

CHAPTER 2 EXPLORING DATA WITH GRAPHS AND NUMERICAL SUMMARIES

2.1 Different Types of Data

- Recall from Chapter 1 that a **variable** is any characteristic observed from the subjects in a study.

Variable



Yes / No

Statistic

Sample



% of Yes, $\hat{p} = \frac{3}{5} = 0.6$

Parameter

Population

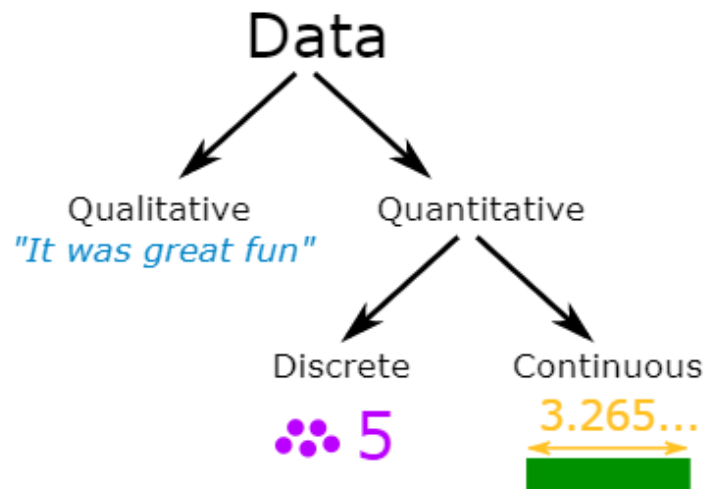


% of Yes, p

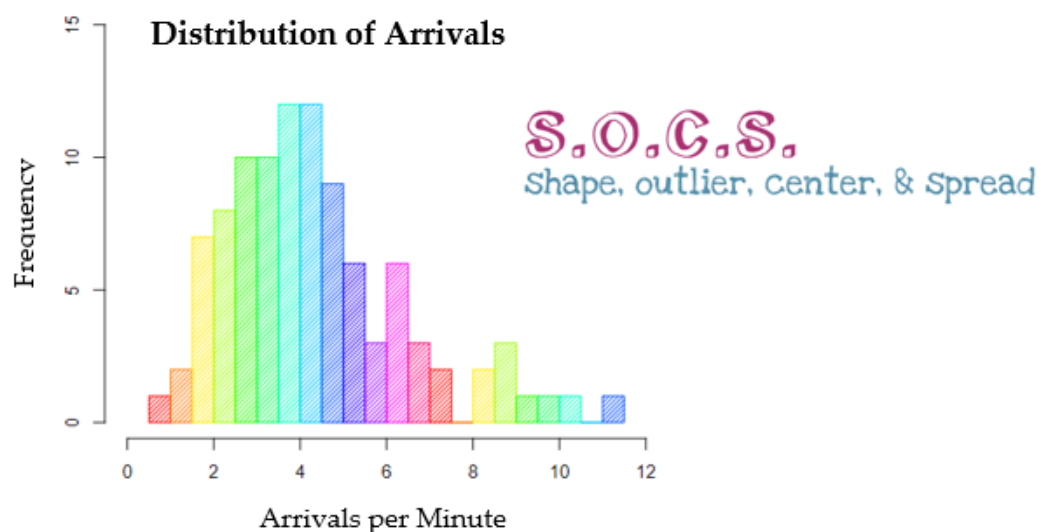
- The data values that we observe for a variable are called **observations**.
- A variable is called **categorical** if each observation belongs to one of a set of distinct categories.
- Example** Give examples of categorical variables.

- A variable is called **quantitative** if observations on it take numerical values that represent different magnitudes of the variable.
 - Example** Give examples of quantitative variables.
-
-
-
-
-
-
-
-
-
-
- We can further classify a quantitative variable as either *discrete* or *continuous*.
 - A quantitative variable is **discrete** if its possible values form a set of separate numbers, such as $0, 1, 2, 3, \dots$
 - A quantitative variable is **continuous** if its possible values form an interval.
 - Example** Classify the quantitative variables given in the previous example as discrete or continuous.

- It is important to know the type of a variable as the method used to analyze a data set will depend on it.

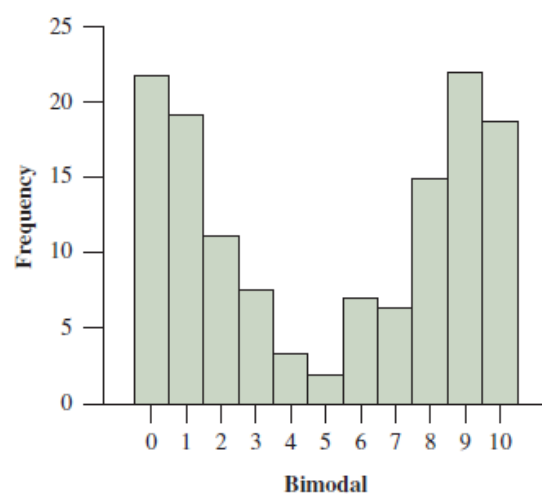
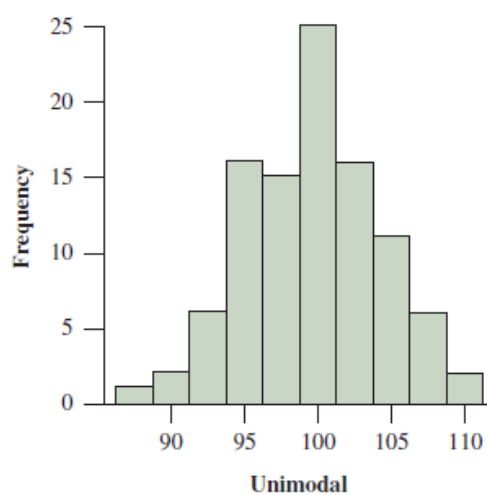
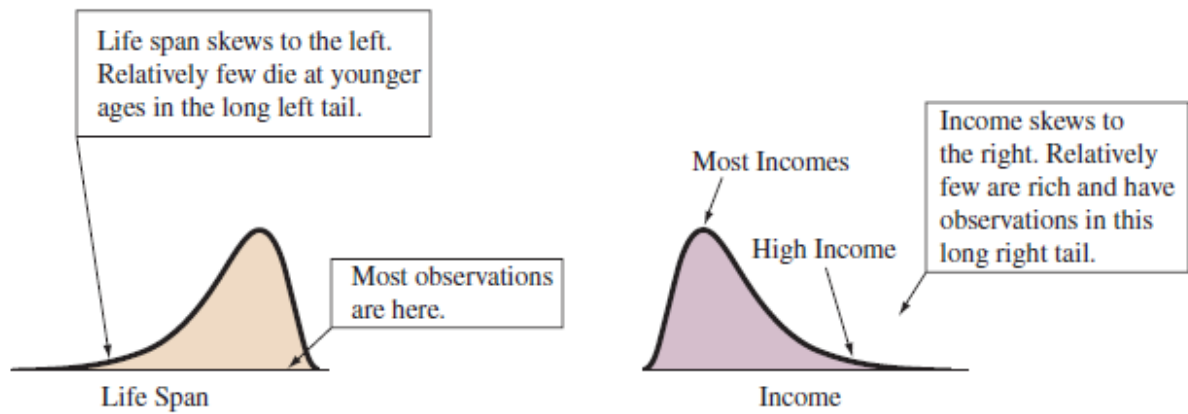
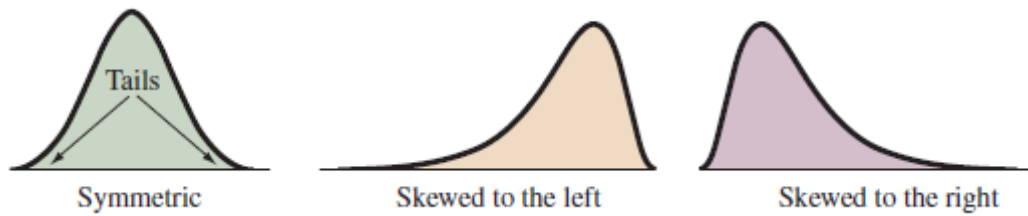


- The first step in analyzing data collected on a variable is to look at the observed values by using graphs and numerical summaries. The goal is to describe key features of the *distribution* of a variable.
- The **distribution** of a variable describes how the observations fall (are distributed) across the range of possible values.

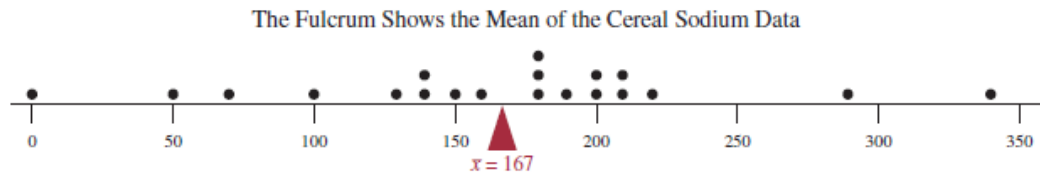


- Features to look for in the distribution of a quantitative variable are:

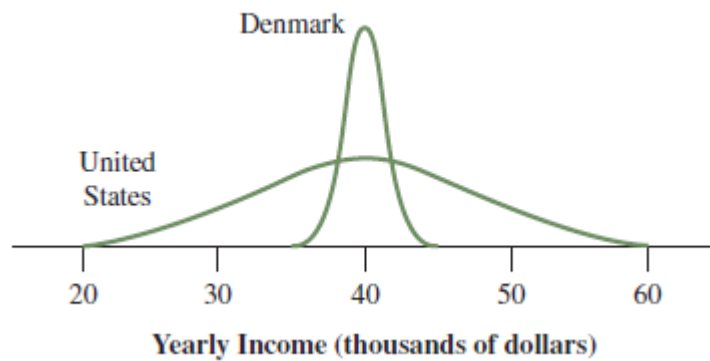
1. **Shape:** Do observations cluster in certain intervals and/or are they spread thin in others?



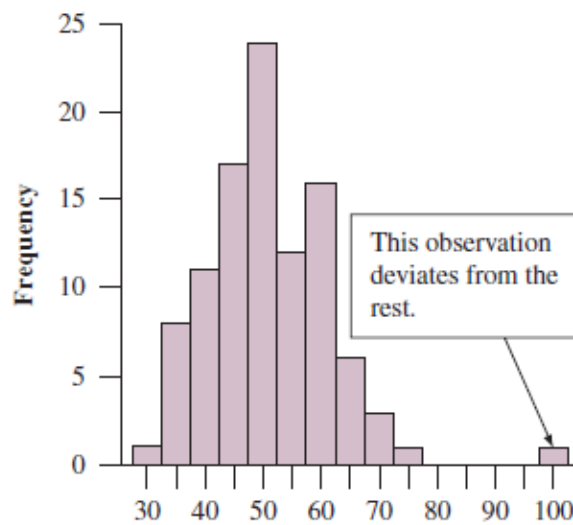
2. **Center:** Where does a typical observation fall?



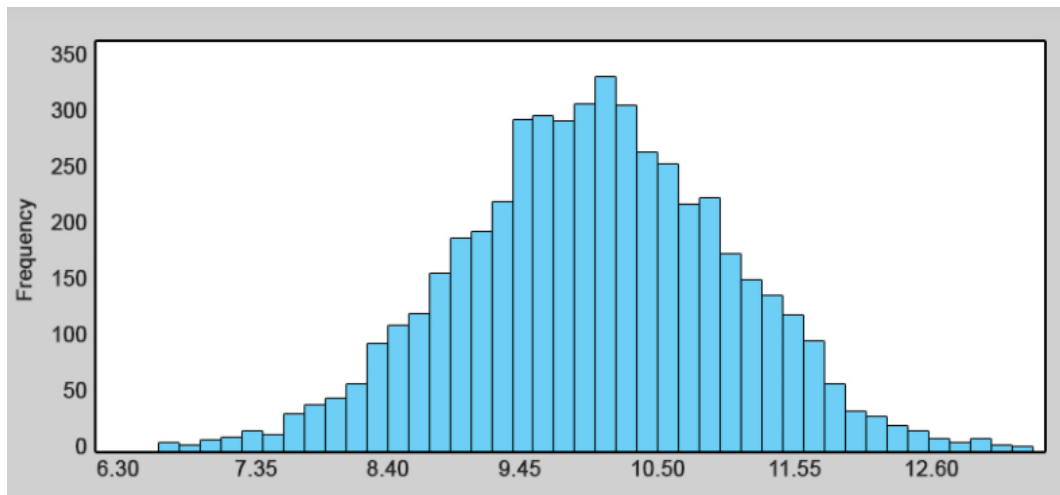
3. **Variability:** How tightly are the observations clustering around a center?



4. **Outlier:** Any observations that are extraordinarily larger or smaller than the rest?

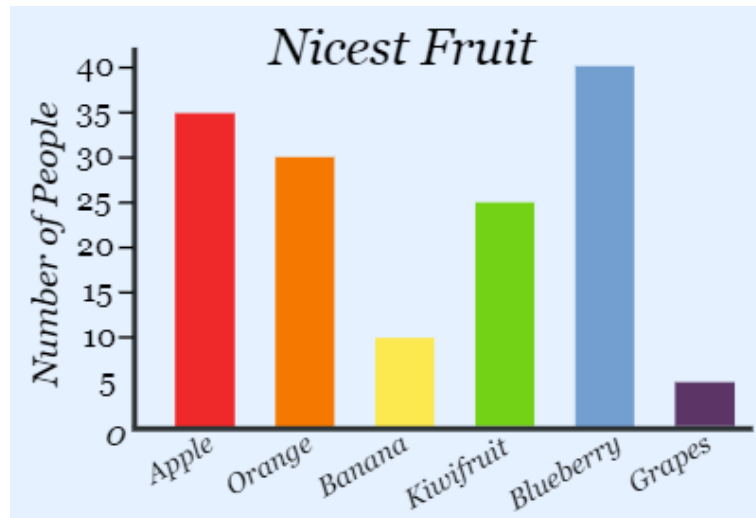


- **Example** Distribution of Amount Spent on Lunch



Describe the distribution of the amount spent on lunch (in \$) by the employees of a company.

- Features to look for in the distribution of a categorical variable are the category with the largest frequency, called the **modal category**, and more generally how frequently each category was observed.
- **Example** Nicest Fruit



- Identify the modal category.
- How many respondents find grapes as the nicest fruit?

