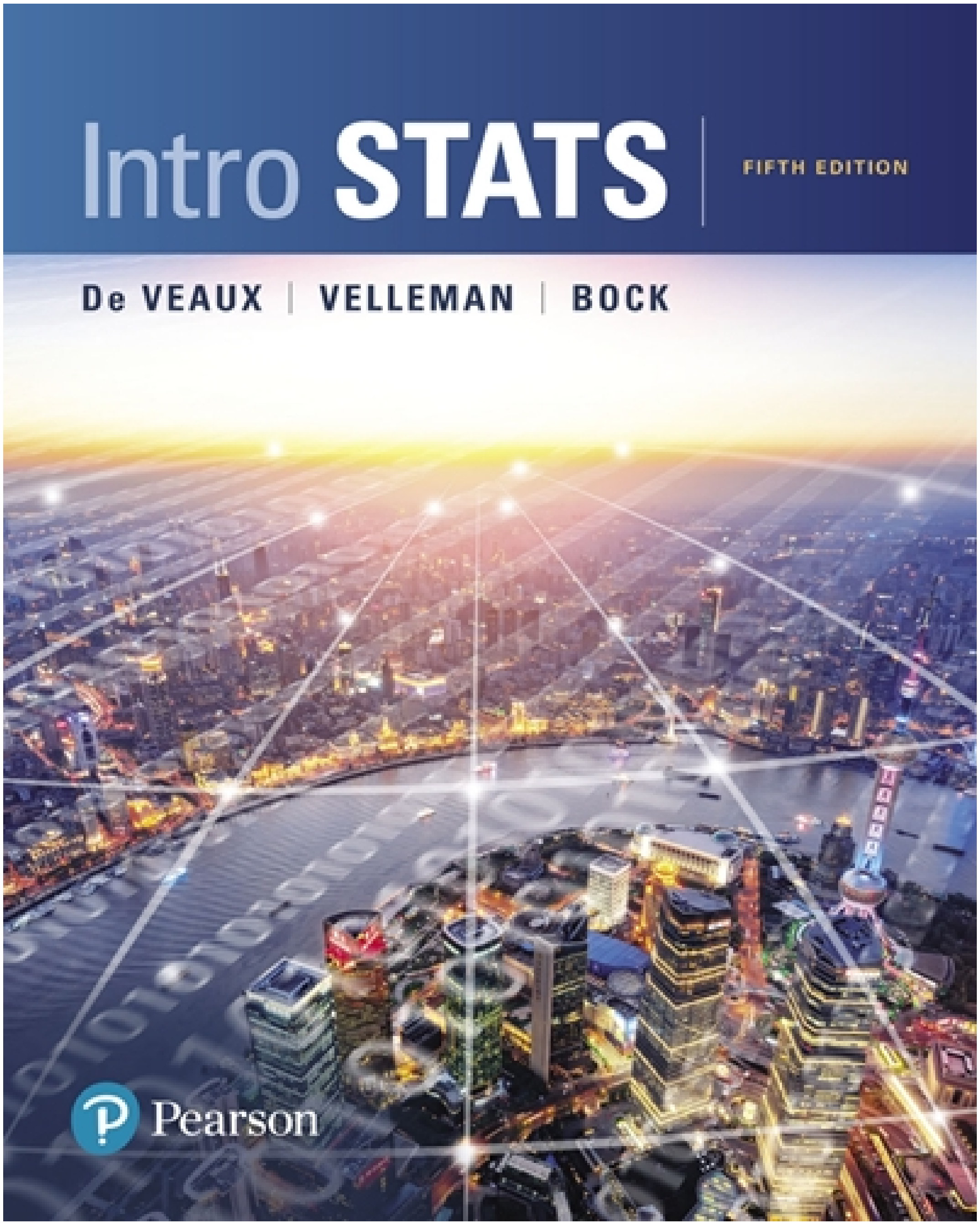
Solutions for Intro Stats 5th Edition by De Veaux



Solutions

Chapter 2 – Displaying and Describing Data

Section 2.1

1. Automobile fatalities.

Subcompact and Mini	0.2658
Compact	0.2084
Intermediate	0.3006
Full	0.2069
Unknown	0.0183

3. Movie genres.

a) A pie chart seems appropriate from the movie genre data. Each movie has only one genre, and the list of all movies constitute a “whole”.

b) “Other” is the least common genre. It has the smallest region in the chart.

5. Movie ratings.

i) C **ii)** A **iii)** D **iv)** B

Section 2.2

7. Traffic Fatalities 2013.

a) The gaps in the histogram for *Year* indicate that we do not have data for those years. This data set contains two variables for each case, and a histogram of the years doesn’t give us much useful information.

b) All of the bars in the *Year* histogram are the same height because each year only appears once in the data set.

c) The distribution of passenger car fatalities has between 17,500 and 25,000 traffic fatalities per year in most years. There were also several years—possibly a second mode—with between 10,000 and 12,500 traffic fatalities.

9. How big is your bicep?

The distribution of the bicep measurements of 250 men is unimodal and symmetric. Based on the height of the tallest points, about 85 of these 250 men have biceps close to 13 inches around. Most are between 12 and 15 inches around. But there are two as small as 10 inches and several that are 16 inches.

11. E-mails.

The distribution of the number of emails received from each student by a professor in a large introductory statistics class during an entire term is skewed to the right, with the number of emails ranging from 1 to 21 emails. The distribution is centered at about 2 emails, with many students only sending 1 email. There is one outlier in the distribution, a student who sent 21 emails. The next highest number of emails sent was only 8.

Section 2.3

13. Biceps revisited.

The distribution of the bicep measurements of 250 men is unimodal and roughly symmetric.

15. Life expectancy.

a) The distribution of life expectancies at birth in 190 countries is skewed to the left.

b) The distribution of life expectancies at birth in 190 countries has one mode, at about 74 to 76 years. The fluctuations from bar to bar don't seem to rise to the level of defining additional modes, although opinions can differ.

17. Life expectancy II.

a) The distribution of life expectancies at birth in 190 countries is skewed to the left, so the median is expected to be larger than the mean. The mean life expectancy is pulled down toward the tail of the distribution.

b) Since the distribution of life expectancies at birth in 190 countries is skewed to the left, the median is the better choice for reporting the center of the distribution. The median is more resistant to the skewed shape of the distribution.

19. How big is your bicep II?

Because the distribution of bicep circumferences is unimodal and symmetric, the mean and the median should be very similar. The usual choice is to report the mean or to report both.

Section 2.5

21. Life expectancy III.

a) We should report the IQR.

b) Since the distribution of life expectancies at birth in 190 countries is skewed to the left, the better measure of spread is the IQR. The skewness of the distribution inflates the standard deviation.

23. How big is your bicep III?

Because the distribution of bicep circumferences is unimodal and roughly symmetric, we should report the standard deviation. The standard deviation is generally more useful whenever it is appropriate. However, it would not be strictly wrong to use the IQR. We just prefer the standard deviation.

Chapter Exercises

25. Graphs in the news. Answers will vary.

27. Tables in the news. Answers will vary.

29. Histogram. Answers will vary.

31. Centers in the news. Answers will vary.

33. Thinking about shape.

a) The distribution of the number of speeding tickets each student in the senior class of a college has ever had is likely to be unimodal and skewed to the right. Most students will have very few speeding tickets (maybe 0 or 1), but a small percentage of students will likely have comparatively many (3 or more?) tickets.

b) The distribution of player's scores at the U.S. Open Golf Tournament would most likely be unimodal and slightly skewed to the right. The best golf players in the game will likely have around the same average score, but some golfers might be off their game and score 15 strokes above the mean. (Remember that high scores are undesirable in the game of golf!)

c) The weights of female babies in a particular hospital over the course of a year will likely have a distribution that is unimodal and symmetric. Most newborns have about the same weight, with some babies weighing more and less than this average. There may be slight skew to the left, since there seems to be a greater likelihood of premature birth (and low birth weight) than post-term birth (and high birth weight).

d) The distribution of the length of the average hair on the heads of students in a large class would likely be bimodal and skewed to the right. The average hair length of the males would be at one mode, and the average hair length of the females would be at the other mode, since women typically have longer hair than men. The distribution would be skewed to the right, since it is not possible to have hair length less than zero, but it is possible to have a variety of lengths of longer hair.

35. Movie genres again.

a) Thriller/Suspense has a higher bar than Adventure, so it is the more common genre.

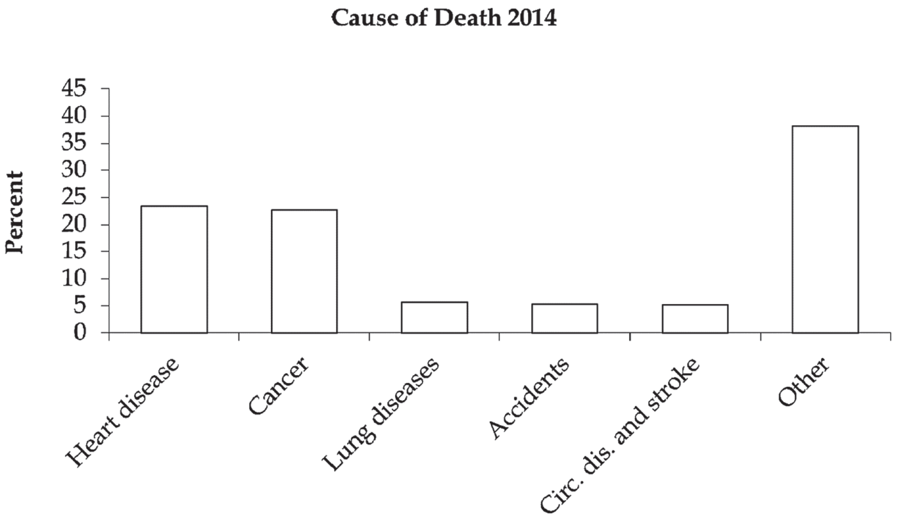
b) It is easy to tell from either chart; sometimes differences are easier to see on the bar chart because slices of the pie chart look too similar in size.

37. Magnet Schools.

There were 1755 qualified applicants for the Houston Independent School District's magnet schools program. 53% were accepted, 17% were wait-listed, and the other 30% were turned away for lack of space.

39. Causes of death 2014.

a) Yes, it is reasonable to assume that heart or lung diseases caused approximately 29% of U.S. deaths in 2014, since there is no possibility for overlap. Each person could only have one cause of death.



b) Since the percentages listed add up to 61.9%, other causes must account for 38.1% of US deaths.

c) A bar chart is a good choice (with the inclusion of the "Other" category). Since causes of US deaths represent parts of a whole, a pie chart would also be a good display.

41. Movie genres once more.

a) There are too many categories to construct an appropriate display. In a bar chart, there are too many bars. In a pie chart, there are too many slices. In each case, we run into difficulty trying to display genres that only represented a few movies.

b) The creators of the bar chart included a category called "Other" for many of the genres that only occurred a few times.

43. Global warming.

Perhaps the most obvious error is that the percentages in the pie chart add up to 141%, when they should, of course, add up to 100%. This means that survey respondents were allowed to choose more than one response, so a pie chart is not an appropriate display. Furthermore, the three-dimensional perspective view distorts the regions in the graph, violating the area principle. The regions corresponding to "Could reduce global warming but unsure if we will" and "Could reduce global warming but people aren't willing to so we won't" look roughly the same size, but at 46% and 30% of respondents, respectively, they should have very different sizes. Always use simple, two-dimensional graphs. Additionally, the graph does not include a title.

45. Cereals.

a) The distribution of the carbohydrate content of breakfast cereals is bimodal, with a cluster of cereals with carbohydrate content around 13 grams of carbs and another cluster of cereals around 22 grams of carbs. The lower cluster shows a bit of skew to the left. Most cereals in the lower cluster have between 10 and 20 grams of carbs. The upper cluster is symmetric, with cereals in the cluster having between 20 and 24 grams of carbs.

b) The cereals with the highest carbohydrate content are Corn Chex, Corn Flakes, Cream of Wheat (Quick), Crispix, Just Right Fruit & Nut, Kix, Nutri-Grain Almond-Raisin, Product 19, Rice Chex, Rice Krispies, Shredded Wheat ʻn' Bran, Shredded Wheat Spoon Size, Total Corn Flakes, and Triples.

47. Heart attack stays.

a) The distribution of length of stays is skewed to the right, so the mean is larger than the median.

b) The distribution of the length of hospital stays of female heart attack patients is bimodal and skewed to the right, with stays ranging from 1 day to 36 days. The distribution is centered around 8 days, with the majority of the hospital stays lasting between 1 and 15 days. There are a relatively few hospital stays longer than 27 days. Many patients have a stay of only one day, possibly because the patient died.

