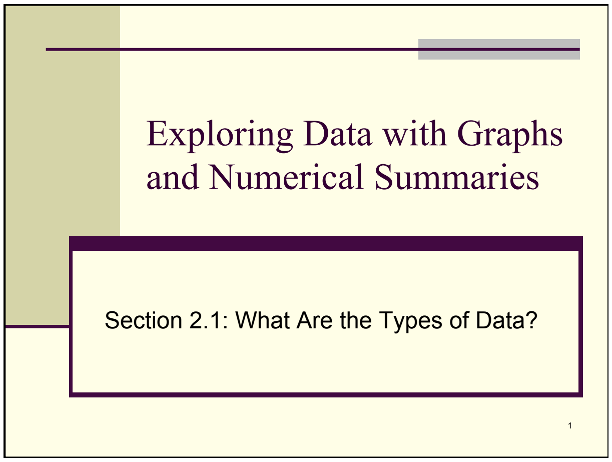
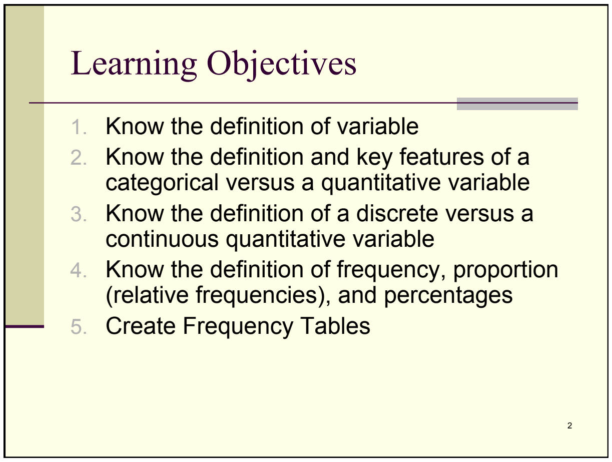
Individuals and Variables



Exploring Data with Graphs and Numerical Summaries

Section 2.1: What Are the Types of Data?

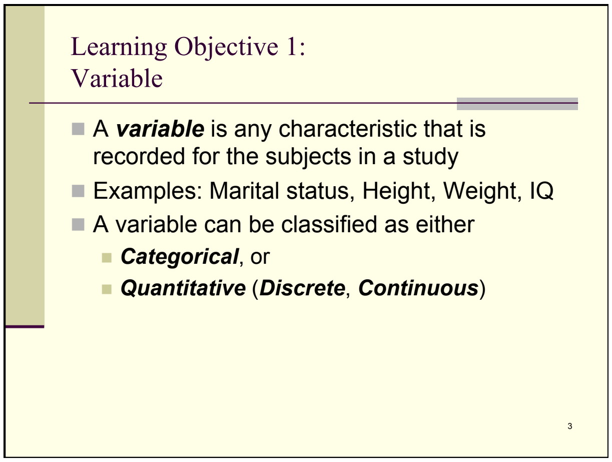
1



Learning Objectives

1. Know the definition of variable
2. Know the definition and key features of a categorical versus a quantitative variable
3. Know the definition of a discrete versus a continuous quantitative variable
4. Know the definition of frequency, proportion (relative frequencies), and percentages
5. Create Frequency Tables

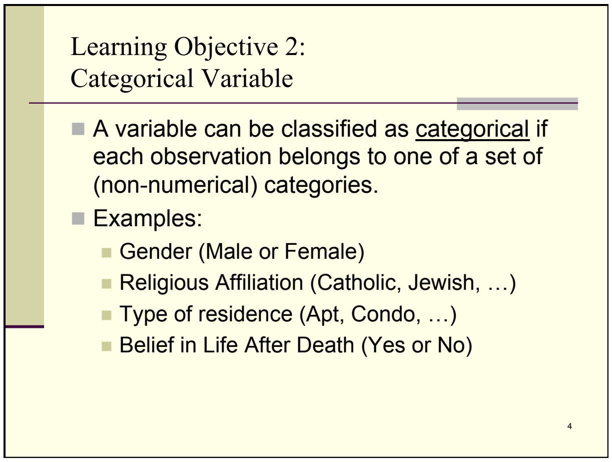
2



Learning Objective 1: Variable

- A ***variable*** is any characteristic that is recorded for the subjects in a study
- Examples: Marital status, Height, Weight, IQ
- A variable can be classified as either
  - ***Categorical***, or
  - ***Quantitative*** (***Discrete***, ***Continuous***)

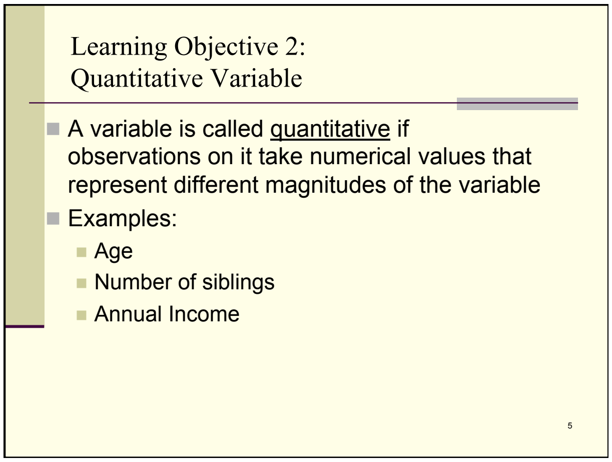
3



Learning Objective 2: Categorical Variable

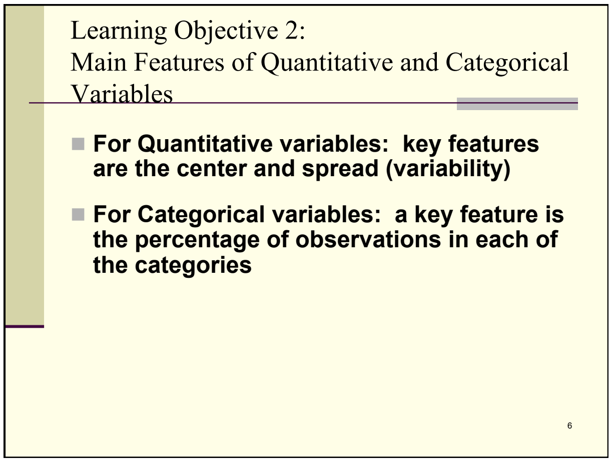
- A variable can be classified as <u>categorical</u> if each observation belongs to one of a set of (non-numerical) categories.
- Examples:
  - Gender (Male or Female)
  - Religious Affiliation (Catholic, Jewish, …)
  - Type of residence (Apt, Condo, …)
  - Belief in Life After Death (Yes or No)

4



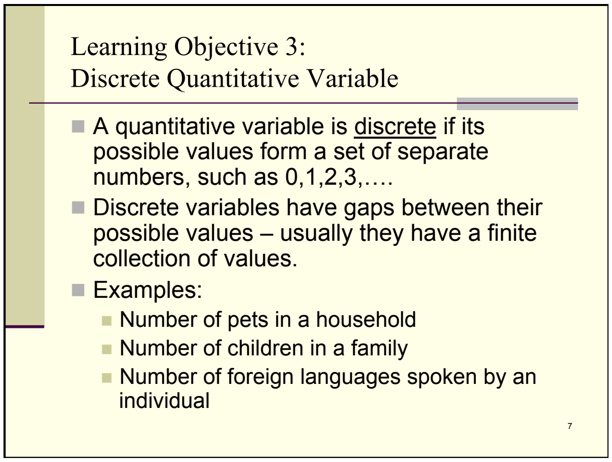
Learning Objective 2: Quantitative Variable

- A variable is called quantitative if observations on it take numerical values that represent different magnitudes of the variable
- Examples:
  - Age
  - Number of siblings
  - Annual Income



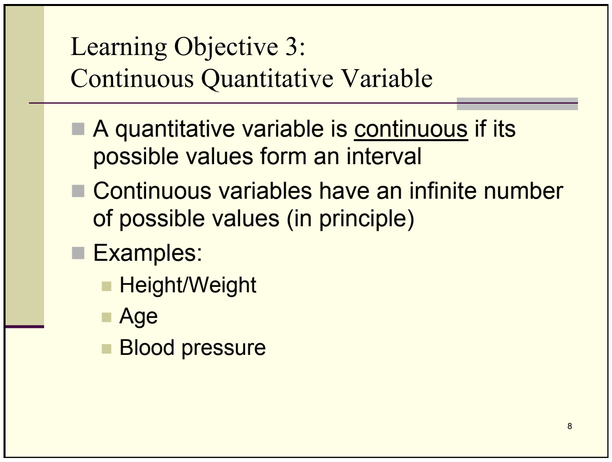
Learning Objective 2: Main Features of Quantitative and Categorical Variables

- **For Quantitative variables: key features are the center and spread (variability)**
- **For Categorical variables: a key feature is the percentage of observations in each of the categories**



Learning Objective 3: Discrete Quantitative Variable

- A quantitative variable is discrete if its possible values form a set of separate numbers, such as 0,1,2,3,….
- Discrete variables have gaps between their possible values – usually they have a finite collection of values.
- Examples:
  - Number of pets in a household
  - Number of children in a family
  - Number of foreign languages spoken by an individual



Learning Objective 3: Continuous Quantitative Variable

- A quantitative variable is continuous if its possible values form an interval
- Continuous variables have an infinite number of possible values (in principle)
- Examples:
  - Height/Weight
  - Age
  - Blood pressure

Learning Objective 4:
Proportion & Percentage (Relative Frequencies)

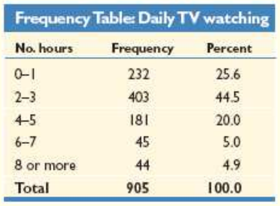
- The number of observations that fall in a certain category is the **frequency** (count) of observations in that category divided by the total number of observations
  - $\frac{\text{Frequency of that class}}{\text{Sum of all frequencies}}$
- The Percentage is the proportion multiplied by 100. Proportions and percentages are also called **relative frequencies**.