- A **frequency table** is a listing of possible values for a variable, together with the number of observations for each value.

| Day | Number of customers | Frequency |
|-----------|---------------------|-----------|
| Monday | | 18 |
| Tuesday | | 13 |
| Wednesday | | 20 |
| Thursday | | 14 |
| Friday | | 21 |
| Saturday | | 27 |
| Sunday | | 26 |

- The **proportion** of observations falling in a certain category is the number of observations in that category divided by the total number of observations.

$$\text{proportion} = \frac{\text{frequency}}{\text{total}}$$

- The **percentage** is the proportion multiplied by 100.

$$\text{percentage} = \text{proportion} \times 100\% = \frac{\text{frequency}}{\text{total}} \times 100\%$$

- Proportions and percentages are also called **relative frequencies** and serve as a way to summarize the distribution numerically.

| X | f | p = f/N | percent = p(100) |
|---|---|------------|------------------|
| 5 | 1 | 1/10 = .10 | 10% |
| 4 | 2 | 2/10 = .20 | 20% |
| 3 | 3 | 3/10 = .30 | 30% |
| 2 | 3 | 3/10 = .30 | 30% |
| 1 | 1 | 1/10 = .10 | 10% |

- **Example Shark Attacks**


The International Shark Attack File (ISAF) collects data on unprovoked shark attacks worldwide. When a shark attack is reported, the region where it took place is recorded.

Table 2.1 Frequency of Shark Attacks in Various Regions for 2004–2013*

| Region | Frequency | Proportion | Percentage |
|----------------|------------|--------------|--------------|
| Florida | 203 | 0.295 | 29.5 |
| Hawaii | 51 | 0.074 | 7.4 |
| South Carolina | 34 | 0.049 | 4.9 |
| California | 33 | 0.048 | 4.8 |
| North Carolina | 23 | 0.033 | 3.3 |
| Australia | 125 | 0.181 | 18.1 |
| South Africa | 43 | 0.062 | 6.2 |
| Réunion Island | 17 | 0.025 | 2.5 |
| Brazil | 16 | 0.023 | 2.3 |
| Bahamas | 6 | 0.009 | 0.9 |
| Other | 138 | 0.200 | 20.0 |
| Total | 689 | 1.000 | 100.0 |

*Source: Data from www.flmnh.ufl.edu/fish/sharks/statistics/statsw.htm. Current as of March 2013.

- What is the variable that was observed? Is it categorical or quantitative?
- How many observations were there? Show how to find the proportion and percentage for Florida.
- Identify the modal category for this variable.
- Describe the distribution of shark attacks.

- We will be using  in this course to analyze the data.
- You can get the software for free from <https://www.r-project.org/>



[\[Home\]](#)

Download

[CRAN](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred [CRAN mirror](#).

- Choose a location close to you.

Japan

<https://cran.ism.ac.jp/>

<https://ftp.yz.yamagata-u.ac.jp/pub/cran/>

The Institute of Statistical Mathematics, Tokyo
Yamagata University

Korea

<https://ftp.harukasan.org/CRAN/>

<https://cran.yu.ac.kr/>

<https://cran.seoul.go.kr/>

<http://healthstat.snu.ac.kr/CRAN/>

<https://cran.biodisk.org/>

Information and Database Systems Laboratory, Pukyong National University
Yeungnam University

Bigdata Campus, Seoul Metropolitan Government

Graduate School of Public Health, Seoul National University, Seoul

The Genome Institute of UNIST (Ulsan National Institute of Science and Technology)

Malaysia

<https://wbc.upm.edu.my/cran/>

Universiti Putra Malaysia

- Download and install R.

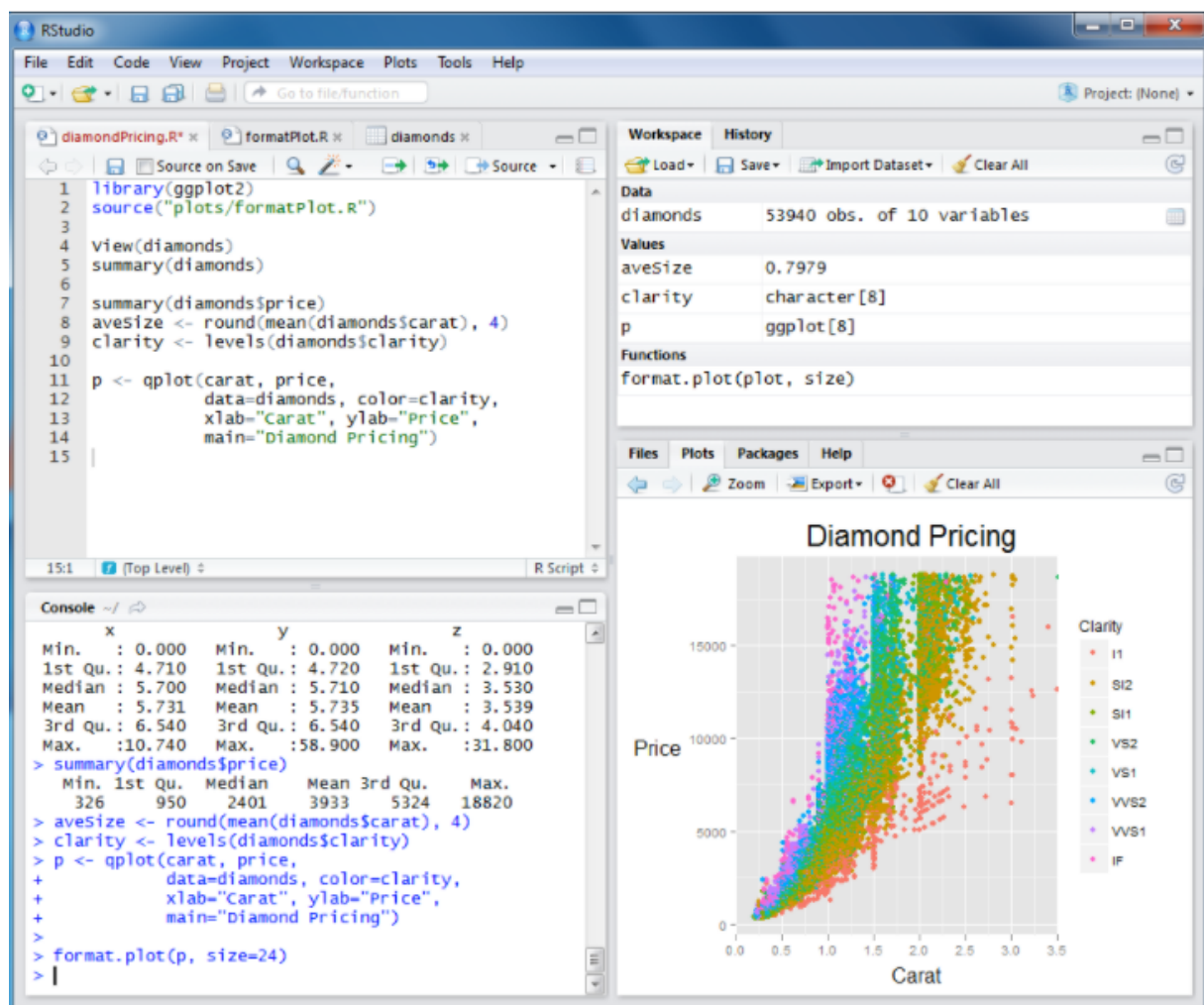
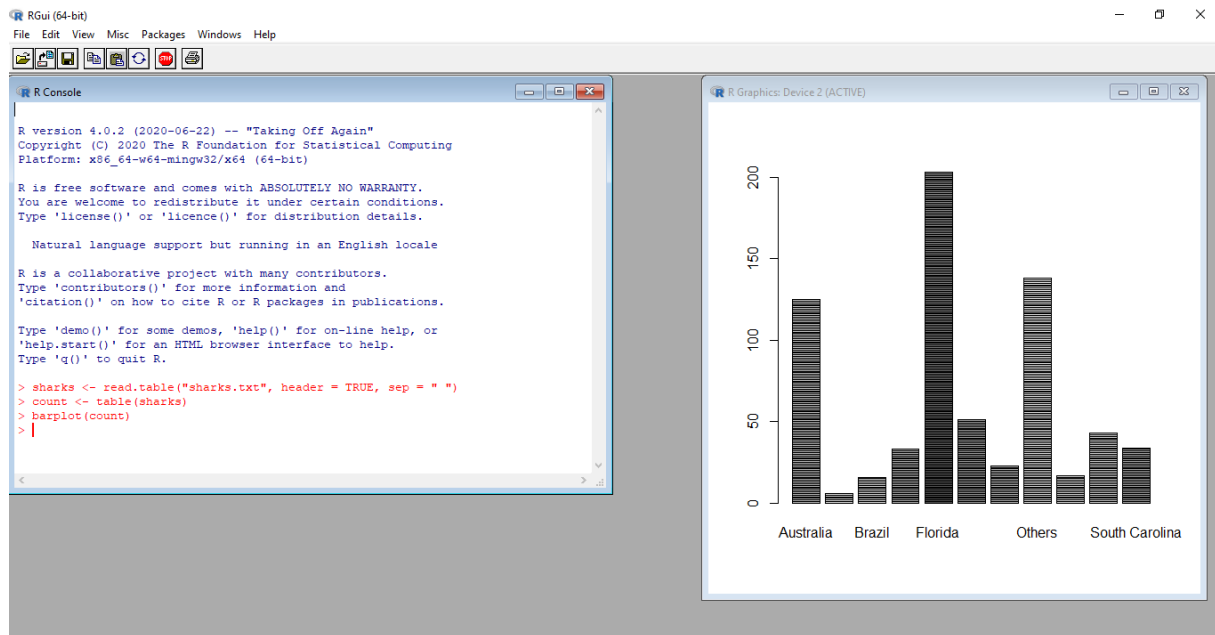
Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

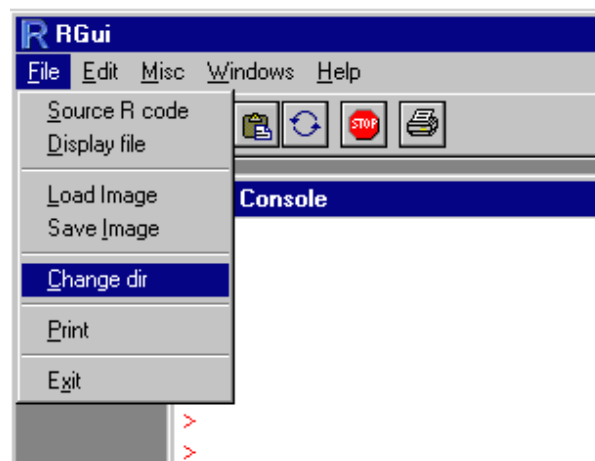
- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

- The R interface.



- You should change your working directory before you start.

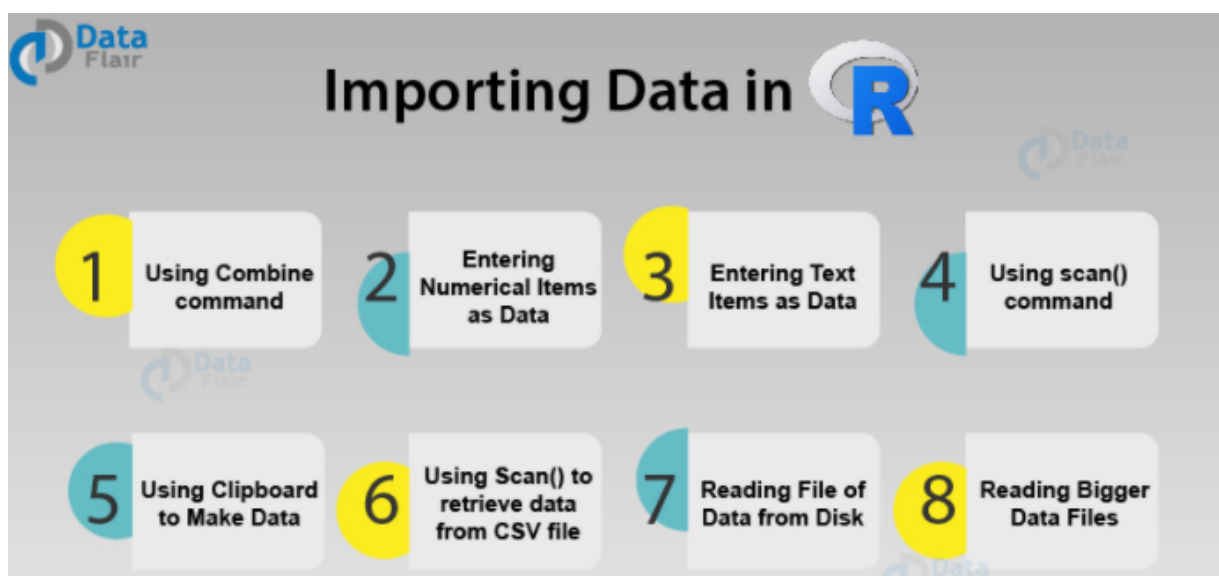


- Some R codes to start off with.

