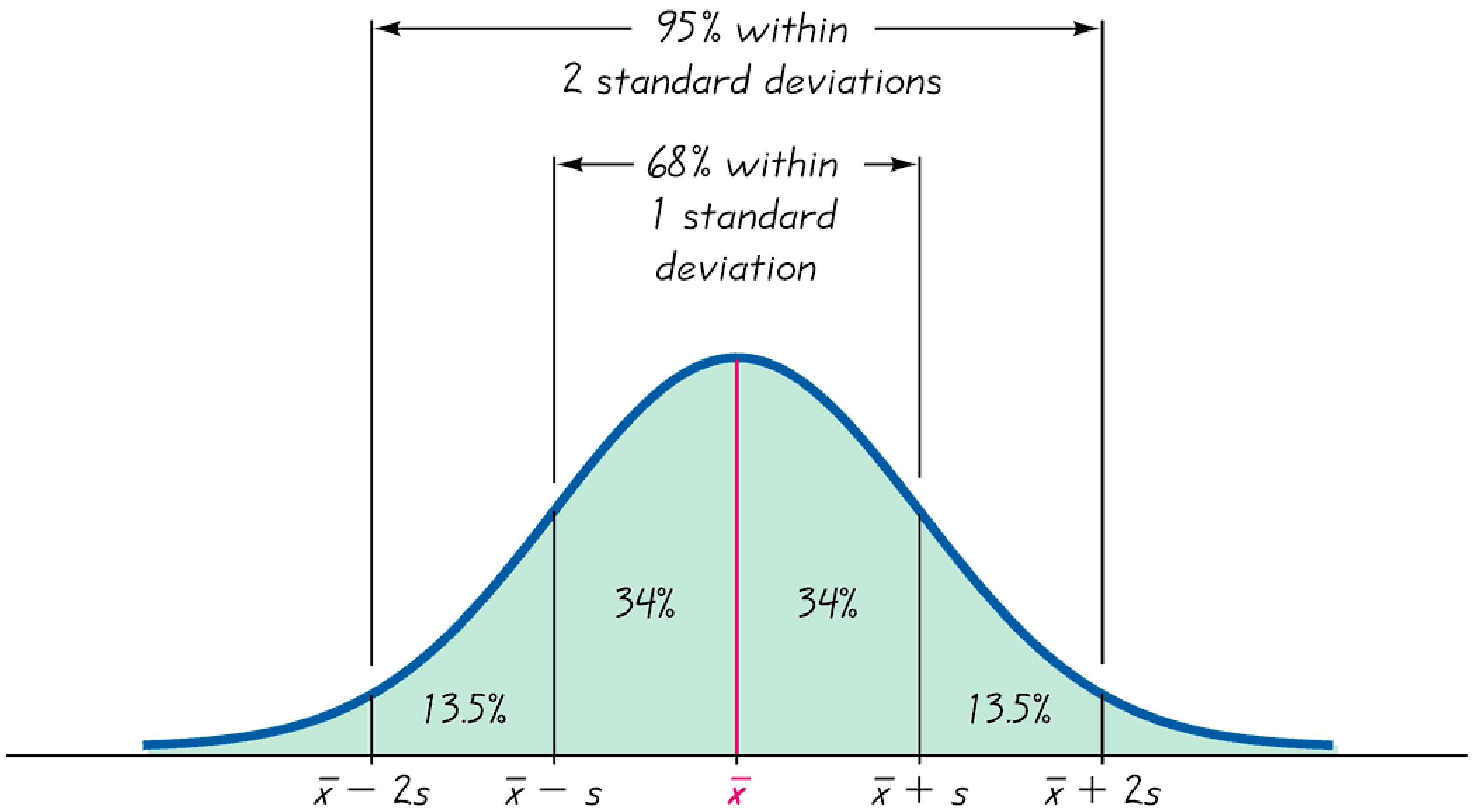
For data sets having a distribution that is approximately bell shaped, the following properties apply:

- **About 68% of all values fall within 1 standard deviation of the mean.**
- **About 95% of all values fall within 2 standard deviations of the mean.**
- **About 99.7% of all values fall within 3 standard deviations of the mean.**