Learning Objective 4:
Frequency, Proportion, & Percentage Example

- If 4 students received an “A” out of 40 students, then,
  - 4 is the frequency
  - 0.10 =4/40 is the proportion and relative frequency
  - 10% is the percentage .1*100=10%

Learning Objective 5:
Frequency Table

- **A frequency table is a listing of possible values for a variable , together with the number of observations and/ or relative frequencies for each value**



Frequency Table: Daily TV watching

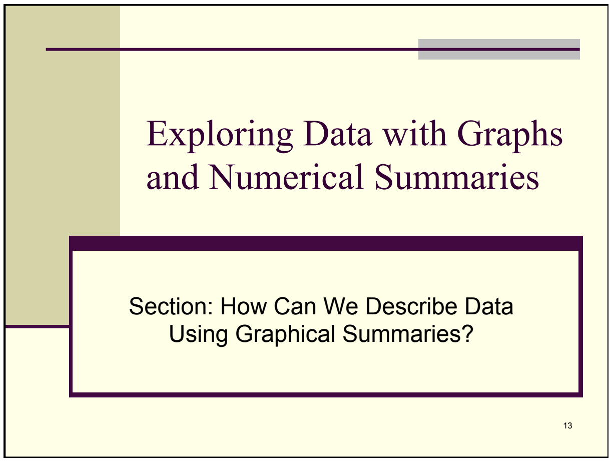
No. hours	Frequency	Percent
0–1	232	25.6
2–3	403	44.5
4–5	181	20.0
6–7	45	5.0
8 or more	44	4.9
Total	905	100.0

Class Problem #3

- A stock broker has been following different stocks over the last month and has recorded whether a stock is up, the same, or down in value. The results were

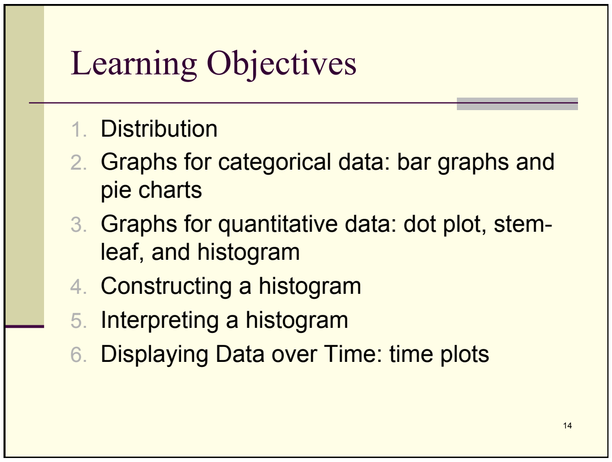
Performance of stock	Up	Same	Down
Count	21	7	12

1. What is the variable of interest
2. What type of variable is it?
3. Add proportions to this frequency table



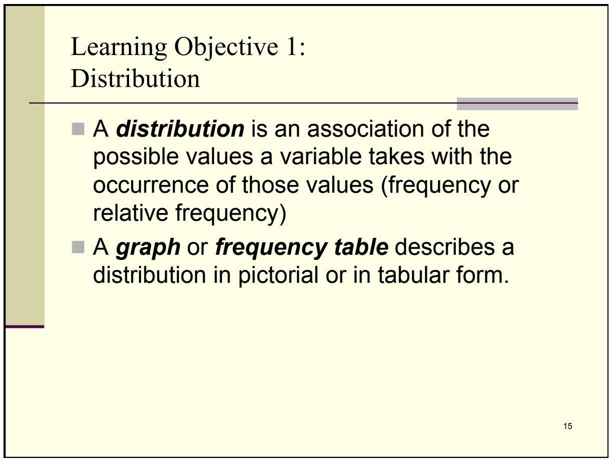
Exploring Data with Graphs and Numerical Summaries

Section: How Can We Describe Data Using Graphical Summaries?



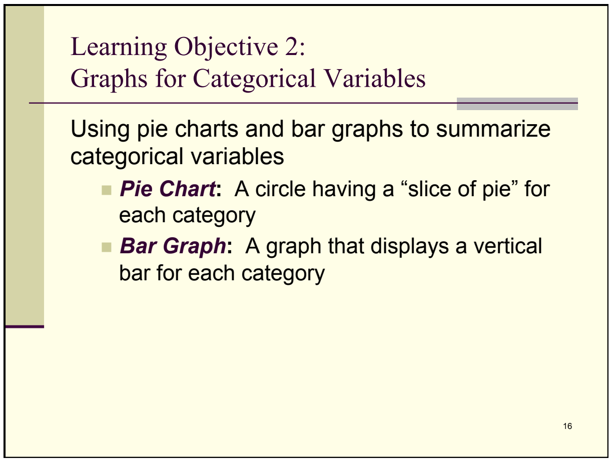
Learning Objectives

1. Distribution
2. Graphs for categorical data: bar graphs and pie charts
3. Graphs for quantitative data: dot plot, stem-leaf, and histogram
4. Constructing a histogram
5. Interpreting a histogram
6. Displaying Data over Time: time plots



Learning Objective 1: Distribution

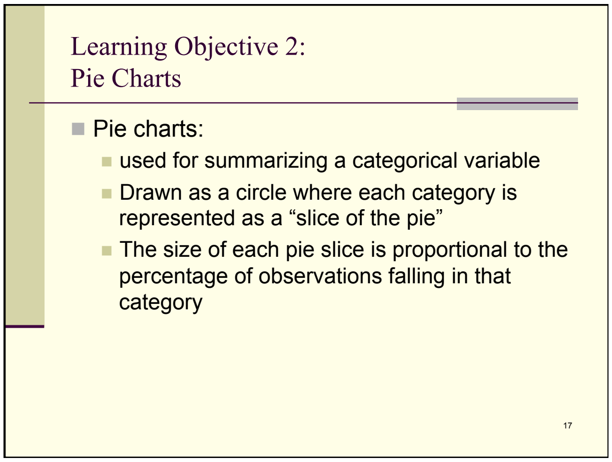
- A ***distribution*** is an association of the possible values a variable takes with the occurrence of those values (frequency or relative frequency)
- A ***graph*** or ***frequency table*** describes a distribution in pictorial or in tabular form.



Learning Objective 2: Graphs for Categorical Variables

Using pie charts and bar graphs to summarize categorical variables

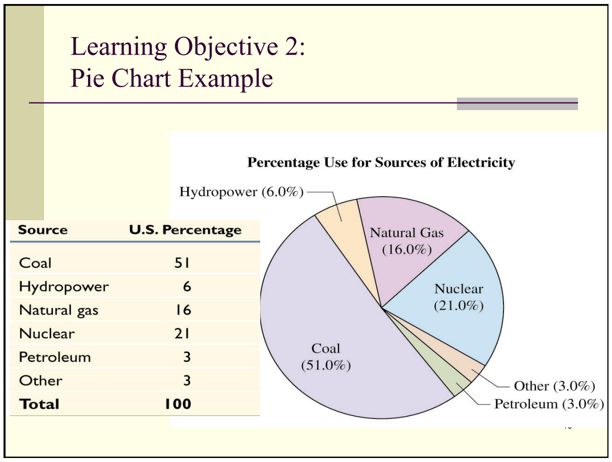
- ***Pie Chart*:** A circle having a “slice of pie” for each category
- ***Bar Graph*:** A graph that displays a vertical bar for each category



Learning Objective 2: Pie Charts

- Pie charts:
  - used for summarizing a categorical variable
  - Drawn as a circle where each category is represented as a “slice of the pie”
  - The size of each pie slice is proportional to the percentage of observations falling in that category

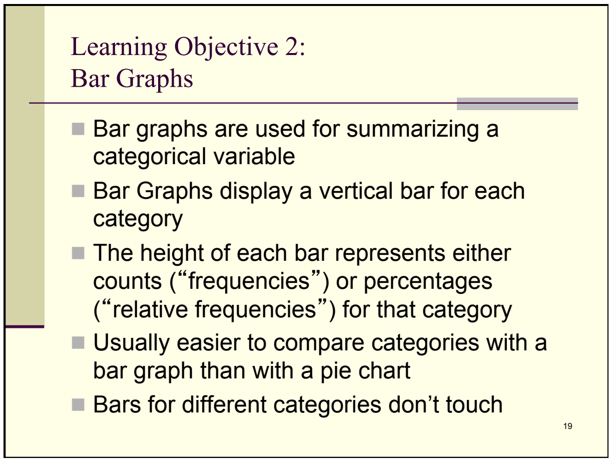
17



Learning Objective 2: Pie Chart Example

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	**100**

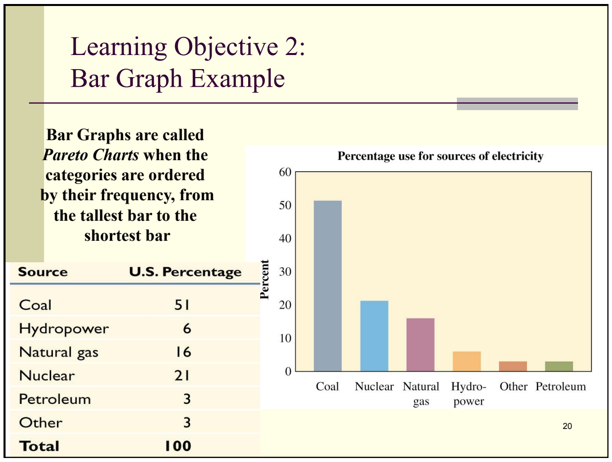
Percentage Use for Sources of Electricity



Learning Objective 2: Bar Graphs

- Bar graphs are used for summarizing a categorical variable
- Bar Graphs display a vertical bar for each category
- The height of each bar represents either counts (“frequencies”) or percentages (“relative frequencies”) for that category
- Usually easier to compare categories with a bar graph than with a pie chart
- Bars for different categories don’t touch

19



Learning Objective 2: Bar Graph Example

Bar Graphs are called *Pareto Charts* when the categories are ordered by their frequency, from the tallest bar to the shortest bar

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	**100**

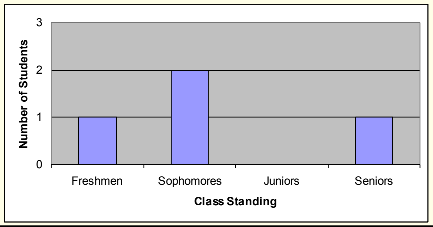
Percentage use for sources of electricity

20

Learning Objective 2: Class Exercise

There are 7 students in a class who are either freshman, sophomores, juniors, or seniors.