- To enter data:

```
> score <- c(8, 10, 5, 7, 6, 7, 6, 6, 9, 5, 8)
```

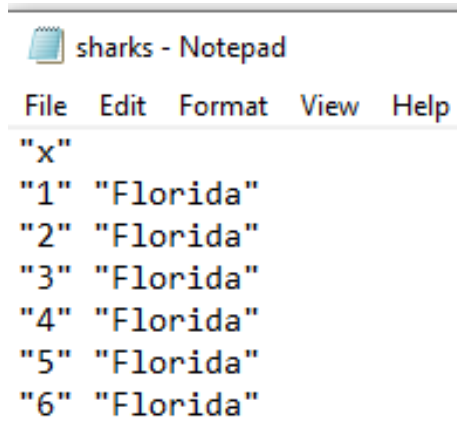
```
> sharks <- c(rep("Florida",203), rep("Hawaii",51),
+ rep("South Carolina",34), rep("California",33),
+ rep("North Carolina",23), rep("Australia",125),
+ rep("South Africa",43), rep("Reunion Island",17),
+ rep("Brazil",16), rep("Bahamas",6), rep("Others",138))
```



- To write a file:

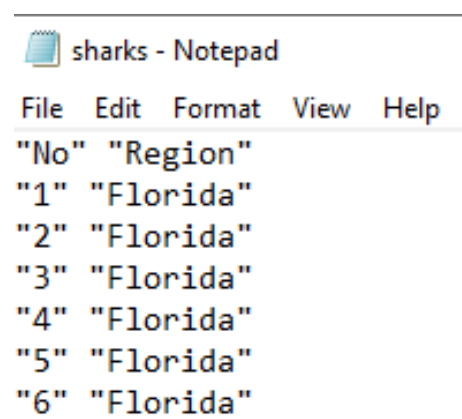
```
> write.table(sharks, "sharks.txt", sep = " ")
```

Need to edit the headers in the .txt file.



sharks - Notepad

| File | Edit | Format | View | Help |
|---------------|------|--------|------|------|
| "x" | | | | |
| "1" "Florida" | | | | |
| "2" "Florida" | | | | |
| "3" "Florida" | | | | |
| "4" "Florida" | | | | |
| "5" "Florida" | | | | |
| "6" "Florida" | | | | |



sharks - Notepad

| File | Edit | Format | View | Help |
|---------------|------|--------|------|------|
| "No" "Region" | | | | |
| "1" "Florida" | | | | |
| "2" "Florida" | | | | |
| "3" "Florida" | | | | |
| "4" "Florida" | | | | |
| "5" "Florida" | | | | |
| "6" "Florida" | | | | |

- To read a file:

```
> sharks <- read.table("sharks.txt", header = TRUE, sep = " ")
```

```
> textbookdata <- read.table("Animals.txt", header = TRUE, sep = "\t")
```

- To view an object:

```
> sharks
```

| | No | Region |
|----|----|---------|
| 1 | 1 | Florida |
| 2 | 2 | Florida |
| 3 | 3 | Florida |
| 4 | 4 | Florida |
| 5 | 5 | Florida |
| 6 | 6 | Florida |
| 7 | 7 | Florida |
| 8 | 8 | Florida |
| 9 | 9 | Florida |
| 10 | 10 | Florida |
| 11 | 11 | Florida |
| 12 | 12 | Florida |

- To create a frequency table:

```
> sharktable <- table(sharks$Region)
> t <- as.data.frame(sharktable)
> names(t)[1] = 'Region'
> t
```

| | Region | Freq |
|----|----------------|------|
| 1 | Australia | 125 |
| 2 | Bahamas | 6 |
| 3 | Brazil | 16 |
| 4 | California | 33 |
| 5 | Florida | 203 |
| 6 | Hawaii | 51 |
| 7 | North Carolina | 23 |
| 8 | Others | 138 |
| 9 | Reunion Island | 17 |
| 10 | South Africa | 43 |
| 11 | South Carolina | 34 |

- To include the proportions and percentages in the frequency table:

```
> Total <- sum(t$Freq)

> Proportion <- t$Freq/Total

> Percentage <- Proportion * 100

> cbind(t, Proportion, Percentage)
```

| | Region | Freq | Proportion | Percentage |
|----|----------------|------|-------------|------------|
| 1 | Australia | 125 | 0.181422351 | 18.1422351 |
| 2 | Bahamas | 6 | 0.008708273 | 0.8708273 |
| 3 | Brazil | 16 | 0.023222061 | 2.3222061 |
| 4 | California | 33 | 0.047895501 | 4.7895501 |
| 5 | Florida | 203 | 0.294629898 | 29.4629898 |
| 6 | Hawaii | 51 | 0.074020319 | 7.4020319 |
| 7 | North Carolina | 23 | 0.033381713 | 3.3381713 |
| 8 | Others | 138 | 0.200290276 | 20.0290276 |
| 9 | Reunion Island | 17 | 0.024673440 | 2.4673440 |
| 10 | South Africa | 43 | 0.062409289 | 6.2409289 |
| 11 | South Carolina | 34 | 0.049346880 | 4.9346880 |

- **Example** Score

**8, 10, 5, 7, 6,
7, 6, 6, 9, 5, 8**

| Score | Tally | Frequency | Cumulative frequency |
|-------|-------|-----------|----------------------|
| 5 | | 2 | 2 |
| 6 | | 3 | 5 |
| 7 | | 2 | 7 |
| 8 | | 2 | 9 |
| 9 | | 1 | 10 |
| 10 | | 1 | 11 |

```
> score <- c(8, 10, 5, 7, 6, 7, 6, 6, 9, 5, 8)
> scoretable <- table(score)
> t <- as.data.frame(scoretable)
> names(t)[1] = 'Score'
> Cumulative <- cumsum(t$Freq)
> cbind(t, Cumulative)
  Score Freq Cumulative
1     5    2          2
2     6    3          5
3     7    2          7
4     8    2          9
5     9    1         10
6    10    1         11
```

- **Example** Waiting Time

**Frequency Table: Waiting Time
Between Two Consecutive Eruptions
of the Old Faithful Geyser**

| Minutes | Frequency | Percentage |
|--------------|------------|--------------|
| < 50 | 21 | 7.7 |
| 50–60 | 56 | 20.6 |
| 60–70 | 26 | 9.6 |
| 70–80 | 77 | 28.3 |
| 80–90 | 80 | 29.4 |
| > 90 | 12 | 4.4 |
| Total | 272 | 100.0 |

Exercises 2.1 2.1, 2.3, 2.5, 2.7, 2.9.

2.2 Graphical Summaries of Data

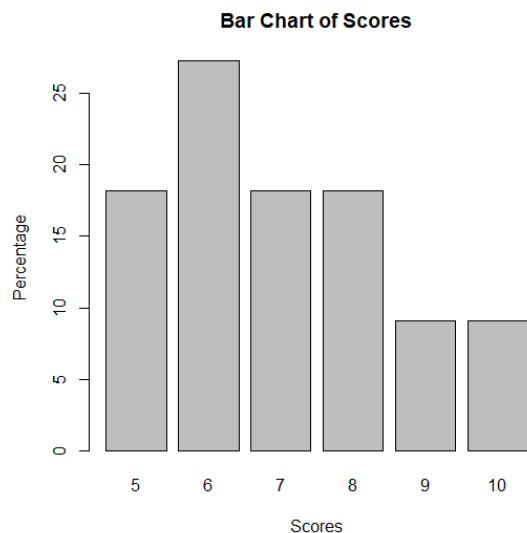
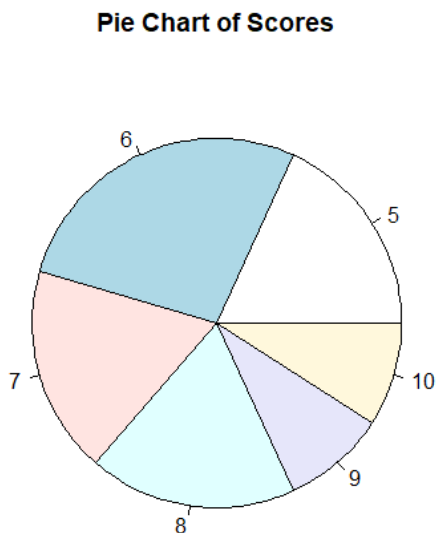
- The two primary graphical displays for summarizing a categorical variable are:

1. **Pie Chart:** A circle having a slice of the pie for each category. The size of a slice corresponds to the percentage of observations in the category.

```
> slice <- t$Freq
> name <- t$Score
> pie(slice, labels = name, main = "Pie Chart of Scores")
```

2. **Bar Graph:** A vertical bar for each category. The height of the bar is the percentage of observations in the category. Typically, the vertical bars for each category are apart, not side by side.

```
> Percentage <- t$Freq / sum(t$Freq) * 100
> barplot(Percentage, names.arg = c("5", "6", "7", "8", "9", "10"),
+ xlab = "Scores", ylab = "Percentage", main = "Bar Chart of Scores")
```



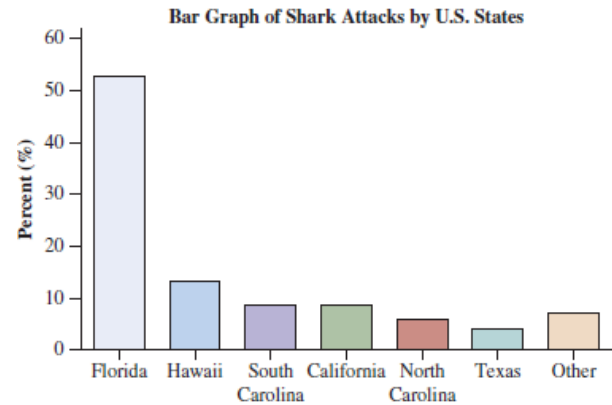
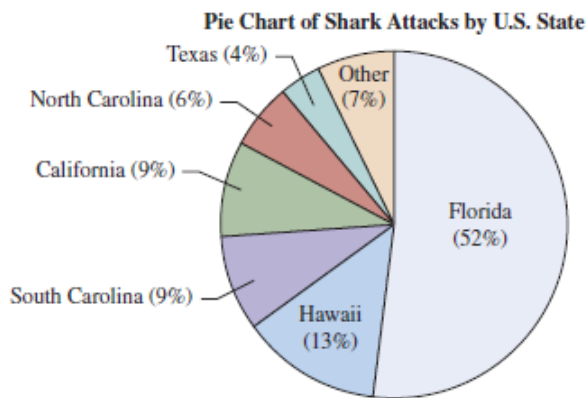
- **Example Shark Attacks in the United States**

For the United States alone, a total of 387 unprovoked shark attacks were reported between 2004 and 2013. Oregon, Alabama, and Georgia with only a few attacks are summarized in the Other category.

Table 2.2 Unprovoked Shark Attacks in the U.S. Between 2004 and 2013*

| U.S. State | Frequency | Proportion | Percentage |
|----------------|------------|--------------|--------------|
| Florida | 203 | 0.525 | 52.5 |
| Hawaii | 51 | 0.132 | 13.2 |
| South Carolina | 34 | 0.088 | 8.8 |
| California | 33 | 0.085 | 8.5 |
| North Carolina | 23 | 0.059 | 5.9 |
| Texas | 16 | 0.041 | 4.1 |
| Other | 27 | 0.070 | 7.0 |
| Total | 387 | 1.000 | 100.0 |

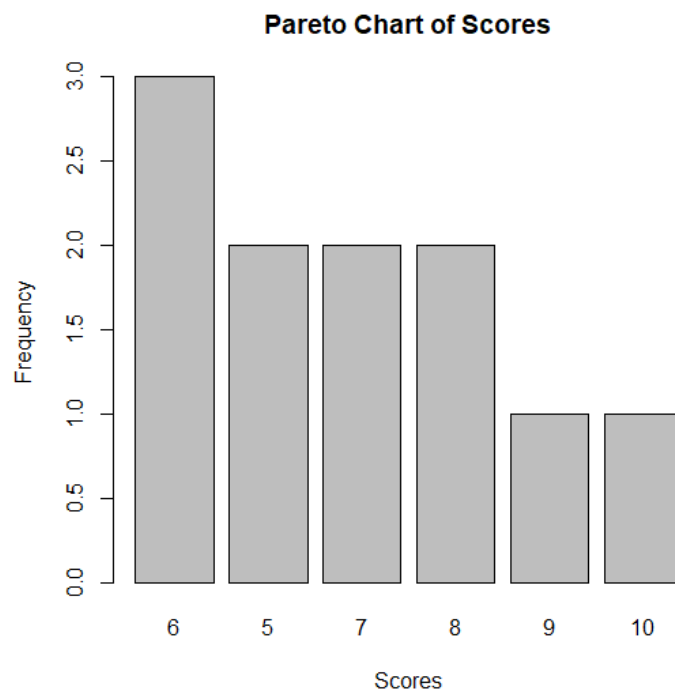
*Source: <http://www.flmnh.ufl.edu/fish/sharks/statistics/statsus.htm>



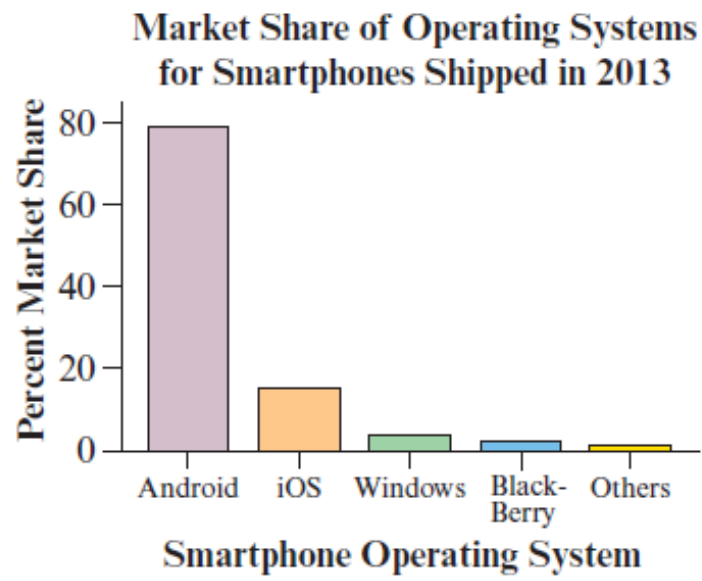
- What percentage of attacks occurred in Florida and the Carolinas?
- Describe the distribution of shark attacks across U.S. states.