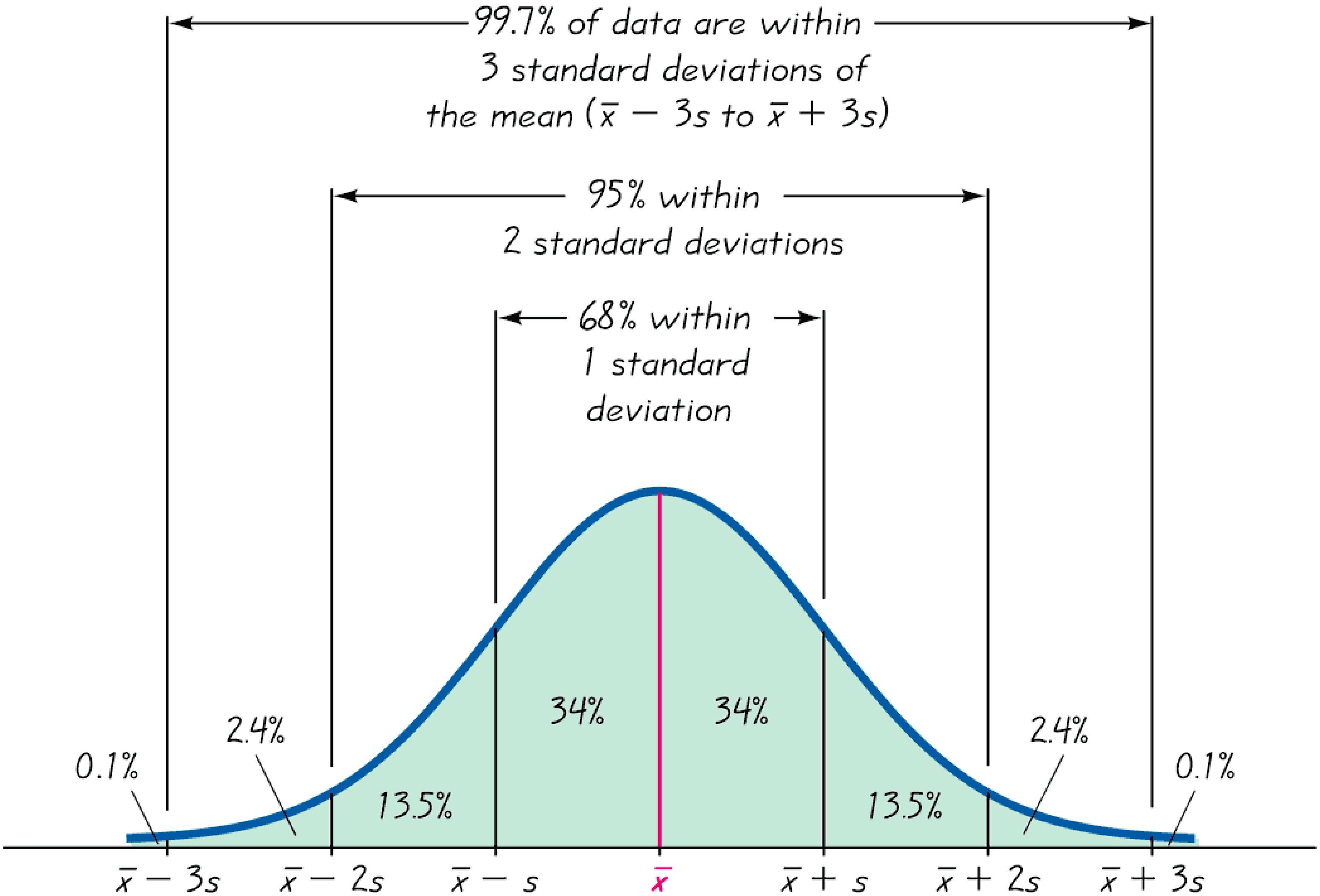
The Empirical Rule



The Empirical Rule



The Empirical Rule



Definition

Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $1-1/K^2$, where K is any positive number greater than 1.

- **For $K = 2$, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.**
- **For $K = 3$, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.**

Rationale for using n-1 versus n

The end of Section 3-3 has a detailed explanation of why n – 1 rather than n is used. The student should study it carefully.

Definition

The coefficient of variation (or CV) for a set of sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

Sample

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

Population

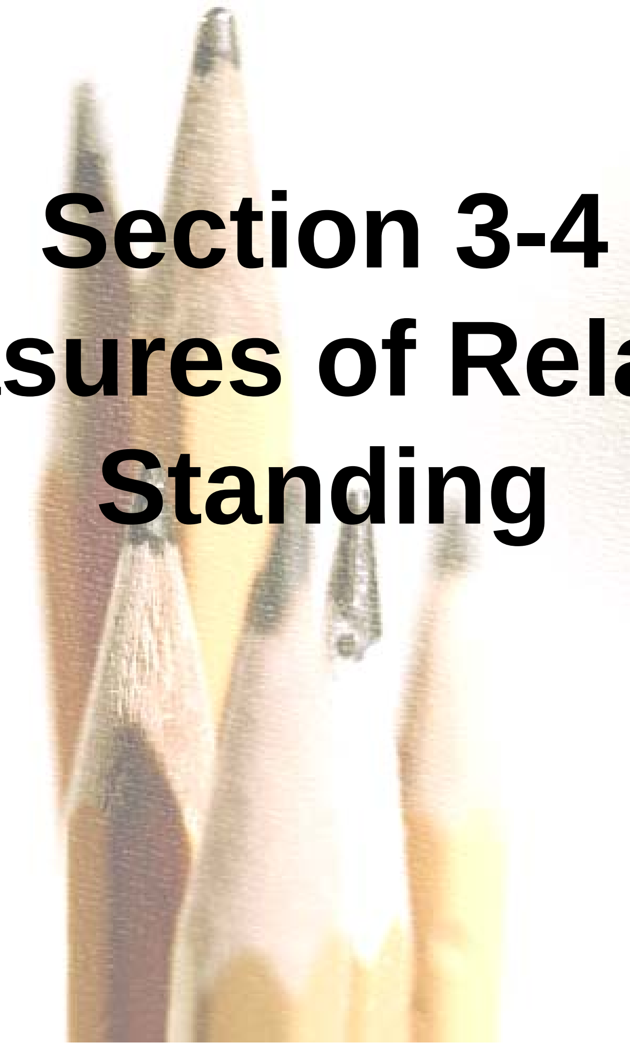
$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Recap