The number of students in this class who are juniors is ______.



Learning Objective 3: Graphs for Quantitative Data

- ***Dot Plot***: shows a dot for each observation placed above its value on a number line
- ***Stem-and-Leaf Plot***: portrays the individual observations
- ***Histogram***: uses bars to portray the data

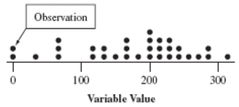
Learning Objective 3: Which Graph?

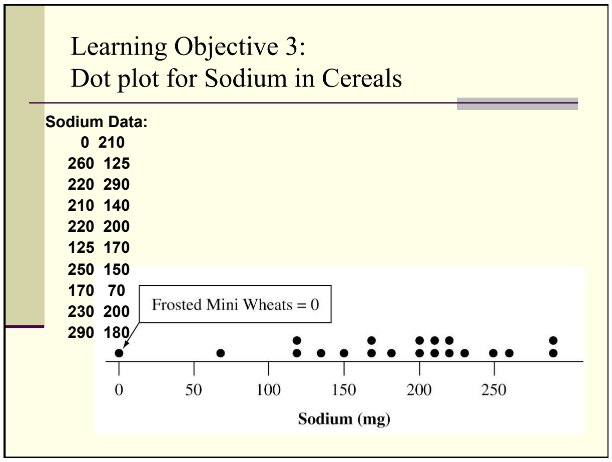
- **Dot-plot and stem-and-leaf plot:**
  - **More useful for small data sets**
  - **Data values are retained**
- **Histogram**
  - **More useful for large data sets**
  - **Most compact display**
  - **More flexibility in defining intervals**

Learning Objective 3: Dot Plots

- Dot Plots are used for summarizing a quantitative variable
- To construct a dot plot

1. Draw a horizontal line
2. Label it with the name of the variable
3. Mark regular values of the variable on it
4. For each observation, place a dot above its value on the number line



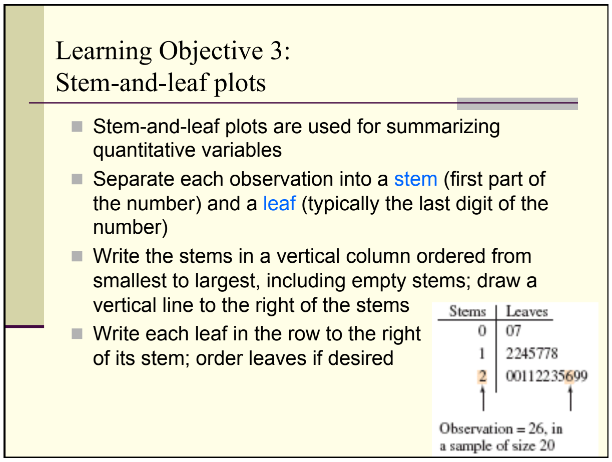


Learning Objective 3: Dot plot for Sodium in Cereals

Sodium Data:

0	210
260	125
220	290
210	140
220	200
125	170
250	150
170	70
230	200
290	180

Frosted Mini Wheats = 0

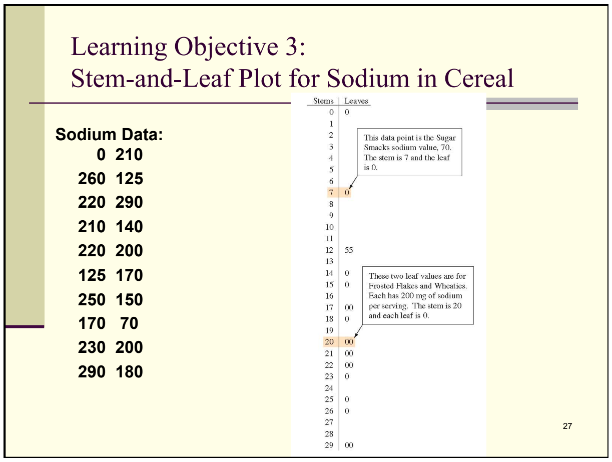


Learning Objective 3: Stem-and-leaf plots

- Stem-and-leaf plots are used for summarizing quantitative variables
- Separate each observation into a stem (first part of the number) and a leaf (typically the last digit of the number)
- Write the stems in a vertical column ordered from smallest to largest, including empty stems; draw a vertical line to the right of the stems
- Write each leaf in the row to the right of its stem; order leaves if desired

Stems	Leaves
0	07
1	2245778
2	00112235699

Observation = 26, in a sample of size 20



Learning Objective 3: Stem-and-Leaf Plot for Sodium in Cereal

Sodium Data:

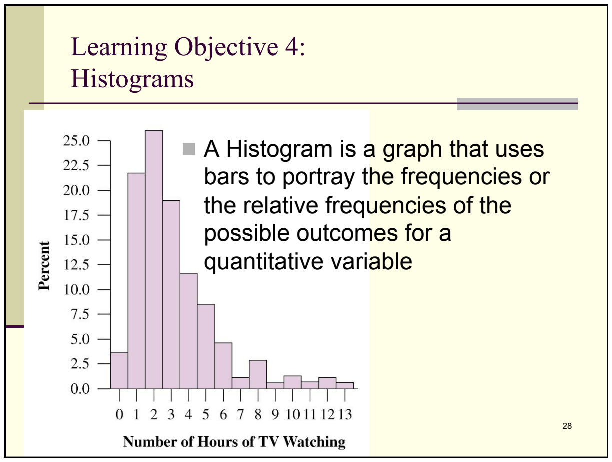
0	210
260	125
220	290
210	140
220	200
125	170
250	150
170	70
230	200
290	180

Stems	Leaves
0	0
1	
2	
3	
4	
5	
6	
7	0
8	
9	
10	
11	
12	55
13	
14	0
15	0
16	
17	00
18	0
19	
20	00
21	00
22	00
23	0
24	
25	0
26	0
27	
28	
29	00

This data point is the Sugar Smacks sodium value, 70. The stem is 7 and the leaf is 0.

These two leaf values are for Frosted Flakes and Wheaties. Each has 200 mg of sodium per serving. The stem is 20 and each leaf is 0.

27



Learning Objective 4: Histograms

- A Histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable

28

Learning Objective 4:
Steps for Constructing a Histogram

1. Divide the range of the data into intervals of equal width
2. Count the number of observations in each interval, creating a frequency table
3. On the horizontal axis, label the values or the endpoints of the intervals.
4. Draw a bar over each value or interval with height equal to its frequency (or percentage), values of which are marked on the vertical axis.
5. Width of all bars the same and all touch
6. Label and title appropriately

29

Learning Objective 4:
Histogram for Sodium in Cereals

Sodium Data:

0 210
260 125
220 290
210 140
220 200
125 170
250 150
170 70
230 200
290 180

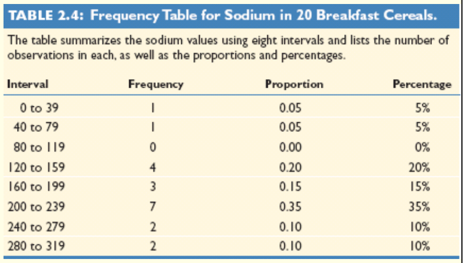


TABLE 2.4: Frequency Table for Sodium in 20 Breakfast Cereals.

The table summarizes the sodium values using eight intervals and lists the number of observations in each, as well as the proportions and percentages.

