Elementary Statistics Lecture 2
Exploring Data with Graphical and Numerical Summaries

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

Different Types of Data

Definition (Variable)

A variable is a characteristic(value) that can change from individual to individual.

Definition (Distribution)

The distribution of a variable describes how the observation fall (are distributed) across the range of possible values.

Two types

- **Categorical**: each observation belongs to a set of distinct categories. e.g., gender, religion affiliation, grades etc.
- **Quantitative**: numerical values that represent different magnitude of the variable. e.g., family income, weights, GPA, number of pets you keep etc.
  1. **Discrete**: the possible values form a set of separate numbers such as $0, 1, 2, 3, \ldots$
  2. **Continuous**: the possible values form an interval

Graphical Summary

- **Categorical**
  - Pie chart
  - Bar graph
- **Quantitative**
  - Dot plots
  - Stem-and-left plots
  - Histograms
  - Time series plots
  - box-plots

Numerical Summary

- **Categorial**: frequency table (contingency table). More interested in the frequency(percentage) for each category when considering one categorical variable. Interested in if there is an association between two categorical variable.

- **Quantitative**: Interested in the center and variability for a quantitative variable when considering merely one variable. As for considering two or more quantitative variables, we'd like to find if there is any linear or quadratic association among them.

Example Alligator

What do alligators eat? For 219 alligators captured in four Florida lakes, researchers classified the primary food choice (in volume) found in the alligator's stomach in one of the categories-fish, invertebrate(snails, insects, crayfish), reptile(turtles, baby alligators), bird or other(amphibian, mammal, plants). Data is available in Pearson statcrunch website.

Tips for making frequency (contingency) table in Statcrunch

- stat $\rightarrow$ tables $\rightarrow$ frequency
- stat $\rightarrow$ tables $\rightarrow$ contingency

Tips for making Pie Chart(Bar Plot) in Statcrunch

- graph $\rightarrow$ Pie Chart $\rightarrow$ with data
- graph $\rightarrow$ Bar Plot $\rightarrow$ with data

Pie Chart

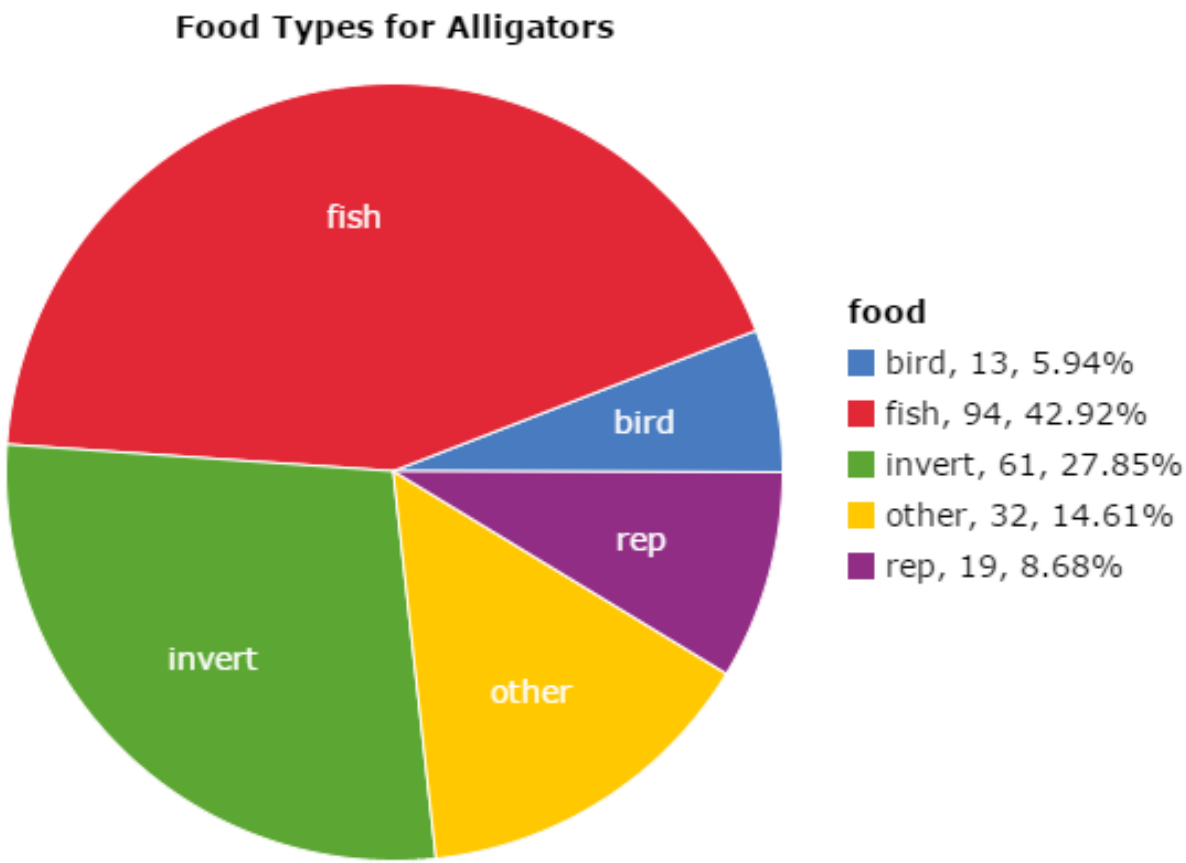


Figure 1: Distribution of food for 219 alligators live in four lakes in Florida, which are George, Hancock, Oklawaha and Trafford, respectively. About 43% of alligators in the sample take fish as the primary food choice.