In this section we have looked at:

- **Range**
- **Standard deviation of a sample and population**
- **Variance of a sample and population**
- **Range rule of thumb**
- **Empirical distribution**
- **Chebyshev's theorem**
- **Coefficient of variation (CV)**

Section 3-4
Measures of Relative Standing



Created by Tom Wegleitner, Centreville, Virginia

Key Concept

This section introduces measures that can be used to compare values from different data sets, or to compare values within the same data set. The most important of these is the *z* score.

Definition

❖ ***z* Score** (or standardized value)

the number of standard deviations that a given value *x* is above or below the mean

Measures of Position *z* score

Sample

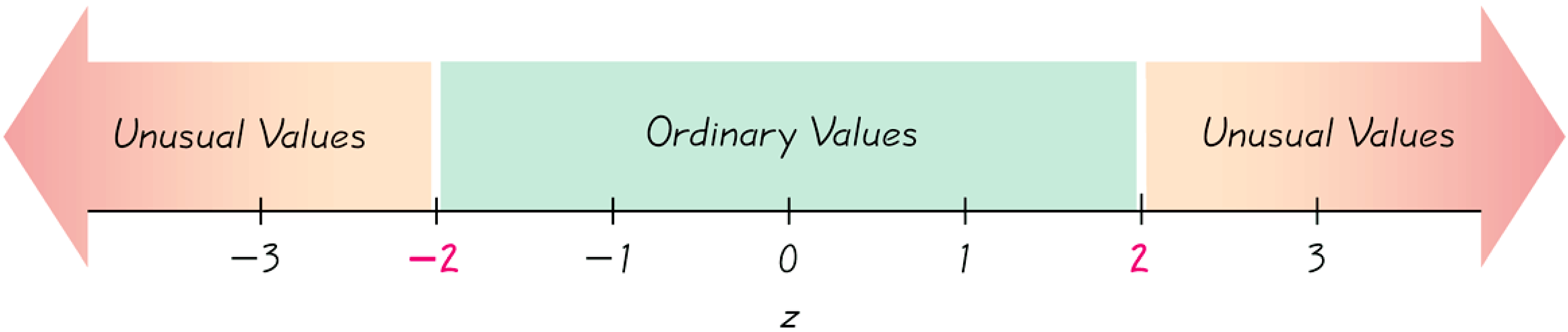
$$z = \frac{x - \bar{x}}{s}$$

Population

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

Round *z* to 2 decimal places

Interpreting *Z* Scores



Whenever a value is less than the mean, its corresponding z score is negative

Ordinary values: z score between –2 and 2
Unusual Values: z score < -2 or z score > 2

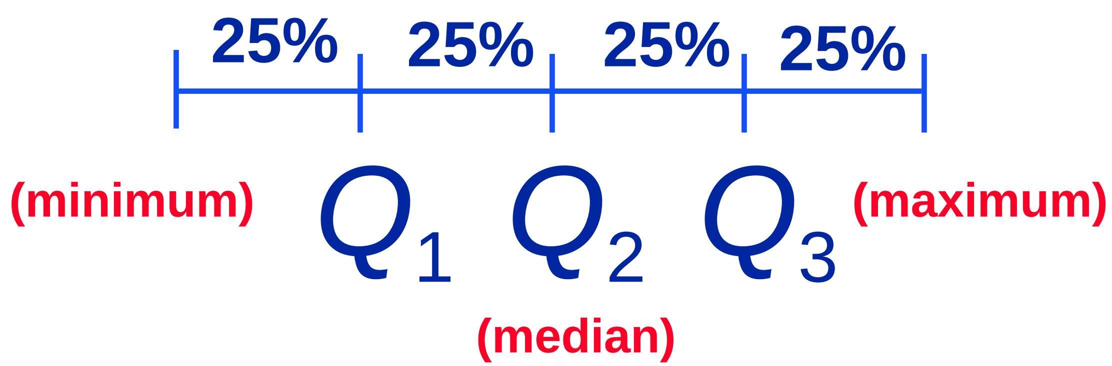
Definition

- Q_1 **(First Quartile)** **separates the bottom 25% of sorted values from the top 75%.**
- Q_2 **(Second Quartile)** **same as the median; separates the bottom 50% of sorted values from the top 50%.**
- Q_1 **(Third Quartile)** **separates the bottom 75% of sorted values from the top 25%.**

Quartiles

$$Q_1, \; Q_2, \; Q_3$$

divide ranked scores into four equal parts



Percentiles

Just as there are three quartiles separating data into four parts, there are 99 percentiles denoted P_1, P_2, . . . P_{99}, which partition the data into 100 groups.