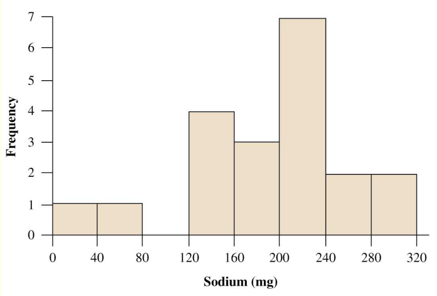
Interval	Frequency	Proportion	Percentage
0 to 39	1	0.05	5%
40 to 79	1	0.05	5%
80 to 119	0	0.00	0%
120 to 159	4	0.20	20%
160 to 199	3	0.15	15%
200 to 239	7	0.35	35%
240 to 279	2	0.10	10%
280 to 319	2	0.10	10%

30

Learning Objective 4:
Histogram for Sodium in Cereals

Sodium Data:

0 210
260 125
220 290
210 140
220 200
125 170
250 150
170 70
230 200
290 180



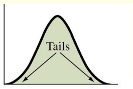
Learning Objective 5:
Interpreting Histograms

- Overall pattern consists of center, spread, and shape
  - Assess where a distribution is **centered** by finding the median (50% of data below median 50% of data above).
  - Assess the **spread** of a distribution.
  - **Shape** of a distribution: roughly symmetric, skewed to the right, or skewed to the left

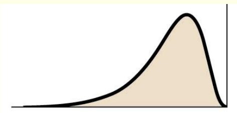
32

Learning Objective 5: Shape

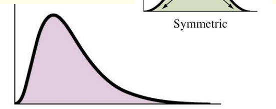
- Symmetric Distributions: if both left and right sides of the histogram are mirror images of each other



Symmetric



Skewed to the left



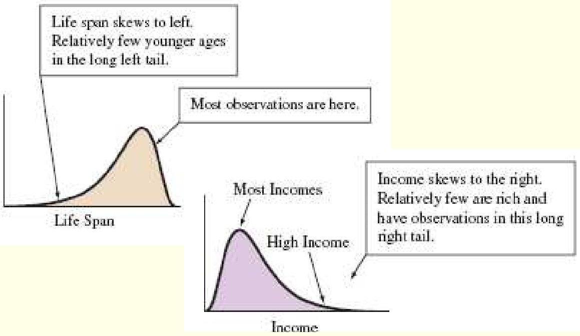
Skewed to the right

A distribution is skewed to the left if the left tail is longer than the right tail

A distribution is skewed to the right if the right tail is longer than the left tail

33

Learning Objective 5: Examples of Skewness



Life span skews to left. Relatively few younger ages in the long left tail.

Most observations are here.

Life Span

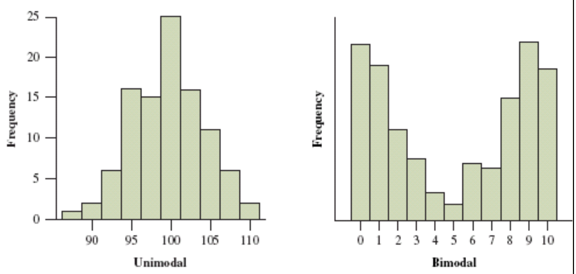
Most Incomes

High Income

Income skews to the right. Relatively few are rich and have observations in this long right tail.

Income

Learning Objective 5: Shape: Type of Mound



Learning Objective 5: Shape and Skewness

- Consider a data set containing IQ scores for the general public:
- What shape would you expect a histogram of this data set to have?

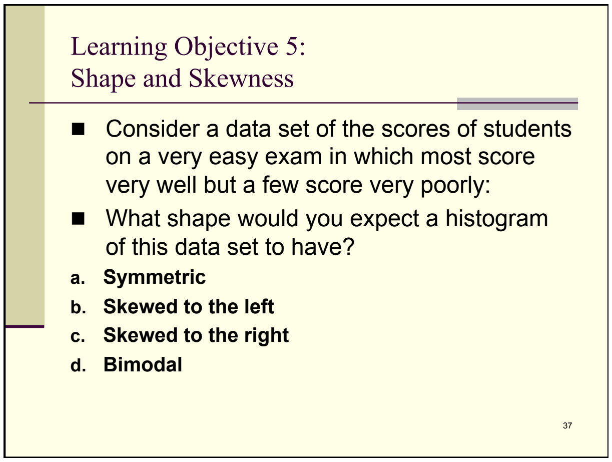
a. **Symmetric**

b. **Skewed to the left**

c. **Skewed to the right**

d. **Bimodal**

36

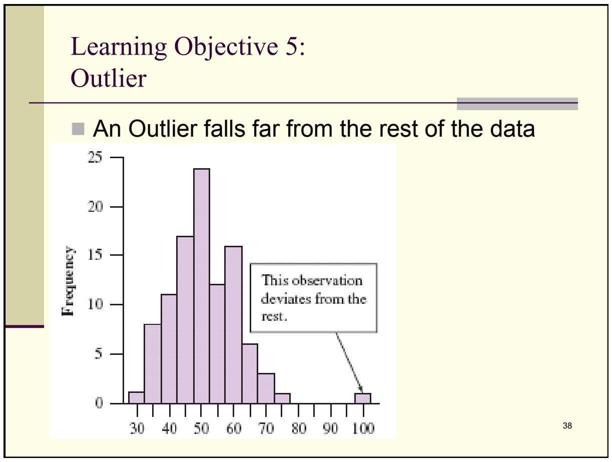


Learning Objective 5: Shape and Skewness

- Consider a data set of the scores of students on a very easy exam in which most score very well but a few score very poorly:
- What shape would you expect a histogram of this data set to have?

a. **Symmetric**
b. **Skewed to the left**
c. **Skewed to the right**
d. **Bimodal**

37

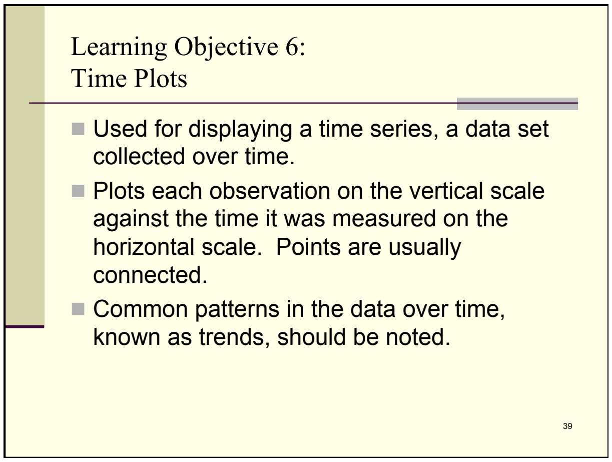


Learning Objective 5: Outlier

- An Outlier falls far from the rest of the data

This observation deviates from the rest.

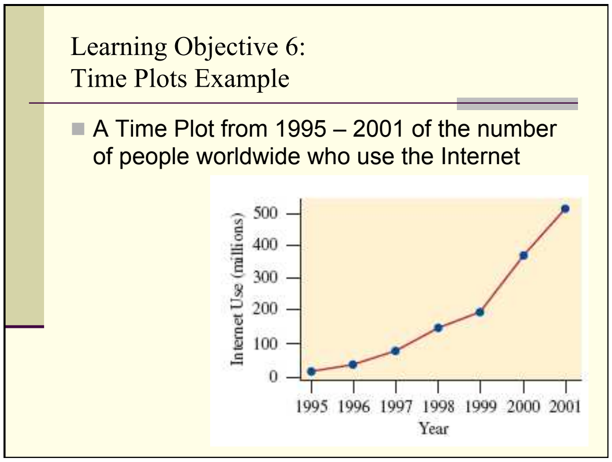
38



Learning Objective 6: Time Plots

- Used for displaying a time series, a data set collected over time.
- Plots each observation on the vertical scale against the time it was measured on the horizontal scale. Points are usually connected.
- Common patterns in the data over time, known as trends, should be noted.

39



Learning Objective 6: Time Plots Example

- A Time Plot from 1995 – 2001 of the number of people worldwide who use the Internet

Exploring Data with Graphs and Numerical Summaries

Section: How Can We Describe the Center of Quantitative Data?

Learning Objectives

1. Calculating the mean
2. Calculating the median
3. Comparing the Mean & Median
4. Definition of Resistant
5. Know how to identify the mode of a distribution

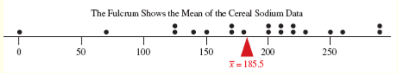
Learning Objective 1: Mean

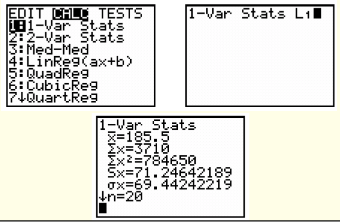
- The mean is the sum of the observations divided by the number of observations
- It is the center of mass

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

Learning Objective 1: Calculate Mean

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	210
All Bran	260
Apple Jacks	125
Capt Crunch	220
Cheerios	290
Cinnamon Toast	210
Crackling Oat Bran	140
Crispix	220
Frosted Flakes	200
Fruit Loops	125
Grape Nuts	170
Honey Nut Cheerios	250
Life	150
Oatmeal Raisin Crisp	170
Sugar Smacks	70
Special K	230
Wheaties	200
Corn Flakes	290
Honeycomb	180


The Fulcrum Shows the Mean of the Cereal Sodium Data


```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```

```
1-Var Stats L1
```

```
1-Var Stats
 x̄=185.5
 Σx=3710
 Σx²=784650
 Sx=71.24642189
 σx=69.44242219
↓n=20
```

Learning Objective 2: Median

- The median is the midpoint of the observations when they are ordered from the smallest to the largest (or from the largest to smallest)
- Order observations
- If the number of observations is:
  - Odd, then the median is the middle observation
  - Even, then the median is the average of the two middle observations

45

Learning Objective 2: Median

1) Sort observations by size.
n = number of observations

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114

2.a) If *n* is **odd,** the median is observation (*n*+1)/2 down the list

← *n* = 9
(*n*+1)/2 = 10/2 = 5
Median = 99

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114
10	121

2.b) If *n* is **even,** the median is the mean of the two middle observations

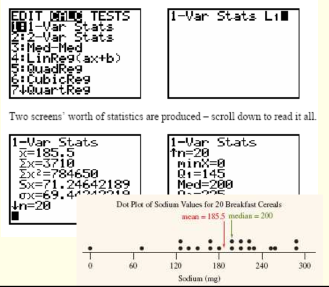
n = 10 →
(*n+1)*/2 = 5.5
Median = (99+101) /2 = 100

46

Learning Objective 1 &2: Calculate Mean and Median

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	210
All Bran	260
Apple Jacks	125
Capt Crunch	220
Cheerios	290
Cinnamon Toast	210
Crackling Oat Bran	140
Crispix	220
Frosted Flakes	200
Fruit Loops	125
Grape Nuts	170
Honey Nut Cheerios	250
Life	150
Oatmeal Raisin Crisp	170
Sugar Smacks	70
Special K	230
Wheaties	200
Corn Flakes	290
Honeycomb	180

Enter data into L1
STAT; CALC; 1:1-Var Stats; Enter
L1;Enter



