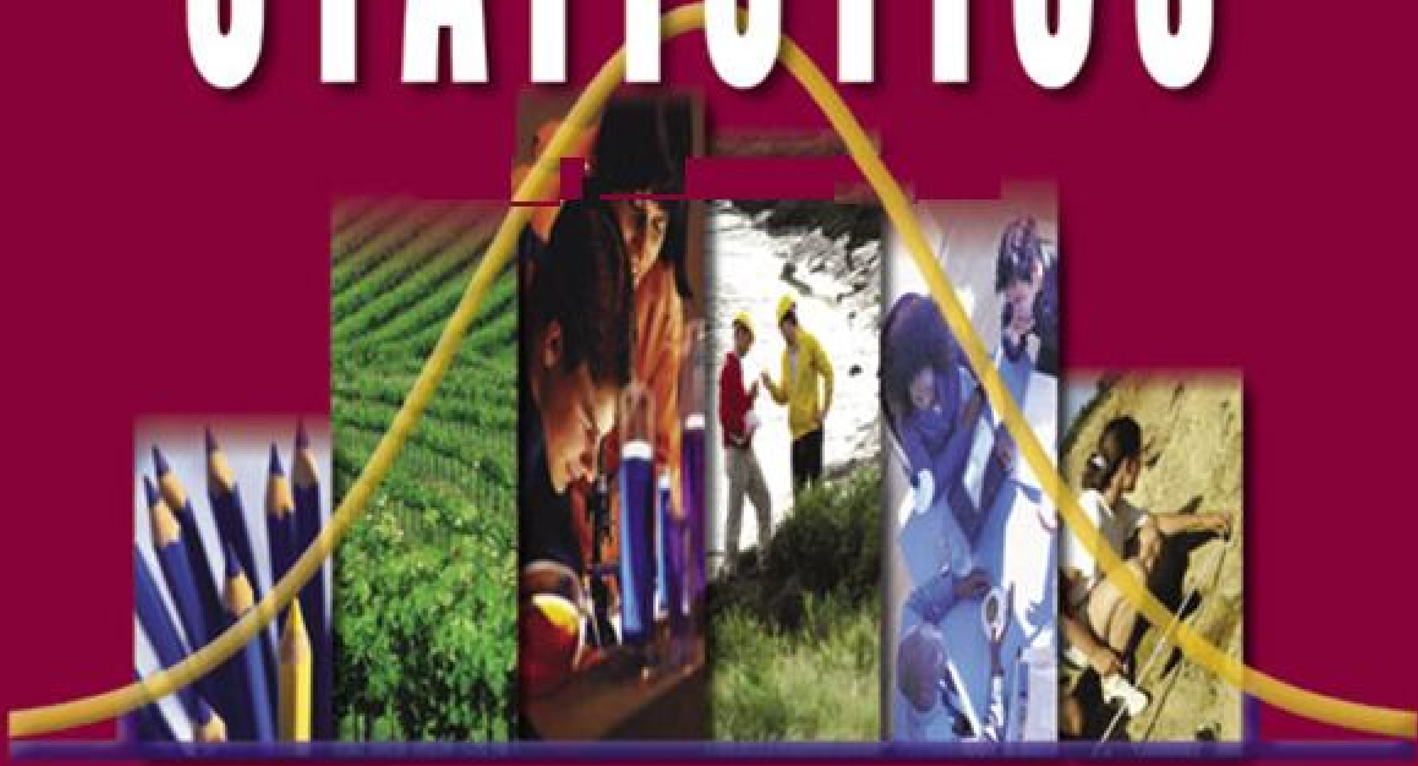
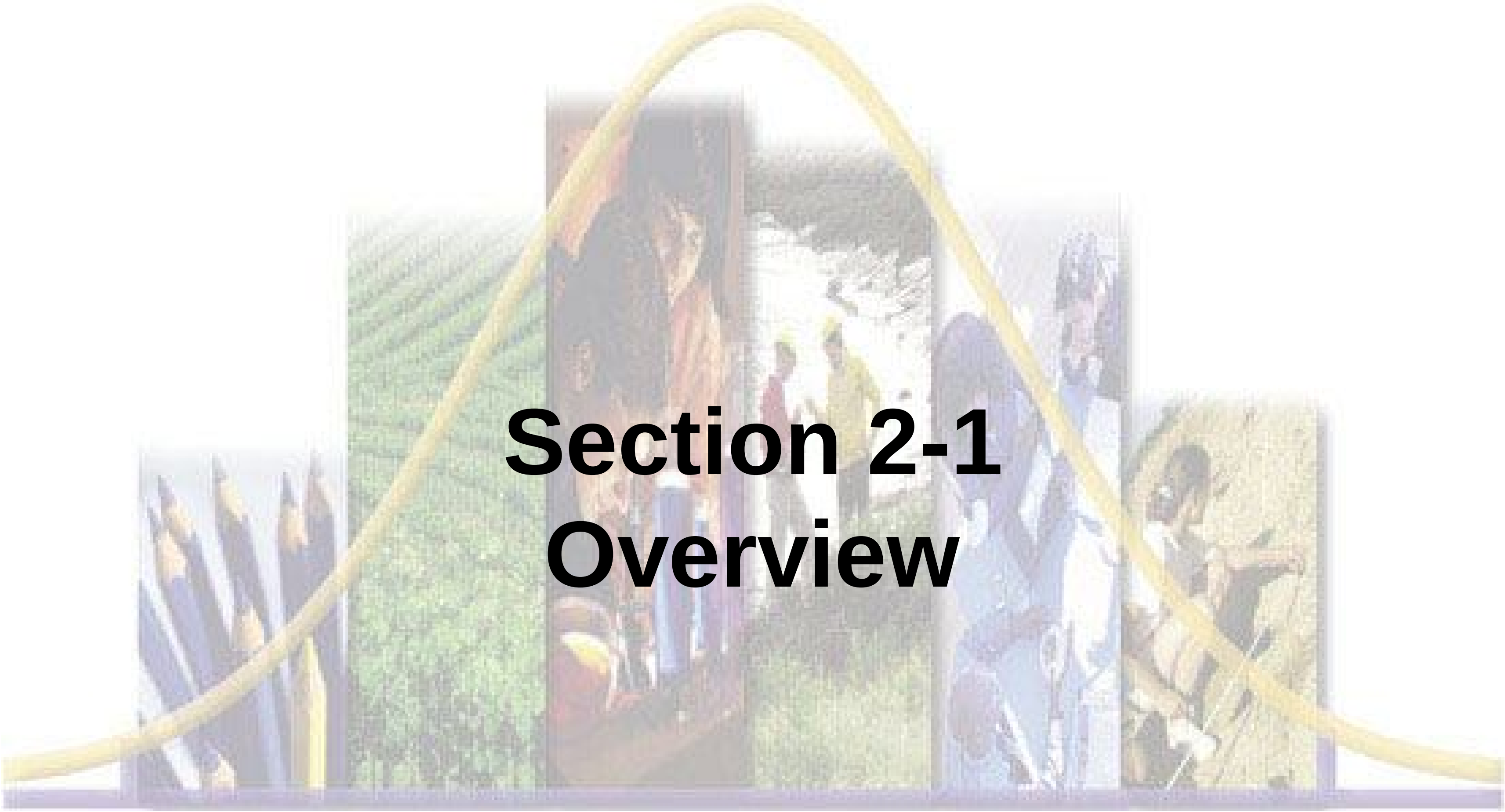
ELEMENTARY

STATISTICS



Dadan Dasari

Chapter 2
Describing, Exploring, and Comparing Data



Section 2-1
Overview

Overview

- **Descriptive Statistics**

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

- **Inferential Statistics**

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

Important Characteristics of Data

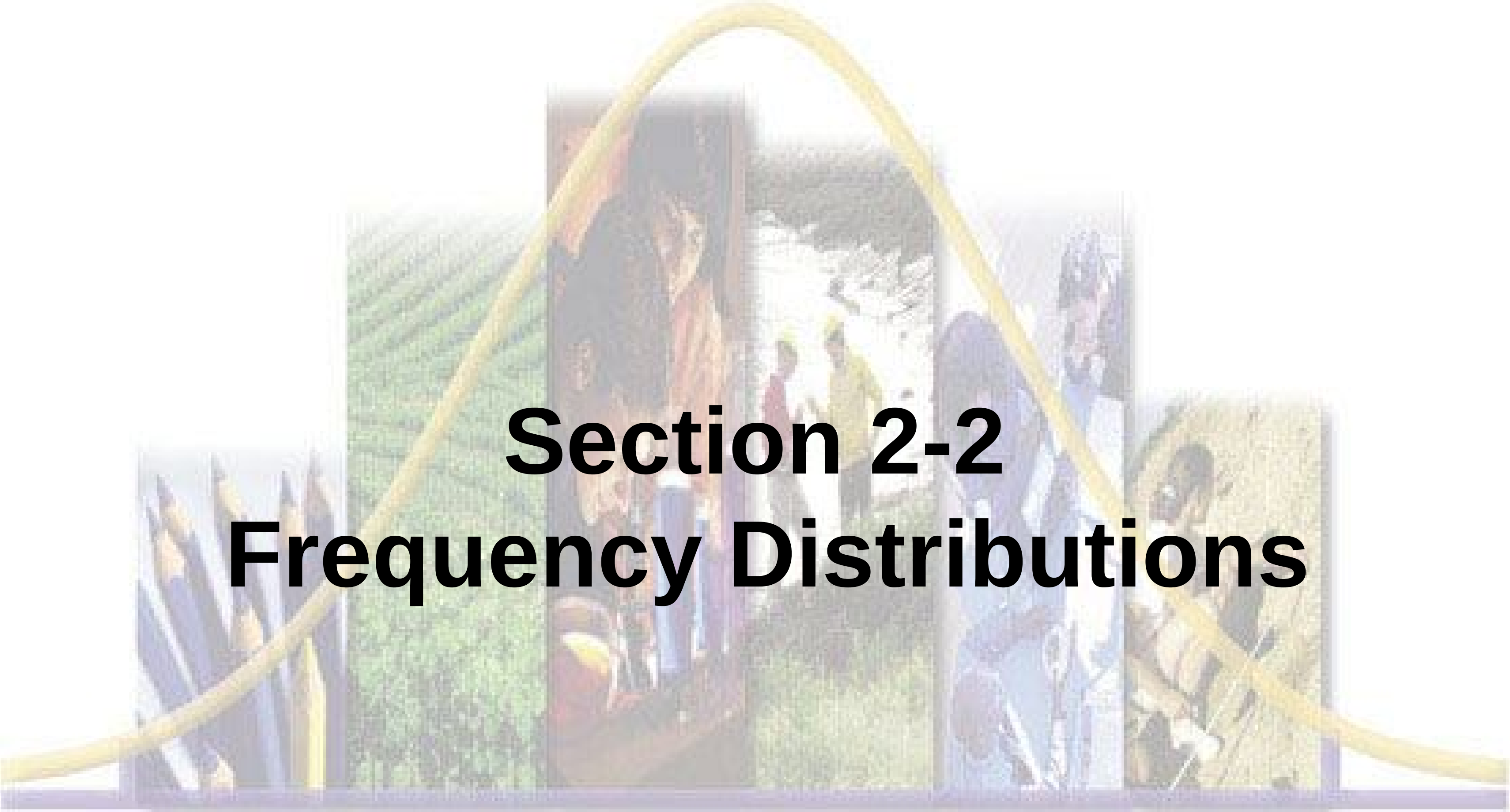
1. **Center**: A representative or average value that indicates where the middle of the data set is located

2. **Variation**: A measure of the amount that the values vary among themselves

3. **Distribution**: The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)

4. **Outliers**: Sample values that lie very far away from the vast majority of other sample values

5. **Time**: Changing characteristics of the data over time



Section 2-2
Frequency Distributions

Frequency Distributions

- **Frequency Distribution**

  lists data values (either individually or by groups of intervals), along with their corresponding frequencies or counts

Table 2-1 Measured Cotinine Levels in Three Groups

Smoker: The subjects report tobacco use.