Finding the Percentile of a Given Score

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Converting from the *k*th Percentile to the Corresponding Data Value

Notation

$$L = \frac{k}{100} \cdot n$$

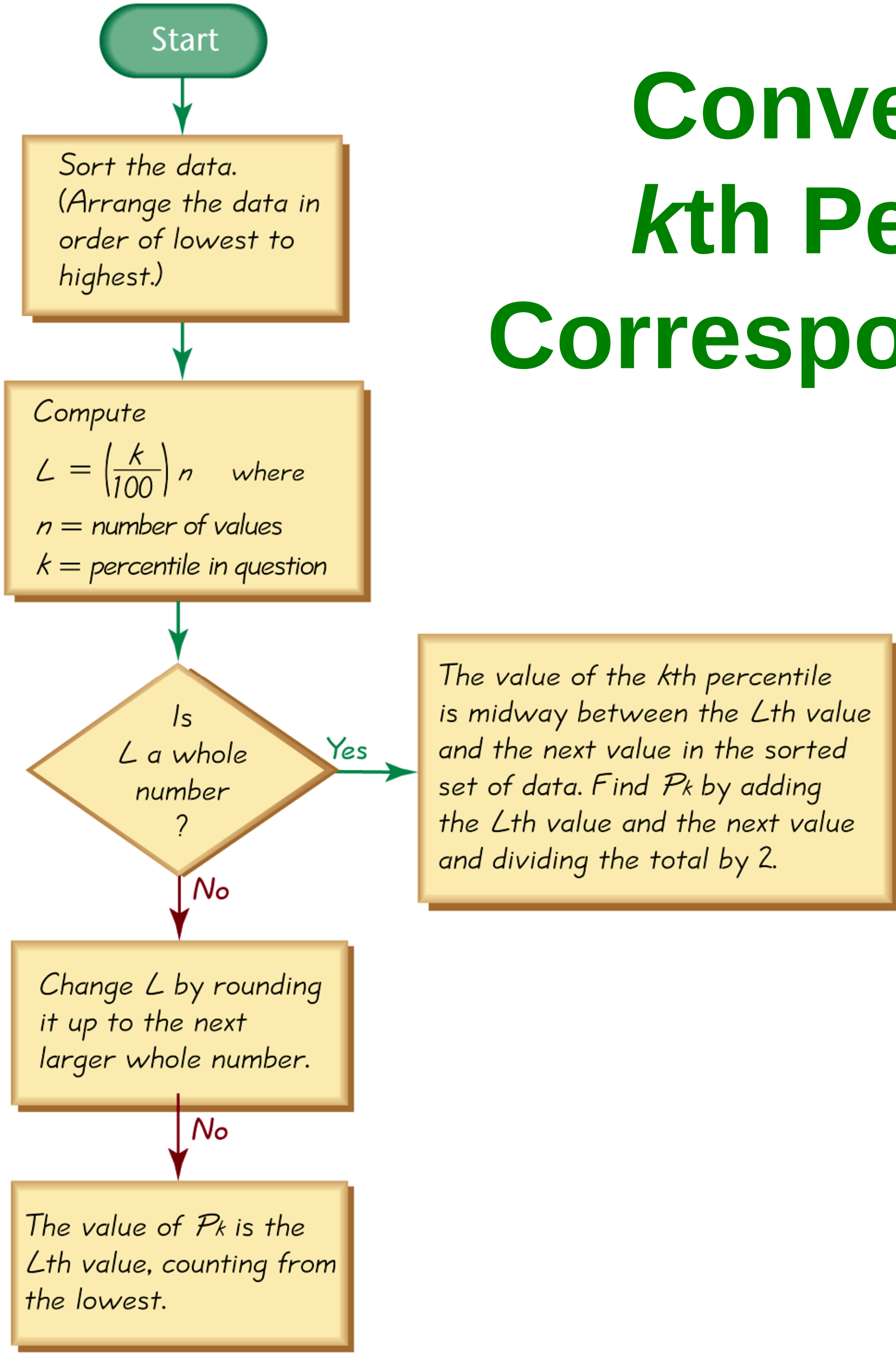
n total number of values in the data set