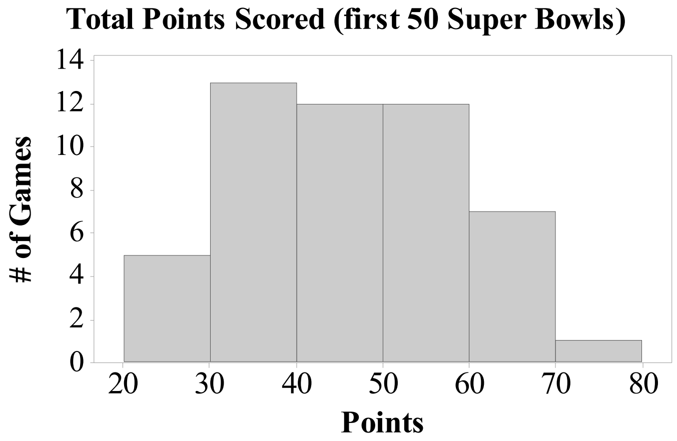
c) The median and IQR would be used to summarize the distribution of hospital stays, since the distribution is strongly skewed.

49. Super Bowl points 2016.

a) The median number of points scored in the first 50 Super Bowl games is 46 points.

b) The first quartile of the number of points scored in the first 50 Super Bowl games is 37 points. The third quartile is 55 points.

c) In the first 50 Super Bowl games, the lowest number of points scored was 21, and the highest number of points scored was 75. The median number of points scored was 46, and the middle 50% of Super Bowls has between 37 and 55 points scored, making the IQR 18 points.



51. Test scores, large class.

a) The distribution of Calculus test scores is bimodal with one mode at about 62 and one at about 78. The higher mode might be math majors, and the lower mode might be non-math majors.

b) Because the distribution of Calculus test scores is bimodal, neither the mean nor the median tells much about a typical score. We should attempt to learn if another variable (such as whether or not the student is a math major) can account for the bimodal character of the distribution.

53. Mistake.

a) As long as the boss's true salary of $200,000 is still above the median, the median will be correct. The mean will be too large, since the total of all the salaries will decrease by $2,000,000 - $200,000 = $1,800,000, once the mistake is corrected.

b) The range will likely be too large. The boss's salary is probably the maximum, and a lower maximum would lead to a smaller range. The IQR will likely be unaffected, since the new maximum has no effect on the quartiles. The standard deviation will be too large, because the $2,000,000 salary will have a large squared deviation from the mean.

55. Floods 2015.

a) The mean annual number of deaths from floods is 81.95.

b) In order to find the median and the quartiles, the list must be ordered.
29 38 38 43 48 49 56 68 76 80 82 82 82 86 87 103 113 118 131 136 176
The median annual number of deaths from floods is 82.
Quartile 1 = 49 deaths, and Quartile 3 = 103 deaths.
(Some statisticians consider the median to be separate from both the lower and upper halves of the ordered list when the list contains an odd number of elements. This changes the position of the quartiles slightly. If median is excluded, Q1 = 48.5, Q3 = 108. In practice, it rarely matters, since these measures of position are best for large data sets.)

c) The range of the distribution of deaths is Max – Min = 176 – 29 = 147 deaths.
The IQR = Q3 – Q1 = 103 – 49 = 54 deaths. (Or, the IQR = 108 – 48.5 = 59.5 deaths, if the median is excluded from both halves of the ordered list.)

57. Floods 2105 II.

The distribution of deaths from floods is slightly skewed to the right and bimodal. There is one mode at about 40 deaths and one at about 80 deaths. There is one extreme value at 180 deaths.

59. Pizza prices.

The mean and standard deviation would be used to summarize the distribution of pizza prices, since the distribution is unimodal and symmetric.

61. Pizza prices again.

a) The mean pizza price is closest to $2.60. That's the balancing point of the histogram.

b) The standard deviation in pizza prices is closest to $0.15, since that is the typical distance to the mean. There are no pizza prices as far as $0.50 or $1.00.

63. Movie lengths 2010.

a) A typical movie would be around 105 minutes long. This is near the center of the unimodal and slightly skewed histogram, with the outlier set aside.

b) You would be surprised to find that your movie ran for 150 minutes. Only 3 movies ran that long.

c) The mean run time would probably be higher, since the distribution of run times is skewed to the right, and also has a high outlier. The mean is pulled towards this tail, while the median is more resistant. However, it is difficult to predict what the effect of the low outlier might be from just looking at the histogram.

65. Movie lengths 2010 II.

a) i) The distribution of movie running times is fairly consistent, with the middle 50% of running times between 98 and 116 minutes. The interquartile range is 18 minutes.

ii) The standard deviation of the distribution of movie running times is 16.6 minutes, which indicates that movies typically varied from the mean running time by 16.6 minutes.

b) Since the distribution of movie running times is skewed to the right and contains an outlier, the standard deviation is a poor choice of numerical summary for the spread. The interquartile range is better, since it is resistant to outliers.

67. Movie budgets.

The industry publication is using the median, while the watchdog group is using the mean. It is likely that the mean is pulled higher by a few very expensive movies.

69. Gasoline 2014.

a) Gasoline Prices

31	1
31	5
32	1233
32	6678
33	
33	9
34	23
34	556

Key : 32 | 1 = $3.21/gal

b) The distribution of gas prices is bimodal, with two clusters, one centered around $3.45 per gallon, and another centered around $3.25 per gallon. The lowest and highest prices were $3.11 and $3.46 per gallon.

c) There is a gap in the distribution of gasoline prices. There were no stations that charged between $3.28 and $3.39.

71. States.

a) There are 50 entries in the stemplot, so the median must be between the 25th and 26th population values. Counting in the ordered stemplot gives median = 4.5 million people. The middle of the lower 50% of the list (25 state populations) is the 13th population, or 2 million people. The middle of the upper half of the list (25 state populations) is the 13th population from the top, or 7 million people. The IQR = Q3 – Q1 = 7 – 2 = 5 million people.

b) The distribution of population for the 50 U.S. States is unimodal and skewed heavily to the right. The median population is 4.5 million people, with 50% of states having populations between 2 and 7 million people. There are two outliers, a state with 37 million people, and a state with 25 million people. The next highest population is only 19 million.

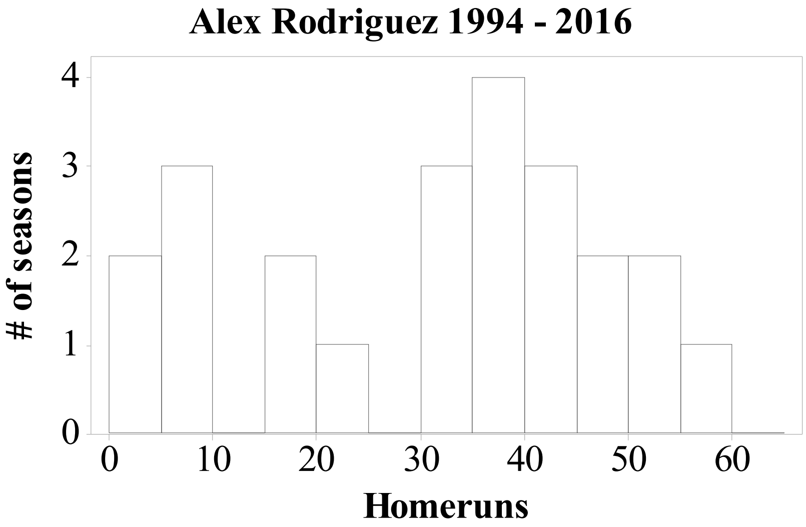
73. A-Rod 2016.

The distribution of the number of homeruns hit by Alex Rodriguez during the 1994 – 2016 seasons is reasonably symmetric, with the exception of a second mode around 10 homeruns. A typical number of homeruns per season was in the high 30s to low 40s. With the exception of 5 seasons in which A-Rod hit 0, 0 , 5, 7, and 9 homeruns, his total number of homeruns per season was between 16 and the maximum of 57.

75. A-Rod again 2016.

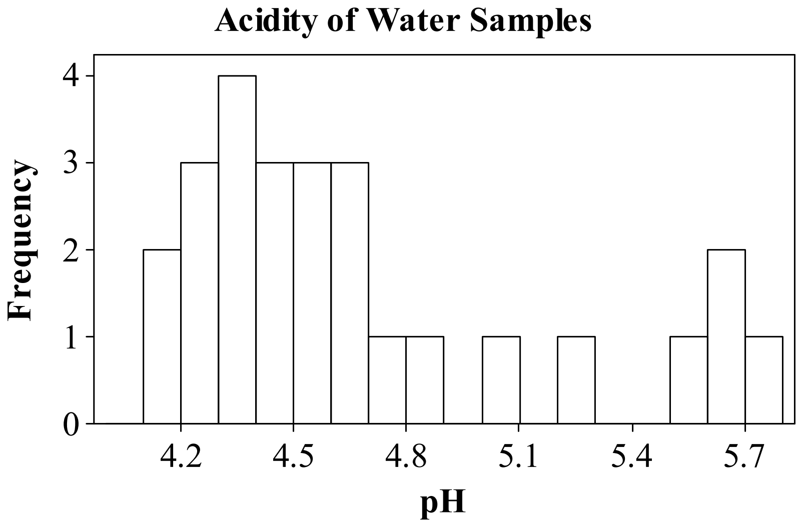
a) This is not a histogram. The horizontal axis should contain the number of home runs per year, split into bins of a convenient width. The vertical axis should show the frequency; that is, the number of years in which A-Rod hit a number of home runs within the interval of each bin. The display shown is a bar chart/time plot hybrid that simply displays the data table visually. It is of no use in describing the shape, center, spread, or unusual features of the distribution of home runs hit per year by A-Rod.

b) The histogram is at the right.



77. Acid rain.

a) The distribution of the pH readings of water samples in Allegheny County, Penn. is bimodal. A roughly uniform cluster is centered around a pH of 4.4. This cluster ranges from pH of 4.1 to 4.9. Another smaller, tightly packed cluster is centered around a pH of 5.6. Two readings in the middle seem to belong to neither cluster.



b) The cluster of high outliers contains many dates that were holidays in 1973. Traffic patterns would probably be different then, which might account for the difference.

79. Final grades.

The width of the bars is much too wide to be of much use. The distribution of grades is skewed to the left, but not much more information can be gathered.

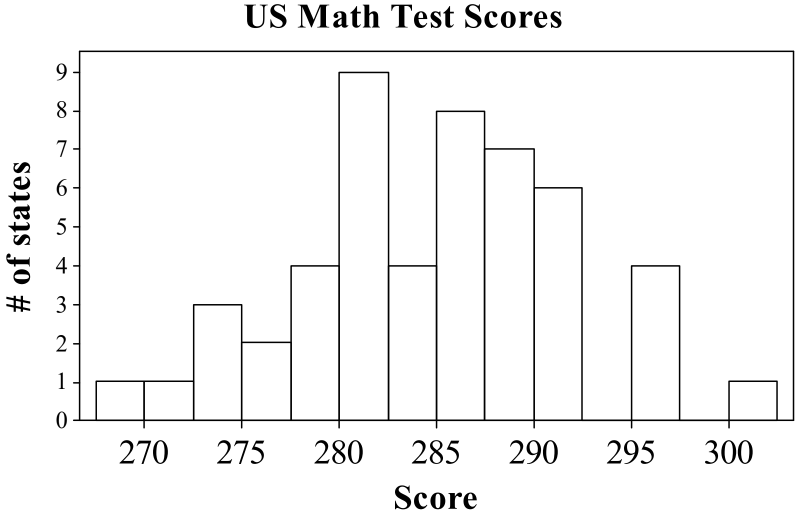
81. Zip codes.

Even though zip codes are numbers, they are not quantitative in nature. Zip codes are categories. A histogram is not an appropriate display for categorical data. The histogram the Holes R Us staff member displayed doesn’t take into account that some 5-digit numbers do not correspond to zip codes or that zip codes falling into the same classes may not even represent similar cities or towns. The employee could design a better display by constructing a bar chart that groups together zip codes representing areas with similar demographics and geographic locations.

83. Math scores 2013.

a) Median: 285
IQR: 9
Mean: 284.36
Standard deviation: 6.84

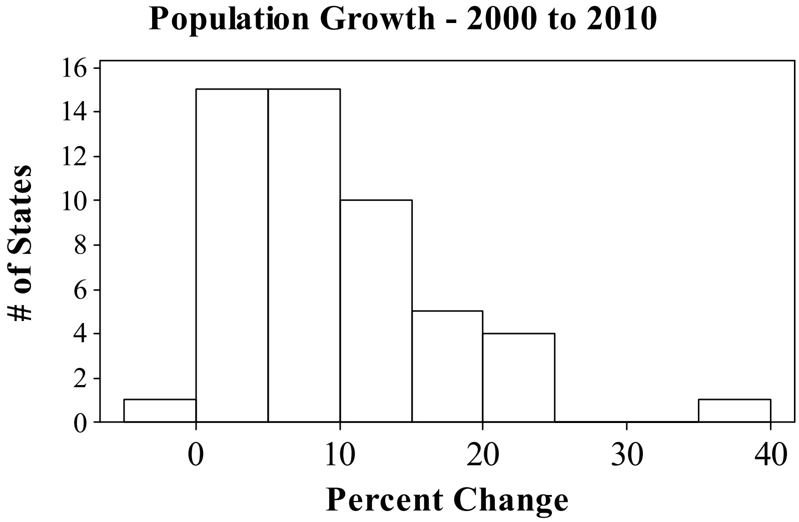
b) Since the distribution of Math scores is skewed to the left, it is probably better to report the median and IQR.



c) The distribution of average math achievement scores for eighth graders in the United States is skewed slightly to the left, and roughly unimodal. The distribution is centered at 285. Scores range from 269 to 301, with the middle 50% of the scores falling between 280 and 289.