ETS: (Environmental Tobacco Smoke) Subjects are nonsmokers who are exposed to environmental tobacco smoke ("secondhand smoke") at home or work.

NOETS: (No Environmental Tobacco Smoke) Subjects are nonsmokers who are not exposed to environmental tobacco smoke at home or work. That is, the subjects do not smoke and are not exposed to secondhand smoke.

Smoker:	1	0	131	173	265	210	44	277	32	3
	35	112	477	289	227	103	222	149	313	491
	130	234	164	198	17	253	87	121	266	290
	123	167	250	245	48	86	284	1	208	173
ETS:	384	0	69	19	1	0	178	2	13	1
	4	0	543	17	1	0	51	0	197	3
	0	3	1	45	13	3	1	1	1	0
	0	551	2	1	1	1	0	74	1	241
NOETS:	0	0	0	0	0	0	0	0	0	0
	0	9	0	0	0	0	0	0	244	0
	1	0	0	0	90	1	0	309	0	0
	0	0	0	0	0	0	0	0	0	0

Table 2-2

Frequency Distribution of Cotinine Levels of Smokers

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Frequency Distributions

Definitions

Lower Class Limits

are the smallest numbers that can actually belong to different classes

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Lower Class Limits

are the smallest numbers that can actually belong to different classes

Lower Class Limits

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Upper Class Limits

are the largest numbers that can actually belong to different classes

Upper Class Limits

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Class Boundaries

are the numbers used to separate classes, but without the gaps created by class limits

Class Boundaries

number separating classes

	Cotinine	Frequency
- 0.5		
	0–99	11
99.5		
	100–199	12
199.5		
	200–299	14
299.5		
	300–399	1
399.5		
	400–499	2
499.5		

Class Boundaries

number separating classes

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Class Boundaries

- 0.5
99.5
199.
5
299.
5
399.
5
499.
5

Class Midpoints

midpoints of the classes

Class midpoints can be found by adding the lower class limit to the upper class limit and dividing the sum by two.

Class Midpoints

midpoints of the classes

Class Midpoints

Cotinine			Frequency
0	49.5	99	11
100	149.5	199	12
200	249.5	299	14
300	349.5	399	1
400	449.5	499	2

Class Width

is the difference between two consecutive lower class limits or two consecutive lower class boundaries

Class Width

	Cotinine	Frequency
100	0–99	11
100	100–199	12
100	200–299	14
100	300–399	1
100	400–499	2

Reasons for Constructing Frequency Distributions

1. Large data sets can be summarized.

2. Can gain some insight into the nature of data.

3. Have a basis for constructing graphs.

Constructing A Frequency Table

1. Decide on the number of classes (should be between 5 and 20) .

2. Calculate (round up).

$$\text{class width} \approx \frac{(\text{highest value}) - (\text{lowest value})}{\text{number of classes}}$$

3. Starting point: Begin by choosing a lower limit of the first class.

4. Using the lower limit of the first class and class width, proceed to list the lower class limits.

5. List the lower class limits in a vertical column and proceed to enter the upper class limits.

6. Go through the data set putting a tally in the appropriate class for each data value.

Relative Frequency Distribution

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

Relative Frequency Distribution

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Total Frequency = 40

Table 2-3
Relative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Relative Frequency
0–99	28%
100–199	30%
200–299	35%
300–399	3%
400–499	5%

11/40 = 28%

12/40 = 40%

etc.

Cumulative Frequency Distribution

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Table 2-4
Cumulative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Cumulative Frequency
Less than 100	11
Less than 200	23
Less than 300	37
Less than 400	38
Less than 500	40

Cumulative Frequencies

Frequency Tables

Table 2-2
Frequency Distribution of Cotinine Levels of Smokers

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Table 2-3
Relative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Relative Frequency
0–99	28%
100–199	30%
200–299	35%
300–399	3%
400–499	5%

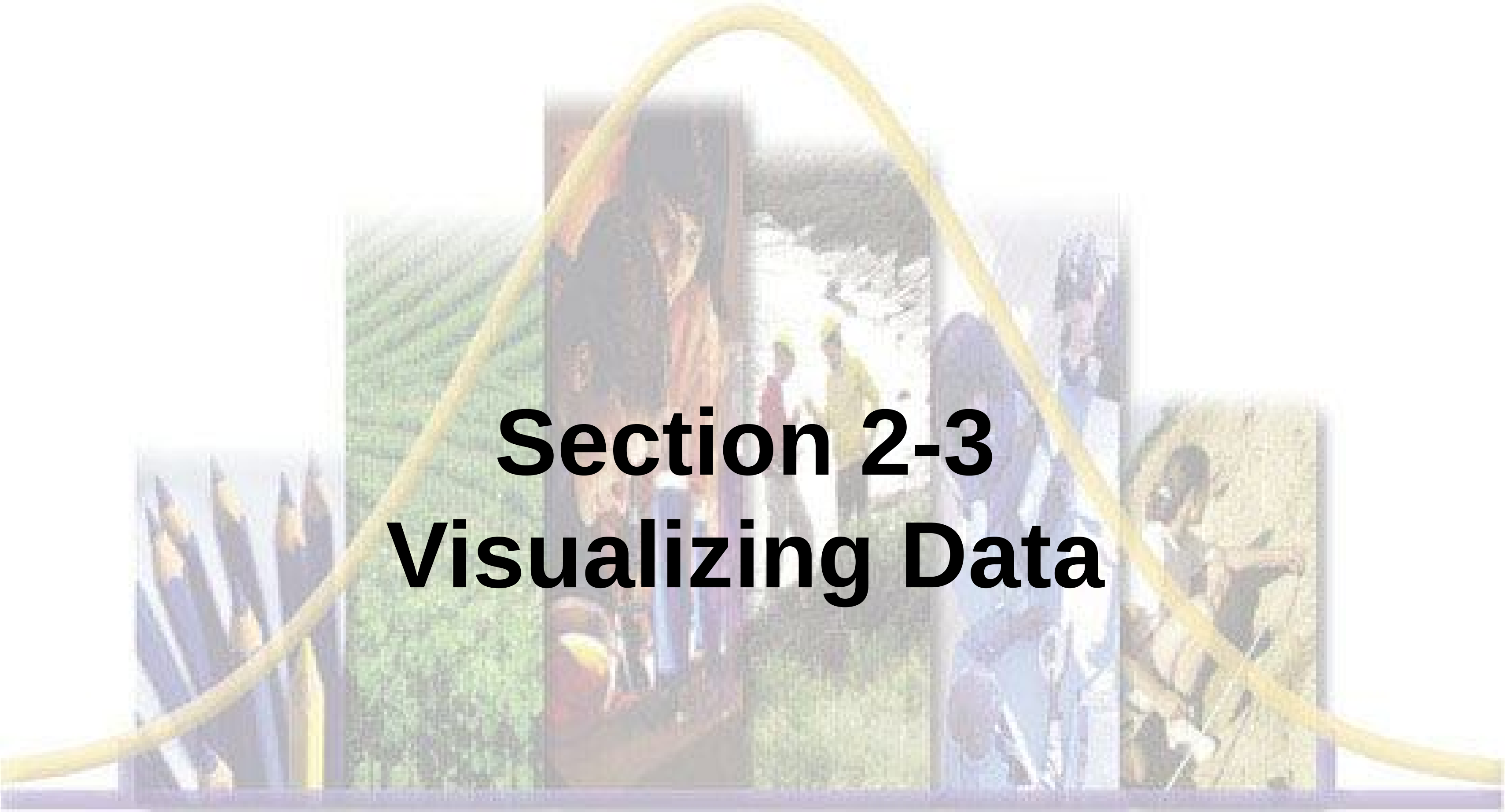
Table 2-4
Cumulative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Cumulative Frequency
Less than 100	11
Less than 200	23
Less than 300	37
Less than 400	38
Less than 500	40

Recap

In this Section we have discussed

- **Important characteristics of data**
- **Frequency distributions**
- **Procedures for constructing frequency distributions**
- **Relative frequency distributions**
- **Cumulative frequency distributions**



Section 2-3
Visualizing Data

Visualizing Data

Depict the nature of shape or shape of the data distribution

Histogram

A bar graph in which the horizontal scale represents the classes of data values and the vertical scale represents the frequencies.

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

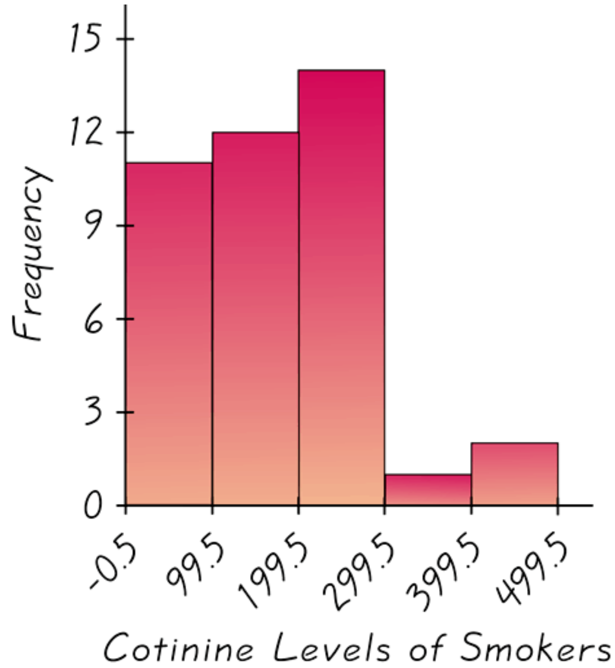


Figure 2-1

Relative Frequency Histogram

Has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies.