k percentile being used

L locator that gives the position of a value

P_k kth percentile

Converting from the *k*th Percentile to the Corresponding Data Value



Some Other Statistics

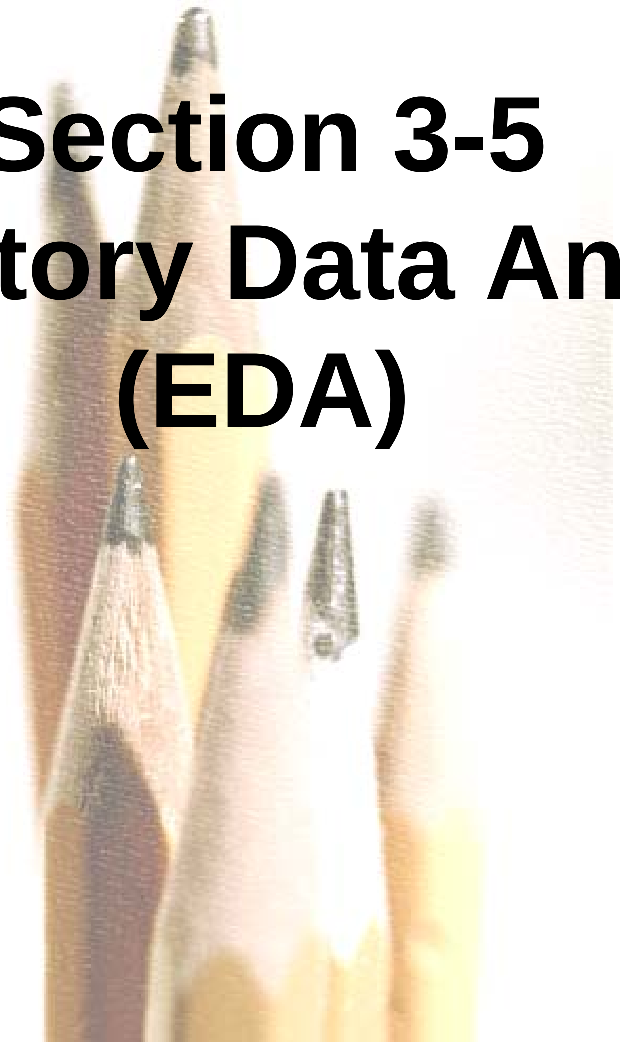
- **Interquartile Range (or IQR):** $Q_3 - Q_1$
- **Semi-interquartile Range:** $\dfrac{Q_3 - Q_1}{2}$
- **Midquartile:** $\dfrac{Q_3 + Q_1}{2}$
- **10 - 90 Percentile Range:** $P_{90} - P_{10}$

Recap

In this section we have discussed:

- ***z*** **Scores**
- ***z*** **Scores and unusual values**
- **Quartiles**
- **Percentiles**
- **Converting a percentile to corresponding data values**
- **Other statistics**

Section 3-5
Exploratory Data Analysis (EDA)



Created by Tom Wegleitner, Centreville, Virginia

Key Concept

This section discusses outliers, then introduces a new statistical graph called a boxplot, which is helpful for visualizing the distribution of data.

Definition

- **Exploratory Data Analysis (EDA)**

  the process of using statistical tools (such as graphs, measures of center, and measures of variation) to investigate data sets in order to understand their important characteristics

Definition

- An **outlier** is a value that is located very far away from almost all of the other values.

Important Principles

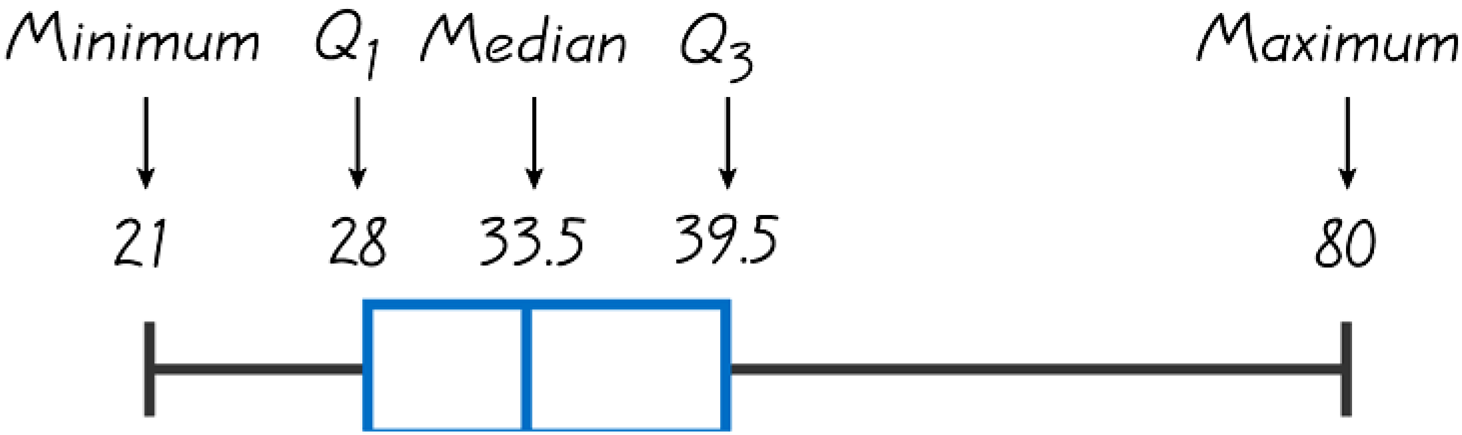
- An outlier can have a dramatic effect on the mean.
- An outlier can have a dramatic effect on the standard deviation.
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Definitions