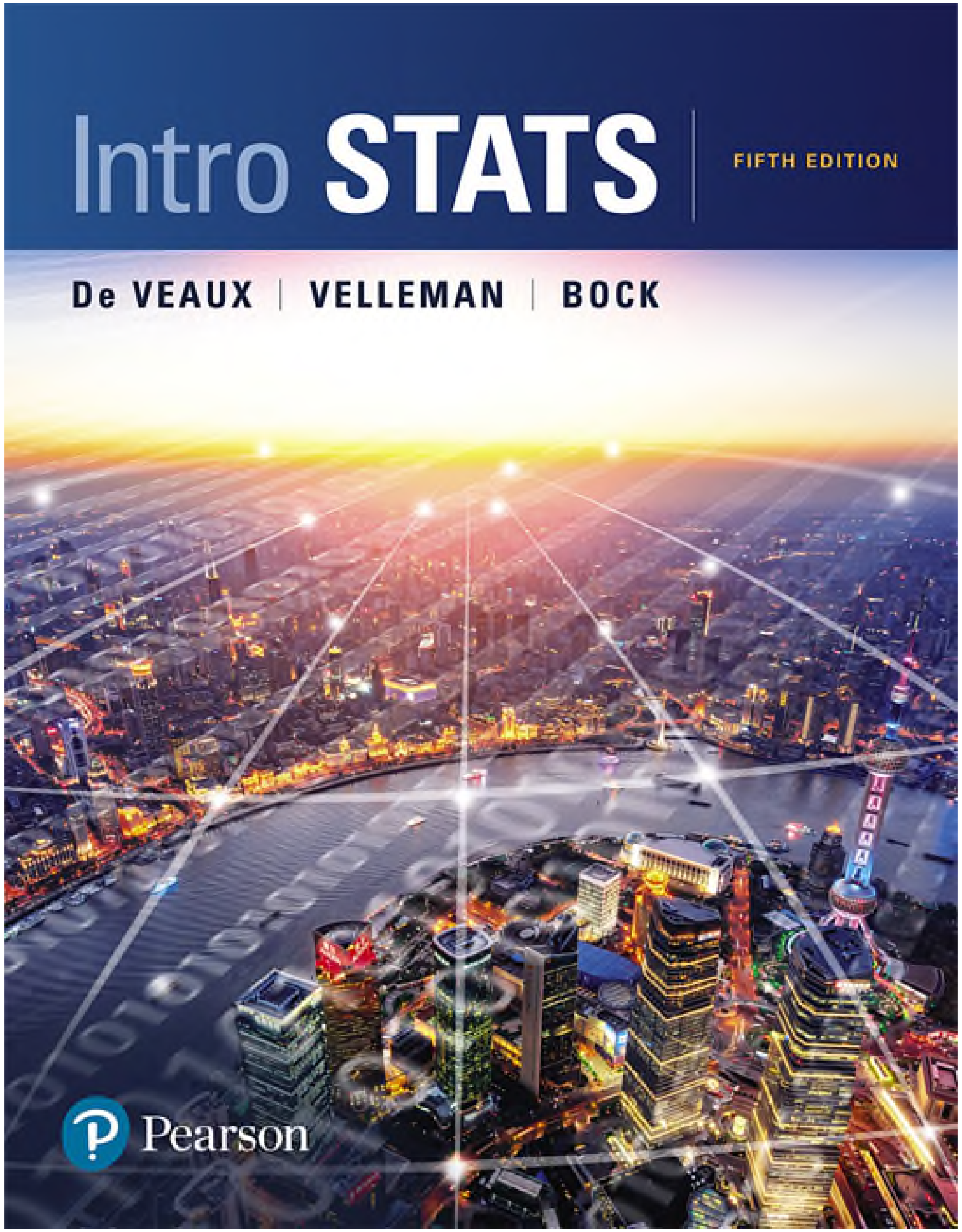
85. Population growth 2010.

The distribution of population growth among the 50 United States and the District of Columbia is unimodal and skewed to the right. Most states experienced modest growth, as measured by percent change in population between 2000 and 2010. Nearly every state experienced positive growth, with the exception of Michigan. The median population growth was 7.8%, with the middle 50% of states experiencing between 4.30% and 14.10% growth, for an IQR of 9.80. The distribution contains one high outlier. Nevada experienced population growth of 35.1%.



Chapter 2

Displaying and Describing Data

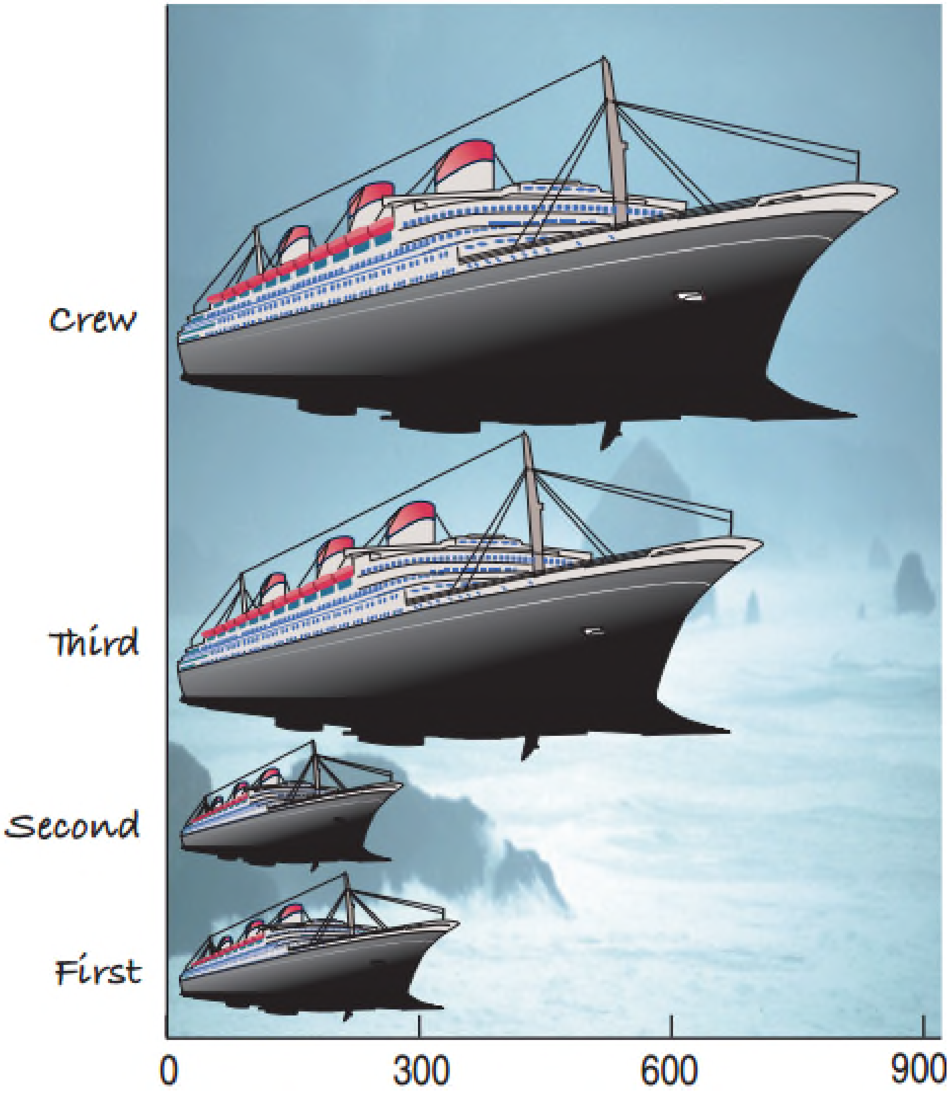


Three Rules of Data Analysis

- Make a Picture: Helps you *Think* clearly about patterns and relationships hidden in the data table.
- Make a Picture: *Shows* the important features of the data.
- Make a Picture: *Tells* others about the data.

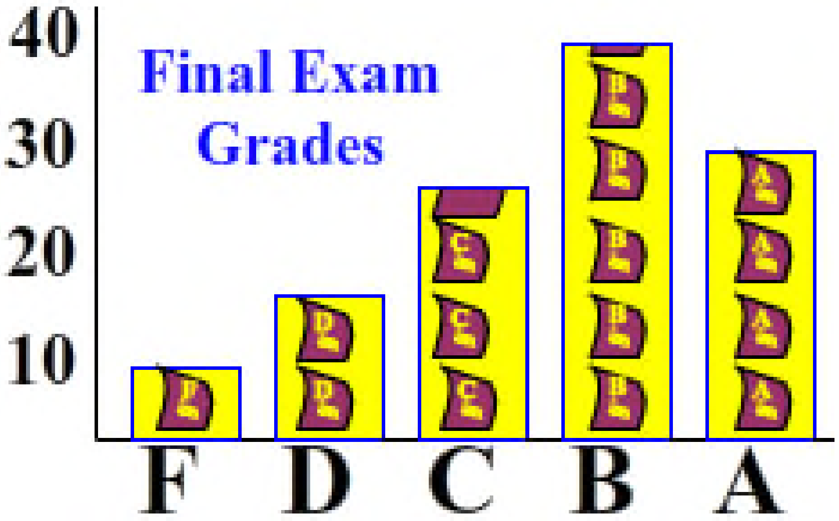
A Titanic Misconception

- Were most members of the Titanic crew members?
- Three times as many crew members as second-class passengers
- The eyes are tricked by the area being nine times as large for the crew.



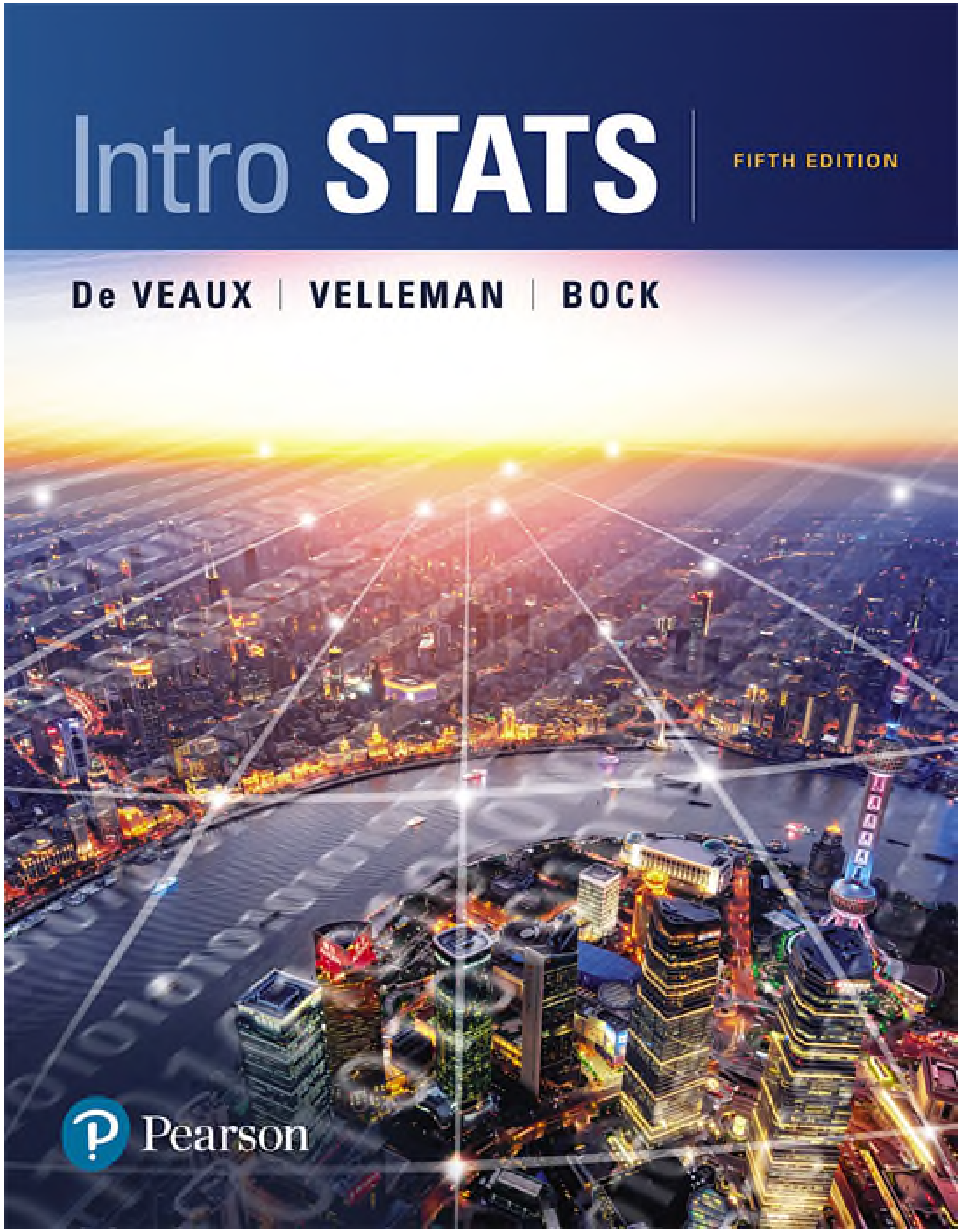
The Area Principle

- The Area Principle: The area occupied by a part of the graph should correspond to the magnitude of the value it represents.
  - Bars should have equal widths in a bar chart.
  - Be cautious when using two-dimensional pictures to exhibit one-dimensional data.