Cotinine	Relative Frequency
[illegible]	[illegible]
100–199	30%
200–299	35%
300–399	3%
400–499	5%

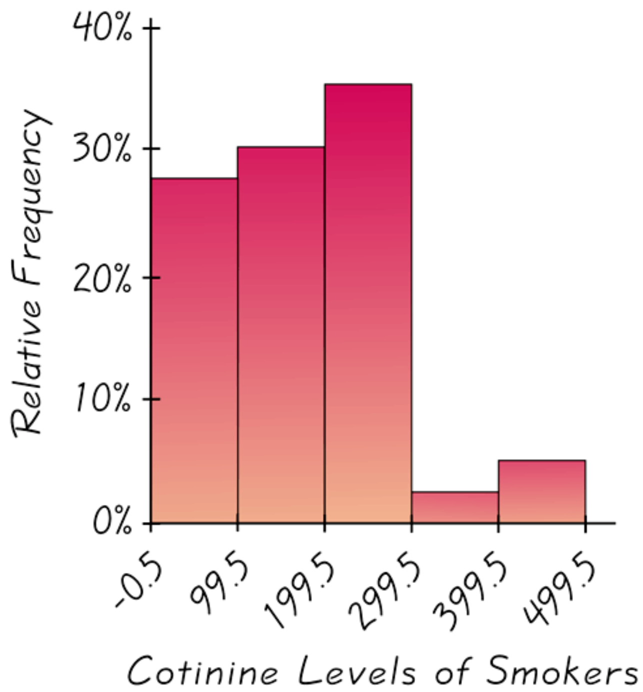


Figure 2-2

Histogram and Relative Frequency Histogram

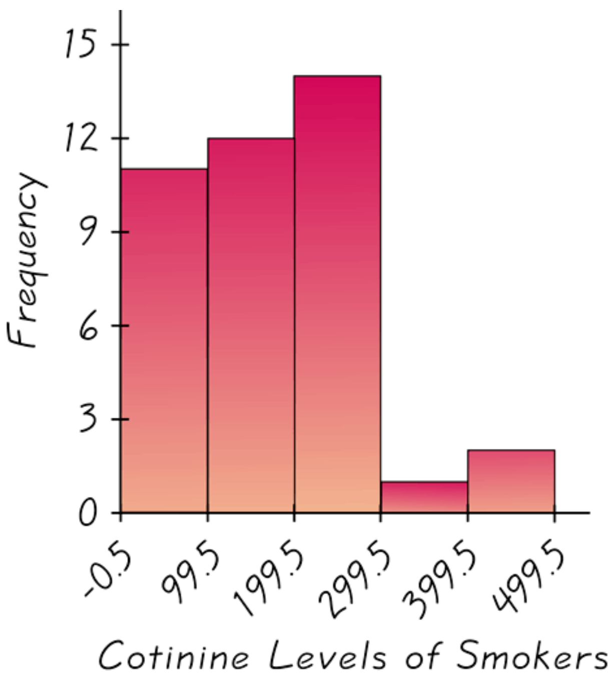


Figure 2-1

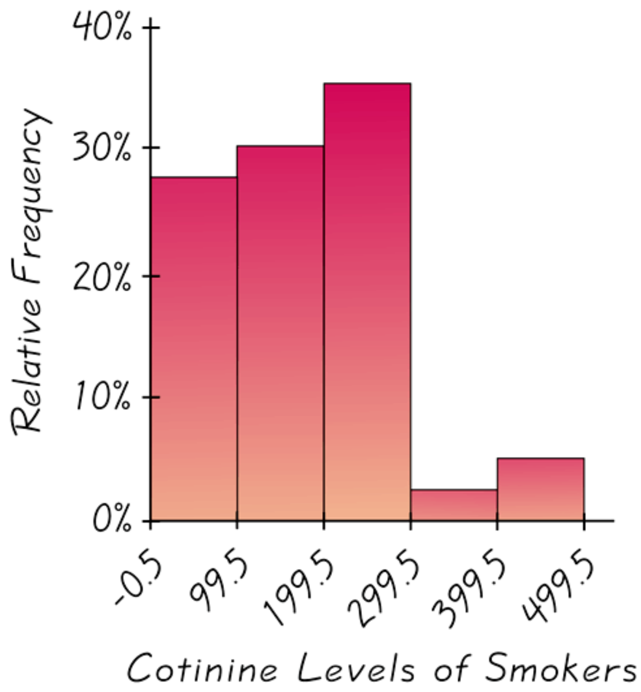


Figure 2-2

Frequency Polygon

Uses line segments connected to points directly above class midpoint values

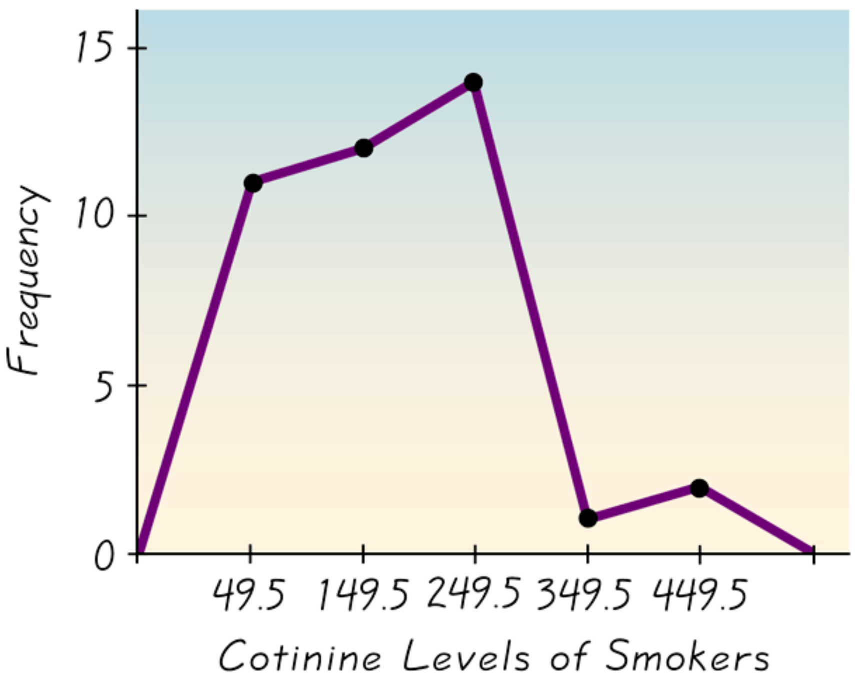


Figure 2-3

Ogive

A line graph that depicts cumulative frequencies

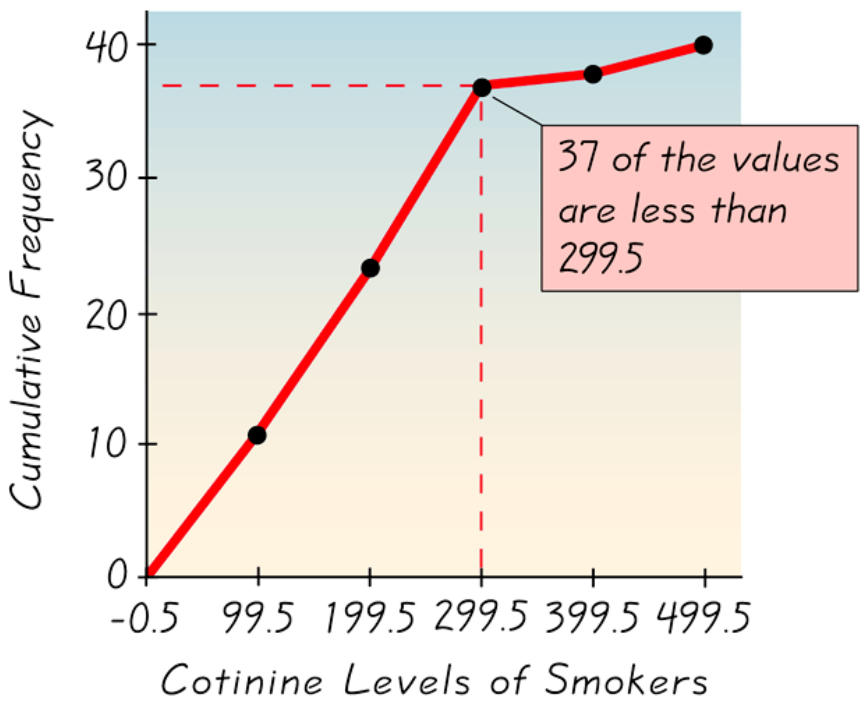


Figure 2-4

Dot Plot

Consists of a graph in which each data value is plotted as a point along a scale of values

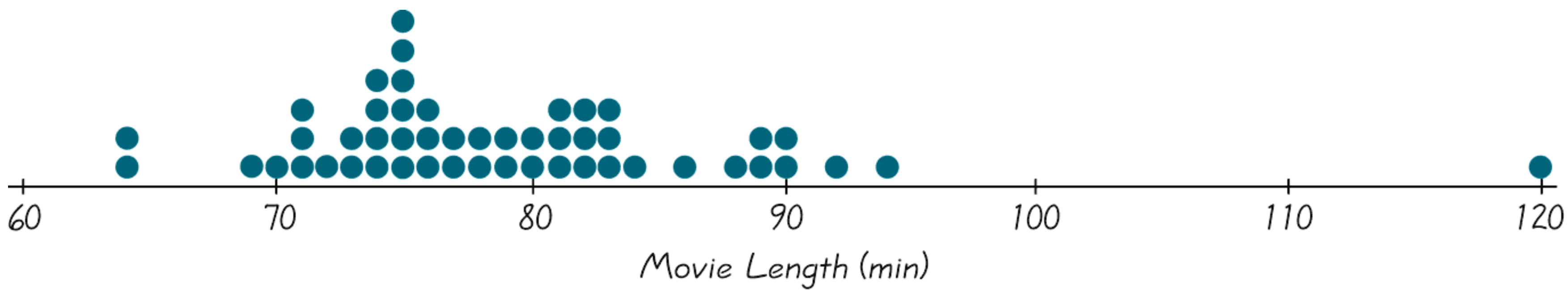


Figure 2-5

Stem-and Leaf Plot

Represents data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

Stem-and-Leaf Plot

Stem (tens)	Leaves (units)	
6	449	← Values are 64, 64, 69.
7	0111233444455555566677889 9	
8	00111222333346899	
9	0024	
10		
11		
12	0	← Value is 120.

Pareto Chart

A bar graph for qualitative data, with the bars arranged in order according to frequencies