- **Pareto Chart** is a bar graph with categories ordered by their frequency, named after Italian economist Vilfredo Pareto (1848 – 1923), who advocate it use.
- The Pareto chart is often used in business applications to identify the most common outcomes, such as identifying products with the highest sales or identifying the most common types of complaints that a customer service center receives.
- The chart helps to portray the **Pareto principle**, which states that a small subset of categories often contains most of the observations.

```
> t[order(-t$Freq),]
  Score Freq
2      6    3
1      5    2
3      7    2
4      8    2
5      9    1
6     10    1
> Pareto <- t[order(-t$Freq),]
> barplot(Pareto$Freq, names.arg = Pareto$Score,
+ xlab = "Scores", ylab = "Frequency",
+ main = "Pareto Chart of Scores")
```



- **Example** Market Share of Operating Systems

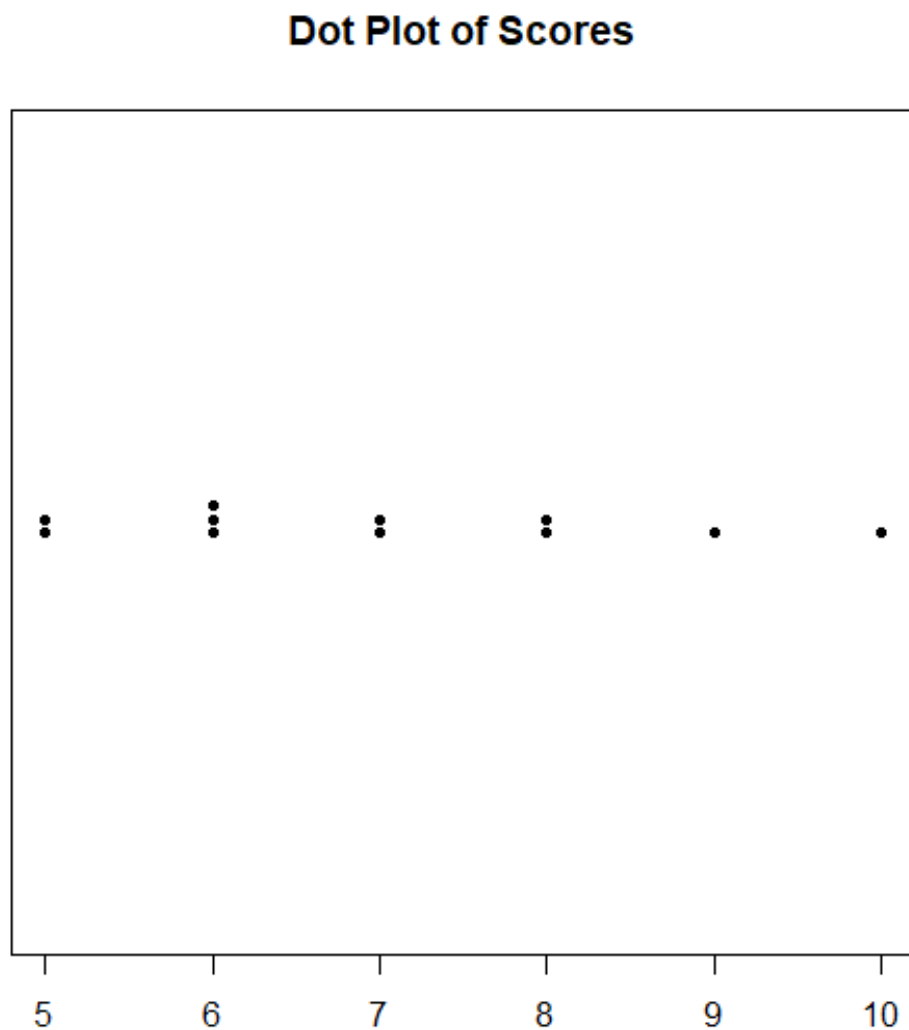


Source: idc.com

- The three primary graphical displays for summarizing a quantitative variable are:

1. **Dot Plot:** A dot plot shows a dot for each observation, placed just above the value on the number line for that observation.

```
> stripchart(score, method = "stack", pch = 20,  
+ main = "Dot Plot of Scores")
```



2. **Stem-and-Leaf Plot:** Each observation is represented by a stem and a leaf.

| Stems | Leaves |
|-------|----------|
| 7 | 69 |
| 8 | 00125699 |
| 9 | 12446 |

Observation = 92, in a sample of 15 test scores

```
> data <- c(76, 79, 80, 80, 81, 82, 85, 86, 89, 89,
+ 91, 92, 94, 94, 96)
> stem(data)

The decimal point is 1 digit(s) to the right of the |

 7 | 69
 8 | 0012
 8 | 5699
 9 | 1244
 9 | 6

> stem(data, scale = 0.5)

The decimal point is 1 digit(s) to the right of the |

 7 | 69
 8 | 00125699
 9 | 12446
```

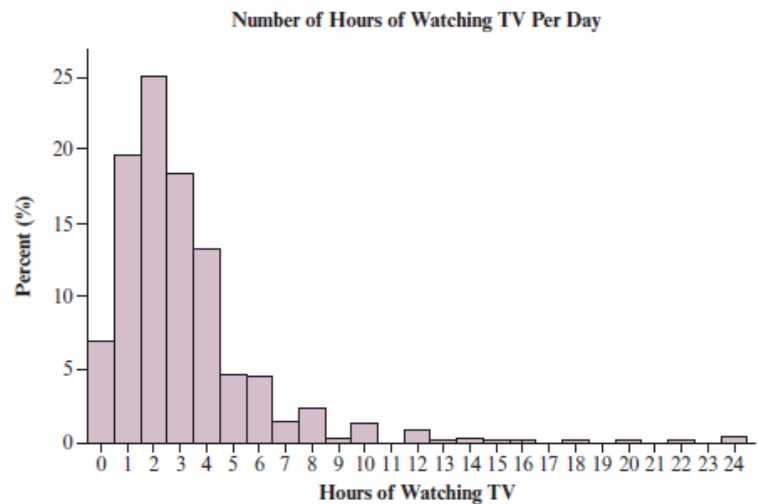
| History | Stem | English |
|-------------|------|-------------|
| | 4 | 7 |
| 9 | 5 | 7 |
| 7 1 | 6 | 2 6 8 9 |
| 3 1 1 | 7 | 2 4 4 6 7 9 |
| 8 5 3 2 2 0 | 8 | 1 3 5 |
| 8 7 2 | 9 | 1 |
| | 10 | |

Leaf unit: 1.0

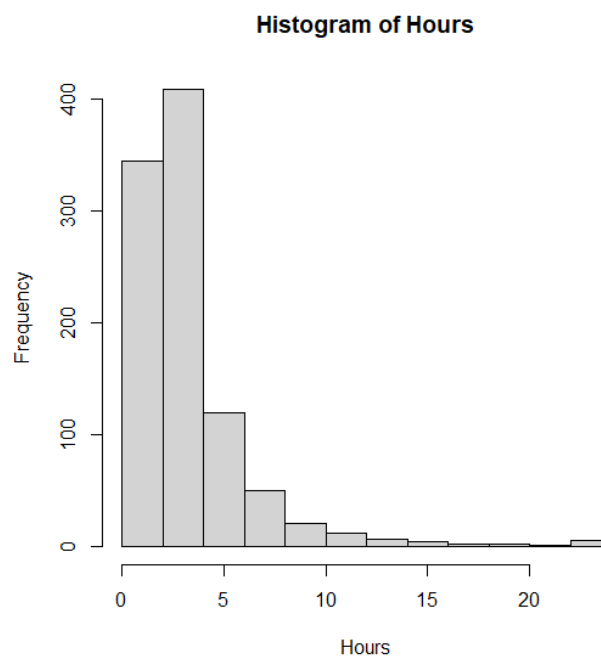
3. **Histogram:** A histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

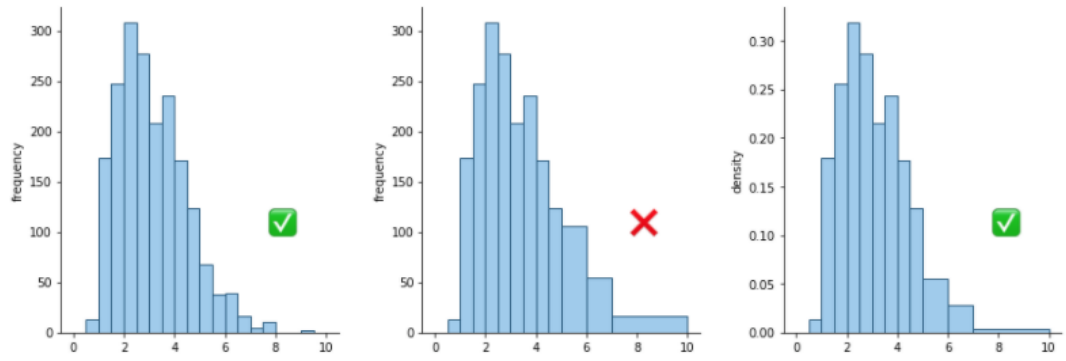
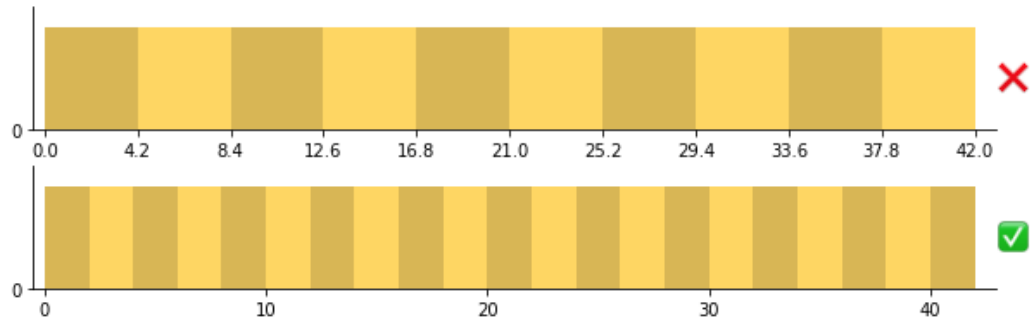
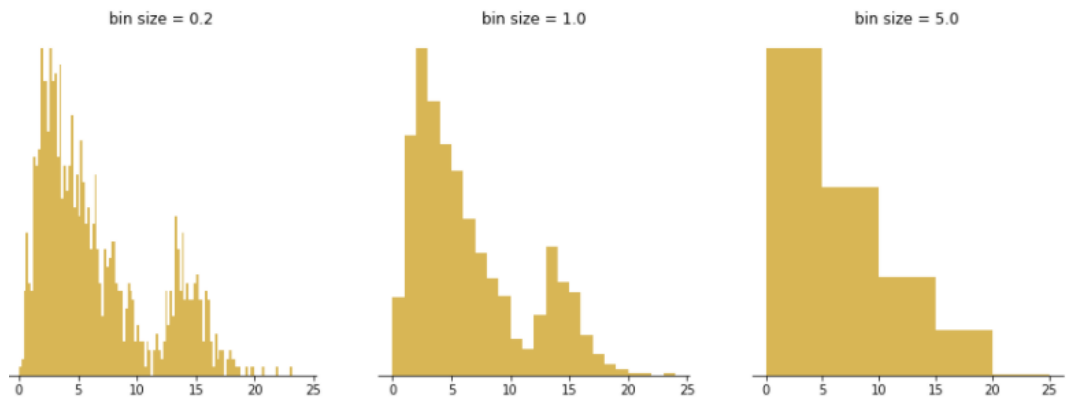
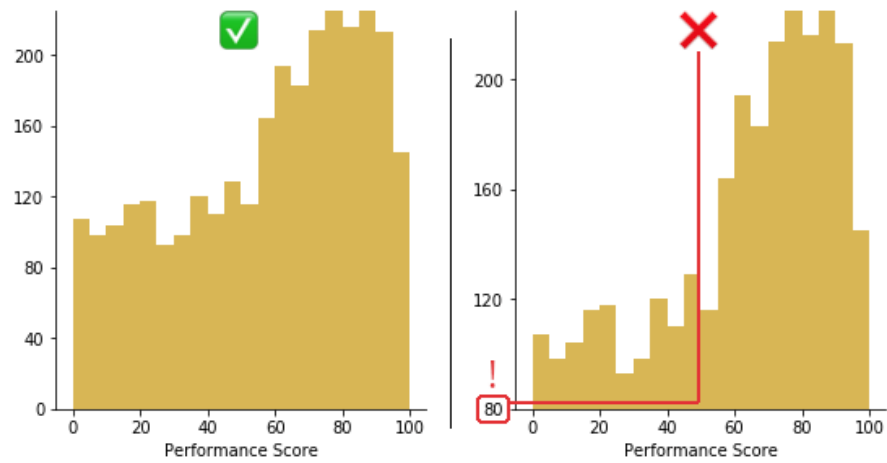
- Example TV Watching**

| Hours | Count | Hours | Count |
|-------|-------|-------|-------|
| 0 | 90 | 13 | 2 |
| 1 | 255 | 14 | 4 |
| 2 | 325 | 15 | 3 |
| 3 | 238 | 16 | 1 |
| 4 | 171 | 17 | 0 |
| 5 | 61 | 18 | 1 |
| 6 | 58 | 19 | 0 |
| 7 | 19 | 20 | 2 |
| 8 | 31 | 21 | 0 |
| 9 | 3 | 22 | 1 |
| 10 | 17 | 23 | 0 |
| 11 | 0 | 24 | 5 |
| 12 | 11 | | |



```
> tv <- c(rep(0,90), rep(1,255), rep(3,238), rep(4,171), rep(5,61),
+ rep(6,58), rep(7,19), rep(8,31), rep(9,3), rep(10,17), rep(11,0),
+ rep(12,11), rep(13,2), rep(14,4), rep(15,3), rep(16,1), rep(17,1),
+ rep(18,1), rep(19,0), rep(20,2), rep(21,0), rep(22,1), rep(23,0),
+ rep(24,5))
> hist(tv, xlab = "Hours", main = "Histogram of Hours")
```



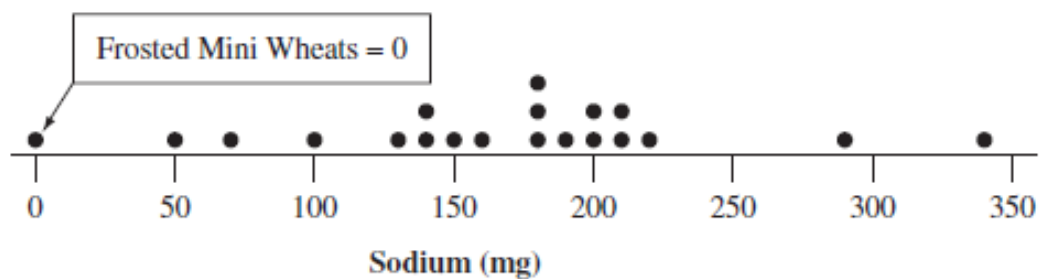


- **Example** Health Value of Cereals

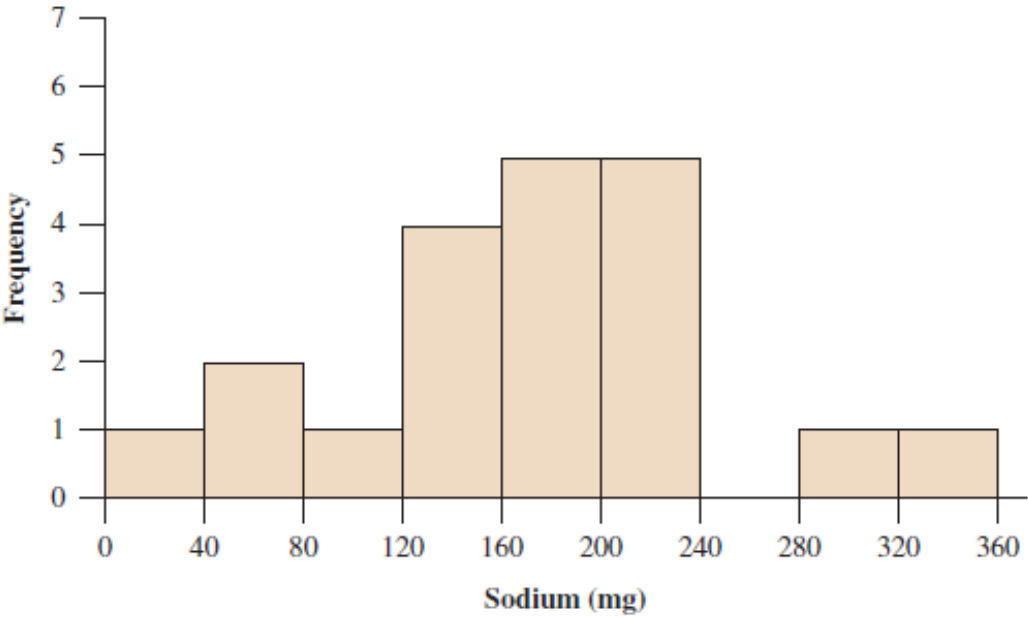
Let us investigate the amount of sugar and salt (sodium) in breakfast cereal.