```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```

```
1-Var Stats L1
```

Two screens' worth of statistics are produced – scroll down to read it all.

```
1-Var Stats
 x̄=185.5
 Σx=3710
 Σx²=784650
 Sx=71.24642189
 σx=69.4
↓n=20
```

```
1-Var Stats
↑n=20
 minX=0
 Q1=145
 Med=200
 Q3=225
```

Dot Plot of Sodium Values for 20 Breakfast Cereals

47

Leaning Objectives 1 & 2: Find the mean and median

CO_2 Pollution levels in 8 largest nations measured in metric tons per person:

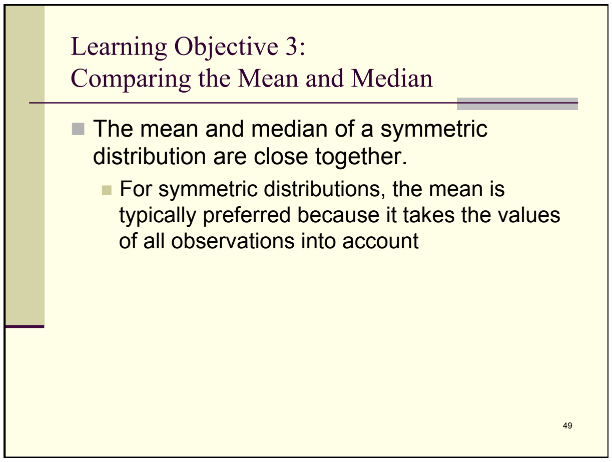
2.3 1.1 19.7 9.8 1.8 1.2 0.7 0.2

a. **Mean = 4.6 Median = 1.5**

b. **Mean = 4.6 Median = 5.8**

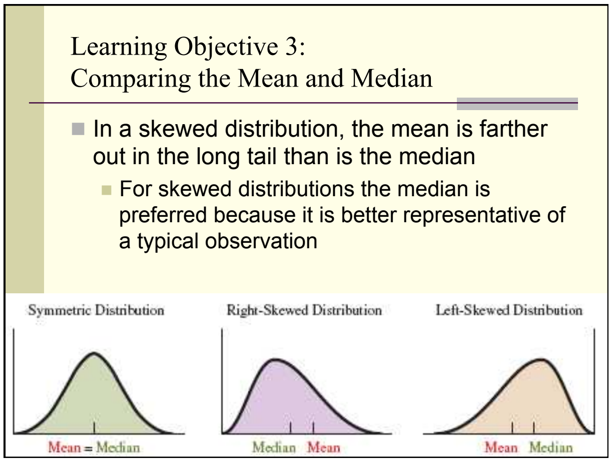
c. **Mean = 1.5 Median = 4.6**

48



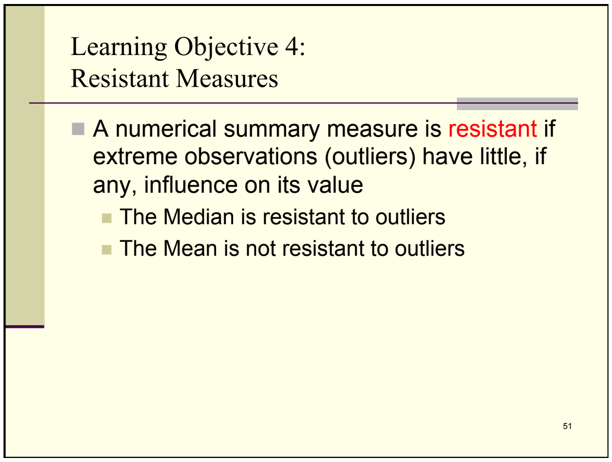
Learning Objective 3: Comparing the Mean and Median

- The mean and median of a symmetric distribution are close together.
  - For symmetric distributions, the mean is typically preferred because it takes the values of all observations into account



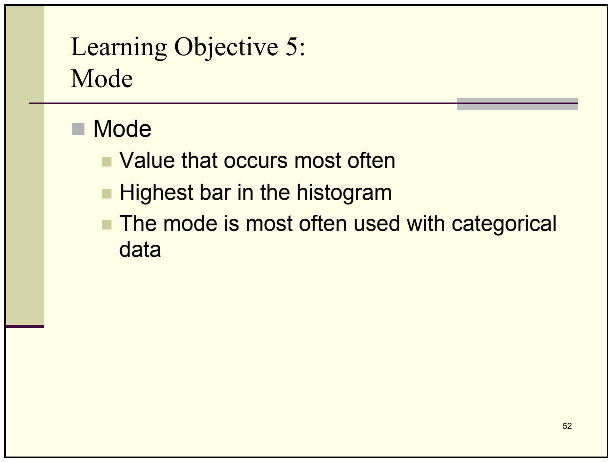
Learning Objective 3: Comparing the Mean and Median

- In a skewed distribution, the mean is farther out in the long tail than is the median
  - For skewed distributions the median is preferred because it is better representative of a typical observation



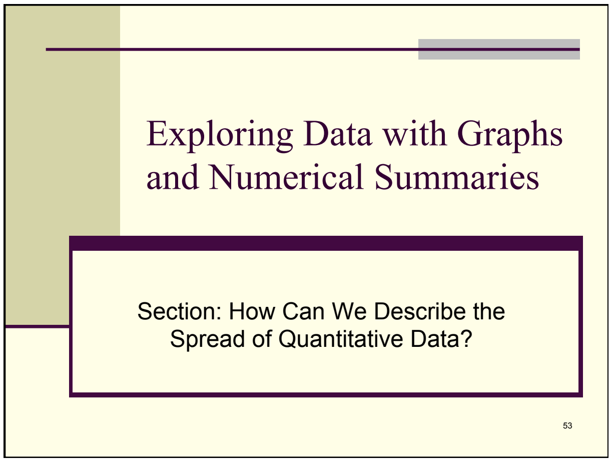
Learning Objective 4: Resistant Measures

- A numerical summary measure is resistant if extreme observations (outliers) have little, if any, influence on its value
  - The Median is resistant to outliers
  - The Mean is not resistant to outliers



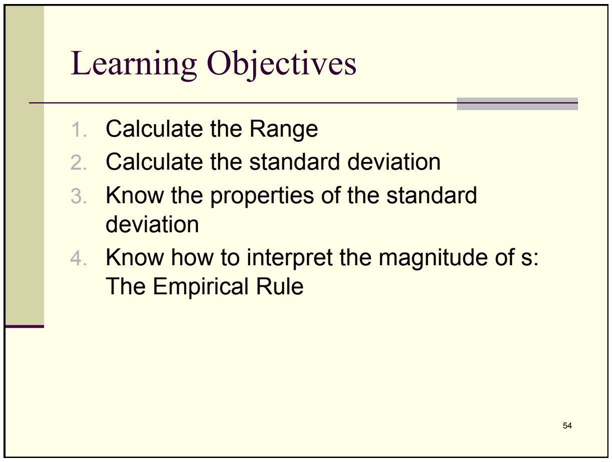
Learning Objective 5: Mode

- Mode
  - Value that occurs most often
  - Highest bar in the histogram
  - The mode is most often used with categorical data



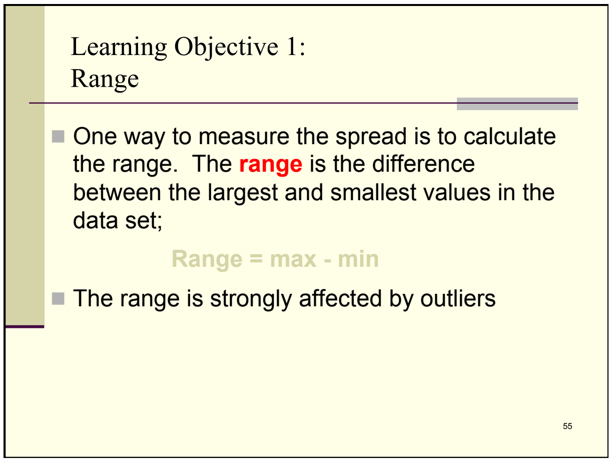
Exploring Data with Graphs and Numerical Summaries

Section: How Can We Describe the Spread of Quantitative Data?