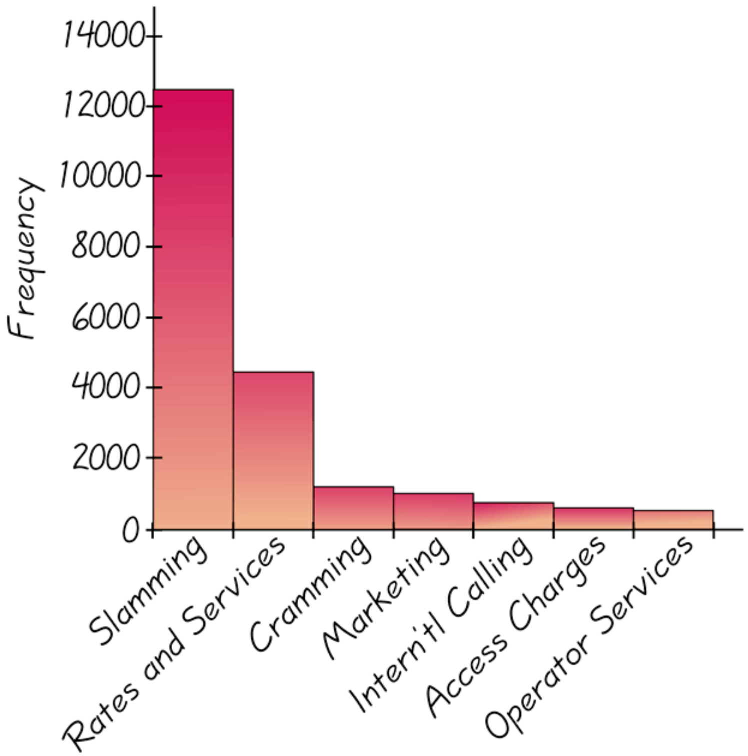


Figure 2-6

Pie Chart

A graph depicting qualitative data as slices pf a pie

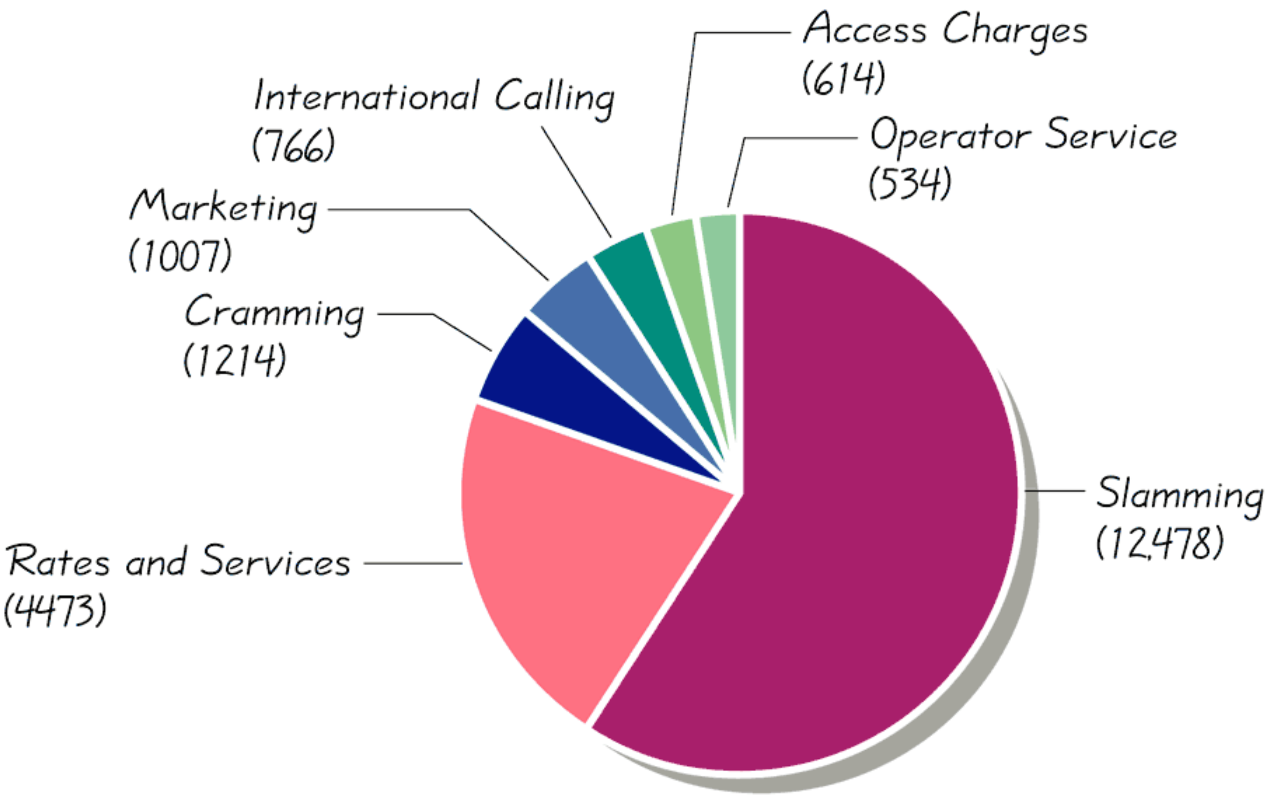
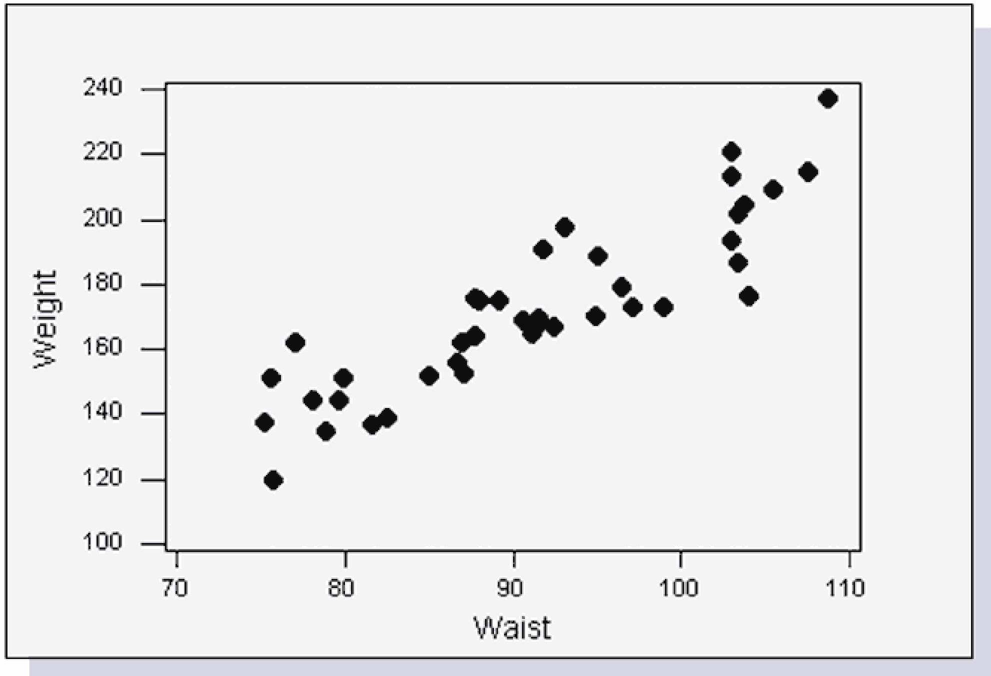


Figure 2-7

Scatter Diagram

A plot of paired (x,y) data with a horizontal x-axis and a vertical y-axis



Time-Series Graph

Data that have been collected at different points in time

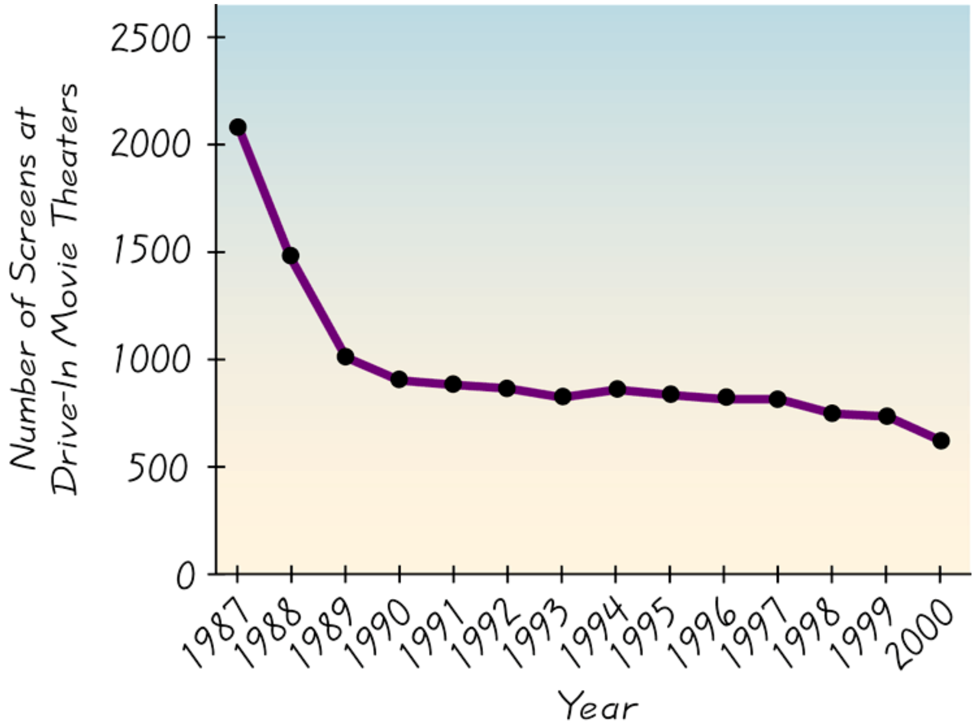
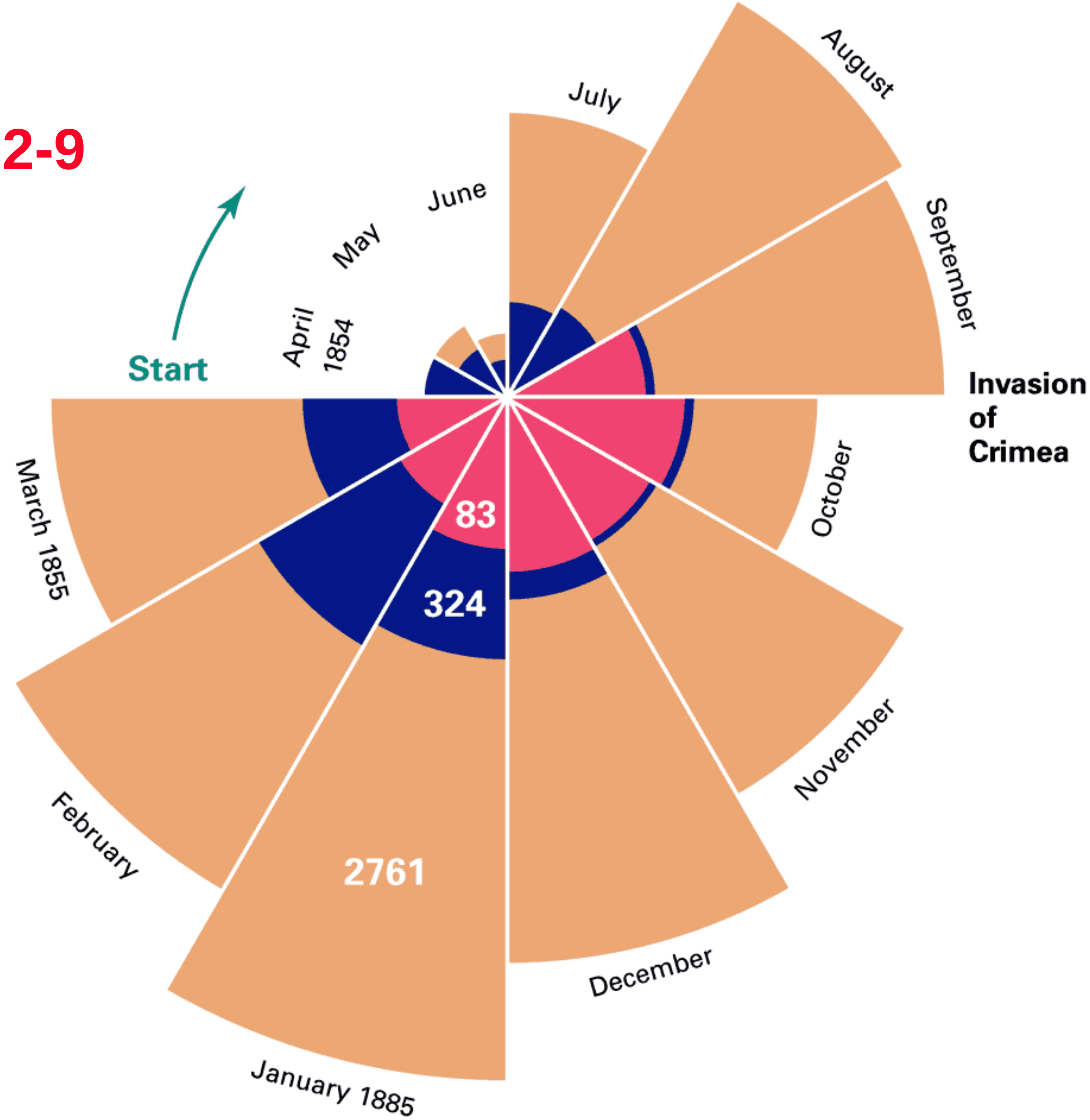