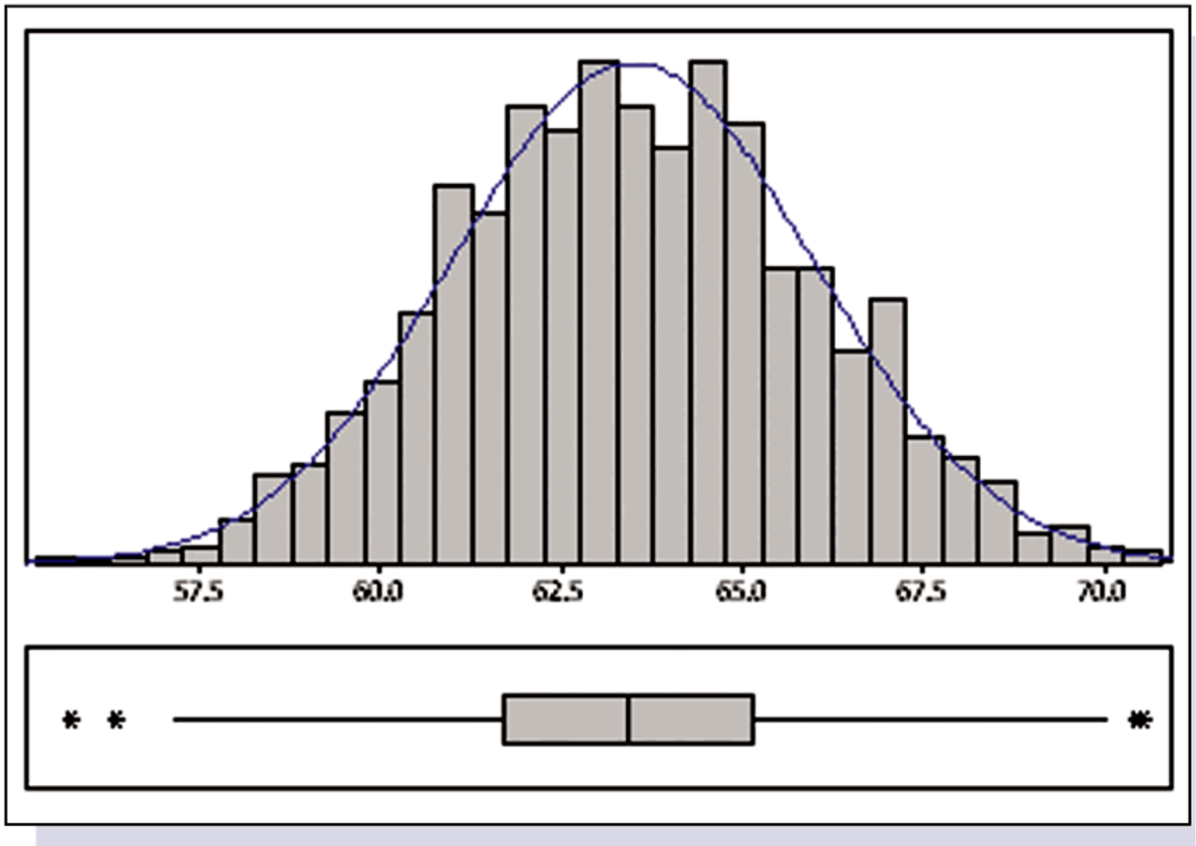
- For a set of data, the **5-number summary** consists of the minimum value; the first quartile Q_1; the median (or second quartile Q_2); the third quartile, Q_3; and the maximum value.