Bar plot

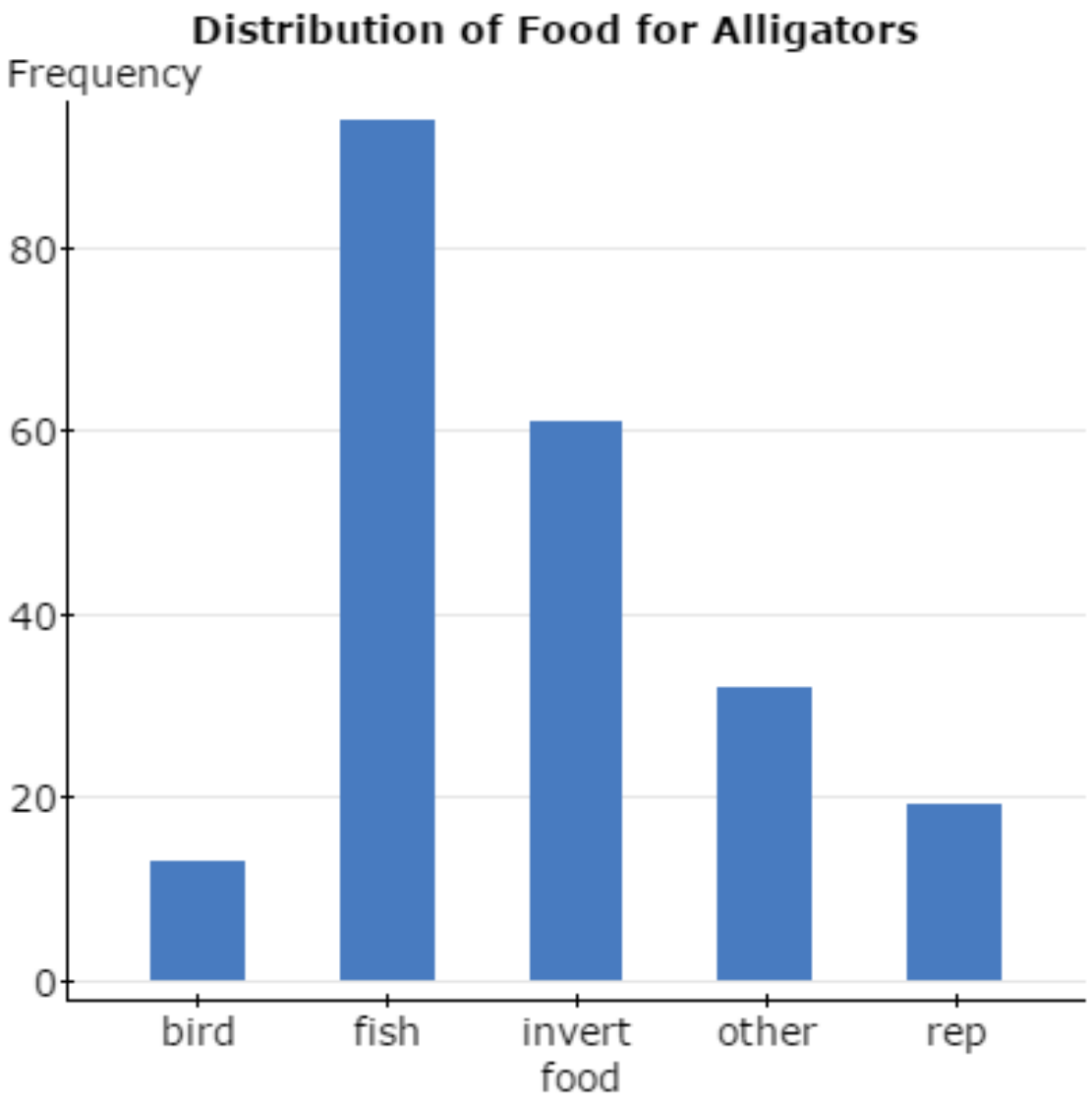


Figure 2: Bar plot for distribution of food for the 219 captured alligators.