Learning Objectives

1. Calculate the Range
2. Calculate the standard deviation
3. Know the properties of the standard deviation
4. Know how to interpret the magnitude of s: The Empirical Rule

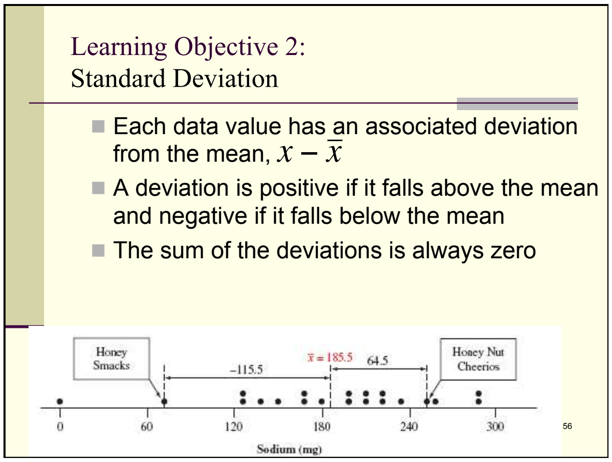


Learning Objective 1: Range

- One way to measure the spread is to calculate the range. The **range** is the difference between the largest and smallest values in the data set;

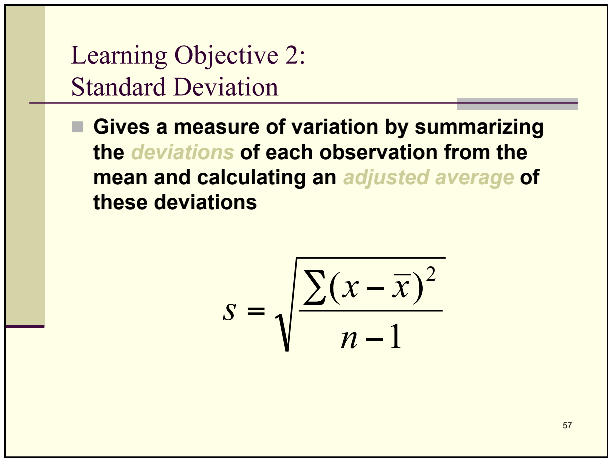
  Range = max - min

- The range is strongly affected by outliers



Learning Objective 2: Standard Deviation

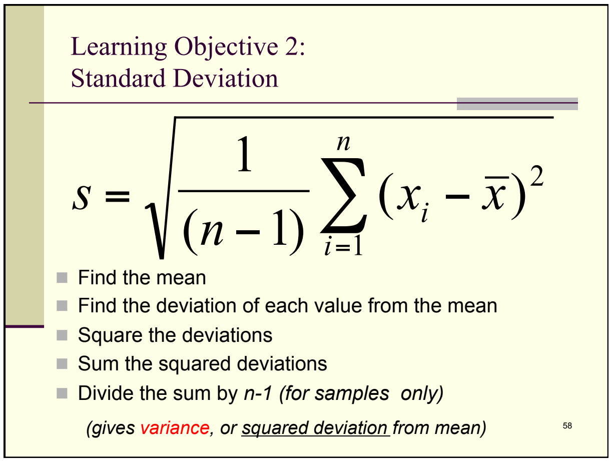
- Each data value has an associated deviation from the mean, $x - \bar{x}$
- A deviation is positive if it falls above the mean and negative if it falls below the mean
- The sum of the deviations is always zero



Learning Objective 2: Standard Deviation

- **Gives a measure of variation by summarizing the *deviations* of each observation from the mean and calculating an *adjusted average* of these deviations**

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

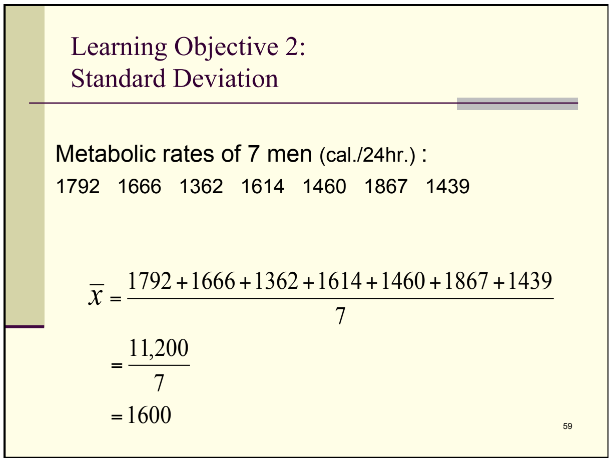


Learning Objective 2: Standard Deviation

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

- Find the mean
- Find the deviation of each value from the mean
- Square the deviations
- Sum the squared deviations
- Divide the sum by *n-1 (for samples only)*

(gives variance, or squared deviation from mean)



Learning Objective 2: Standard Deviation

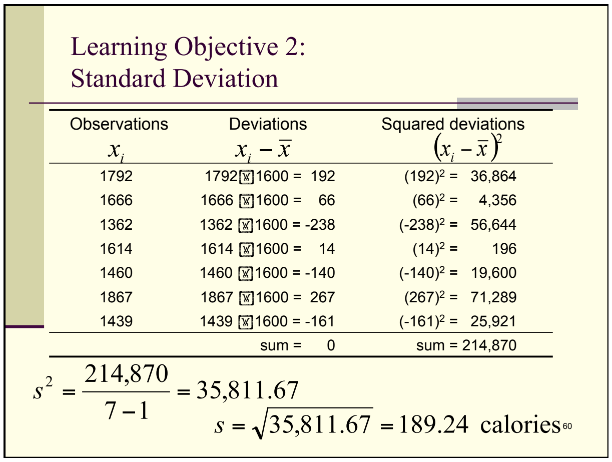
Metabolic rates of 7 men (cal./24hr.) :

1792 1666 1362 1614 1460 1867 1439

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7}$$

$$= \frac{11{,}200}{7}$$

$$= 1600$$

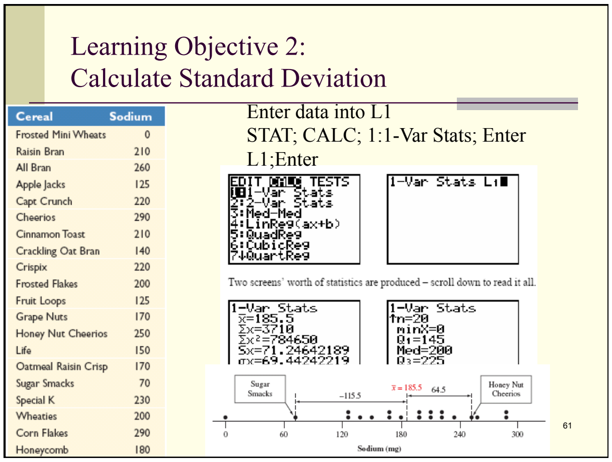


Learning Objective 2: Standard Deviation

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	1792☒1600 = 192	$(192)^2$ = 36,864
1666	1666 ☒1600 = 66	$(66)^2$ = 4,356
1362	1362 ☒1600 = -238	$(-238)^2$ = 56,644
1614	1614 ☒1600 = 14	$(14)^2$ = 196
1460	1460 ☒1600 = -140	$(-140)^2$ = 19,600
1867	1867 ☒1600 = 267	$(267)^2$ = 71,289
1439	1439 ☒1600 = -161	$(-161)^2$ = 25,921
	sum = 0	sum = 214,870

$$s^2 = \frac{214{,}870}{7-1} = 35{,}811.67$$

$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$



Learning Objective 2: Calculate Standard Deviation

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	210
All Bran	260
Apple Jacks	125
Capt Crunch	220
Cheerios	290
Cinnamon Toast	210
Crackling Oat Bran	140
Crispix	220
Frosted Flakes	200
Fruit Loops	125
Grape Nuts	170
Honey Nut Cheerios	250
Life	150
Oatmeal Raisin Crisp	170
Sugar Smacks	70
Special K	230
Wheaties	200
Corn Flakes	290
Honeycomb	180

Enter data into L1
STAT; CALC; 1:1-Var Stats; Enter
L1;Enter

```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```

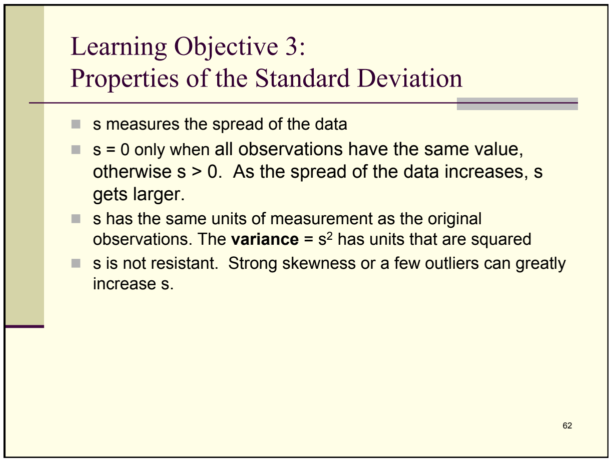
```
1-Var Stats L1
```

Two screens' worth of statistics are produced – scroll down to read it all.

```
1-Var Stats
 x̄=185.5
 Σx=3710
 Σx²=784650
 Sx=71.24642189
 σx=69.44242219
```