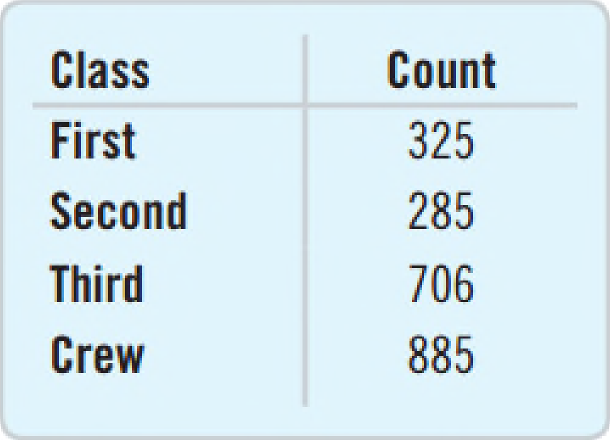
2.1

Summarizing and Displaying a Categorical Variable



Frequency Tables

- A frequency table is a table whose first column displays each distinct outcome and second column displays that outcome's frequency.



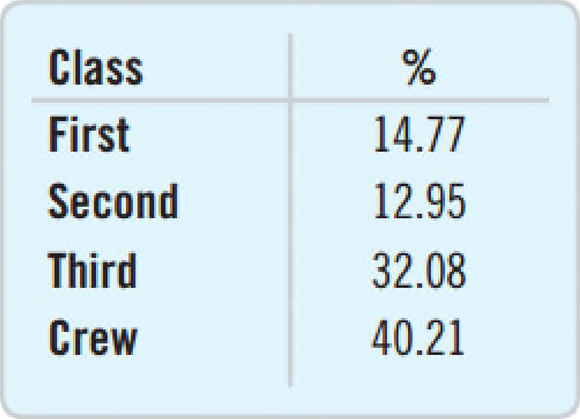
Class	Count
First	325
Second	285
Third	706
Crew	885

- If there are many distinct outcomes, then combining them into a few categories is recommended.

Relative Frequency Tables

- A relative frequency table is a table whose first column displays each distinct outcome and second column displays that outcome's relative frequency.

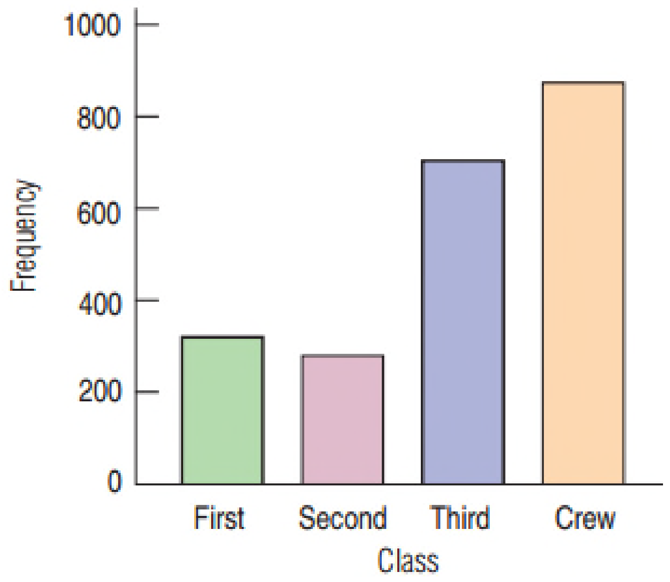
- The relative frequency table is similar to the frequency table, but it displays relative frequencies rather than frequencies.



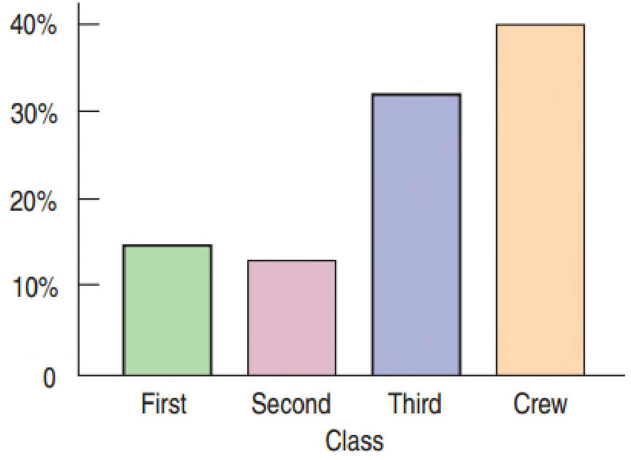
Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Bar Charts

- A bar chart displays the frequency or relative frequency of each category.
- All bars must have the same width.
- Good for general audience

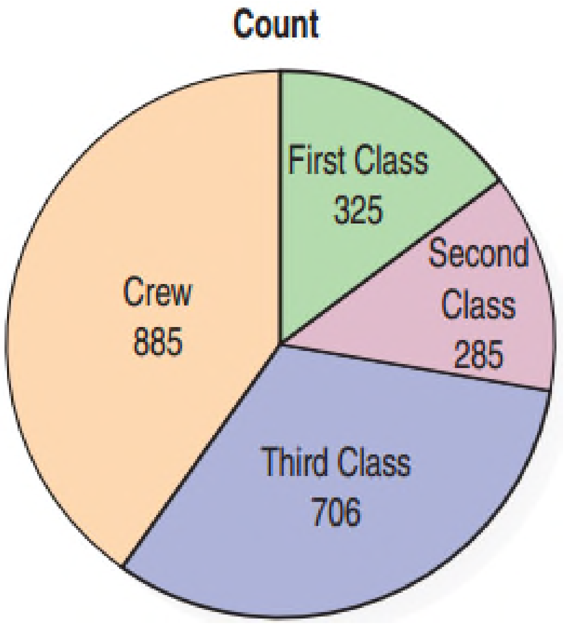


Frequency Bar Chart



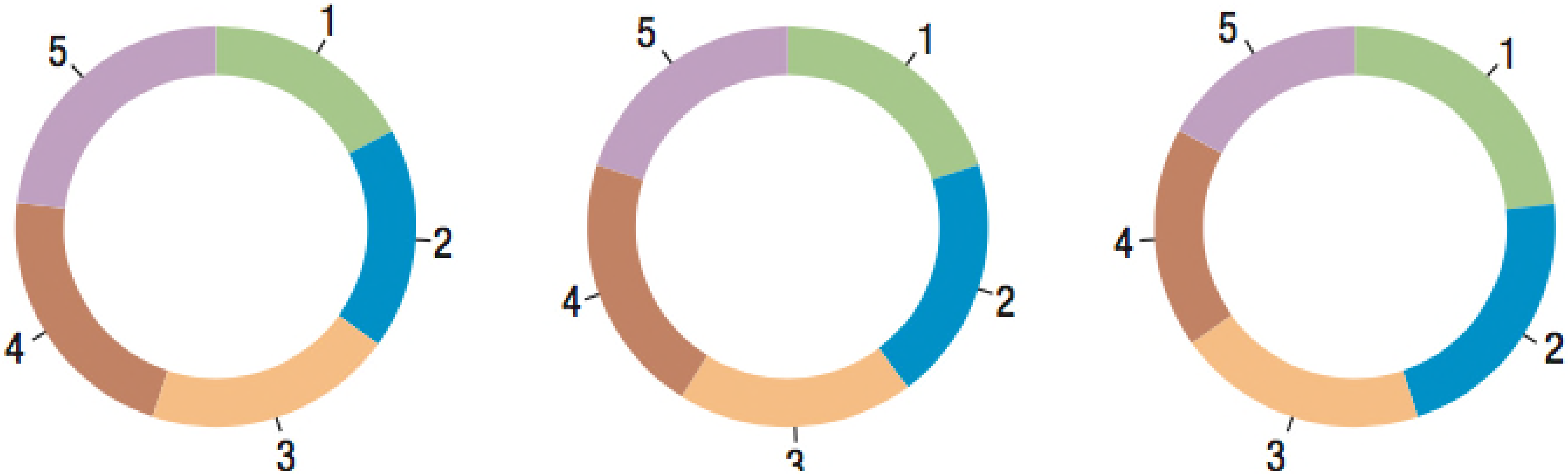
Relative Frequency Bar Chart

Pie Charts



- A pie chart presents each category as a slice of a circle so that each slice has a size that is proportional to the whole in each category.
- Pie charts are also good for a general audience.
- Pie charts help to display the fraction of the whole that each category represents.

Ring Charts



- A ring chart presents each category as a partition of a ring that is proportional in area to the value of each category.

- Ring charts are also good for a general audience.

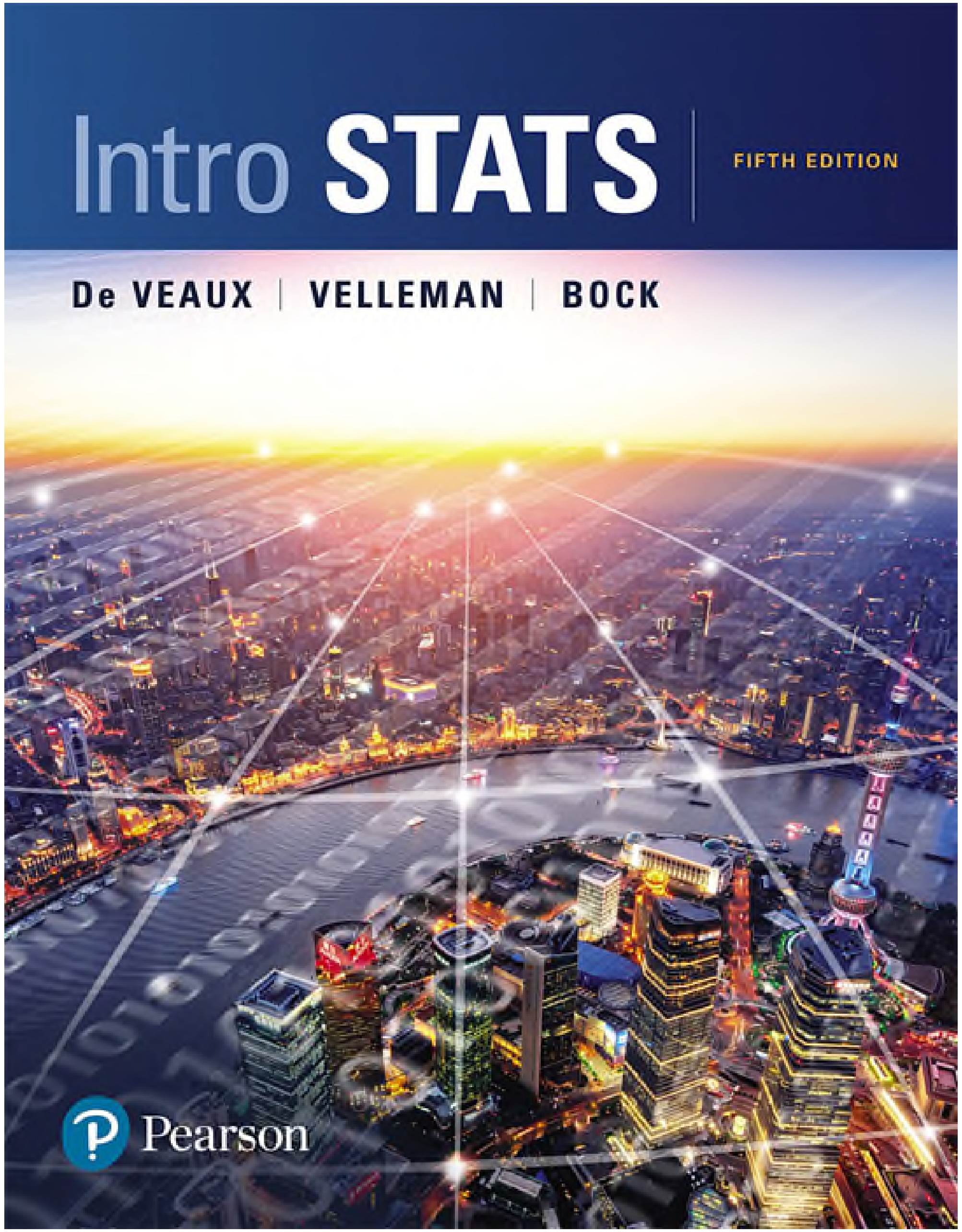
- Ring charts may be easier to read.