Side-By-Side Pie Chart

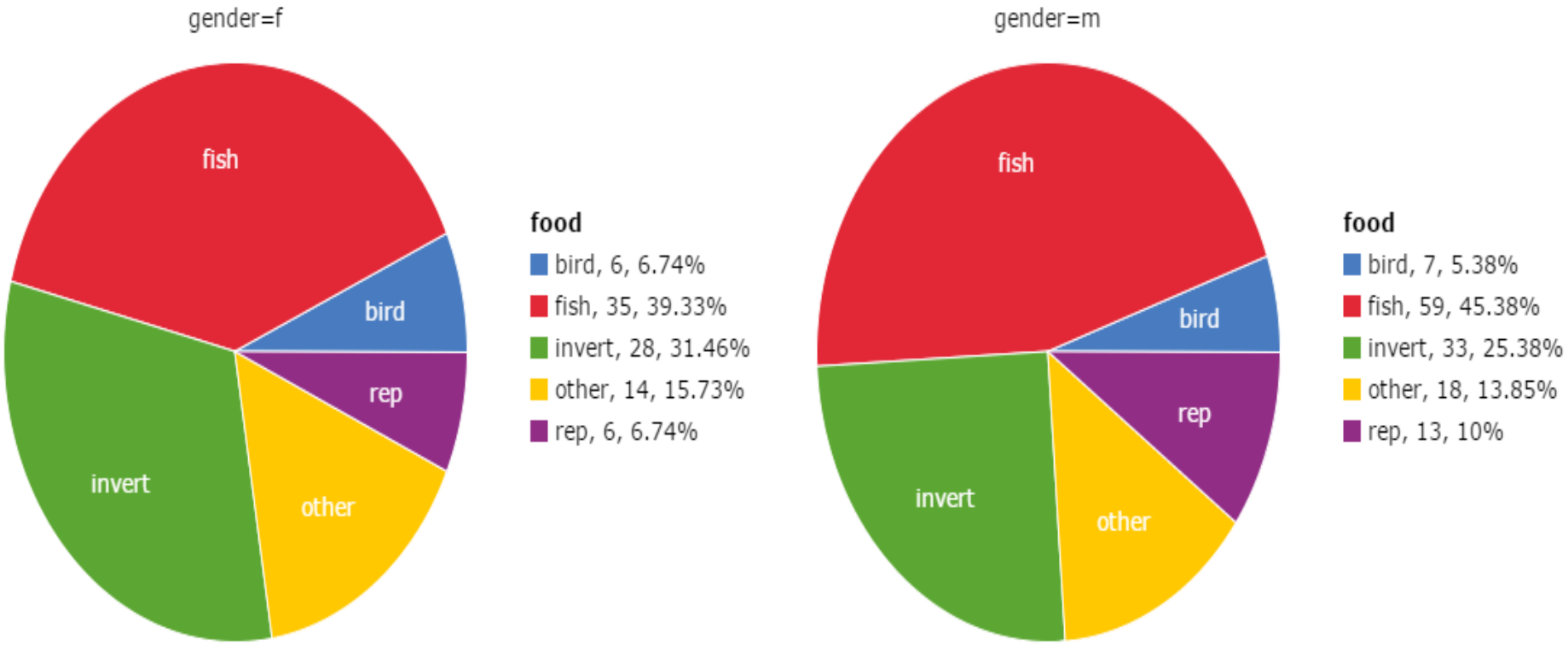


Figure 3: Side-by-side Pie charts for the 219 alligators based on food by size. It shows that about 5% more male alligators have fish as the primary food than female and the alligators who have fish as primary food is more dominant in male population than female.

Frequency(Contingency) Table

Frequency table results for lake:
Count = 219

lake	Frequency	Relative Frequency
george	63	0.28767123
hancock	55	0.25114155
oklawaha	48	0.21917808
trafford	53	0.24200913

Figure 4: 28.8% of the 219 captured alligators in Florida live in George lake.

Contingency table results:
Rows: lake
Columns: size

	<2.3	>2.3	Total
george	41	22	63
hancock	39	16	55
oklawaha	20	28	48
trafford	24	29	53
Total	124	95	219

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	3	13.550804	0.0036

Figure 5: It indicates that there are more smaller alligators in George and Hancock lakes and the opposite to Oklawaha and Trafford lake.

Example Cereal

Cereal	Sodium	Sugar	Type
Frosted Mini Wheats	0	11	A
Raisin Bran	340	18	A
All Bran	70	5	A
Apple Jacks	140	14	C
Cap'n Crunch	200	12	C
Cheerios	180	1	C
Cinnamon Toast Crunch	210	10	C
Crackling Oat Bran	150	16	A
Fiber One	100	0	A
Frosted Flakes	130	12	C
Froot Loops	140	14	C
Honey Bunches of Oats	180	7	A
Honey Nut Cheerios	190	9	C
Life	160	6	C
Rice Krispies	290	3	C
Honey Smacks	50	15	A
Special K	220	4	A
Wheaties	180	4	A
Corn Flakes	200	3	A
Honeycomb	210	11	C

Figure 6: 20 popular cereals and the amounts of sodium and sugar contained in a single serving.

Dot Plot

Tips for making Dot Plot in Statcrunch

- graph $\rightarrow$ Dot Plot $\rightarrow$ select the variable of interest

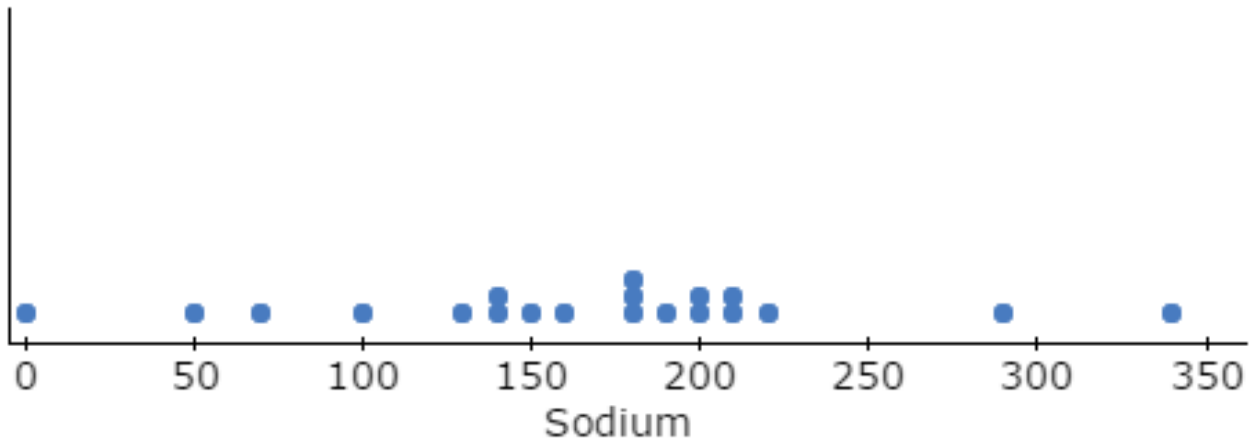
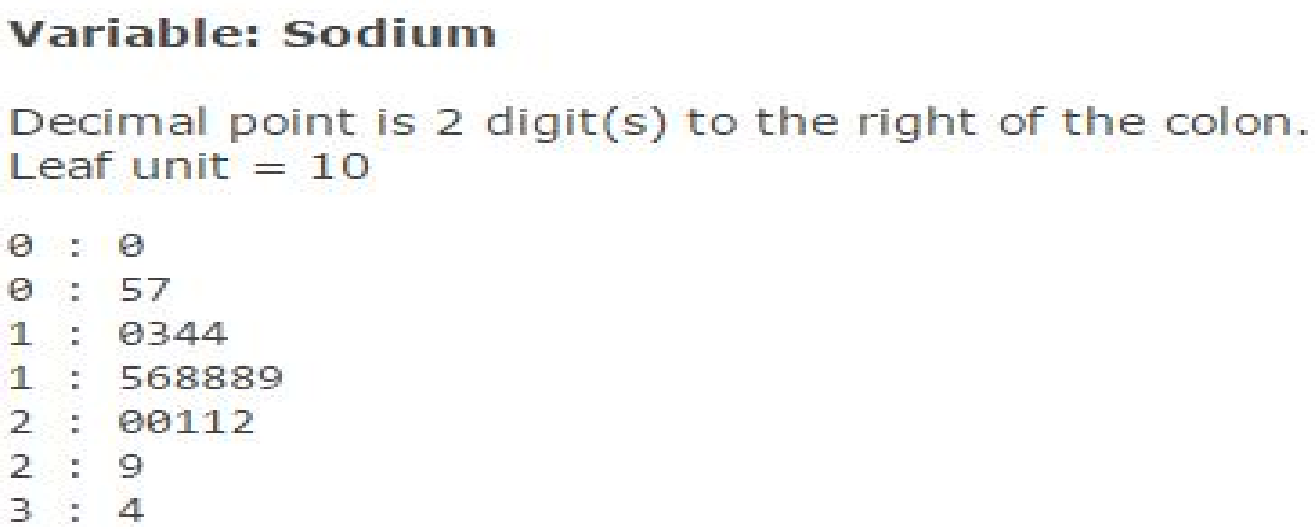


Figure 7: Distribution of sodium in a single serving for 20 popular cereals. Most of the 20 cereals has amount of sodium between 50 and 230 mg in a single serving.