- A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

Boxplots

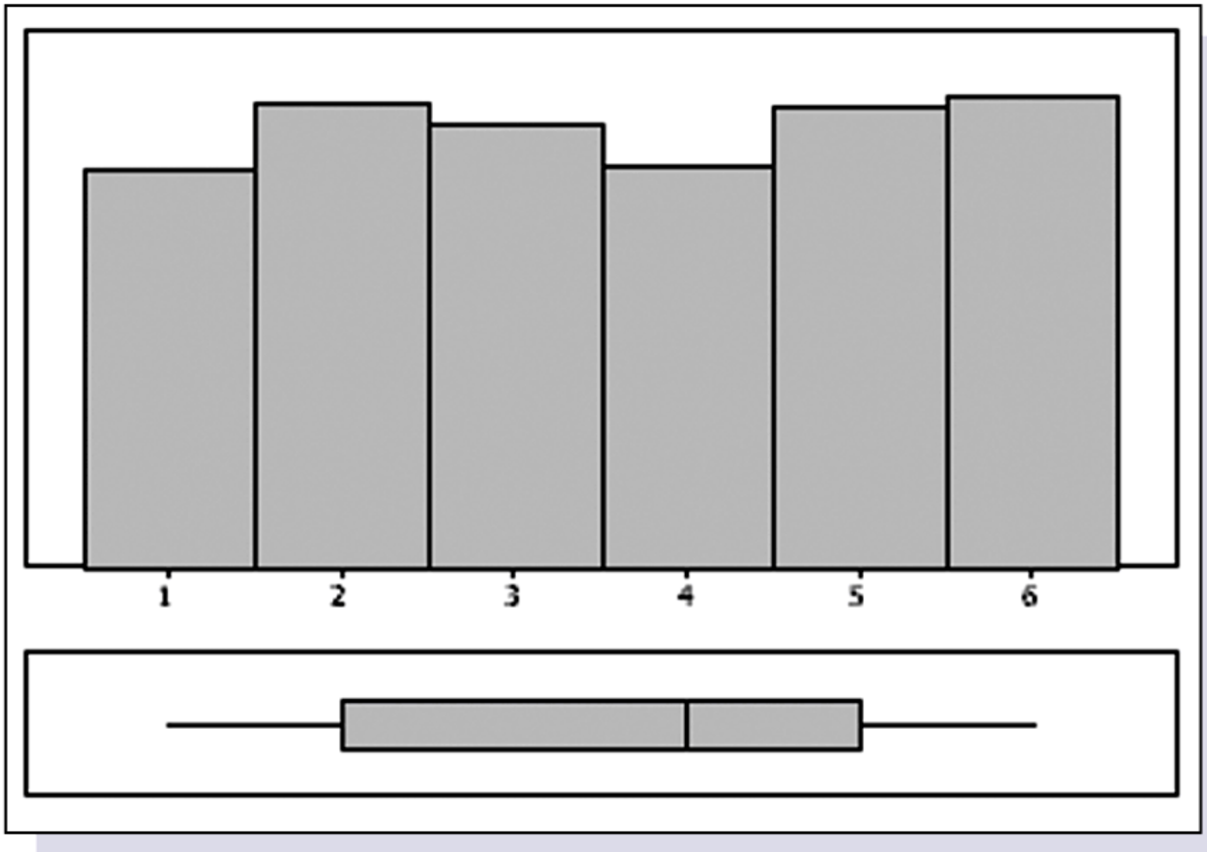


Boxplots - cont



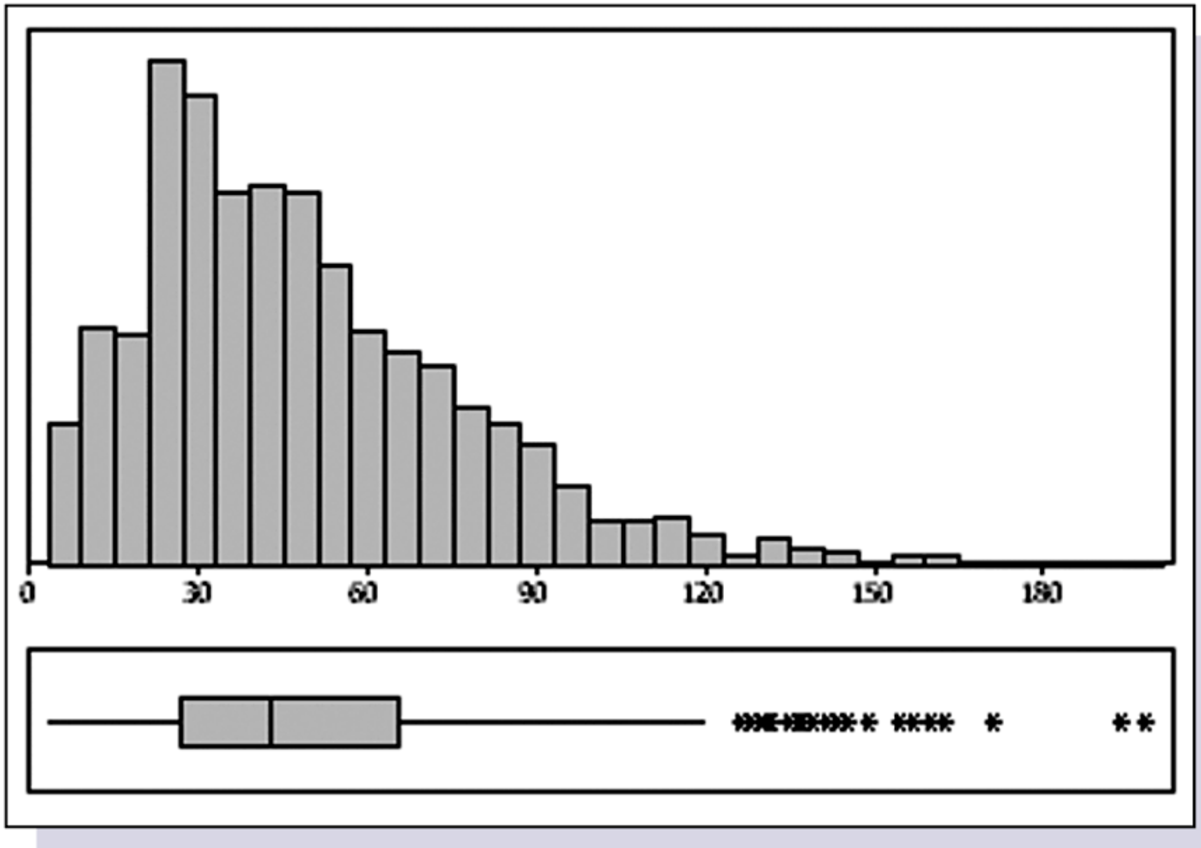
(a) Normal (bell-shaped) distribution
1000 heights (in.) of women

Boxplots - cont



(b) Uniform distribution
1000 rolls of a die

Boxplots - cont



(c) Skewed distribution

Incomes (thousands of dollars) of 1000 statistics professors

Modified Boxplots

Some statistical packages provide modified boxplots which represent outliers as special points.