Think Before You Draw

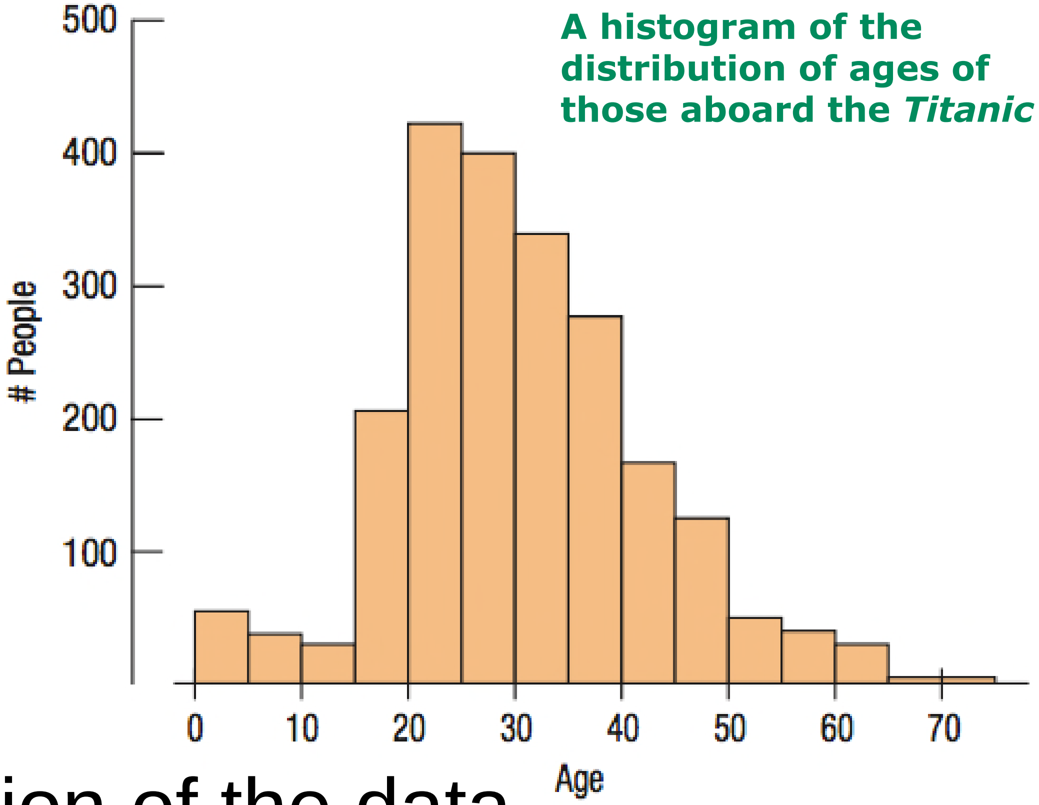
- Choose the chart that best tells the story of your data.
- Think about the intended audience to select a chart that is best for them.
- Charts often work better when the categories do not overlap.
- Don’t try to fool your audience, just give a chart that honestly expresses the interesting features of the data.

2.2

Displaying a Quantitative Variable



Histograms

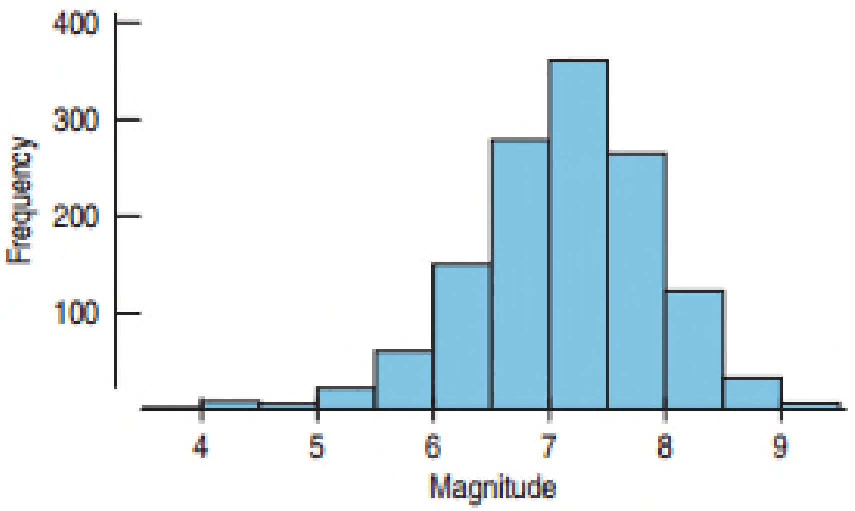


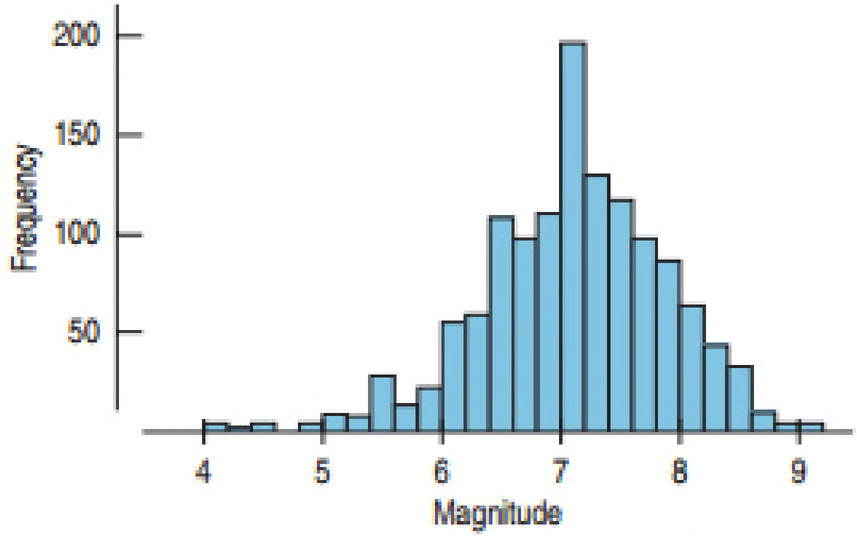
A histogram of the distribution of ages of those aboard the *Titanic*

- Histogram: A chart that displays quantitative data
- Great for seeing the distribution of the data
- Most populous age group 20- to 24 year-olds
- Youngest were infants. Oldest were over 70.
- Fewer and fewer people in the advancing ages, above 25
- More infants and toddlers than pre-teens

Choosing the Bin Width

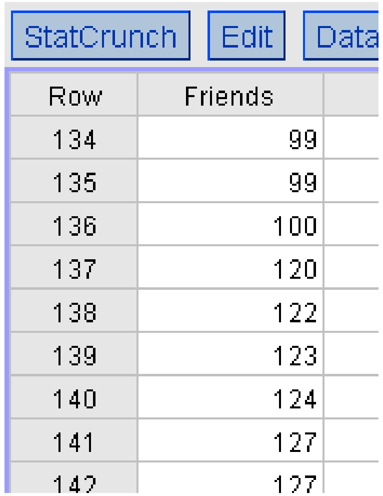
- Different bin widths tell different stories.
- Choose the width that best shows the important features.
- Presentations can feature two histograms that present the same data in different ways.
- A gap in the histogram means that there were no occurrences in that range.



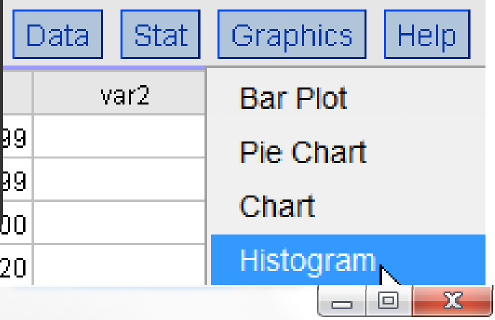


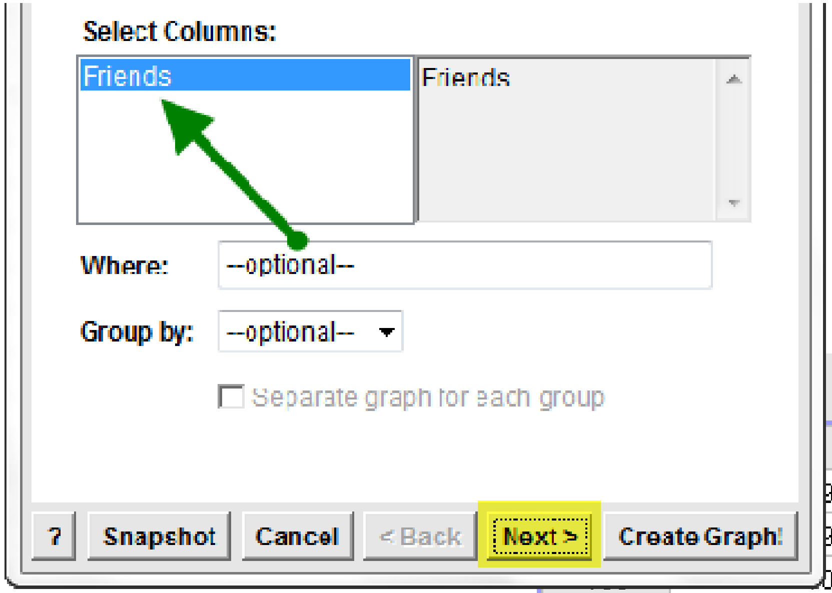
Histograms and StatCrunch

- Enter Data.
- Graphics → Histogram
- Click on the data variable and Next.
- Select Frequency or Relative Frequency.
- Put in starting value and/or Binwidth if desired.
- Click Next twice, and type in labels. Click Create Graph.



Row	Friends
134	99
135	99
136	100
137	120
138	122
139	123
140	124
141	127
142	127



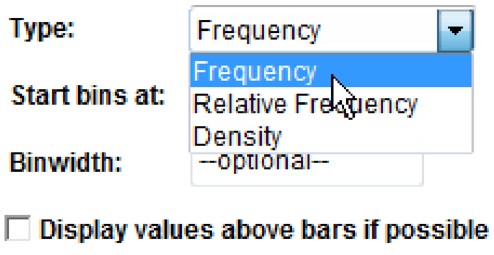

Select Columns: Friends

Friends

Where: --optional--

Group by: --optional--

□ Separate graph for each group

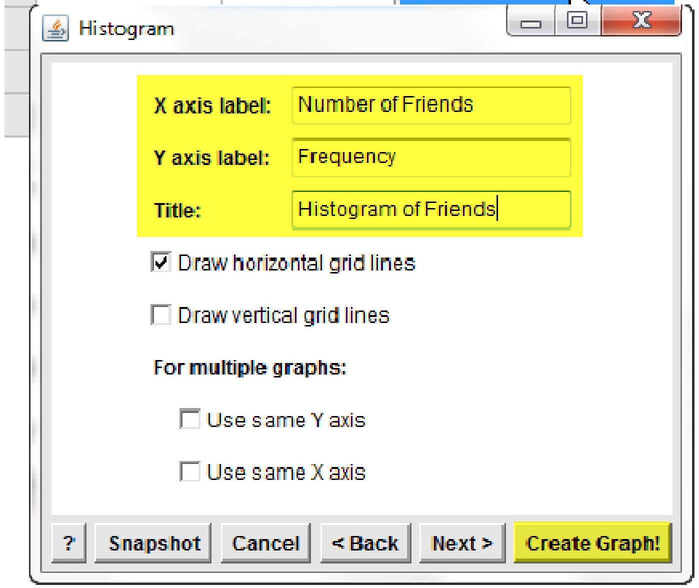

Type: Frequency

Frequency
Relative Frequency
Density

Start bins at:

Binwidth: --optional--

□ Display values above bars if possible


Histogram

X axis label: Number of Friends

Y axis label: Frequency

Title: Histogram of Friends

☑ Draw horizontal grid lines

□ Draw vertical grid lines

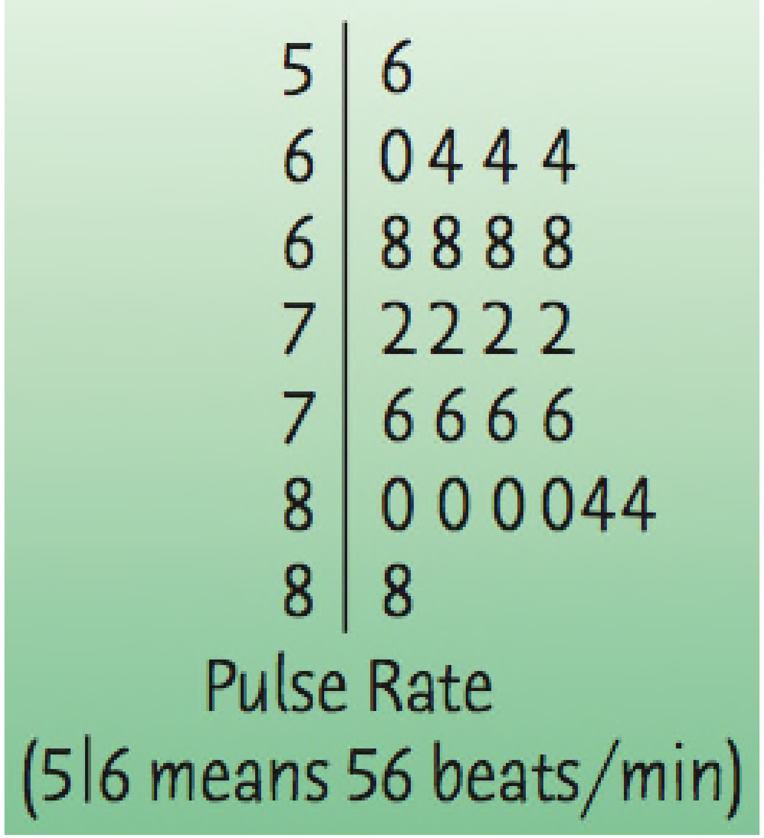
For multiple graphs:

□ Use same Y axis

□ Use same X axis

Stem-and-Leaf Displays

- Stem-and-Leaf: Shows both the shape of the distribution and all of the individual values
- Not as visually pleasing as a histogram; more technical looking



5	6
6	0 4 4 4
6	8 8 8 8
7	2 2 2 2
7	6 6 6 6
8	0 0 0 0 4 4
8	8

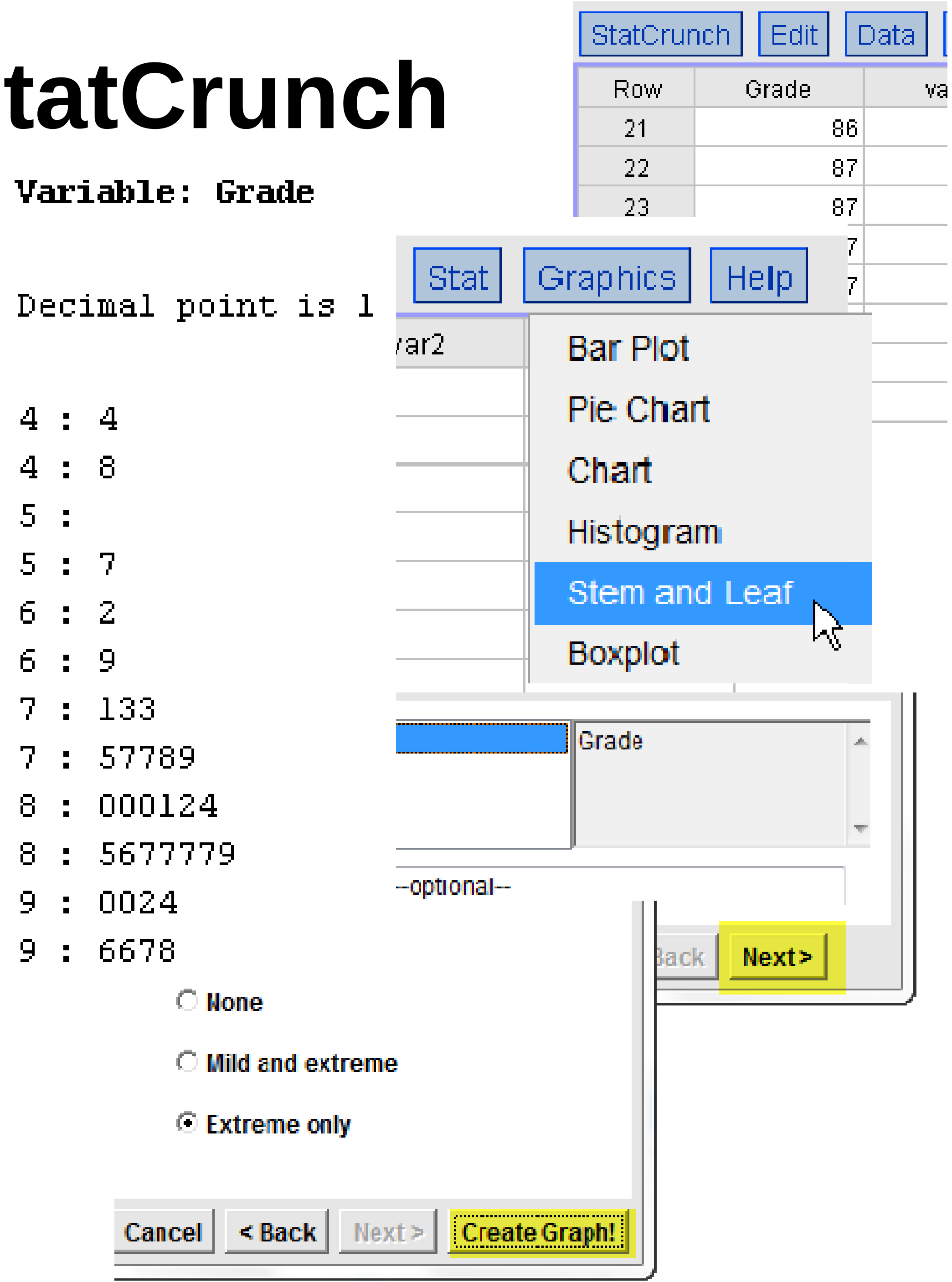
Pulse Rate

(5|6 means 56 beats/min)

- Can only be used for small collections of data
- The first column (stems) represents the leftmost digit.
- The second column (leaves) shows the remaining digit(s).

Stem and Leaf with StatCrunch

- Enter Data
- Graphics → Stem and Leaf
- Click on the variable name and Next
- Select Outlier Trimming Type and Create Graph!



```
Variable: Grade

Decimal point is 1

4 : 4
4 : 8
5 :
5 : 7
6 : 2
6 : 9
7 : 133
7 : 57789
8 : 000124
8 : 5677779
9 : 0024
9 : 6678
```

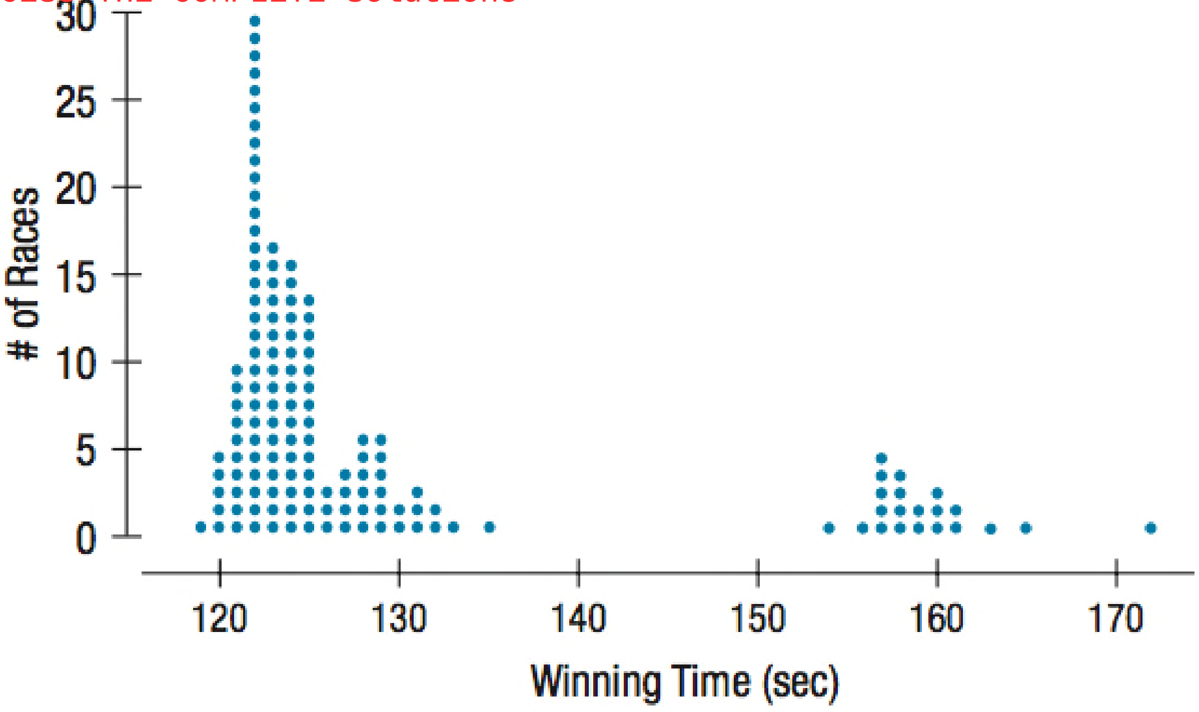
Row	Grade	va
21	86	
22	87	
23	87	

○ None

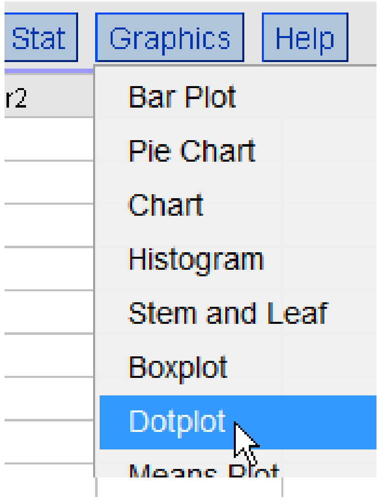
○ Mild and extreme

● Extreme only

Dotplots

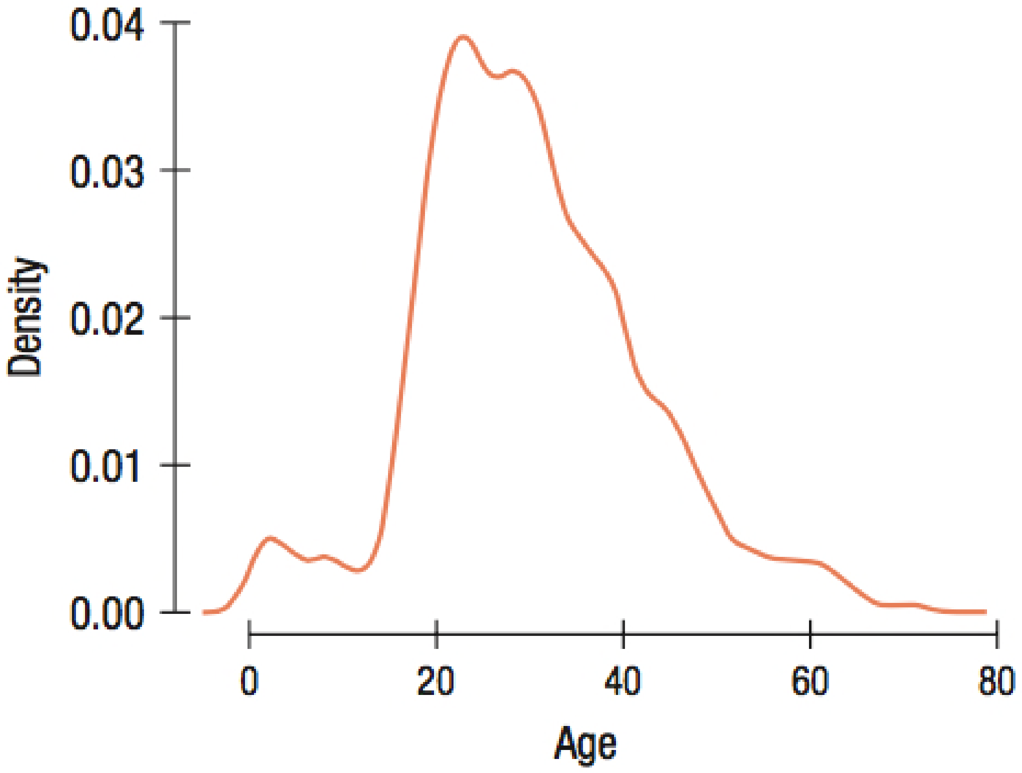


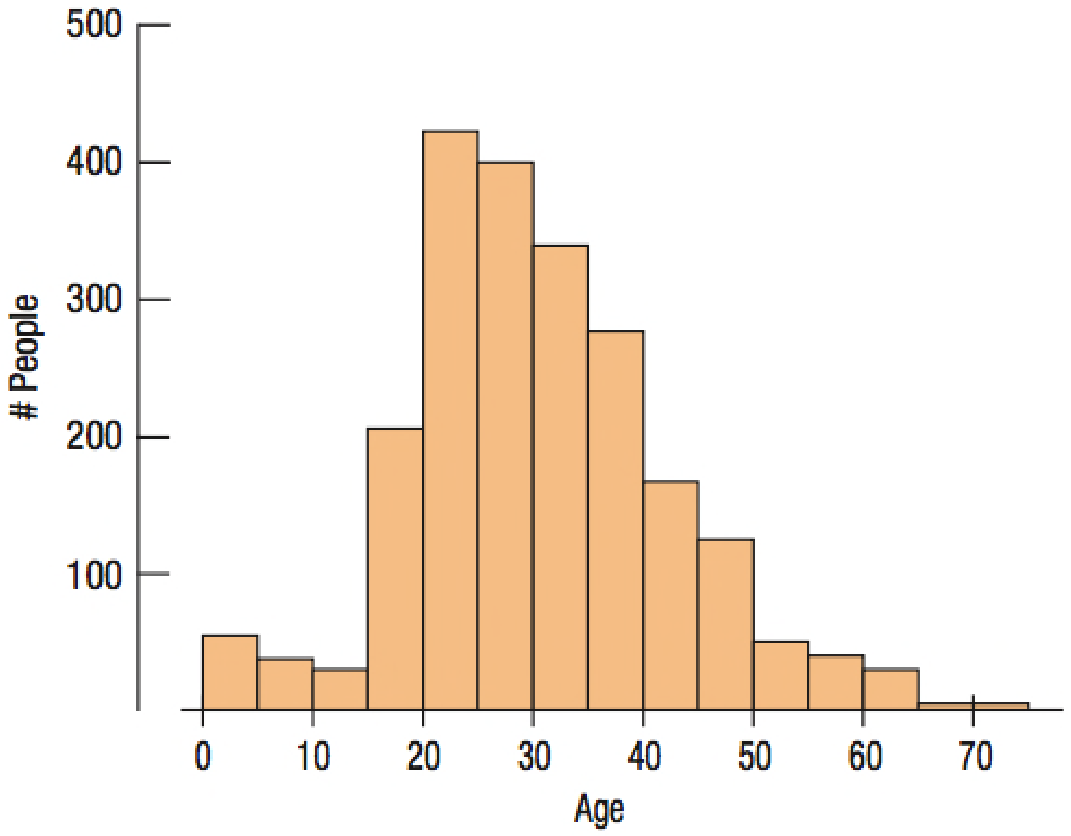
- Dotplot: Displays dots to describe the shape of the distribution
- There were 30 races with a winning time of 122 seconds.
- Good for smaller data sets
- Visually more appealing than stem-and-leaf
- In StatCrunch: Graphics → Dotplot