Stem-and-Leaf Plot

Tips for making Dot Plot in Statcrunch

- graph $\rightarrow$ Stem and Leaf $\rightarrow$ select the variable of interest



Variable: Sodium

Decimal point is 2 digit(s) to the right of the colon.
Leaf unit = 10

```
0 : 0
0 : 57
1 : 0344
1 : 568889
2 : 00112
2 : 9
3 : 4
```

Figure 8: The stem-and-leaf plot provides more information by stacking close values together for us to understand the distribution of the variable(sodium). However, both stem-and-leaf and dot plot are appropriate for a small data.

Histogram

Histogram

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

The shape of a histogram(distribution)

- **modal**: unimodal, bimodal or multi-modal
- **skewness**: symmetric, left-skewed or right-skewed

How to decide the skewness

- **left-skewed**: the left tail is longer than the right tail, e.g. life span
- **right-skewed**: the right tail is longer than the left tail, e.g. income

Example STAT 201 Survey

Tips for making Histograms in Statcrunch

- graph $\rightarrow$ Histogram $\rightarrow$ select the variable of interest

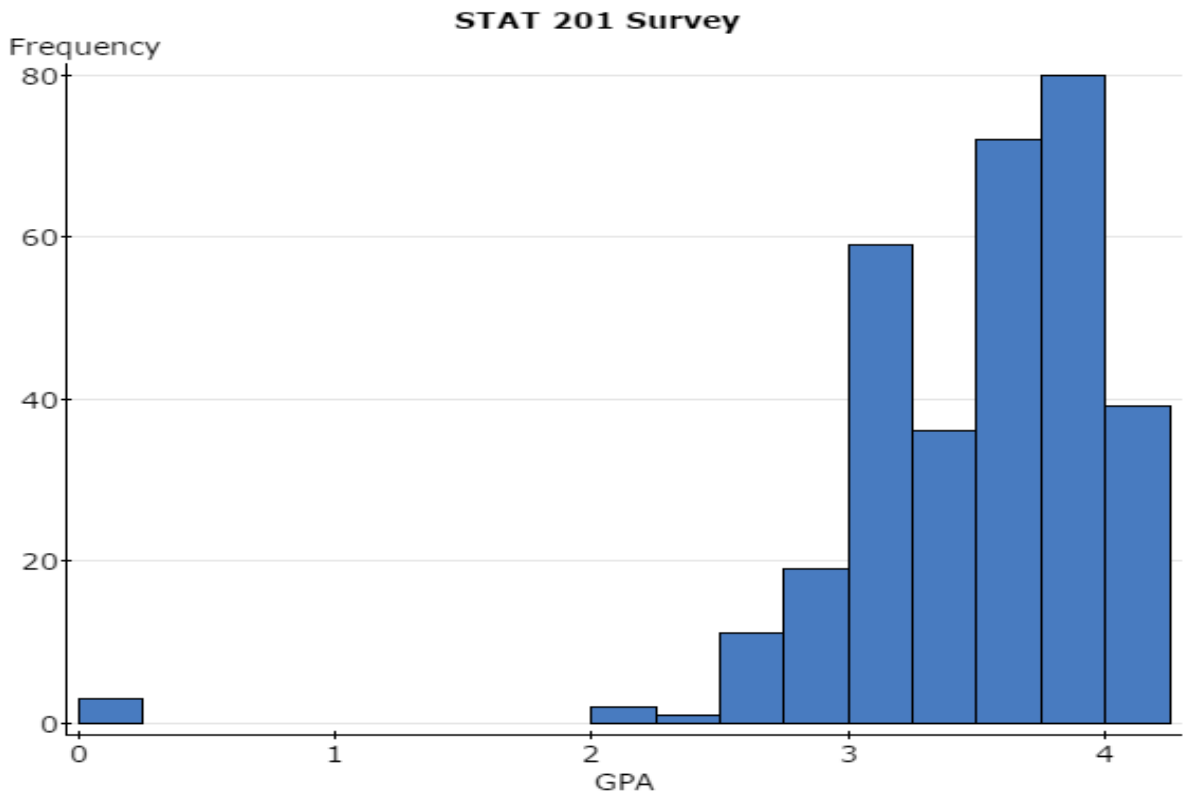


Figure 9: Distribution of STAT 201 students' GPA. It shows us that the histogram is unimodal and left skewed.

Continue . . .