| Cereal | Sodium (mg) | Sugar (g) | 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 |

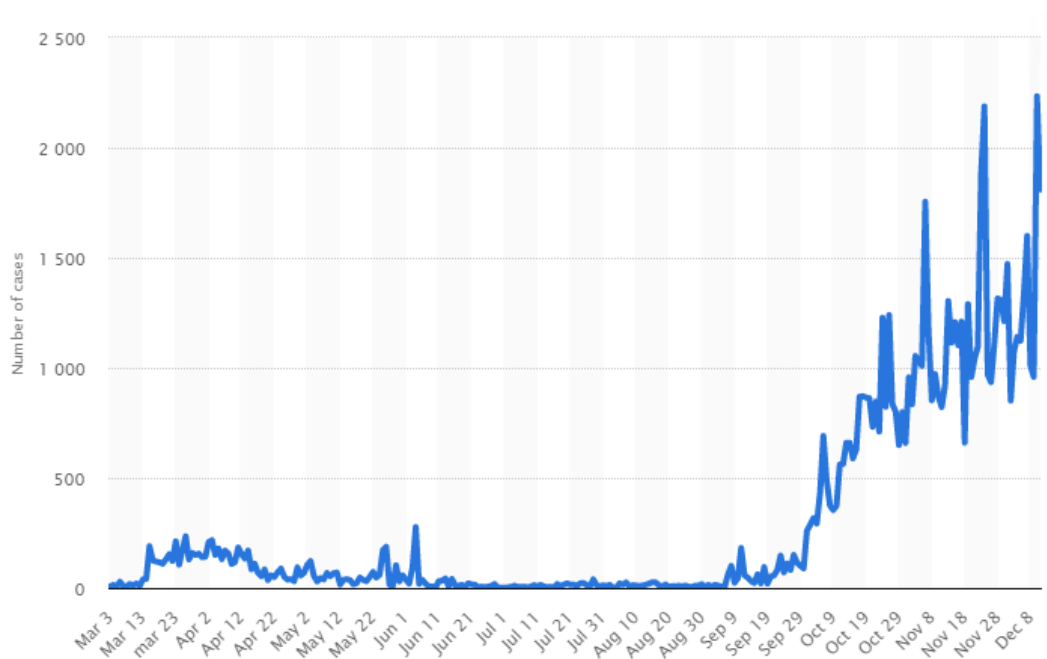
Source: www.weightchart.com (click Nutrition).



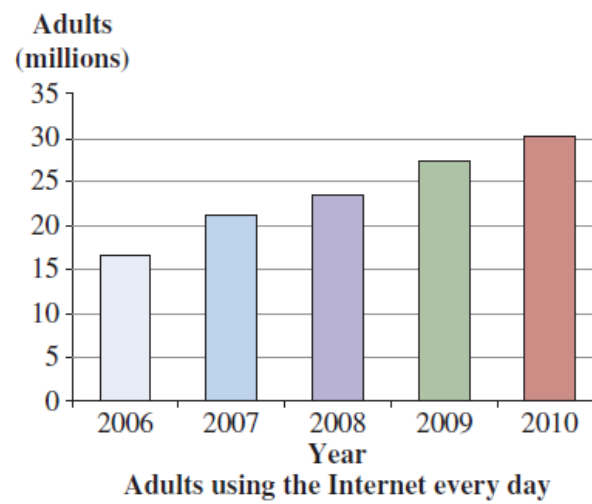
| | | | | |
|---|--|---|-----------|--------|
| | | 0 | | 0 |
| | | 0 | | 57 |
| | | 1 | | 0344 |
| | | 1 | | 568889 |
| 0 | | 0 | 57 | |
| 1 | | 0 | 344568889 | |
| 2 | | 0 | 01129 | |
| 3 | | 4 | | |



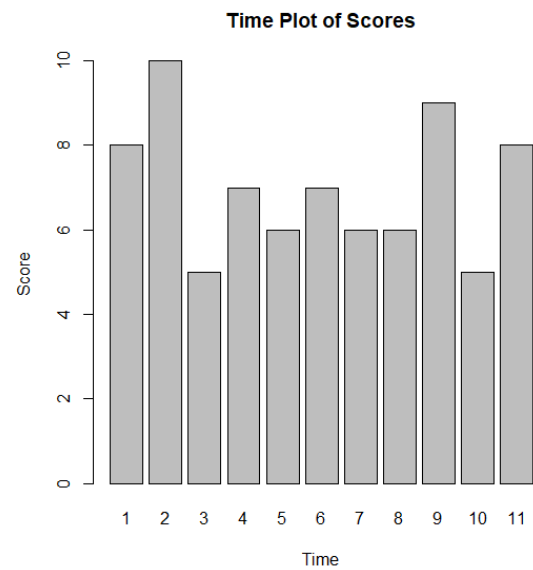
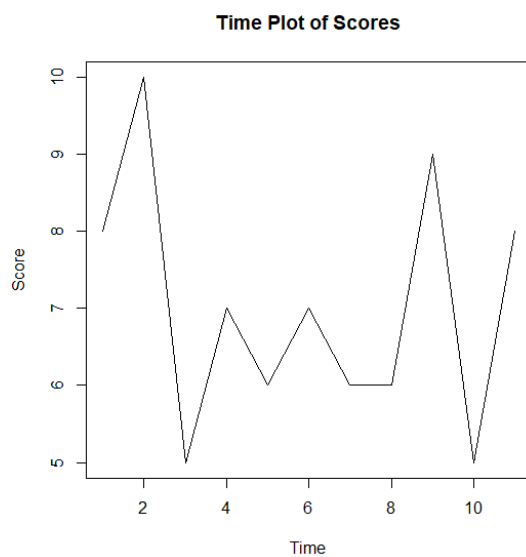
- For some variables, observations occur over time. Examples include the daily closing price of a stock and the population of a country measured every decade in a census. A data set collected over time is called a **time series**.
- We can display time-series data graphically using a **time plot**. This charts each observation, on the vertical scale, against the time it was measured, on the horizontal scale. A common pattern to look for is a **trend** over time, indicating a tendency of the data to either rise or fall. To see a trend more clearly, it is beneficial to connect the data points in their time sequence.
- **Example** Number of daily confirmed cases of the novel coronavirus infection COVID-19 in Malaysia from March to December 2020



- Another way time series data is displayed is with a type of bar graph.



```
> score
[1] 8 10 5 7 6 7 6 6 9 5 8
> ts.plot(score, ylab = "Score", main = "Time Plot of Scores")
> barplot(score, names.arg = c(1:11), xlab = "Time",
+ ylab = "Score", main = "Time Plot of Scores")
```



- **Article Reading** Florence Nightingale, Statistics and the Crimean War



On the Shoulders of...Florence Nightingale

Graphical Displays Showing Deaths From Disease Versus Military Combat

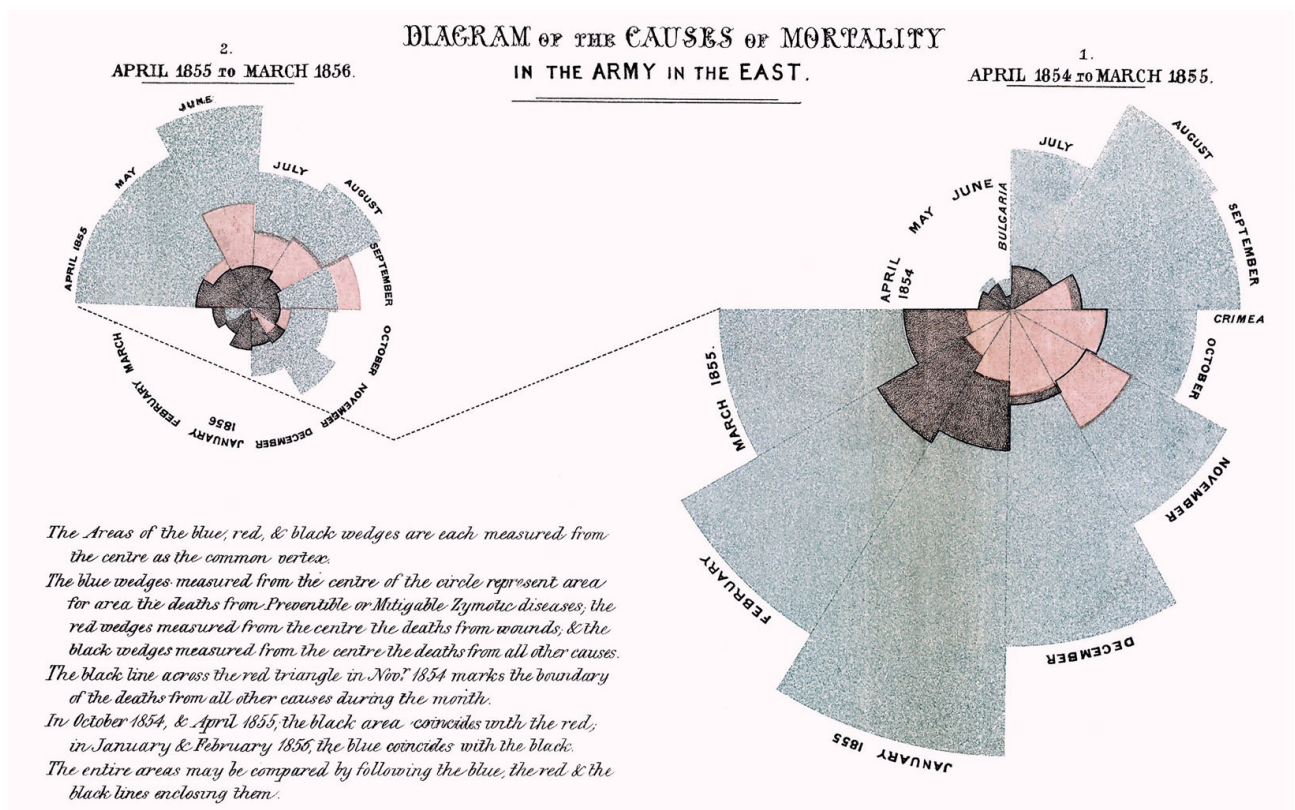
Florence Nightingale

During the Crimean War in 1854, the British nurse Florence Nightingale (1820–1910) gathered data on the number of soldiers who died from various causes. She prepared graphical displays such as time plots and pie charts for policy makers. The graphs showed that more soldiers were dying from contagious diseases than from war-related wounds. The plots were

revolutionary for her time. They helped promote her national cause of improving hospital conditions.

After implementing sanitary methods, Nightingale showed with time plots that the relative frequency of soldiers' deaths from contagious disease decreased sharply and no longer exceeded that of deaths from wounds.

Throughout the rest of her life, Nightingale promoted the use of data for making informed decisions about public health policy. For example, she used statistical arguments to campaign for improved medical conditions in the United States during the Civil War in the 1860s (Franklin, 2002).



Exercises 2.2 2.11, 2.13, 2.15, 2.17, 2.19, 2.21, 2.23, 2.25, 2.27.

2.3 Measuring the Center of Quantitative Data

- The best-known and most frequently used measure of the center of a distribution of a quantitative variable is the *mean*. It is found by averaging the observations. Another popular measure is the *median*. Half the observations are smaller than it, and half are larger.
- The **mean** is the sum of the observations divided by the number of observations. It is interpreted as the balance point of the distribution.

$$\text{Population Mean, } \mu = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Sample Mean, } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- The **median** is the middle value of the observations when the observations are ordered from the smallest to the largest (or from the largest to the smallest). If there are two observations in the middle, the average will be taken.
- **Example** Health Value of Cereal

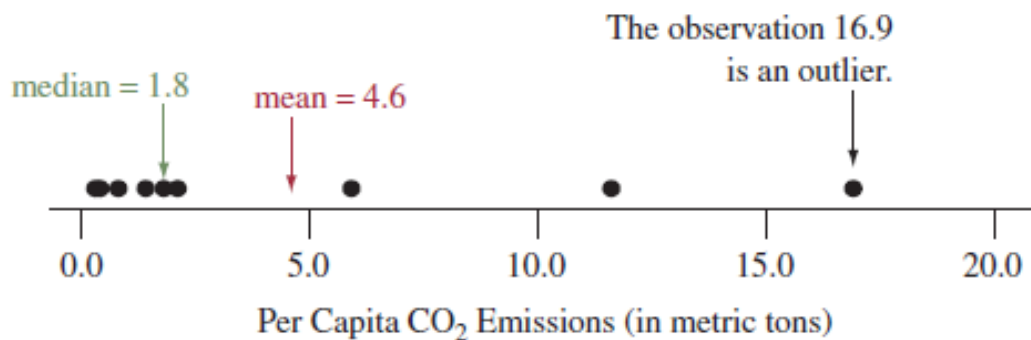
```
0 340 70 140 200 180 210 150 100 130
140 180 190 160 290 50 220 180 200 210
```

```
> cereal <- c(0,340,70,140,200,180,210,150,100,130,
+ 140,180,190,160,290,50,220,180,200,210)
> mean(cereal)
[1] 167
> sort(cereal)
[1] 0 50 70 100 130 140 140 150 160 180 180 180 190 200 200
[16] 210 210 220 290 340
> median(cereal)
[1] 180
```

- **Example CO₂ Pollution**

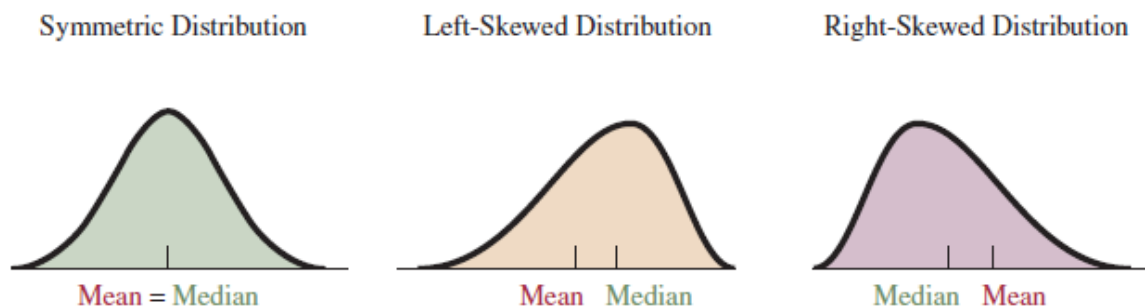
The Pew Center on Global Climate Change reports that global warming is largely a result of human activity that produces carbon dioxide (CO₂) emissions and other greenhouse gases. The CO₂ emissions from fossil fuel combustion are the result of electricity, heating, industrial processes, and gas consumption in automobiles. The International Energy Agency reported the per capita CO₂ emissions by country (that is, the total CO₂ emissions for the country divided by the population size of that country) for 2011. For the nine largest countries in population size (which make up more than half the world's population), the values were, in metric tons per person:

| | | | | | |
|---------------|------|-----------|-----|--------------------|------|
| China | 5.9 | Indonesia | 1.8 | Nigeria | 0.3 |
| India | 1.4 | Brazil | 2.1 | Bangladesh | 0.4 |
| United States | 16.9 | Pakistan | 0.8 | Russian Federation | 11.6 |



Using this data set, explain the effect an outlier can have on the mean.