Density Plots

- Density Plots: smooth the bins in a histogram
- Do not provide hard cut-offs to the bins

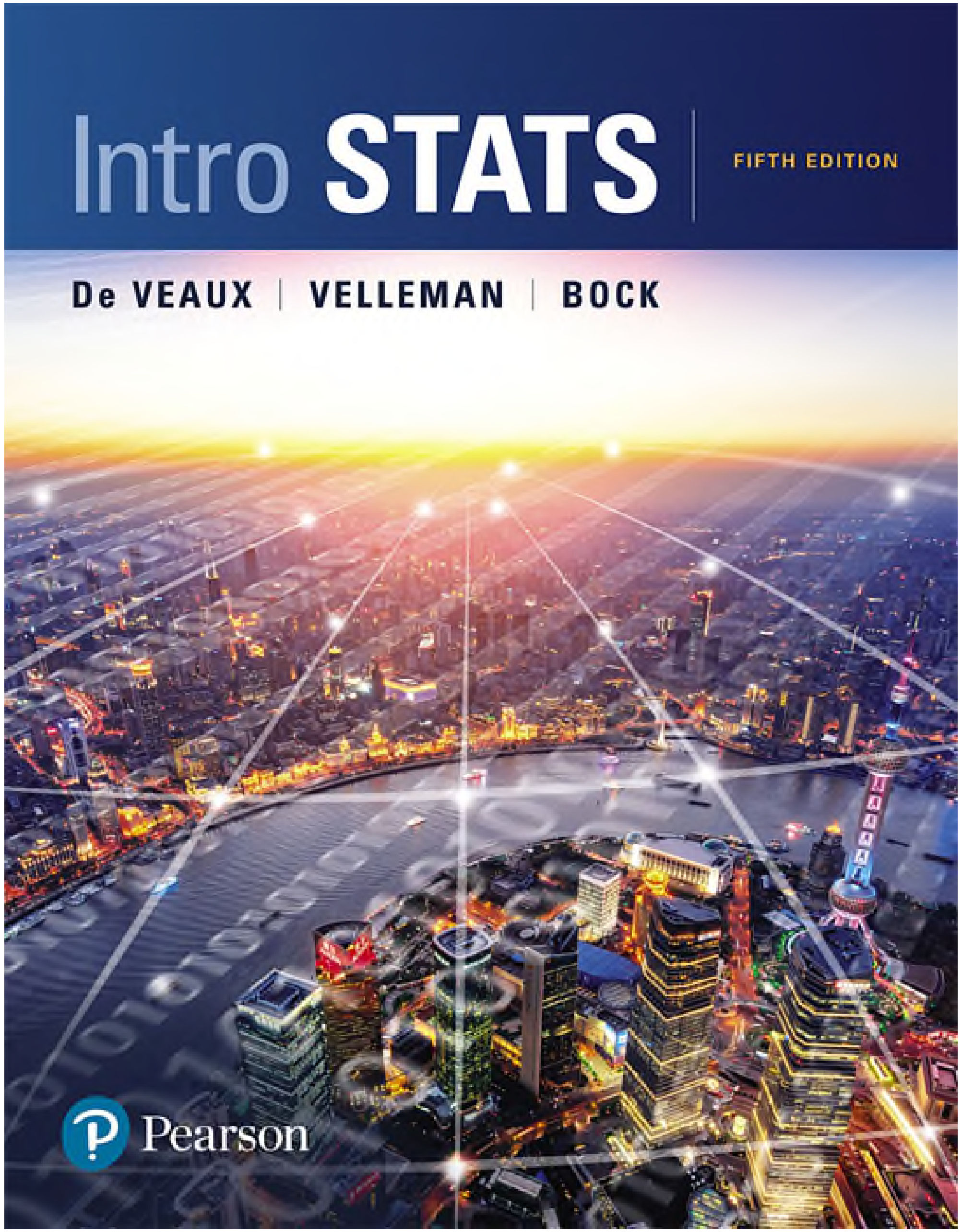




Think Before you Draw

- Shape, Center, and Spread
- Is the variable quantitative? Is the answer to the survey question or result of the experiment a number whose units are known?
- Bar and pie charts display categorical data.
- Histograms, stem-and-leaf diagrams, and dotplots can only display quantitative data.

2.3

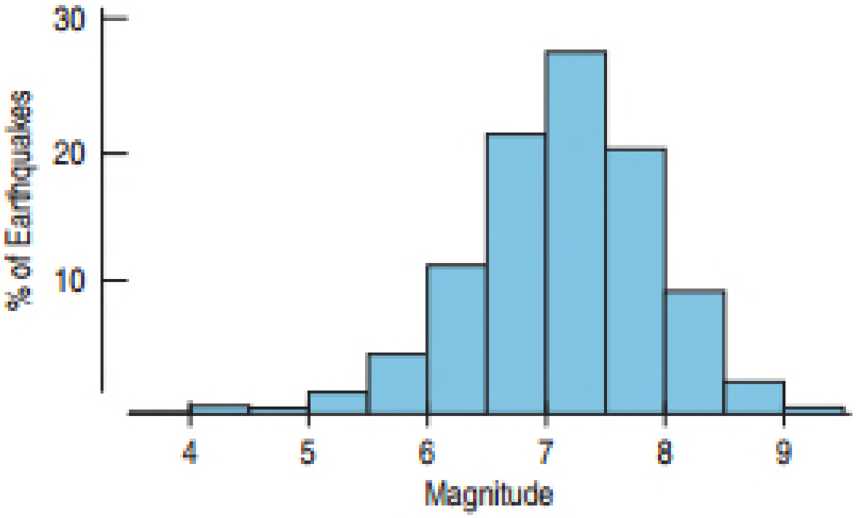


Shape

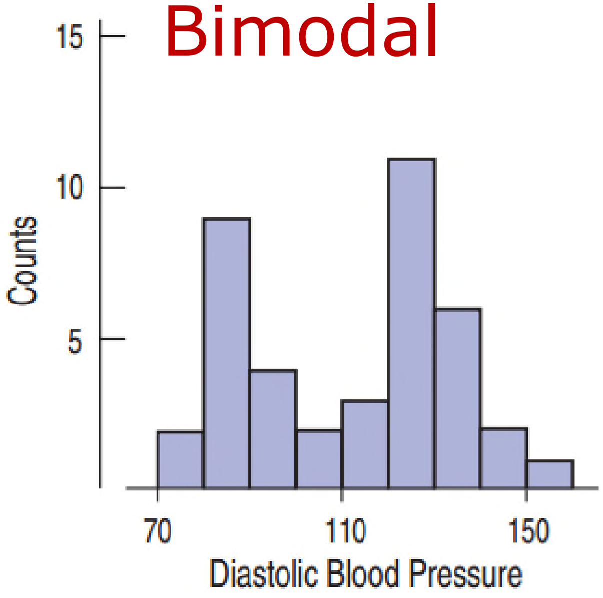
Modes

- A Mode of a histogram is a hump or high-frequency bin.
  - One mode → Unimodal
  - Two modes → Bimodal
  - 3 or more → Multimodal