Figure 2-8

Other Graphs

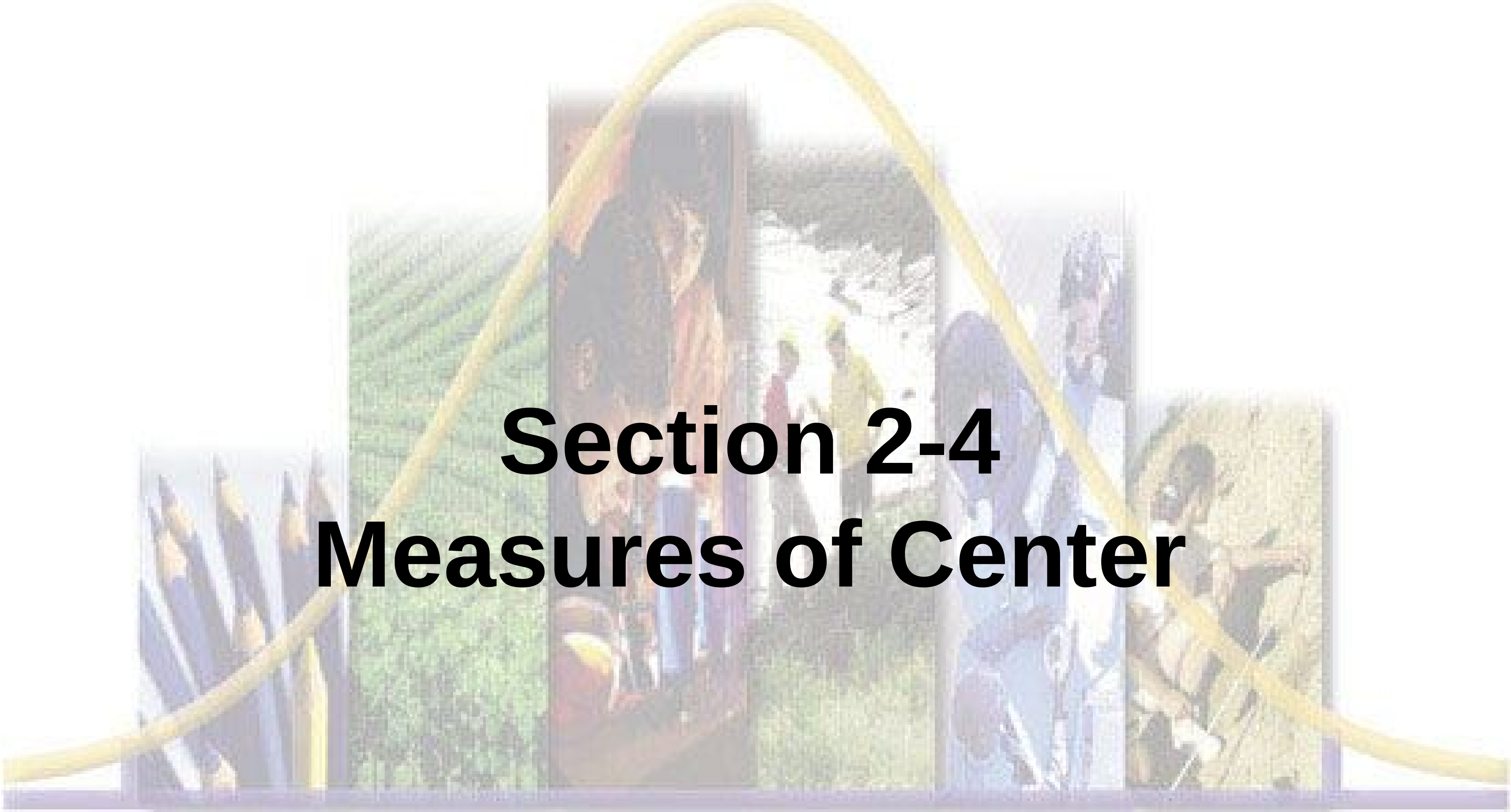
Figure 2-9



Recap

In this Section we have discussed graphs that are pictures of distributions.

Keep in mind that the object of this section is not just to construct graphs, but to learn something about the data sets – that is, to understand the nature of their distributions.



Section 2-4
Measures of Center

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 **addition** 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**

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

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

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

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

(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

The mode is not always unique. A data set may be:

Bimodal

Multimodal

No Mode

- **denoted by M**

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

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

Definitions

- **Midrange**

  the value midway between the highest and lowest values in the original data set

$$\text{Midrange} = \frac{\text{highest score} + \text{lowest score}}{2}$$

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*

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

Formula 2-2

x = class midpoint

f = frequency

$\Sigma f = n$

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

Table 2-10 Comparison of Mean, Median, Mode, and Midrange

Measure of Center	Definition	How Common?	Existence	Takes Every Value into Account?	Affected by Extreme Values?	Advantages and Disadvantages
Mean	$\bar{x} = \frac{\Sigma x}{n}$	most familiar "average"	always exists	yes	yes	used throughout this book; works well with many statistical methods
Median	middle value	commonly used	always exists	no	no	often a good choice if there are some extreme values
Mode	most frequent data value	sometimes used	might not exist; may be more than one mode	no	no	appropriate for data at the nominal level
Midrange	$\frac{\text{high} + \text{low}}{2}$	rarely used	always exists	no	yes	very sensitive to extreme values

General comments:

- For a data collection that is approximately symmetric with one mode, the mean, median, mode, and midrange tend to be about the same.
- For a data collection that is obviously asymmetric, it would be good to report both the mean and median.
- The mean is relatively *reliable.* That is, when samples are drawn from the same population, the sample means tend to be more consistent than the other measures of center (consistent in the sense that the means of samples drawn from the same population don't vary as much as the other measures of center).

Definitions

- **Symmetric**

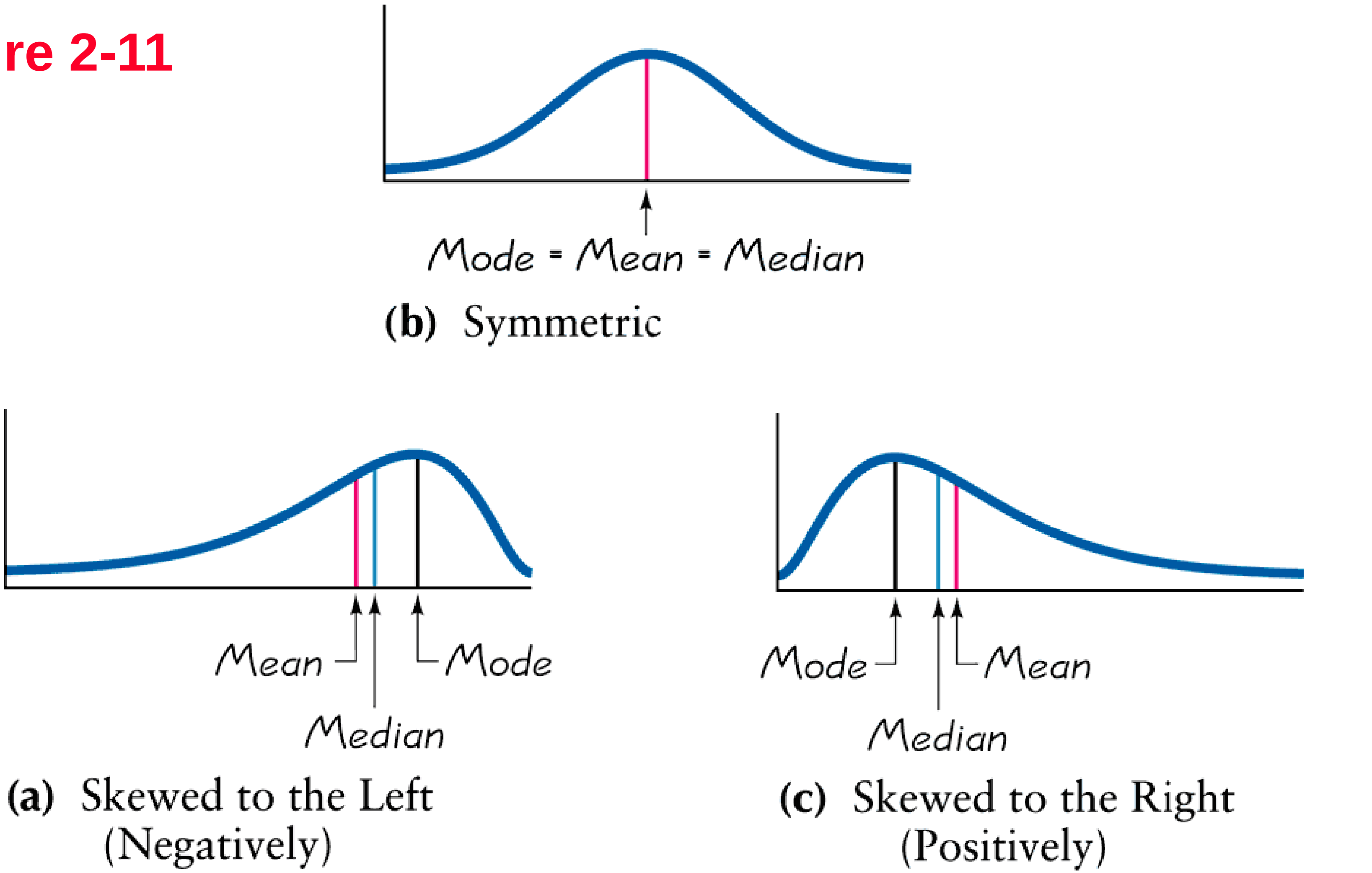
  Data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

- **Skewed**

  Data is skewed if it is not symmetric and if it extends more to one side than the other.