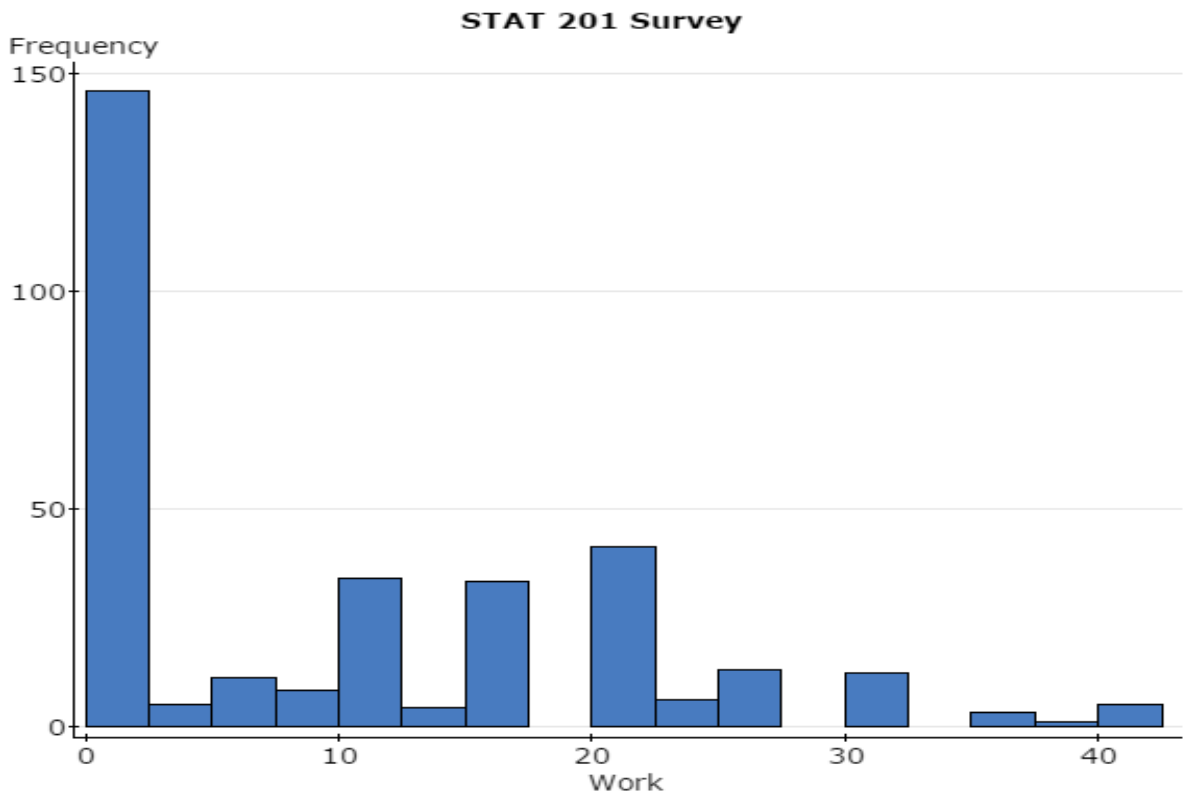


Figure 10: Distribution of the number of hours per week for working during the semester in STAT 201 students. It indicates that the histogram is bimodal where one mode is at 0 and the other is at around 20.

Time Series Plot

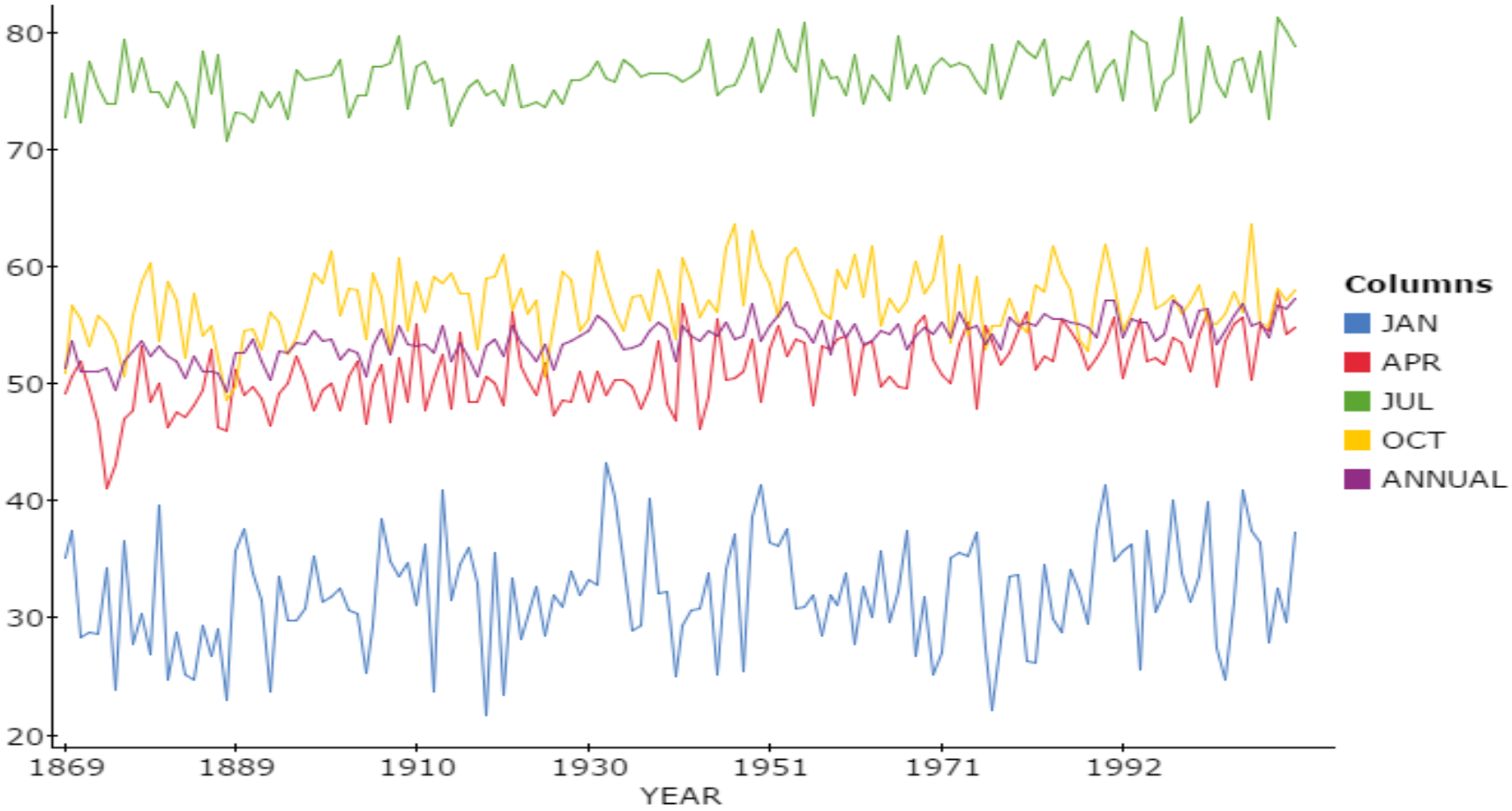


Figure 11: Time series plot of month-average temperature in January, April, July, October and annual average temperature in New York Central Park in 1869-2012.