A data value is an outlier if it is …

above Q_3 by an amount greater than 1.5 X IQR

or

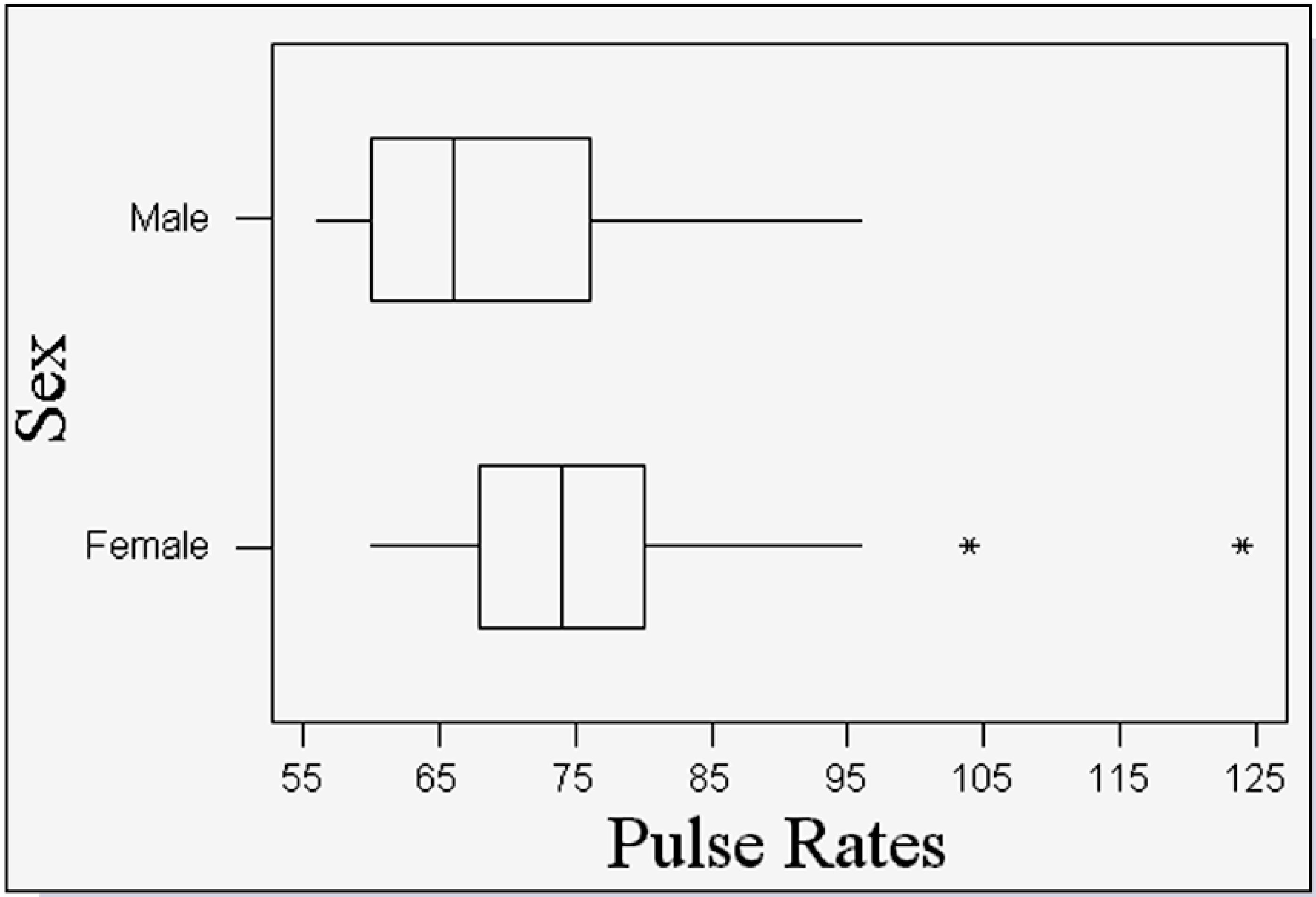
below Q_1 by an amount greater than 1.5 X IQR

Modified Boxplot Construction

A modified boxplot is constructed with these specifications:

- **A special symbol (such as an asterisk) is used to identify outliers.**
- **The solid horizontal line extends only as far as the minimum data value that is not an outlier and the maximum data value that is not an outlier.**

Modified Boxplots - Example



Recap

In this section we have looked at:

- **Exploratory Data Analysis**
- **Effects of outliers**
- **5-number summary**
- **Boxplots and modified boxplots**