Skewness

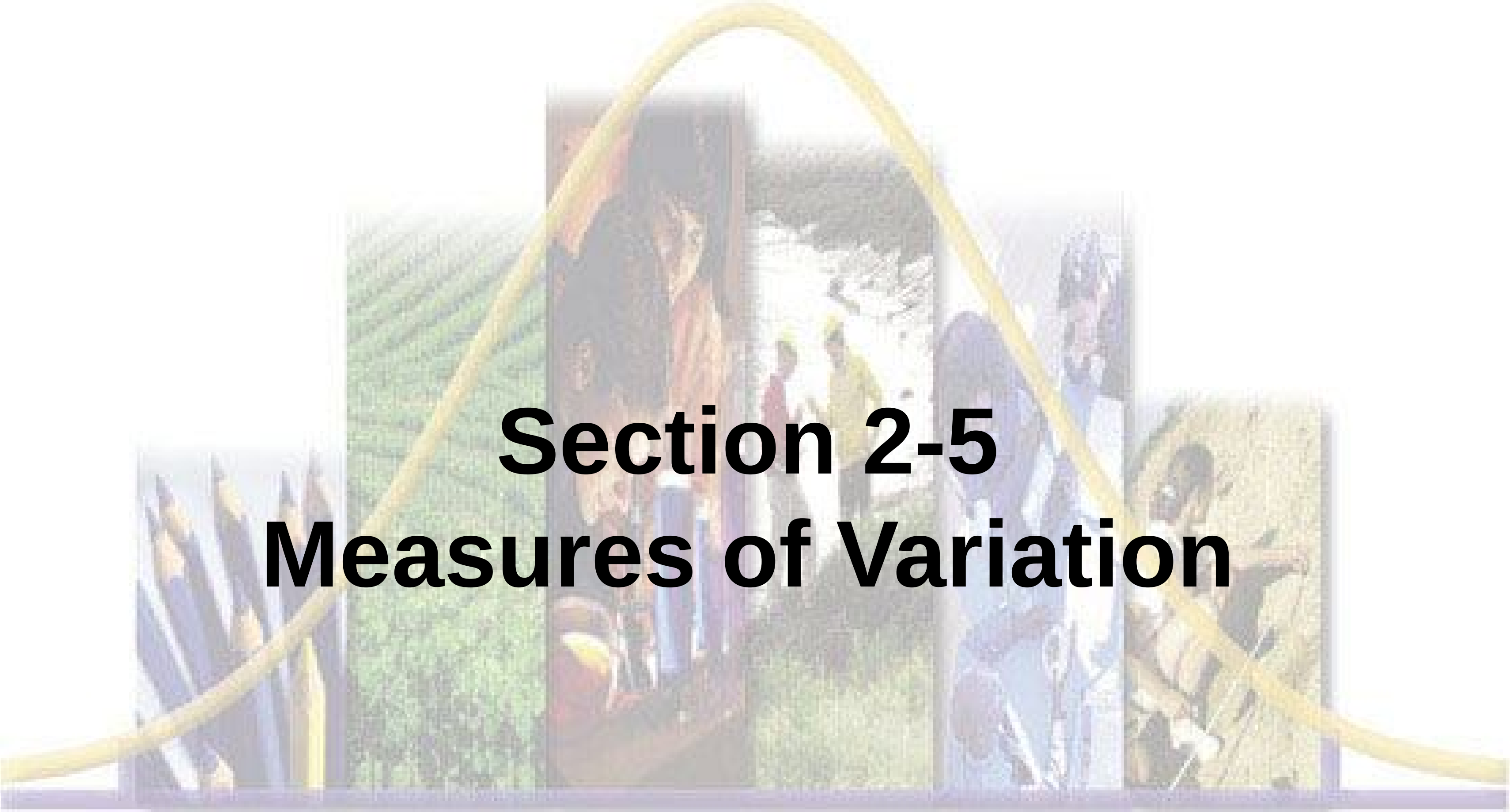
Figure 2-11



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 2-5
Measures of Variation

Measures of Variation

Because this section introduces the concept of variation, this is one of the most important sections in the entire book

Definition

The range of a set of data is the difference between the highest value and the lowest value

highest value – lowest 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

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

Formula 2-4

Sample Standard Deviation (Shortcut Formula)

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

Formula 2-5

Standard Deviation - Key Points

- 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

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

This formula is similar to Formula 2-4, 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

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

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.

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

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

Population

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

Standard Deviation from a Frequency Distribution

Formula 2-6

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

Use the class midpoints as the x values

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 = (highest value) – (lowest 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" values by using:

Minimum "usual" value $\approx$ (mean) – 2 X (standard deviation)

Maximum "usual" value $\approx$ (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