Center and Variability

Center: The mean and median

- **mean**: $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$
- **median**: $\tilde{x}$ is the middle value of the observations when observations are ordered from the smallest to the largest. e.g.,
  $x_1, x_2, \ldots, x_{n-1}, x_n \overset{ordered}{\longrightarrow} x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)}$

Variability: range, interquantile(IQR), and standard deviation(s)

- **range**: difference between the largest and the smallest, i.e. $x_{(n)} - x_{(1)}$
- **IQR**: the distance between the third and first quartiles, i.e. $Q_3 - Q_1$
- **s**: the square root of the variance s^2, which is an average of the squares of the deviations from their mean, i.e.

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

Mean Vs. Median

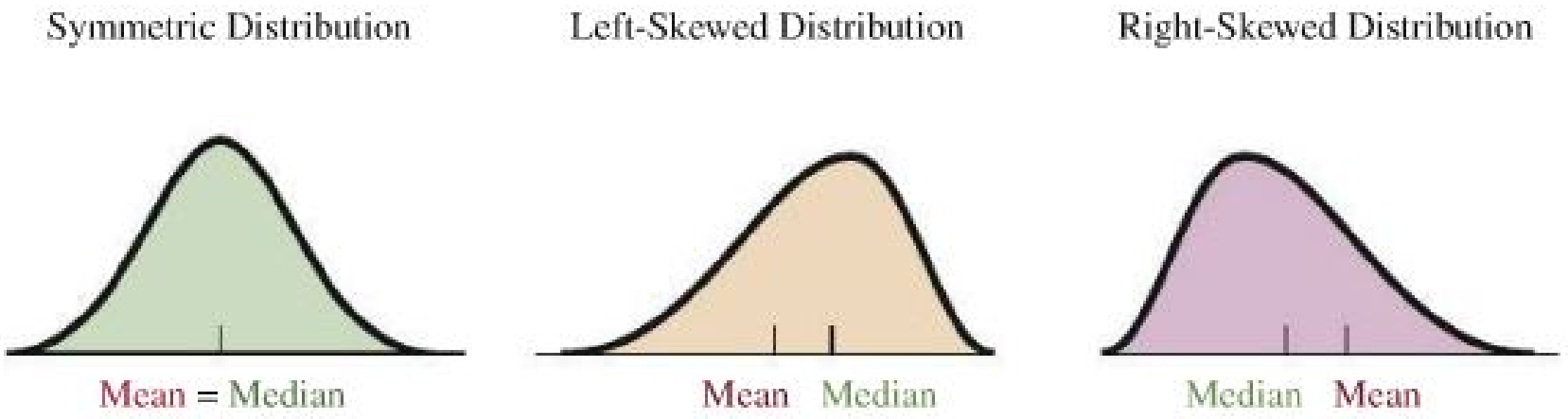


Figure 12: Relationship between the mean and median. The median is usually preferred over the mean if the distribution is highly skewed because it better describes what is typical; conversely, the mean is usually preferred. It is a good idea to report both of the mean and median when describing the center of a distribution.

Continue . . .

Resistant

A numerical summary of the observations is called resistant if extreme observations have little, if any, influence on its value.

Outlier

An outlier is an observation that falls well above or well below the overall bulk of the data.

Remark: The median is more resistant than the mean, because the median is solely determined by having an equal number of observations above it and below it.

Example Cereal continue

In the example cereal in slide 10, the sodium in a single serving for 20 brands are

$$0, 340, 70, 140, 200, 180, 210, 150, 100, 130$$
$$140, 180, 190, 160, 290, 50, 220, 180, 200, 210$$

Calculate the mean and median for the variable of sodium.

- **mean**: $\bar{x} = \frac{0+340+\cdots+200+210}{20} = 167$
- **median**: $\tilde{x} = 180$

  Sort the values of sodium ascending

  $$0, 50, 70, 100, 130, 140, 140, 150, 160, 180$$
  $$180, 180, 190, 200, 200, 210, 210, 220, 290, 340$$

Example Cereal continue

In the example cereal in slide 10, the sodium in a single serving for 20 brands are

$$0, 340, 70, 140, 200, 180, 210, 150, 100, 130$$
$$140, 180, 190, 160, 290, 50, 220, 180, 200, 210$$

Calculate the mean and median for the variable of sodium.

- **mean**: $\bar{x} = \frac{0+340+\cdots+200+210}{20} = 167$
- **median**: $\tilde{x} = 180$

  Sort the values of sodium ascending

  $$0, 50, 70, 100, 130, 140, 140, 150, 160, 180$$
  $$180, 180, 190, 200, 200, 210, 210, 220, 290, 340$$

Example Cereal continue

In the example cereal in slide 10, the sodium in a single serving for 20 brands are

$$0, 340, 70, 140, 200, 180, 210, 150, 100, 130$$
$$140, 180, 190, 160, 290, 50, 220, 180, 200, 210$$

Calculate the mean and median for the variable of sodium.

- **mean**: $\bar{x} = \frac{0+340+\cdots+200+210}{20} = 167$
- **median**: $\tilde{x} = 180$
  Sort the values of sodium ascending

$$0, 50, 70, 100, 130, 140, 140, 150, 160, 180$$
$$180, 180, 190, 200, 200, 210, 210, 220, 290, 340$$

CO_2 emissions 1

Global warming is largely a result of human activity that produces carbon dioxide(CO_2) emissions and other greenhouse gases. The CO_2 emissions from fossil fuel combustion are the result of electricity, heating, industrial processes, and gas consumption in automobiles. The International Energy Agency reported the per capital CO_2 emissions by country for 2011.

Bangladish 0.4	Brazil 2.1	China 5.9
India 1.4	Indonesia 1.8	Nigeria 0.3
Pakistan 0.8	Russia 11.6	United States 16.9

Table 1: For the nine largest countries in population size, the values were in metric tons per person.