- The shape of a distribution influences whether the mean is larger or smaller than the median. For instance, an extremely large value out in the right-hand tail pulls the mean to the right. The mean then usually falls above the median.
- Generally, if the shape is
 - perfectly symmetric, the mean equals the median.
 - skewed to the left, the mean is smaller than the median.
 - skewed to the right, the mean is larger than the median.



- The calculation of the mean uses all the numerical values. It depends on how far observations fall from the middle. Because the mean is the balance point, an extreme value on the right side pulls the mean toward the right tail.
- The median is not affected by an outlier. How far an outlier falls from the middle of the distribution does not influence the median. The median is determined solely by having an equal number of observations above it and below it. Because the median is not affected, it is said to be **resistant** to the effect of extreme observations.

- From these properties, you might think that it's always better to use the median rather than the mean. That's not true. The mean has other useful properties that we'll learn about and take advantage of in later chapters.
- It is a good idea to report both the mean and the median when describing the center of a distribution.
- If a distribution is highly skewed, the median is usually preferred over the mean because it better represents what is typical.
- If the distribution is close to symmetric or only mildly skewed, the mean is usually preferred because it uses the numerical values of all the observations.
- The **mode** is the value that occurs most frequently. It describes a typical observation in terms of the most common outcome. The concept of the mode is most often used to describe the category of a categorical variable that has the highest frequency (the modal category).
- With quantitative variables, the mode is most useful with discrete variables taking a small number of possible values. For continuous observations, it is usually not meaningful to look for a mode because there can be multiple modes or no mode at all.
- The mode need not be near the center of the distribution. It may be the largest or the smallest value. Thus, it is somewhat inaccurate to call the mode a measure of center, but often it is useful to report the most common outcome.

Exercises 2.3 2.29, 2.31, 2.33, 2.35, 2.37, 2.39, 2.41, 2.43, 2.45.

2.4 Measuring the Variability of Quantitative Data

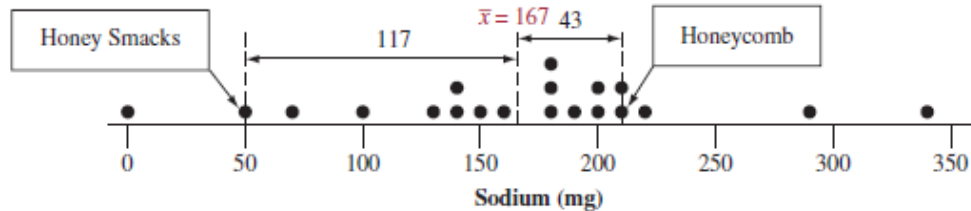
- A measure of the center is not enough to describe a distribution for a quantitative variable adequately. It tells us nothing about the variability of the data.
- The simplest way to describe the variability is with the *range*.
- The **range** is the difference between the largest and the smallest observations.

$$\text{range} = \text{maximum} - \text{minimum}$$

```
> sort(cereal)
 [1]  0  50  70 100 130 140 140 150 160 180 180 180 190 200 200
[16] 210 210 220 290 340
> range(cereal)
 [1]  0 340
> max(cereal)
 [1] 340
> min(cereal)
 [1] 0
> max(cereal) - min(cereal)
 [1] 340
```

- The range is simple to compute and easy to understand, but it uses only the extreme values and ignores the other values. Therefore, it's affected severely by outliers.
- The range is not a resistant statistic. It shares the worst property of the mean, not being resistant, and the worst property of the median, ignoring the numerical values of nearly all the data.

- A much better numerical summary of variability uses all the data, and it describes a typical distance of how far the data falls from the mean. It does this by summarizing deviations from the mean.
- The deviation of an observation x from the mean is $(x - \bar{x})$, the difference between the observation and the sample mean.
- Each observation has a deviation from the mean.
- A deviation is positive when the observation falls above the mean. A deviation is negative when the observation falls below the mean.



- The interpretation of the mean as the balance point implies that the positive deviations counterbalance the negative deviations. Because of this, the sum (and therefore the mean) of the deviations always equals zero, regardless of the actual data values.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- Hence, summary measures of variability from the mean use either the squared deviations or their absolute values.

$$\sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or} \quad \sum_{i=1}^n |x_i - \bar{x}|$$

- The average of the squared deviations is called the **variance**. Because the variance uses the square of the units of measurement for the original data, its square root is easier to interpret. This is called the **standard deviation**.

$$\text{Population Variance, } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$\text{Population Standard Deviation, } \sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

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

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

```
> var(cereal)
[1] 5969.474
> sd(cereal)
[1] 77.26237
```

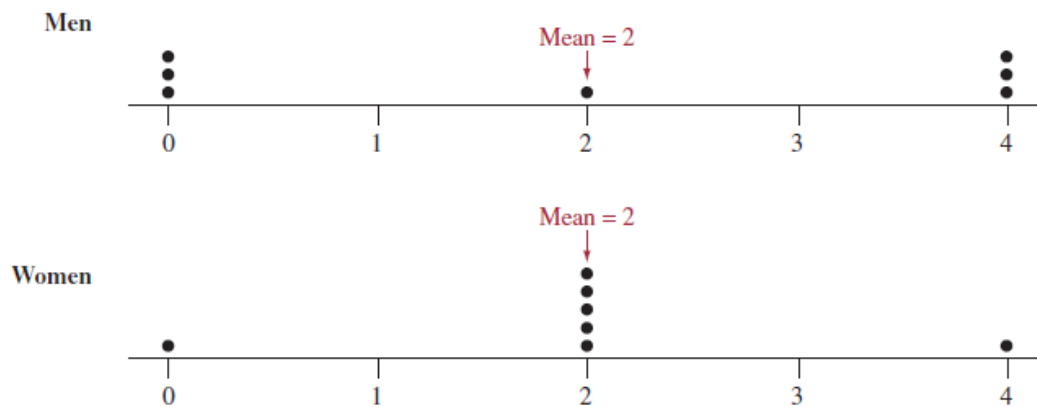
- Roughly, the standard deviation represents a typical distance or a type of average distance of an observation from the mean.
- The larger the standard deviation, the greater the variability of the data.
- You may wonder why the denominators use $n - 1$ instead of n for the sample. Basically it is because the deviations provide only $n - 1$ pieces of information about variability. That is, $n - 1$ of the deviations determine the last one, because the deviations sum to 0. This is called the *degrees of freedom*.

- Example Women's and Men's Ideal Number of Children**

Students in a class were asked on a questionnaire at the beginning of the course, “How many children do you think is ideal for a family?” The observations, classified by student's gender, were

Men: 0 0 0 2 4 4 4

Women: 0 2 2 2 2 2 4



Range Standard Deviation

Men: 4 2.0

Women: 4 1.2 (rounded to 1 decimal place)

Do the distributions of data have the same amount of variability around the mean? If not, which distribution has more variability?

- **Example** Exam Scores

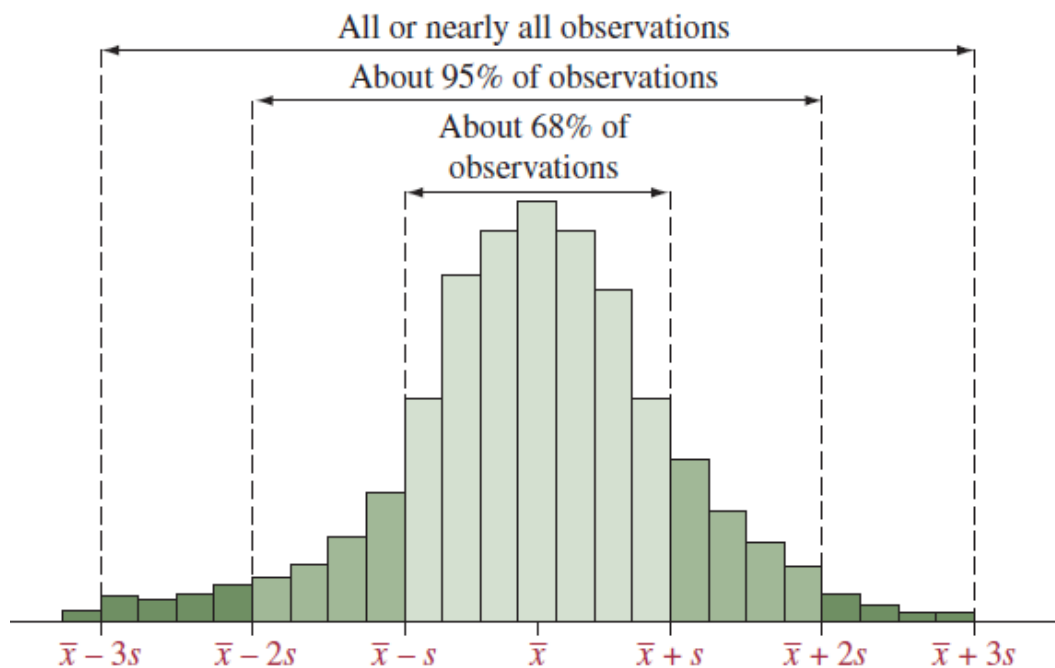
The first exam in your statistics course is graded on a scale of 0 to 100. Suppose that the mean score in your class is 80. Which value is most plausible for the standard deviation s : 0, 0.5, 10, or 50?

- Suppose that a distribution is unimodal and approximately symmetric with a bell shape. The value of s then has a more precise interpretation. Using the mean and standard deviation alone, we can form intervals that contain certain percentages (approximately) of the data.

Bell-shaped Distribution

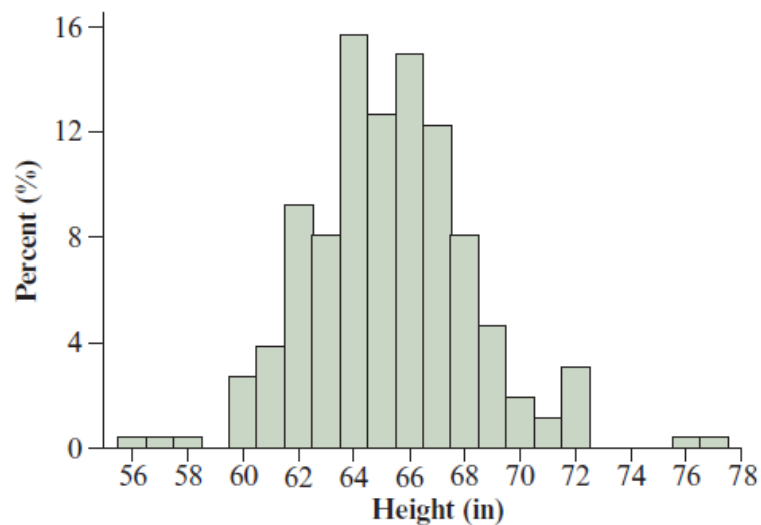


- **Empirical Rule** If a distribution of data is bell shaped, then approximately
 - 68% of the observations fall within 1 standard deviation of the mean.
 - 95% of the observations fall within 2 standard deviations of the mean.
 - 99.7% of the observations (*all or nearly all*) fall within 3 standard deviations of the mean.



- **Example Female Student Heights**

Many human physical characteristics have bell-shaped distributions. Let us explore height. The figure below shows a histogram of the heights from responses a survey by 261 female students at the University of Georgia. (The data are in the Heights data file on the book's website. Note that the height of 92 inches was omitted from the analysis.)



And, the table below presents some descriptive statistics, using MINITAB.

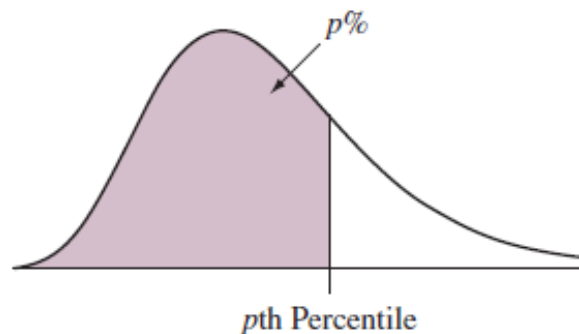
| Variable | N | Mean | Median | StDev | Minimum | Maximum |
|----------|-----|--------|--------|-------|---------|---------|
| HEIGHT | 261 | 65.284 | 65.000 | 2.953 | 56.000 | 77.000 |

Can we use the empirical rule to describe the variability from the mean of these data? If so, how?

Exercises 2.4 2.47, 2.49, 2.51, 2.53, 2.55, 2.57, 2.59, 2.61.

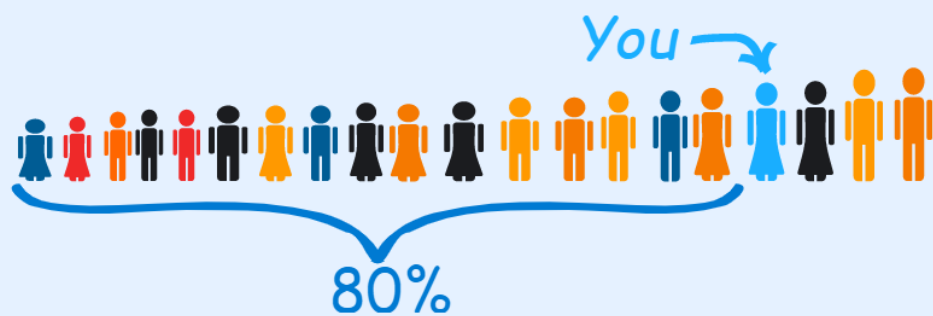
2.5 Using Measures of Position to Describe Variability

- We'll now learn about some other ways of describing a distribution using measures of position.
- The ***p*th percentile** is a value such that p percent of the observations fall below or at that value.