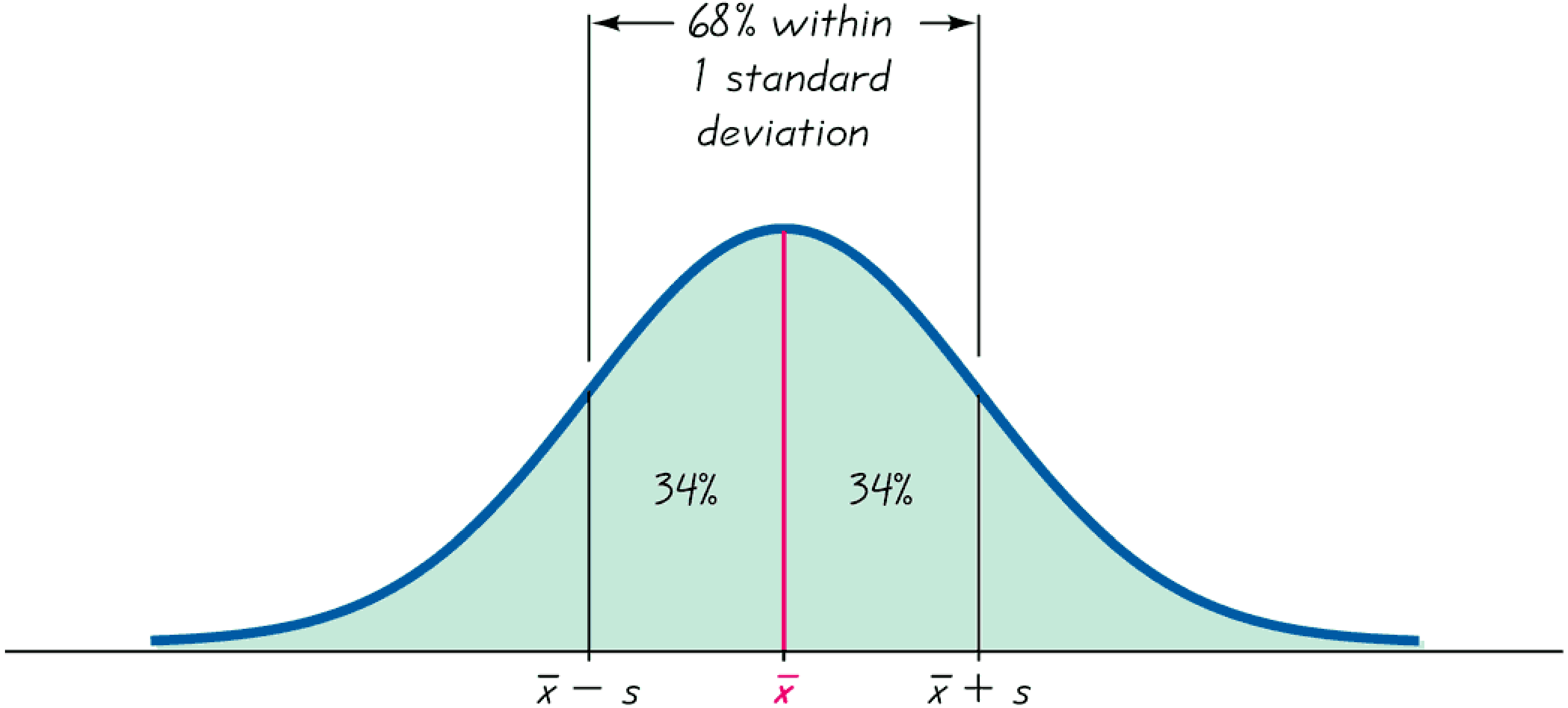


FIGURE 2-13

The Empirical Rule

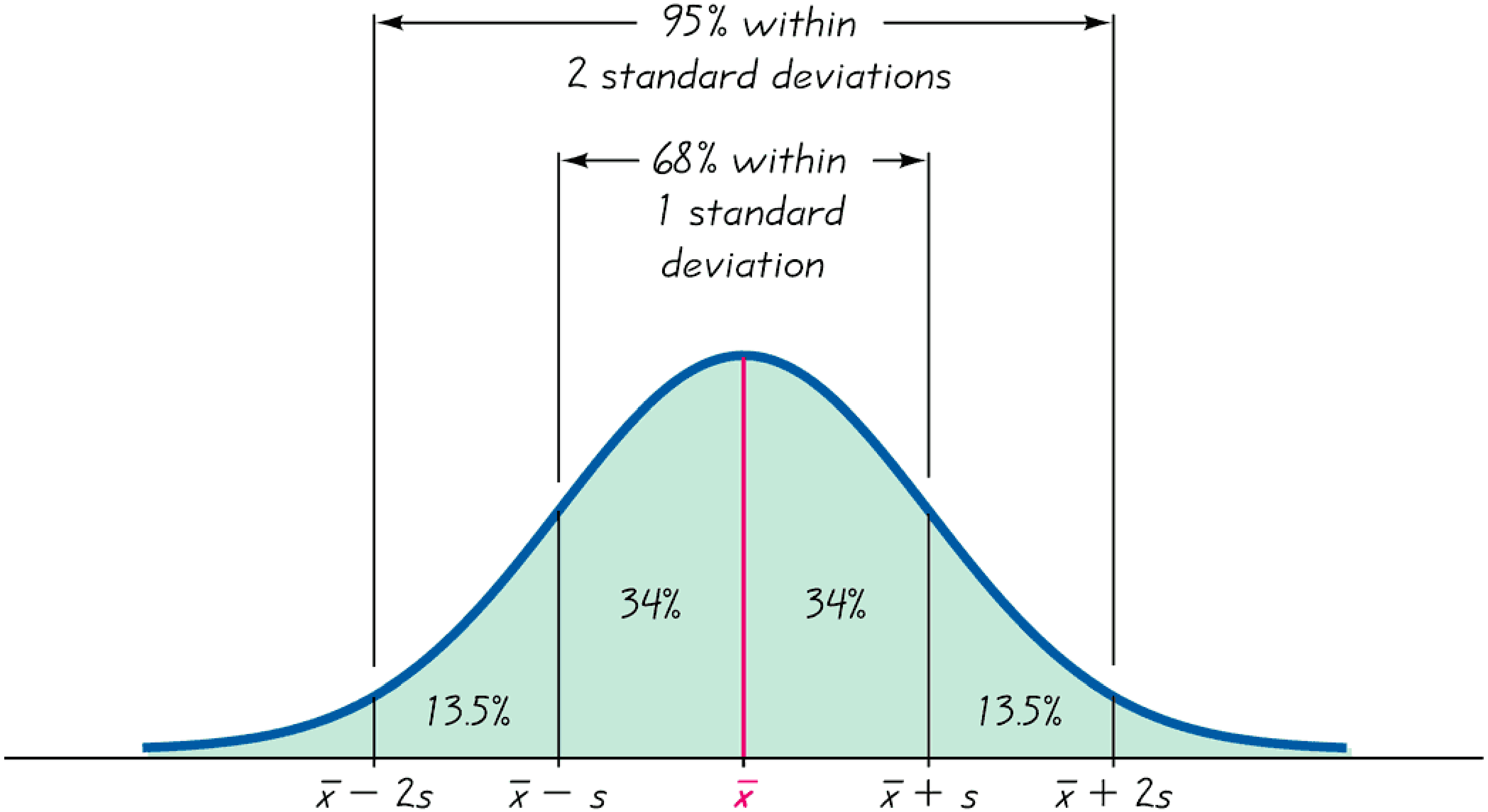


FIGURE 2-13

The Empirical Rule

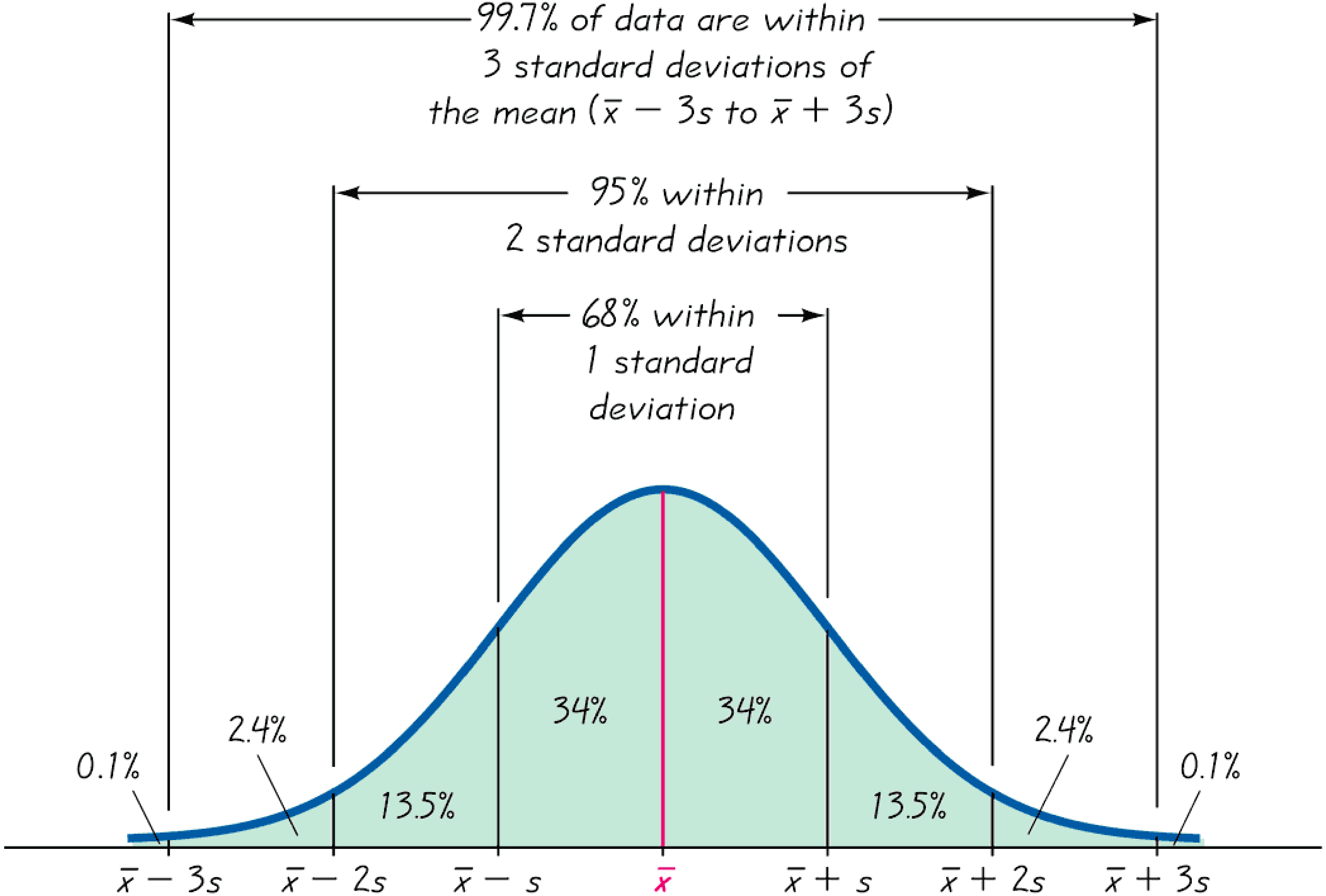


FIGURE 2-13

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**

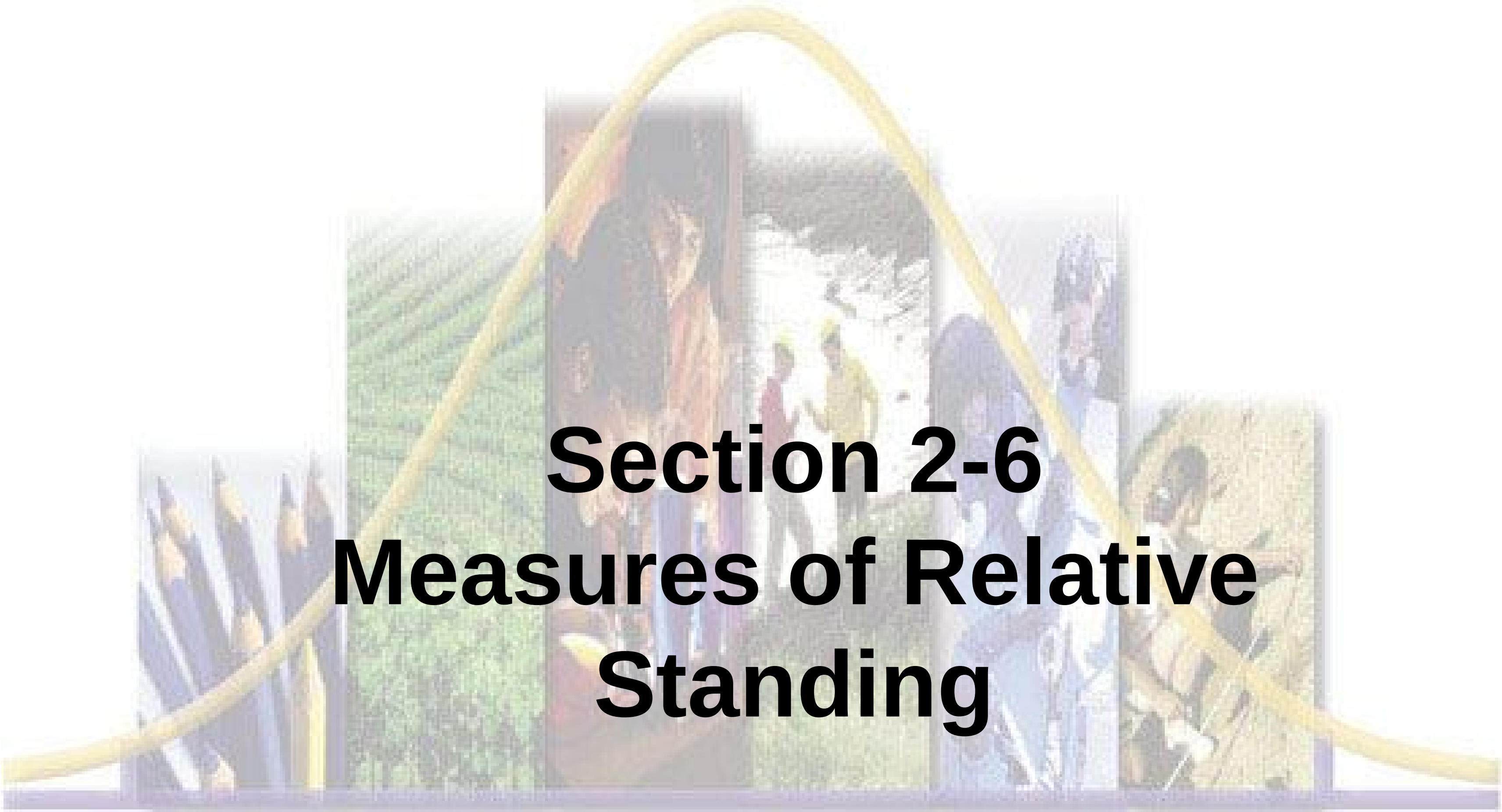
Rationale for Formula 2-4

The end of Section 2- 5 has a detailed explanation of why Formula 2- 4 is employed instead of other possibilities and, specifically, why n – 1 rather than n is used. The student should study it carefully

Recap

In this section we have looked at:

- **Range**
- **Standard deviation of a sample and population**
- **Variance of a sample and population**
- **Coefficient of Variation (CV)**
- **Standard deviation using a frequency distribution**
- **Range Rule of Thumb**
- **Empirical Distribution**
- **Chebyshev's Theorem**



Section 2-6
Measures of Relative Standing

Definition

- ***z*** **Score** **(or standard score)**

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

Measures of Position
z score

Sample

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

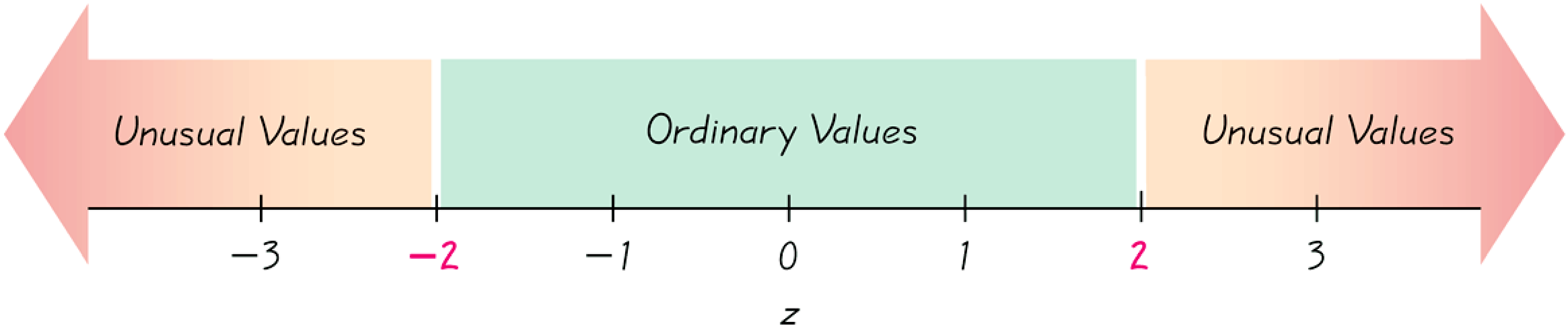
Population

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

Round to 2 decimal places

Interpreting Z Scores

FIGURE 2-14



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

Ordinary values: z **score between –2 and 2 sd**

Unusual Values: z **score < -2 or** z **score > 2 sd**

Definition

- Q_1 **(First Quartile)** **separates the bottom 25% of sorted values from the top 75%.**

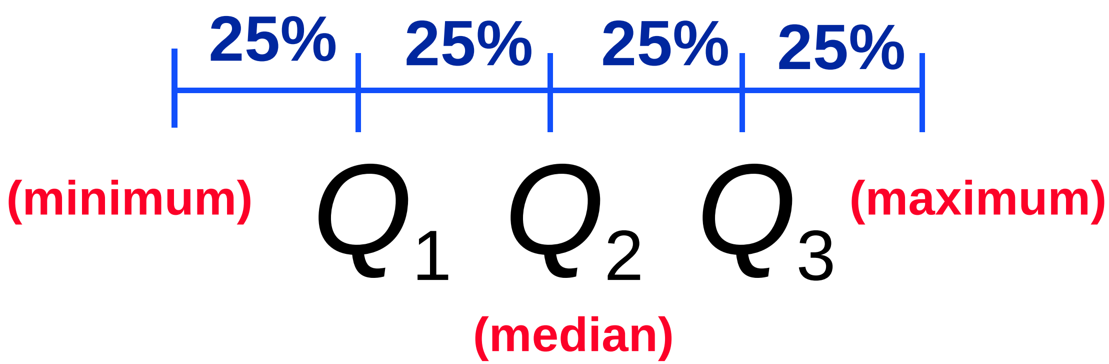
- Q_2 **(Second Quartile)** **same as the median; separates the bottom 50% of sorted values from the top 50%.**

- Q_1 **(Third Quartile)** **separates the bottom 75% of sorted values from the top 25%.**

Quartiles

$Q_1,\ Q_2,\ Q_3$

divides ranked scores into four equal parts



Percentiles

Just as there are 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 *k*th percentile