Question What's the mean and median? Any outliers?

$$\text{mean } \bar{x} = 4.58$$

$$\text{median } \tilde{x} = 1.8$$

CO_2 emissions 1

Global warming is largely a result of human activity that produces carbon dioxide(CO_2) emissions and other greenhouse gases. The CO_2 emissions from fossil fuel combustion are the result of electricity, heating, industrial processes, and gas consumption in automobiles. The International Energy Agency reported the per capital CO_2 emissions by country for 2011.

Bangladish 0.4	Brazil 2.1	China 5.9
India 1.4	Indonesia 1.8	Nigeria 0.3
Pakistan 0.8	Russia 11.6	United States 16.9

Table 1: For the nine largest countries in population size, the values were in metric tons per person.

Question What's the mean and median? Any outliers?

$$\text{mean } \bar{x} = 4.58$$

$$\text{median } \tilde{x} = 1.8$$

CO_2 emissions 2

The International Energy Agency also reported the CO_2 emissions (measured in gigatons,Gt) from fossil fuel combustion for the top 9 countries in 2011.

Canada 0.5	China 8	India 1.8
Iran 0.4	Germany 0.8	Japan 1.2
Korea 0.6	Russia 1.7	United States 5.3

Table 2: CO_2 emissions for the top 10 countries in 2011 and the values are in Gt.

Question What's the mean and median? Any outliers?

$$\text{mean } \bar{x} = 2.26$$

$$\text{median } \tilde{x} = 1.2$$

CO_2 emissions 2

The International Energy Agency also reported the CO_2 emissions (measured in gigatons,Gt) from fossil fuel combustion for the top 9 countries in 2011.

Canada 0.5	China 8	India 1.8
Iran 0.4	Germany 0.8	Japan 1.2
Korea 0.6	Russia 1.7	United States 5.3

Table 2: CO_2 emissions for the top 10 countries in 2011 and the values are in Gt.

Question What's the mean and median? Any outliers?

$$\text{mean } \bar{x} = 2.26$$

$$\text{median } \tilde{x} = 1.2$$

Variability-Standard Deviation

Name	x	$x - \bar{x}$	$(x - \bar{x})^2$
Canada	0.5	-1.76	3.08
China	8.0	5.74	32.99
India	1.8	0.46	0.21
Iran	0.4	-1.86	3.44
Germany	0.8	-1.46	2.12
Japan	1.2	-1.06	1.11
Korea	0.6	-1.66	2.74
Russia	1.7	-0.56	0.31
U.S.	5.3	-3.04	9.27
Total	20.3	0	55.28

Table 3: The mean is 2.26.The standard deviation $s = 2.63$.

$$s = \sqrt{\frac{\sum_{i=1}^{9}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{55.28}{9-1}} = 2.63$$

Five-number-summary

Definition (Percentile)

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

The five-number-summary consists of minimum($x_{(1)}$), Q1, Q2($\tilde{x}$), Q3, and maximum($x_{(n)}$).

How to find quartiles:

- Arrange the data in order.
- Consider the median first, which is the second quartile **Q2**.
- Consider the lower half of the observations (excluding the median itself if n is odd). The median of these observations is the first quartile **Q1**.
- Consider the upper half of the observations (excluding the median itself if n is odd). The median of these observations is the third quartile **Q3**.

Box Plot

Constructing a Box Plot

- Draw a box going from Q1 to Q3.
- Draw the median line inside the box.
- Draw a line from the lower end of the box to the smallest observation that is not a potential outlier. Draw a separate line from the upper end of the box to the largest observation that is not a potential outlier.
- Draw the potential outliers with special symbols(e.g. a dot or a star).

Remark: An observation is called a potential outlier if it is more than 1.5IQR below the first quartile(Q1) or above the thrid quartile(Q3)

CO_2 Box Plot

The sorted CO_2 emission in total for the top 9 countries are

$$0.4, 0.5, 0.6, 0.8, 1.2, 1.7, 1.8, 5.3, 8.0$$

The five-number-summary is

```
> summary(x2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.400   0.600   1.200   2.256   1.800   8.000
```

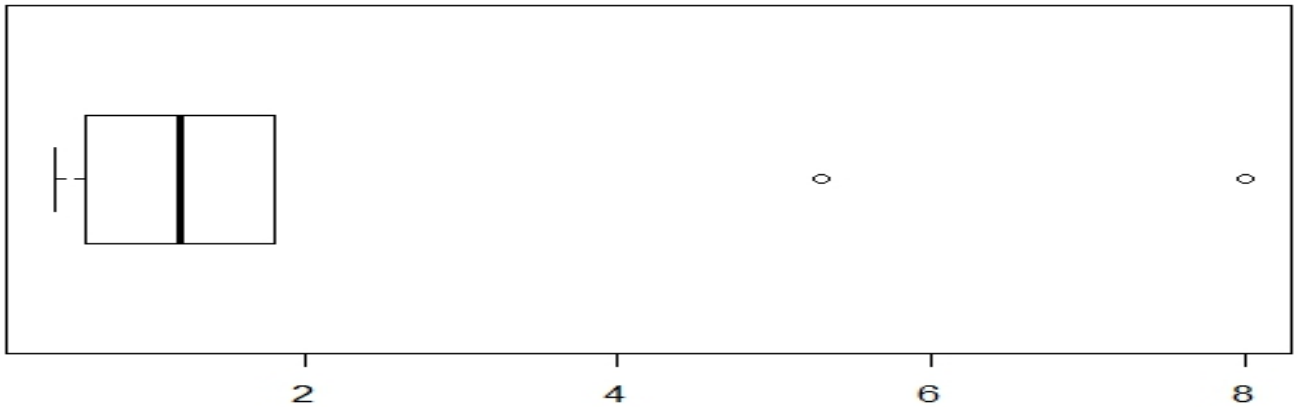


Figure 13: The boxplot of the CO_2 in total for the top 9 countries. It indicates there 2 countries are potential outliers.