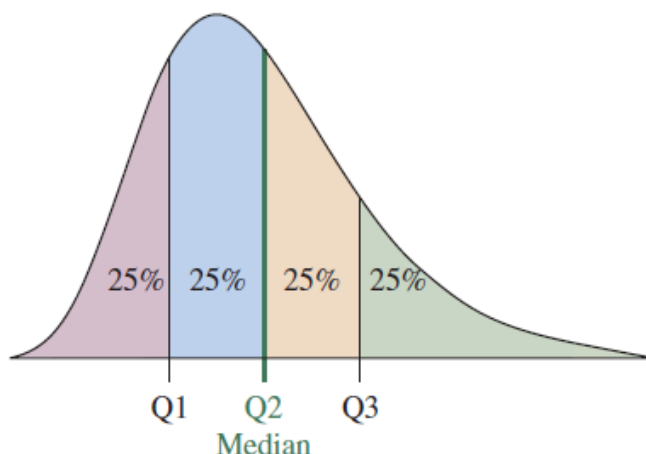
- Suppose you're informed that your height falls at the 80th percentile in your class. Set $p = 80$ in this definition. Then, 80% of your classmates have heights between the minimum height and your height. Only 20% of them are taller than you.

80% of people are shorter than you:



That means you are at the **80th percentile**.

- Three useful percentiles are the **quartiles**.



| | | | |
|------------------------|----------------|-------|-----------------|
| First Quartile | Lower Quartile | Q_1 | 25th Percentile |
| Second Quartile | Median | Q_2 | 50th Percentile |
| Third Quartile | Upper Quartile | Q_3 | 75th Percentile |

- The quartiles are also used to define a measure of variability that is more resistant than the range and the standard deviation.
- This measure summarizes the range for the middle half of the data. The middle 50% of the observations fall between the first quartile and the third quartile – 25% from Q_1 to Q_2 and 25% from Q_2 to Q_3 .
- The distance from Q_1 to Q_3 is called the **interquartile range**, denoted by **IQR**.

$$\text{IQR} = Q_3 - Q_1$$

- **Example** Health Value of Cereal

```

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

```

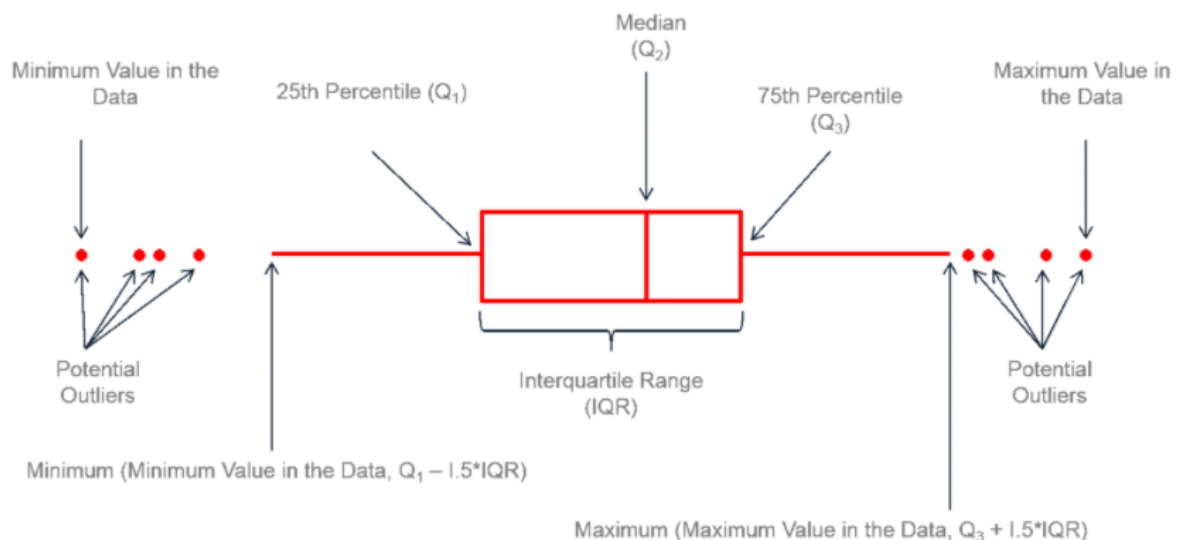
```

> sort(cereal)
[1] 0 50 70 100 130 140 140 150 160 180 180 180 190 200 200
[16] 210 210 220 290 340
> quantile(cereal)
 0%   25%   50%   75%  100%
0.0 137.5 180.0 202.5 340.0
> quantile(cereal)[2]
 25%
137.5
> quantile(cereal)[3]
 50%
180
> quantile(cereal)[4]
 75%
202.5
> summary(cereal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.0   137.5   180.0   167.0   202.5   340.0
> IQR(cereal)
[1] 65

```

Interpret the quartiles in the context of the cereal data.

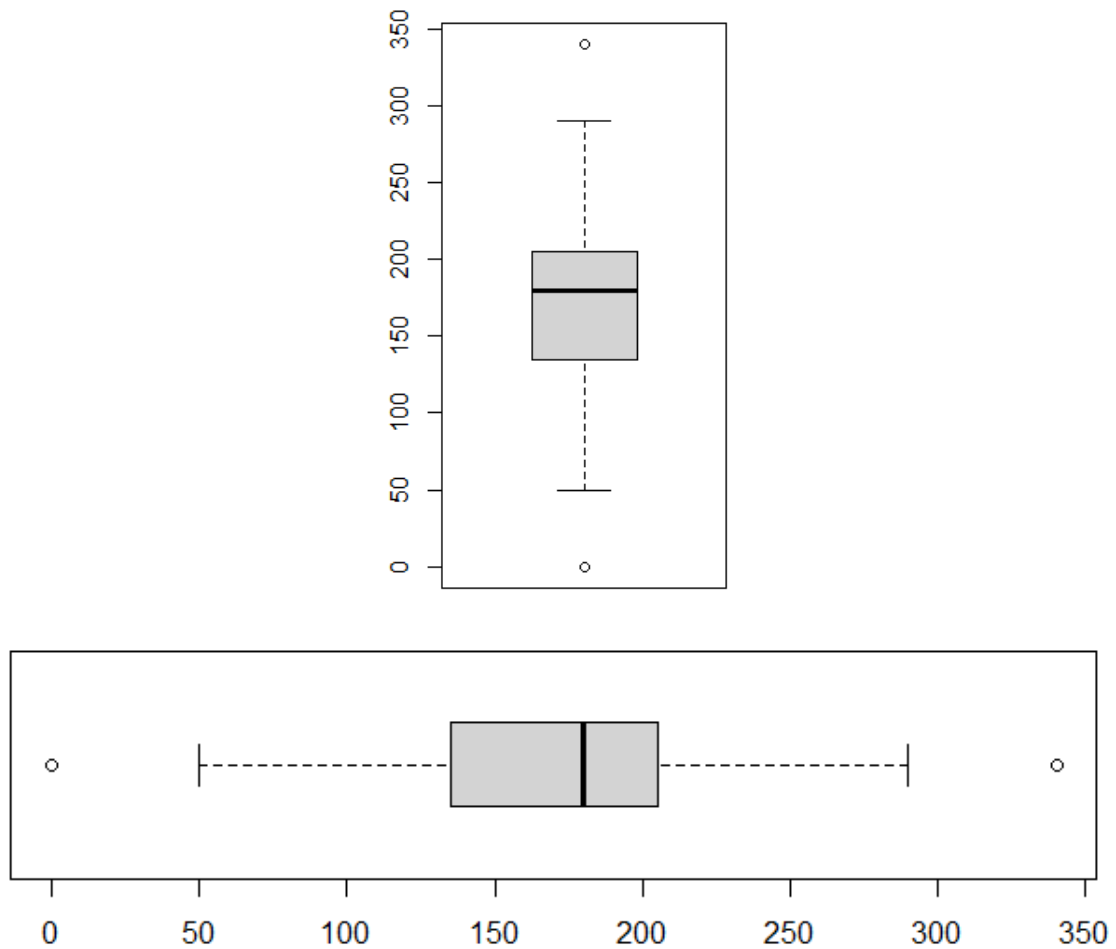
- Recall that an outlier is an unusual observation and it is important to detect it in any statistical analysis.
- We can use the interquartile range to detect an outlier. An observation is a *potential* outlier if it falls a distance of more than $1.5 \times \text{IQR}$ below the first quartile or a distance of more than $1.5 \times \text{IQR}$ above the third quartile.
- We identify an observation as a *potential outlier* rather than calling it a *definite outlier* because when a distribution has a long tail, some observations may be more than $1.5 \times \text{IQR}$ below the first quartile or above the third quartile even if they are not outliers, in the sense that they are not separated far from the bulk of the data.
- The **five-number summary** of a data set consists of the minimum value, first quartile, median, third quartile, and the maximum value.
- These five numbers are the basis of a graphical display called the **Box-and-Whiskers Plot** (in short **Box Plot**).



- **Example** Health Value of Cereal

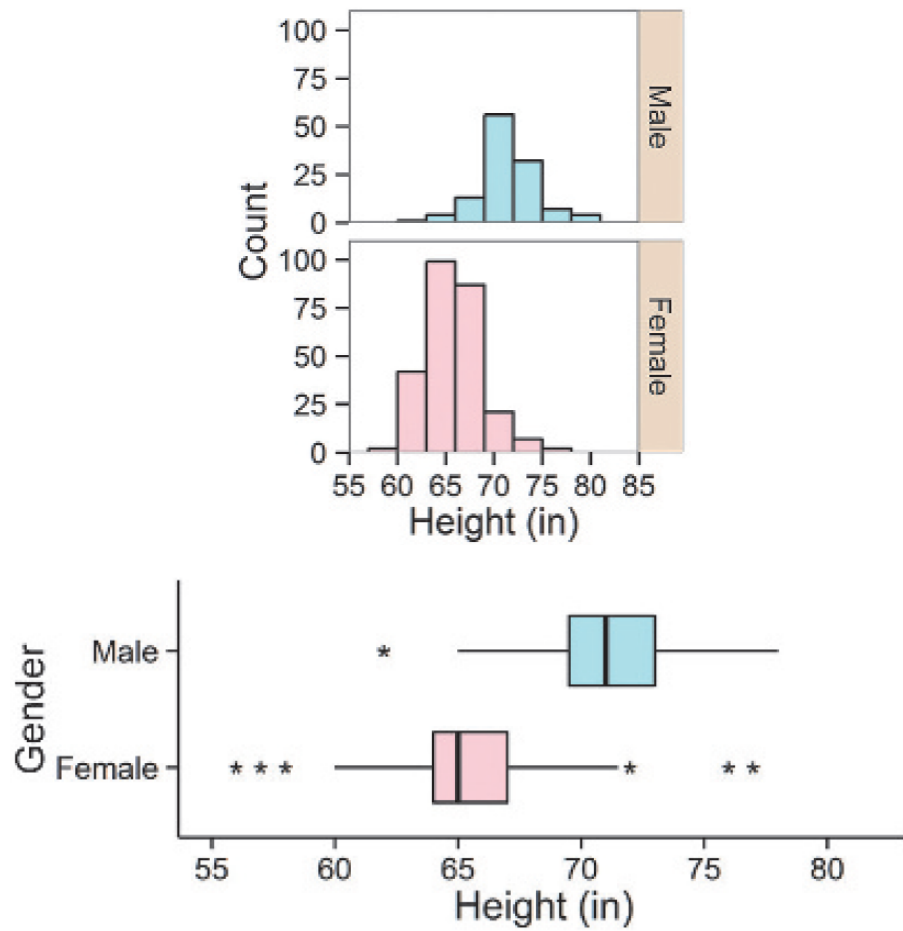
```
0 340 70 140 200 180 210 150 100 130
140 180 190 160 290 50 220 180 200 210
```

```
> boxplot(cereal)
> boxplot(cereal, horizontal = TRUE)
```



Use the box plot to describe the distribution of the cereal data.

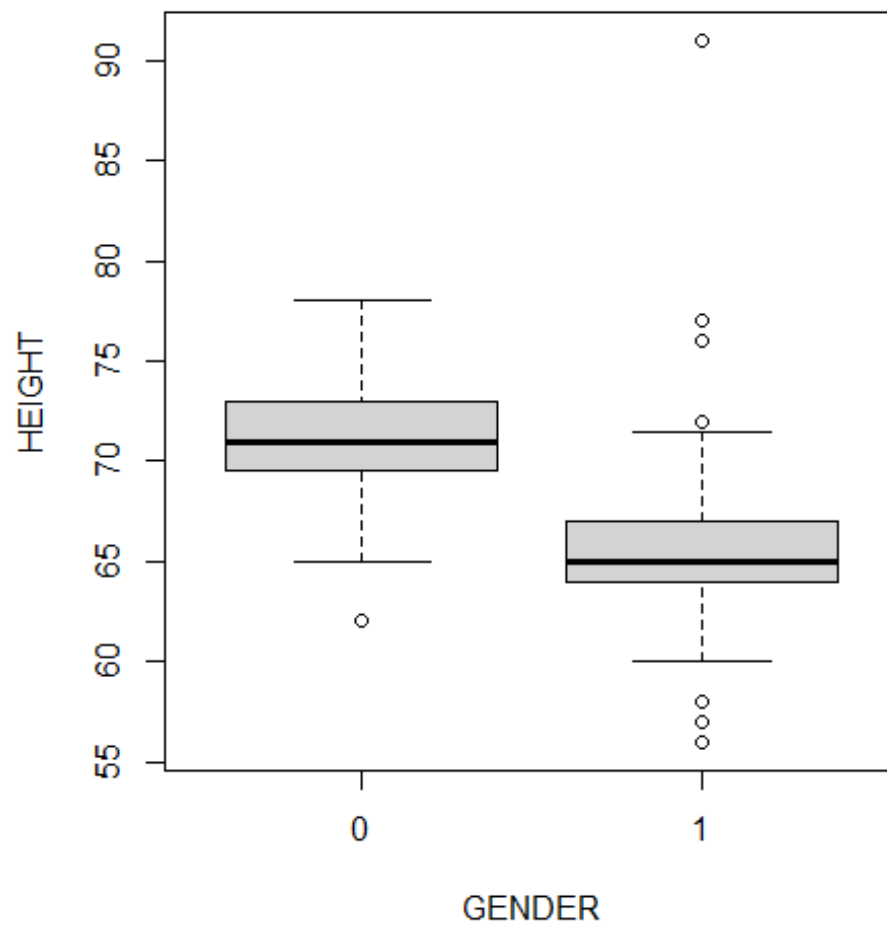
- **Example** Male and Female College Student Heights