```
1-Var Stats
↑n=20
 minX=0
 Q1=145
 Med=200
 Q3=225
```

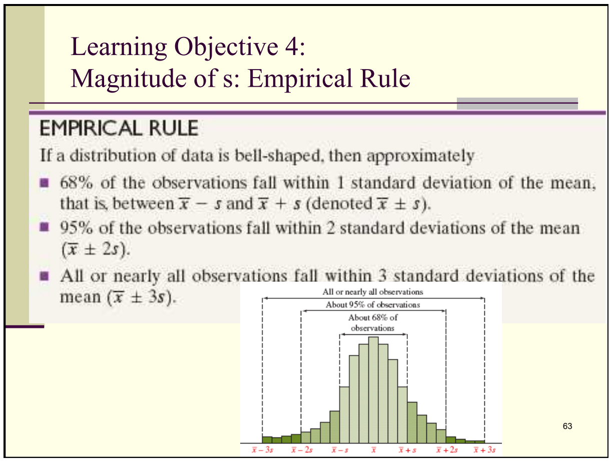
61



Learning Objective 3: Properties of the Standard Deviation

- s measures the spread of the data
- s = 0 only when all observations have the same value, otherwise s > 0. As the spread of the data increases, s gets larger.
- s has the same units of measurement as the original observations. The **variance** = s^2 has units that are squared
- s is not resistant. Strong skewness or a few outliers can greatly increase s.

62



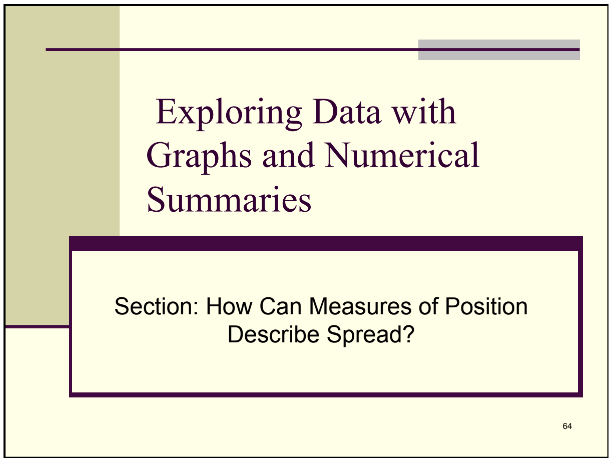
Learning Objective 4: Magnitude of s: Empirical Rule

EMPIRICAL RULE

If a distribution of data is bell-shaped, then approximately

- 68% of the observations fall within 1 standard deviation of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$ (denoted $\bar{x} \pm s$).
- 95% of the observations fall within 2 standard deviations of the mean ($\bar{x} \pm 2s$).
- All or nearly all observations fall within 3 standard deviations of the mean ($\bar{x} \pm 3s$).

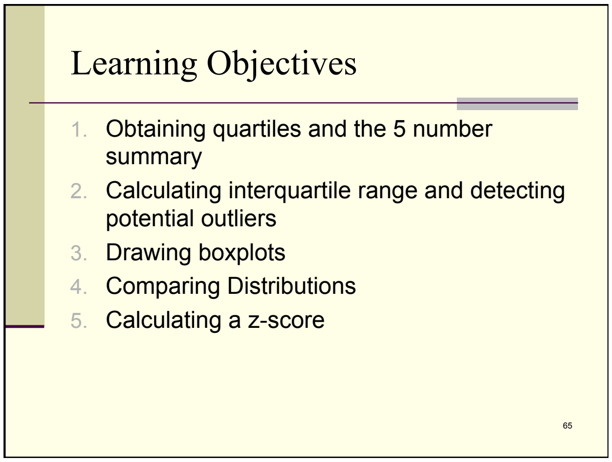
63



Exploring Data with Graphs and Numerical Summaries

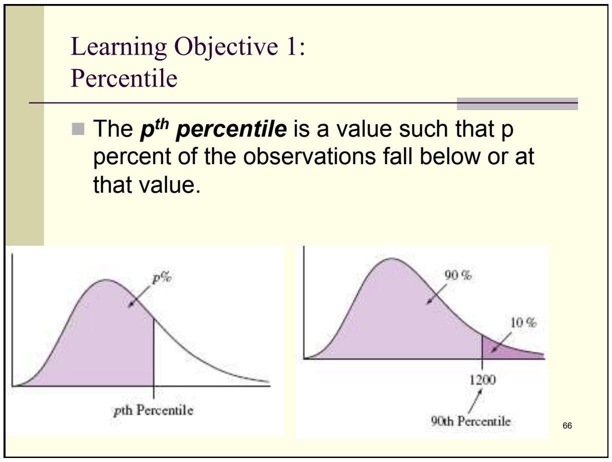
Section: How Can Measures of Position Describe Spread?

64



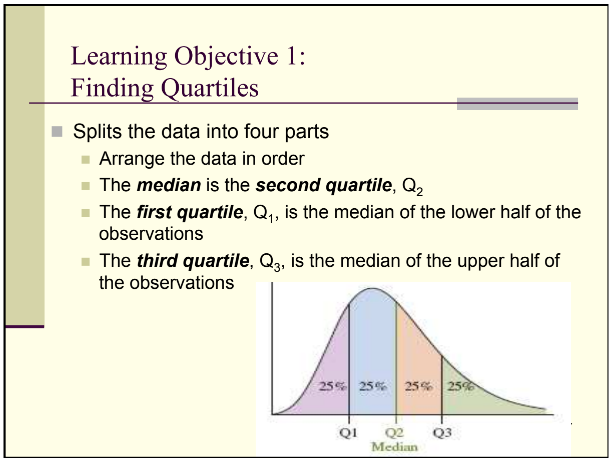
Learning Objectives

1. Obtaining quartiles and the 5 number summary
2. Calculating interquartile range and detecting potential outliers
3. Drawing boxplots
4. Comparing Distributions
5. Calculating a z-score



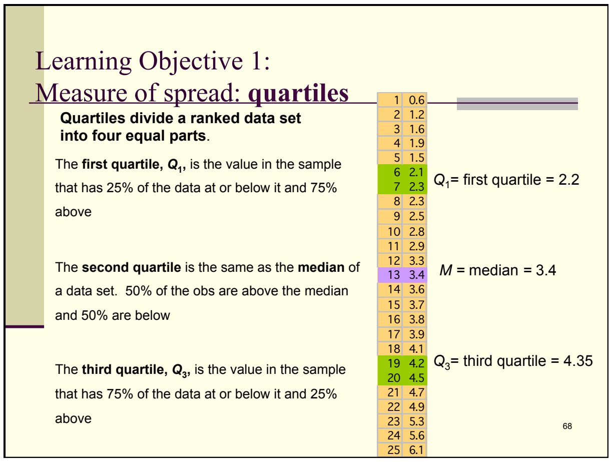
Learning Objective 1: Percentile

- The ***p^th^ percentile*** is a value such that p percent of the observations fall below or at that value.



Learning Objective 1: Finding Quartiles

- Splits the data into four parts
  - Arrange the data in order
  - The ***median*** is the ***second quartile***, Q_2
  - The ***first quartile***, Q_1, is the median of the lower half of the observations
  - The ***third quartile***, Q_3, is the median of the upper half of the observations



Learning Objective 1: Measure of spread: **quartiles**

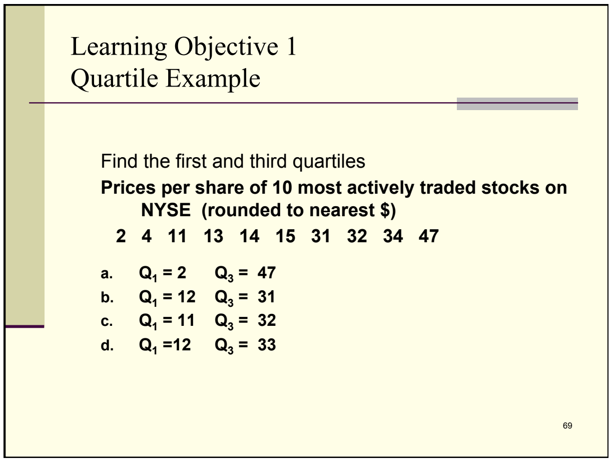
Quartiles divide a ranked data set into four equal parts.

The **first quartile, Q_1,** is the value in the sample that has 25% of the data at or below it and 75% above

The **second quartile** is the same as the **median** of a data set. 50% of the obs are above the median and 50% are below

The **third quartile, Q_3,** is the value in the sample that has 75% of the data at or below it and 25% above

1	0.6	
2	1.2	
3	1.6	
4	1.9	
5	1.5	
6	2.1	Q_1= first quartile = 2.2
7	2.3	
8	2.3	
9	2.5	
10	2.8	
11	2.9	
12	3.3	
13	3.4	*M* = median = 3.4
14	3.6	
15	3.7	
16	3.8	
17	3.9	
18	4.1	
19	4.2	Q_3= third quartile = 4.35
20	4.5	
21	4.7	
22	4.9	
23	5.3	
24	5.6	
25	6.1	



Learning Objective 1
Quartile Example

Find the first and third quartiles

Prices per share of 10 most actively traded stocks on NYSE (rounded to nearest $)

2 4 11 13 14 15 31 32 34 47

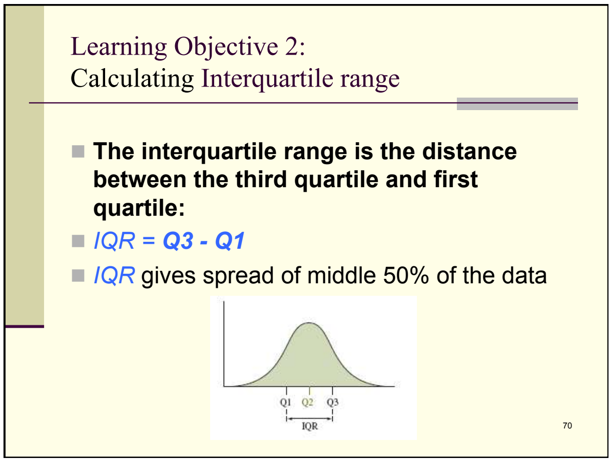
a. $Q_1 = 2$ $Q_3 = 47$

b. $Q_1 = 12$ $Q_3 = 31$

c. $Q_1 = 11$ $Q_3 = 32$

d. $Q_1 = 12$ $Q_3 = 33$

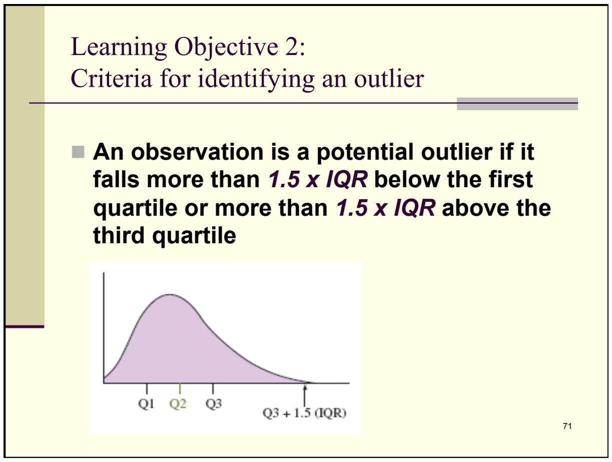
69



Learning Objective 2:
Calculating Interquartile range

- **The interquartile range is the distance between the third quartile and first quartile:**
- *IQR* = ***Q3 - Q1***
- *IQR* gives spread of middle 50% of the data

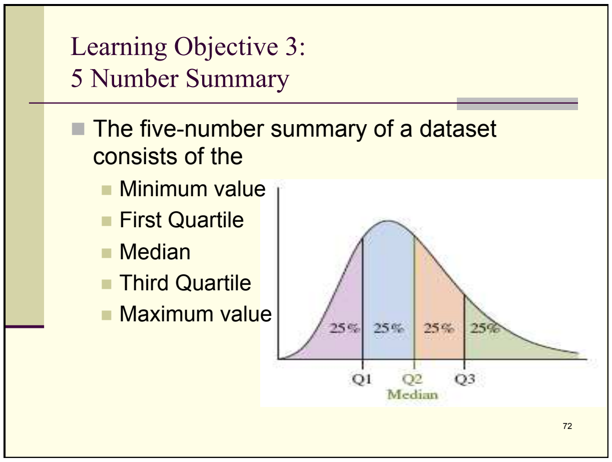
70



Learning Objective 2:
Criteria for identifying an outlier

- **An observation is a potential outlier if it falls more than *1.5 x IQR* below the first quartile or more than *1.5 x IQR* above the third quartile**

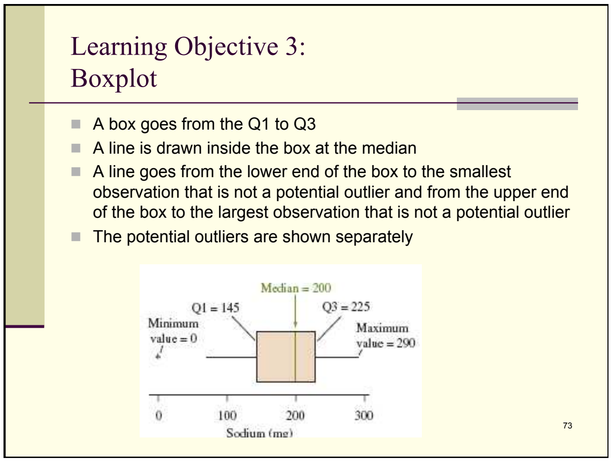
71



Learning Objective 3:
5 Number Summary

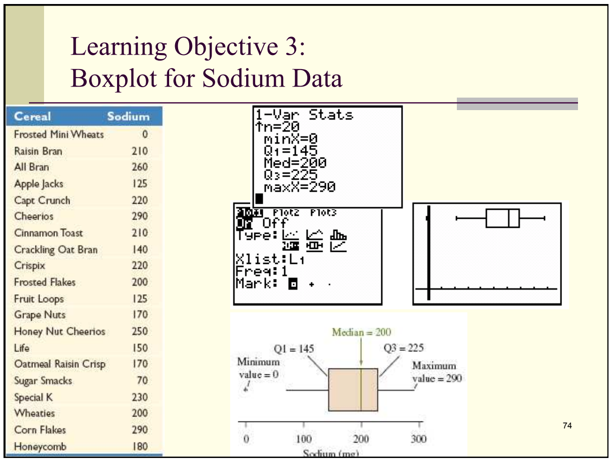
- The five-number summary of a dataset consists of the
  - Minimum value
  - First Quartile
  - Median
  - Third Quartile
  - Maximum value

72



Learning Objective 3: Boxplot

- A box goes from the Q1 to Q3
- A line is drawn inside the box at the median
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier and from the upper end of the box to the largest observation that is not a potential outlier
- The potential outliers are shown separately

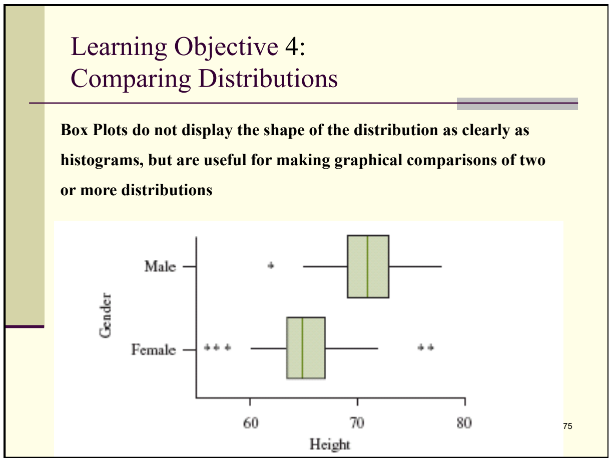


Learning Objective 3: Boxplot for Sodium Data

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	210
All Bran	260
Apple Jacks	125
Capt Crunch	220
Cheerios	290
Cinnamon Toast	210
Crackling Oat Bran	140
Crispix	220
Frosted Flakes	200
Fruit Loops	125
Grape Nuts	170
Honey Nut Cheerios	250
Life	150
Oatmeal Raisin Crisp	170
Sugar Smacks	70
Special K	230
Wheaties	200
Corn Flakes	290
Honeycomb	180