CO_2 Box Plot

Example

Use the data of CO_2 per capital for the 9 countries, calculate the mean, five-number-summary and draw a boxplot. The data are

`0.4  2.1  5.9  1.4  1.8  0.3  0.8 11.6 16.9`

Empirical Rule

Empirical Rule

If a distribution of data is bell shaped, the approximately

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

Remark: If an observation falls beyond 3 standard deviations of the mean, we say it a potental outlier.

Empirical Rule

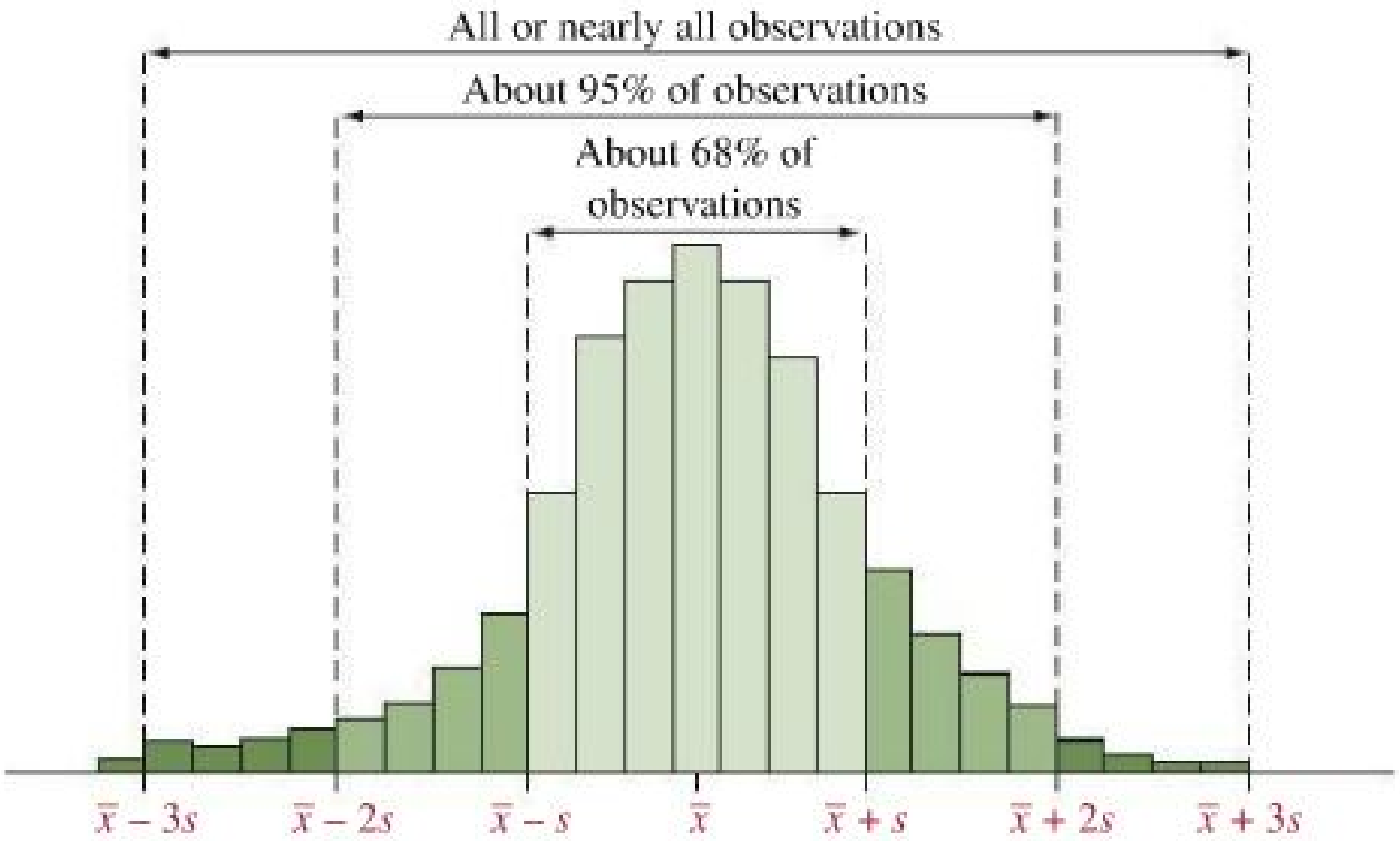


Figure 14: **The Empirical Rule**. For bell-shaped distributions, this tells us approximately how much of the data fall within 1, 2, and 3 standard deviations of the mean.

Z-Score

Z-Score for an observation

The z-score for an observation is the number of standard deviations that it falls from the mean. A positive z-score indicates that the observation is above the mean and a negative z-score indicates that the observation is below the mean. The formula is as

$$\begin{aligned} \text{z-score} &= \frac{\text{observation - mean}}{\text{standard deviation}} \\ &= \frac{x_{obs} - \bar{x}}{s} \end{aligned}$$

Detecting Potential Outliers

- **1.5∗IQR**: more than 1.5IQR below the first quartile(Q1) or above the thrid quartile(Q3).
- **Empirical Rule**: z-score to fall more than 3 standard deviations from the mean where

$$\text{z-score} = \frac{observation - mean}{standarddeviation}$$

i.e., $|\text{Z-score} > 3| \Leftrightarrow$ potential outlier

Exercise 1

College Students Heights For the 262 female and 117 male college student heights, the average height for female is 65.4 inches and the standard deviation is 3.3 inches; the average height for male is 70.9 inches and the standard deviation 2.9 inches. The tallest female in this sample is 91 inches and the shortest male in this sample is 62 inches.

1. Calculate the z-scores for the hight of 91 inches in the female group and the hight of 62 in the male group.
2. Are they potential outliers in their corresponding group?

Exercise 2

Male heights According to a recent report from the U.S. National Center for Health Statistics, for males ages 25-34 years, 2% of their heights are 64 inches or less, 27% are 68 inches or less, 54% are 70 inches or less, 80% are 72 inches or less, 93% are 74 inches or less and 98% are 76 inches or less. These are called cumulative percentages.

1. Which category has the median height?
2. Nearly all the heights fall between 60 and 80 inches, with fewer than 1% falling outside that range. If the heights are approximately bell-shaped, give a rough approximation for the standard deviation of the heights.