Compare the two distributions using the plots provided.

```
> heights <- read.table("heights.txt", header = TRUE, sep = "\t")
> heights
  HEIGHT GENDER
1   70.0      0
2   75.0      0
3   67.0      0
4   67.0      1
5   73.0      0
6   65.0      0
7   73.0      0
8   70.0      0
9   65.0      1
10  73.0      0
11  71.0      0
12  65.5      1
13  74.0      0
14  69.0      1
15  64.0      1

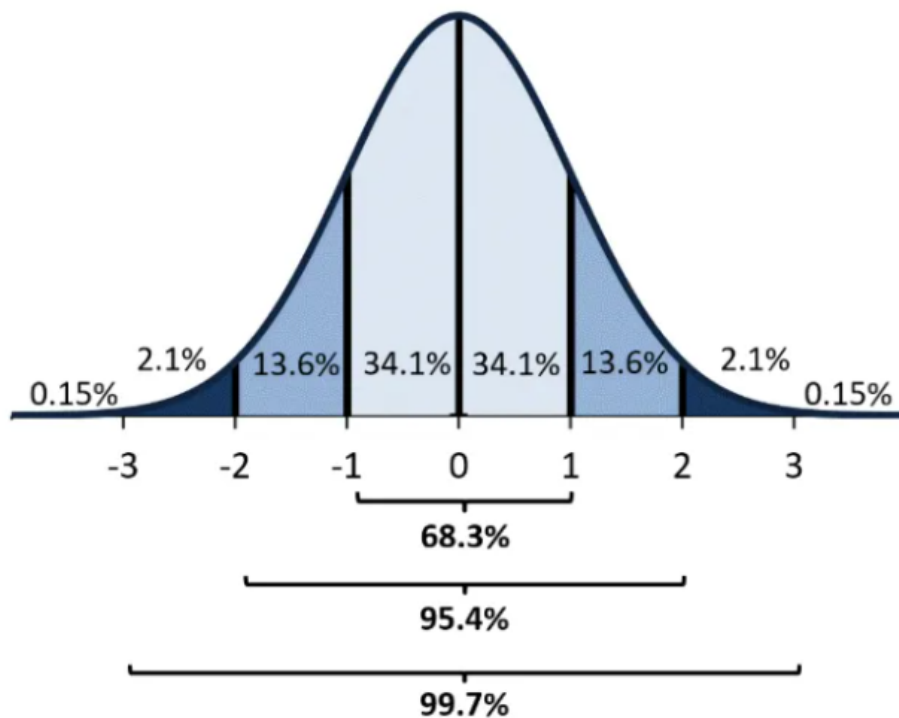
> boxplot(HEIGHT ~ GENDER, data = heights)
```



- The empirical rule tells us that for a bell-shaped distribution, it is unusual for an observation to fall more than 3 standard deviations from the mean. An alternative criterion for identifying potential outliers uses the standard deviation.
- The ***z*-score** for an observation is the number of standard deviations that it falls from the mean. A positive *z*-score indicates the observation is above the mean. A negative *z*-score indicates the observation is below the mean. For sample data, the *z*-score is calculated as

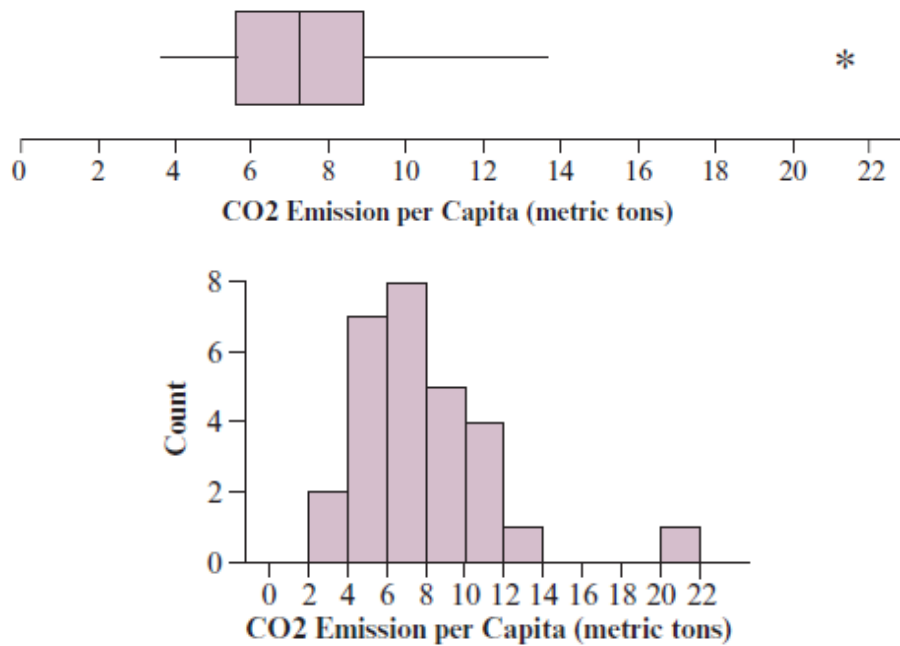
$$z = \frac{x - \bar{x}}{s}$$

- An observation in a bell-shaped distribution is regarded as a potential outlier if its *z*-score is below -3 or above 3 .



- **Example Pollution Outliers**

Let's consider air pollution data for the European Union (EU). The Energy-EU data file on the book's website contains data on per capita carbon dioxide (CO_2) emissions, in metric tons, for the 28 nations in the EU. The mean was 7.9 and the standard deviation was 3.6. In the box plot of the data, the maximum of 21.4, representing Luxembourg, is highlighted as a potential outlier. The histogram is also provided.



- How many standard deviations from the mean was the CO_2 value of 21.4 for Luxembourg?
- The CO_2 value for United States was 16.9. According to the three standard deviation criterion, is the United States an outlier on carbon dioxide emissions relative to the EU?

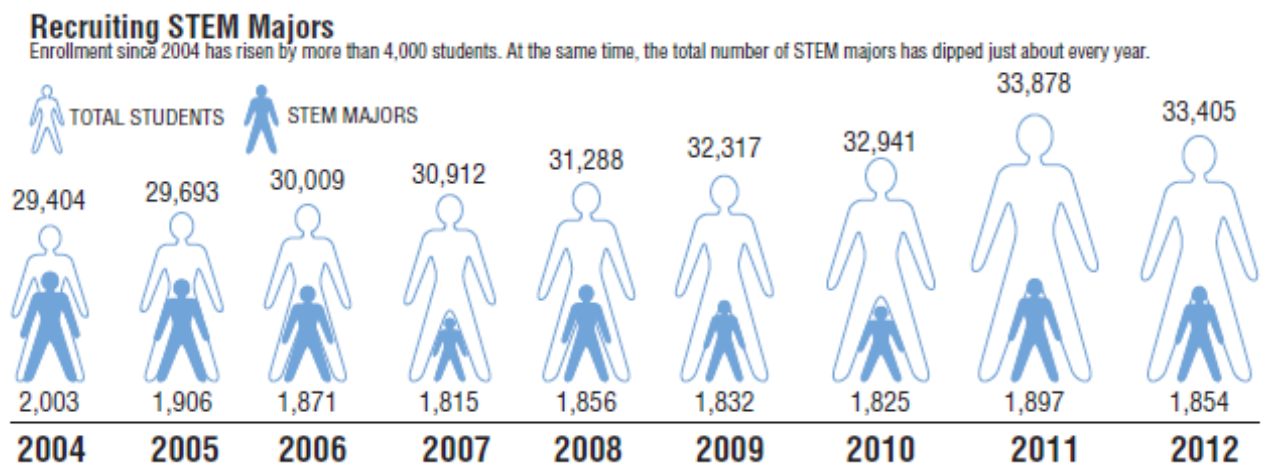
Exercises 2.5 2.63, 2.65, 2.67, 2.69, 2.71, 2.73, 2.75, 2.77, 2.79, 2.81.

2.6 Recognizing and Avoiding Misuses of Graphical Summaries

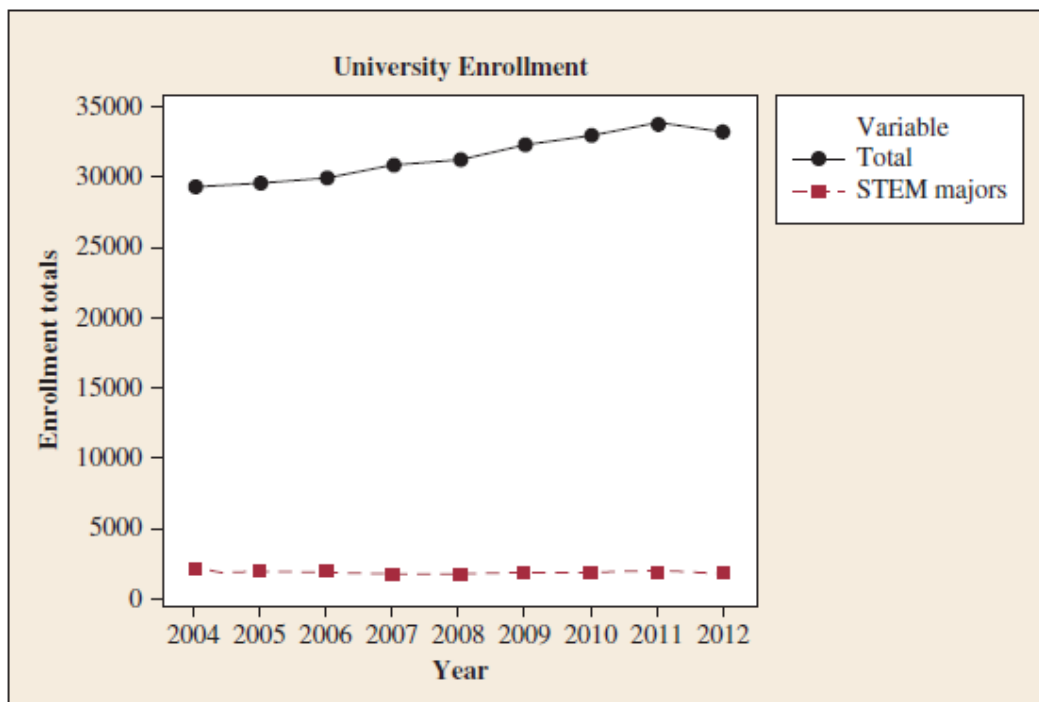
- With modern computer-graphic capabilities, web sites, newspapers, and other periodicals use graphs in an increasing variety of ways to portray information. The graphs are not always well designed, however, and you should look at them skeptically.
- Guidelines for constructing effective graphs:
 1. Label both axes and provide a heading to make clear what the graph is intended to portray.
 2. To help our eyes visually compare relative sizes accurately, the vertical axis should usually start at 0.
 3. Be cautious in using figures, such as people, in place of the usual bars or points. It can make a graph more attractive, but it is easy to get the relative percentages that the figures represent incorrect.
 4. It can be difficult to portray more than one group on a single graph when the variable values differ greatly. Consider instead using separate graphs or plotting relative sizes such as ratios or percentages.

- **Example Recruiting STEM Majors**

Look at the figure below. According to the title and the two-sentence caption, the graph is intended to display how total enrollment has risen at a United States (U.S.) university in recent years while the number of STEM (science, technology, engineering, and mathematics) students has “dipped just about every year.” A graphic designer used a software program to construct a graph for use in a local newspaper. The Sunday headline story was about the decline of STEM majors. The graph is a time plot showing the enrollment between 2004 and 2012, using outlined human figures to portray total enrollment and blue human figures to portray STEM enrollment.

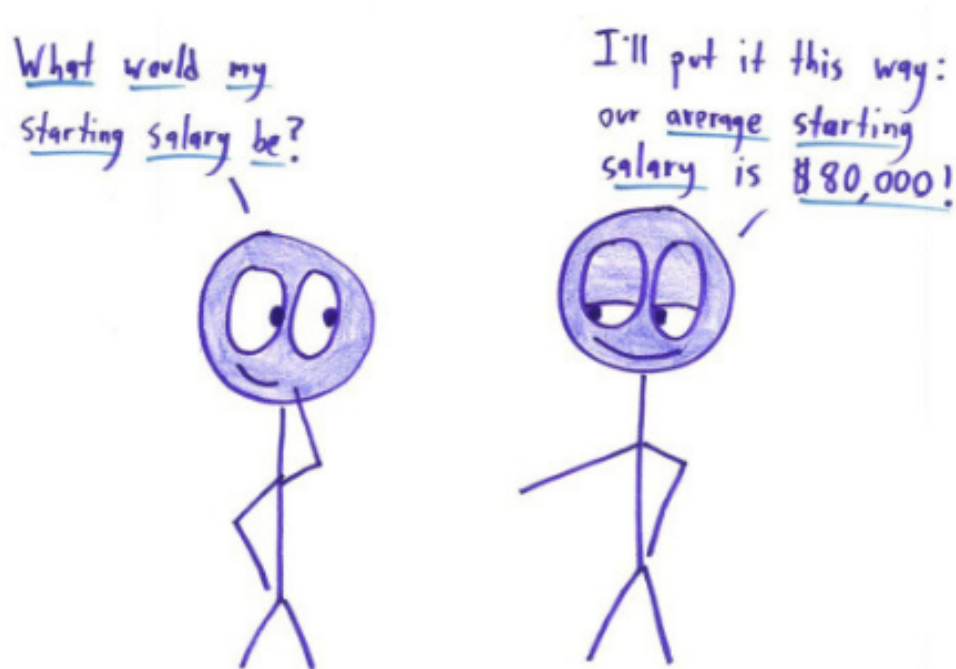


- Do the heights of the human figures accurately represent the counts? Are the areas of the figures in accurate proportions to each other?
- Is an axis shown and labeled as a reference line for the enrollments displayed?
- What other design problems do you see?



What problems do you see in this graphical presentation? Suggest a better way to graphically portray the same information.

- **Example** Starting Salary



How should the potential employee interpret the answer?



On the Shoulders of...John Tukey

"The best thing about being a statistician is that you get to play in everyone's backyard."

—John Tukey (1915–2000)

John Tukey

In the 1960s, John Tukey of Princeton University was concerned that statisticians were putting too much emphasis on complex data analyses and ignoring simpler ways to examine and learn from the data. Tukey developed new descriptive methods, under the title of **exploratory data analysis (EDA)**. These methods make few assumptions about the structure of the data and emphasize data display and ways of searching for

patterns and deviations from those patterns. Two graphical tools that Tukey invented were the stem-and-leaf plot and the box plot.

Initially, few statisticians promoted EDA. However, in recent years, some EDA methods have become common tools for data analysis. Part of this acceptance has been inspired by the availability of computer software and calculators that can implement some of Tukey's methods more easily.

Tukey's work illustrates that statistics is an evolving discipline. Almost all of the statistical methods used today were developed in the past century, and new methods continue to be created, largely because of increasing computer power.

Exercises 2.6 2.83, 2.85, 2.87, 2.89.