```
1-Var Stats
↑n=20
 minX=0
 Q1=145
 Med=200
 Q3=225
 maxX=290
```

```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: ▫ + ·
```



Learning Objective 4: Comparing Distributions

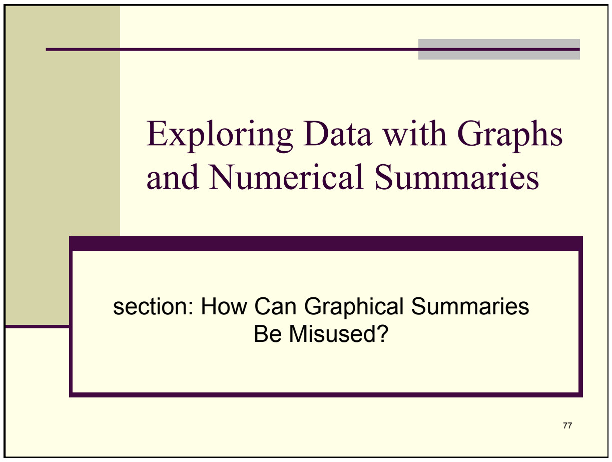
Box Plots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions

Learning Objective 5: Z-Score

- **The *z-score* for an observation is the *number of standard deviations* away from the mean:**

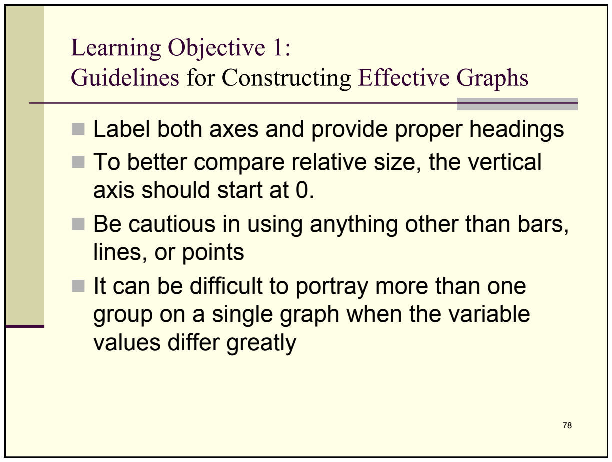
$$z = \frac{\text{observation - mean}}{\text{standard deviation}}$$

- **An observation from a bell-shaped distribution is a potential outlier if its z-score < -3 or > +3**



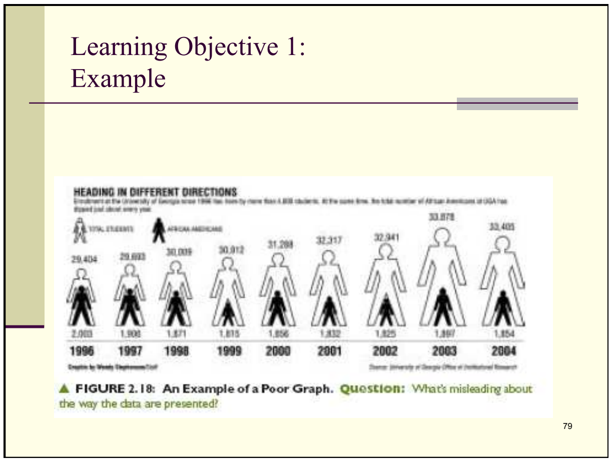
Exploring Data with Graphs and Numerical Summaries

section: How Can Graphical Summaries Be Misused?



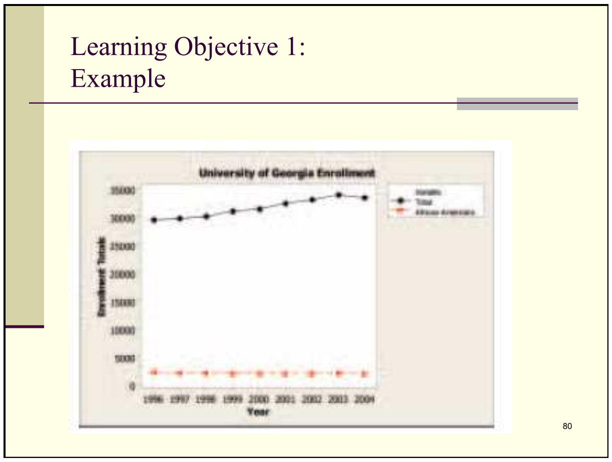
Learning Objective 1: Guidelines for Constructing Effective Graphs

- Label both axes and provide proper headings
- To better compare relative size, the vertical axis should start at 0.
- Be cautious in using anything other than bars, lines, or points
- It can be difficult to portray more than one group on a single graph when the variable values differ greatly



Learning Objective 1: Example

▲ FIGURE 2.18: An Example of a Poor Graph. Question: What's misleading about the way the data are presented?



Learning Objective 1: Example