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

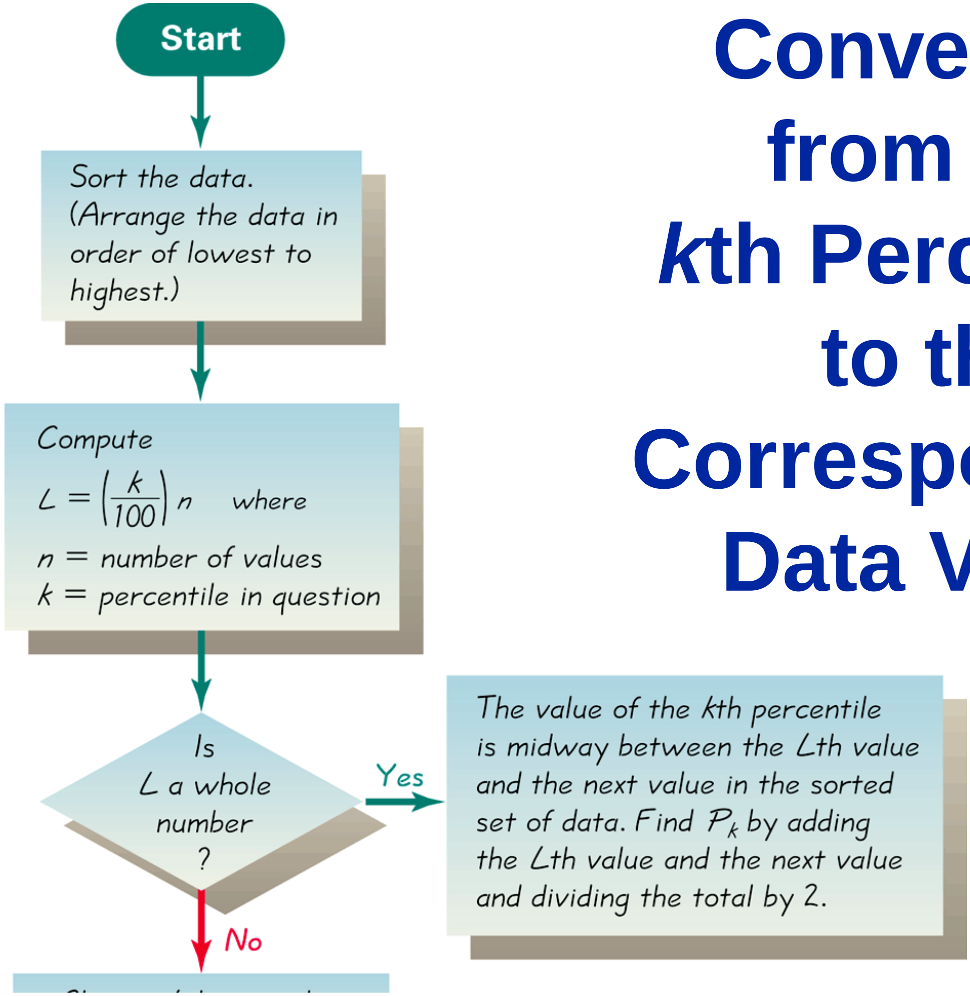


Figure 2-15

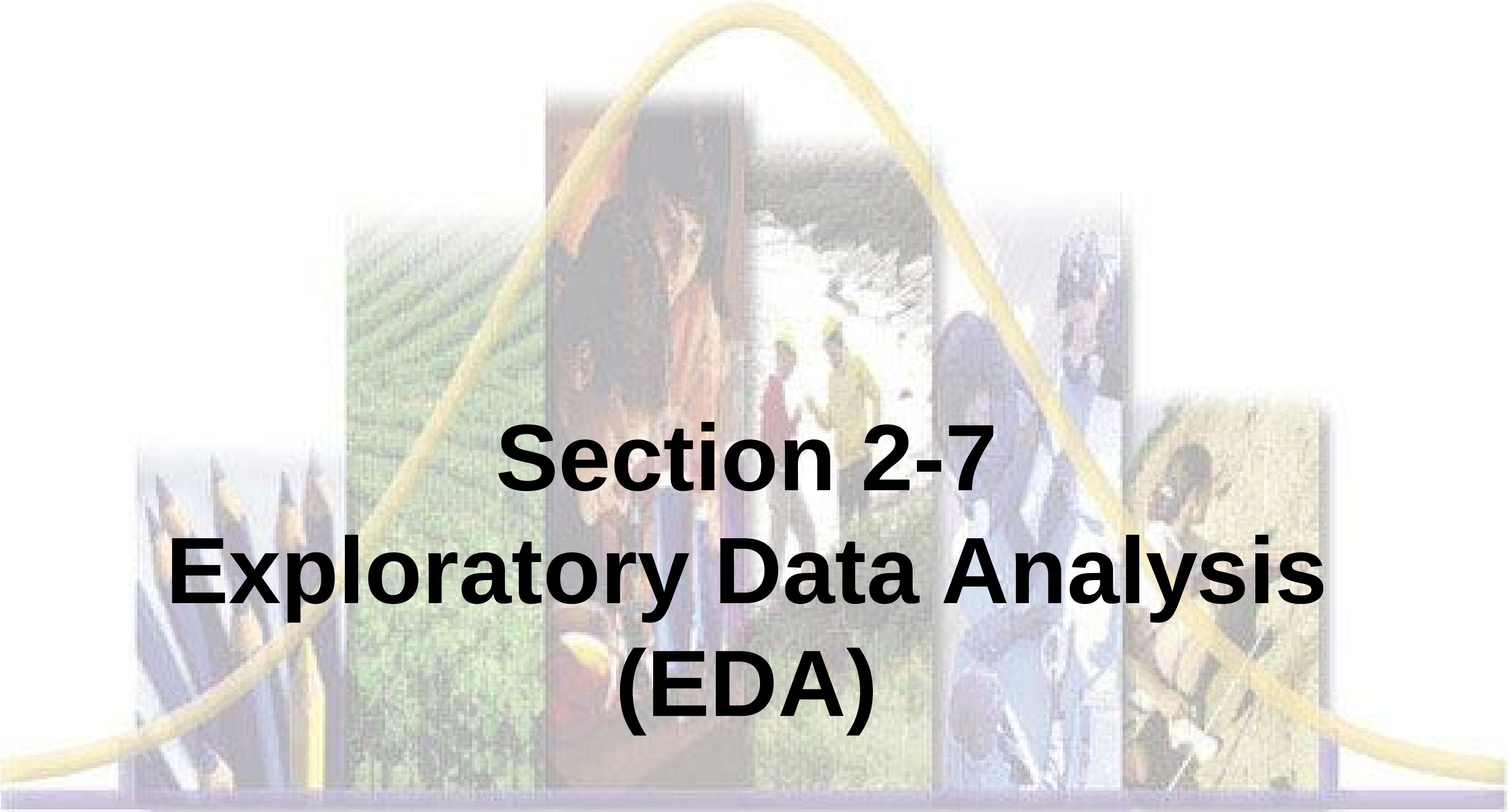
Some Other Statistics

- **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 2-7
Exploratory Data Analysis (EDA)

Definition

- **Exploratory Data Analysis is 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 the other values

Important Principles

- An outlier can have a dramatic effect on the mean

- An outlier 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

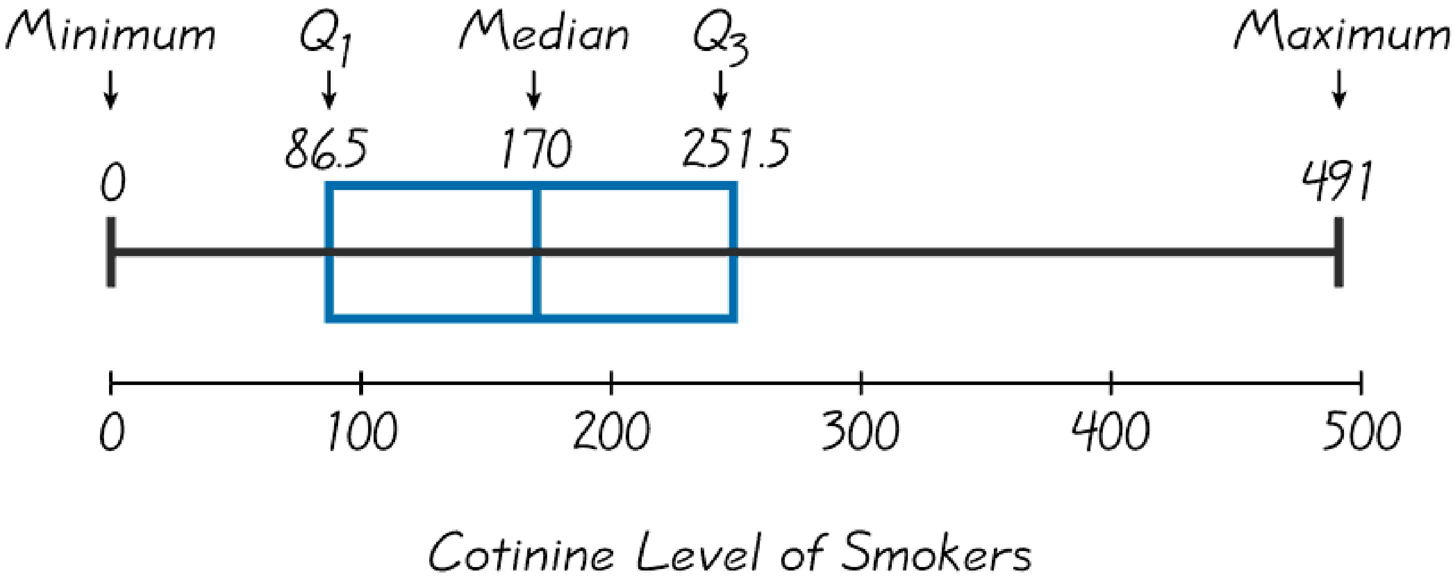


Figure 2-16

Boxplots

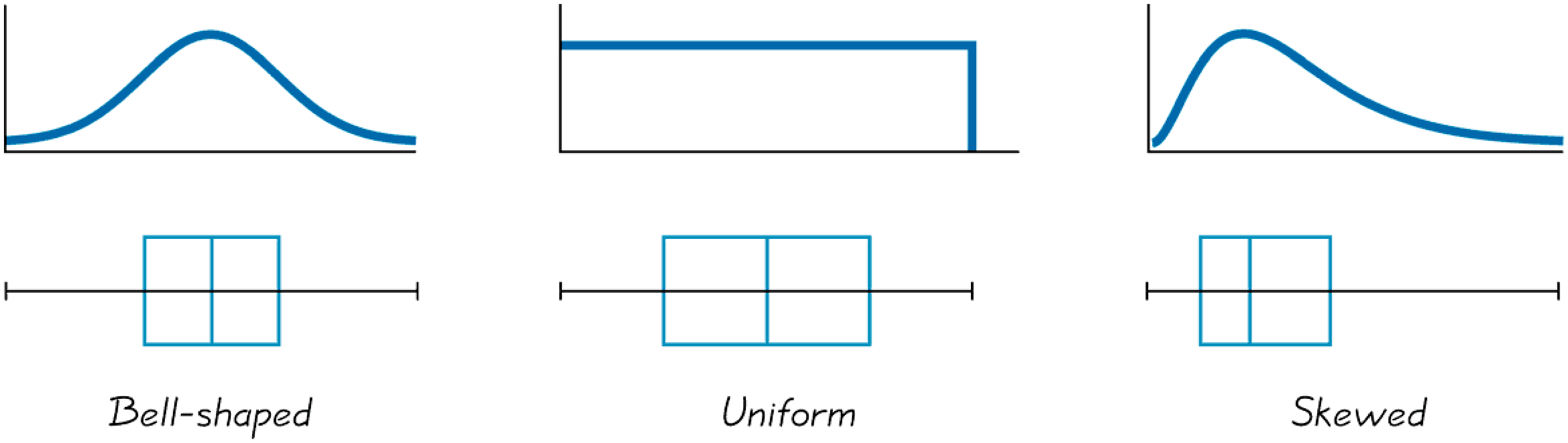


Figure 2-17

Recap

In this section we have looked at:

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


Ukuran
Pemusatan dan
Sebaran Data
15-Feb-10
Dadan Dasari
JURDIKMAT FPMIPA UPI
1