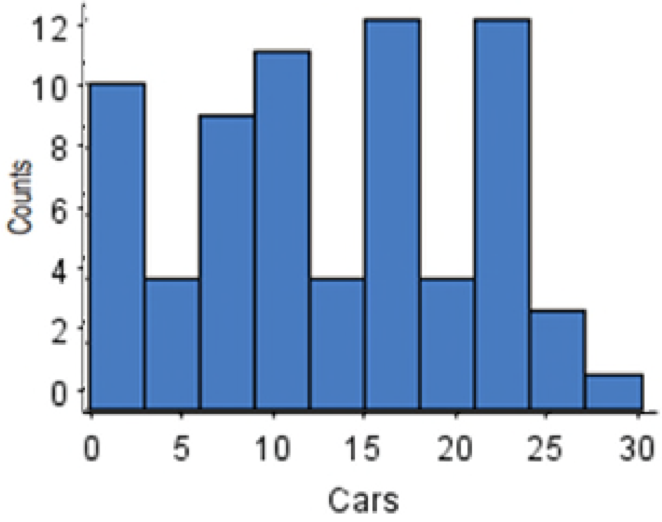
Unimodal



Bimodal

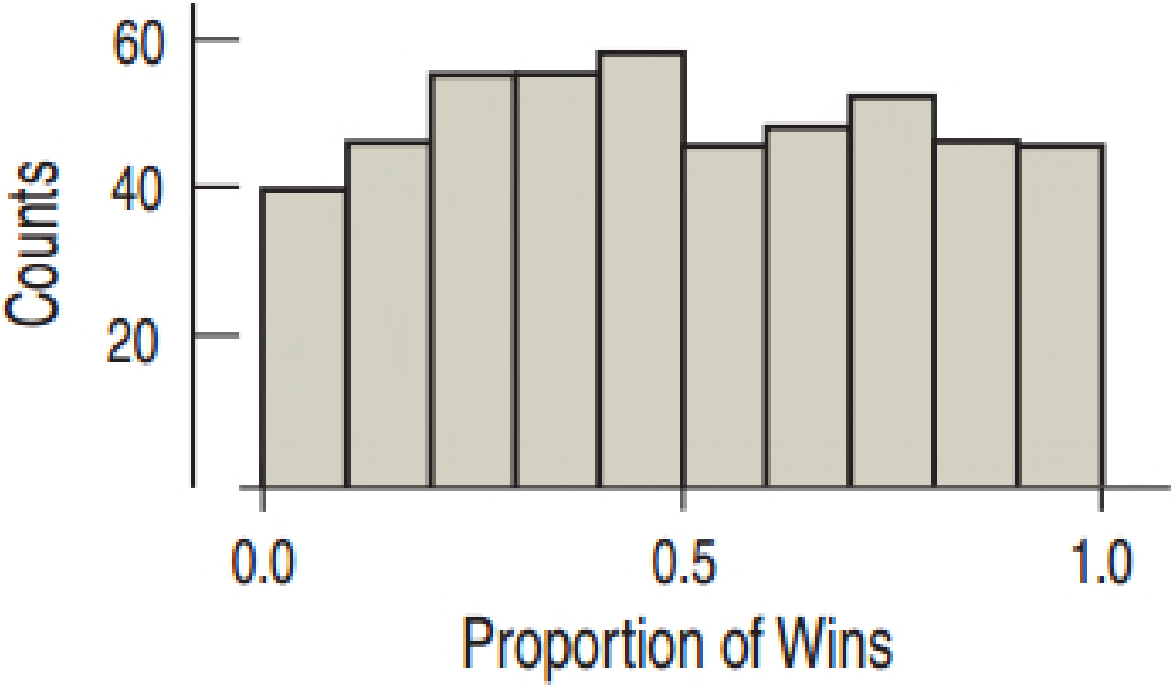


Multimodal



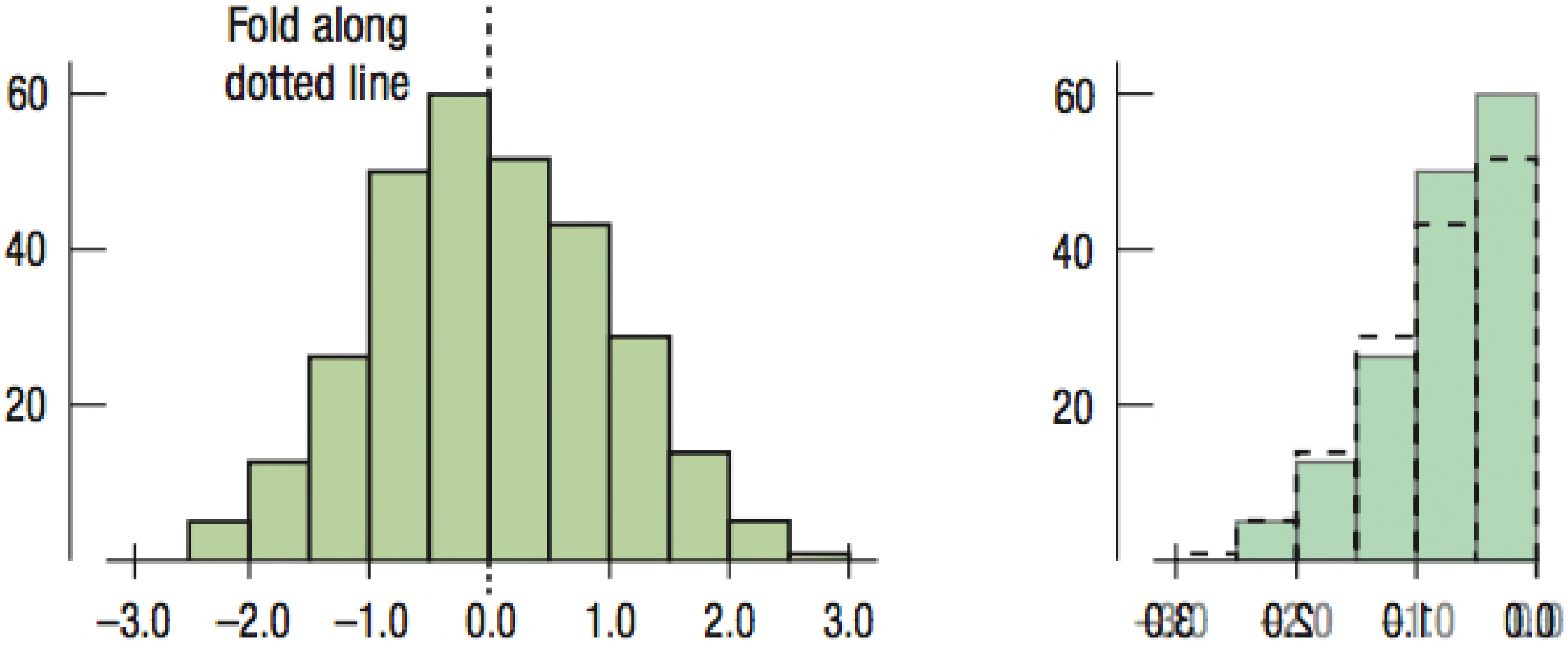
Uniform Distributions

- Uniform Distribution: All the bins have the same frequency, or at least close to the same frequency.
- The histogram for a uniform distribution will be nearly **flat**.



Symmetry

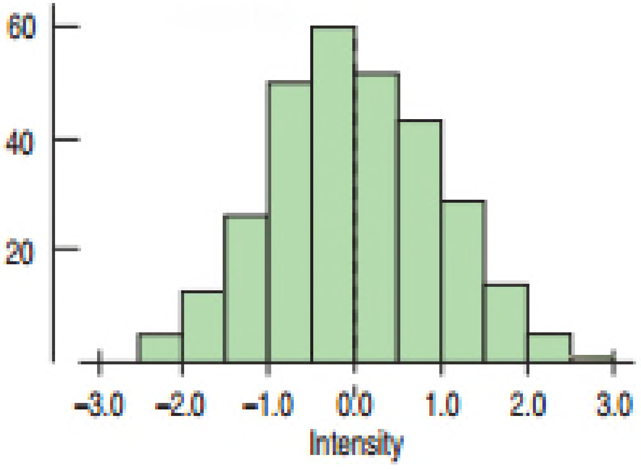
- The histogram for a symmetric distribution will look the same on the left and the right of its center.



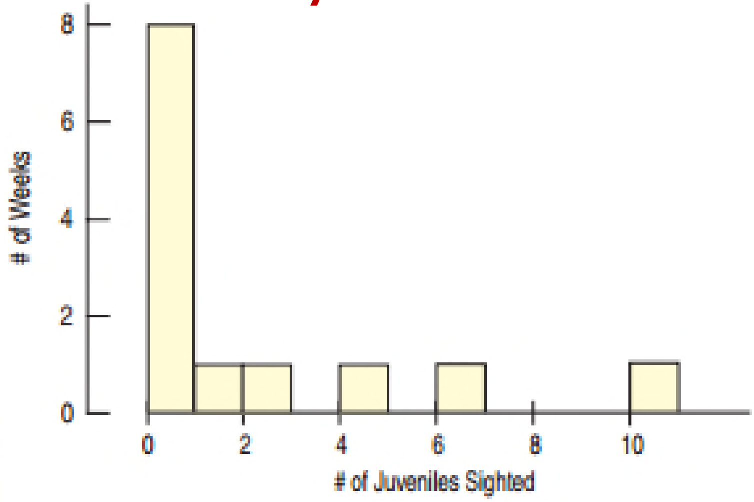
Symmetry

- A symmetric histogram is not necessarily bell-shaped.

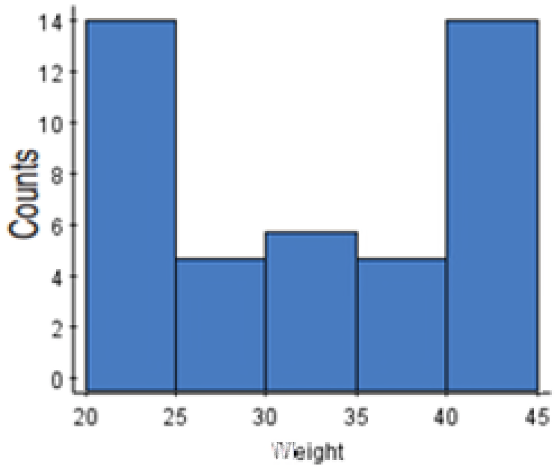
Symmetric



Not Symmetric

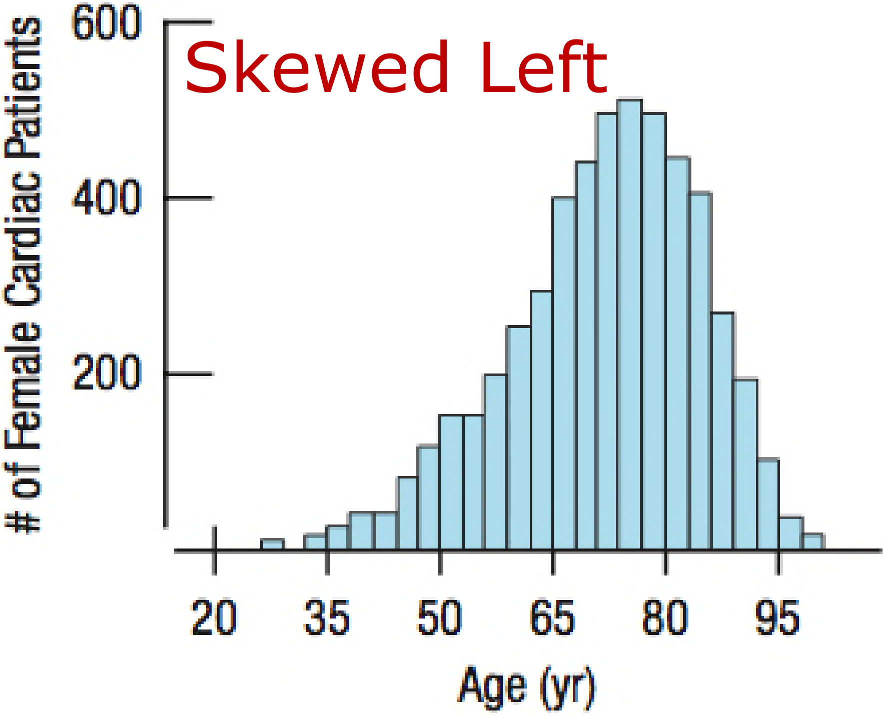


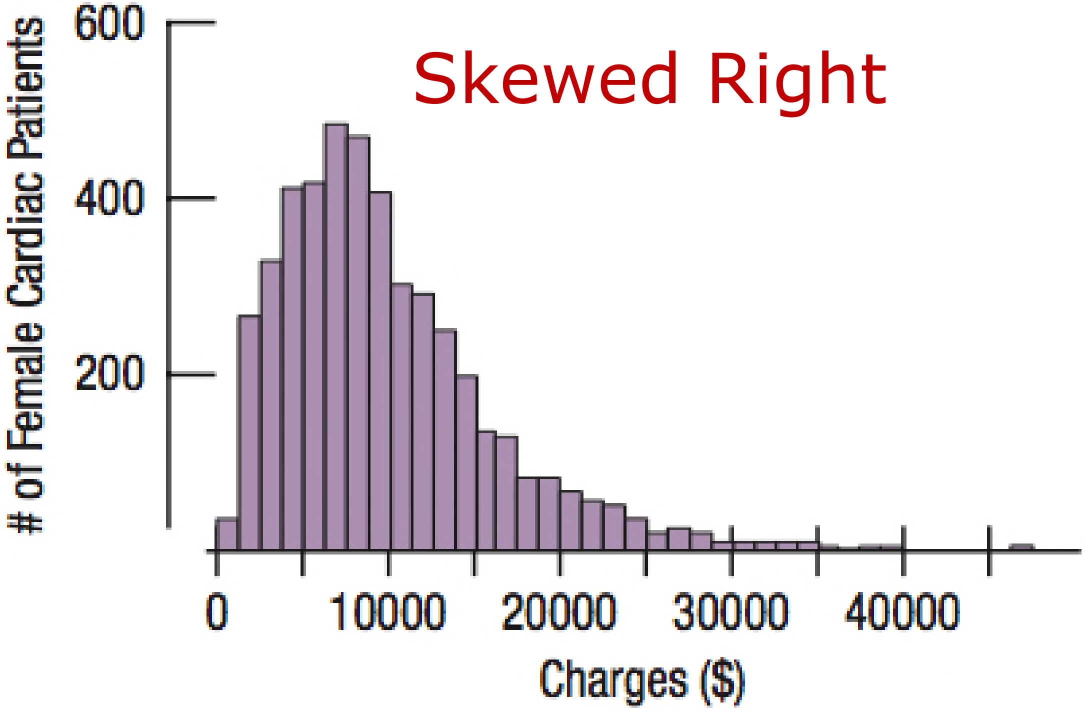
Symmetric



Skew

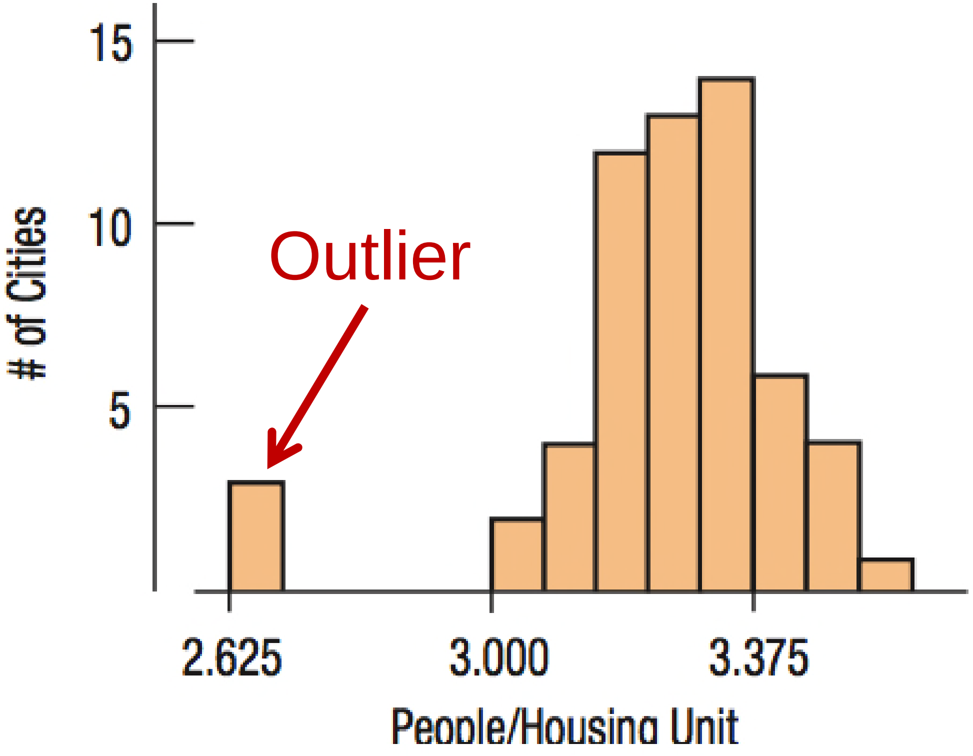
- A histogram is skewed right if the longer tail is on the right side of the mode.
- A histogram is skewed left if the longer tail is on the left side of the mode.





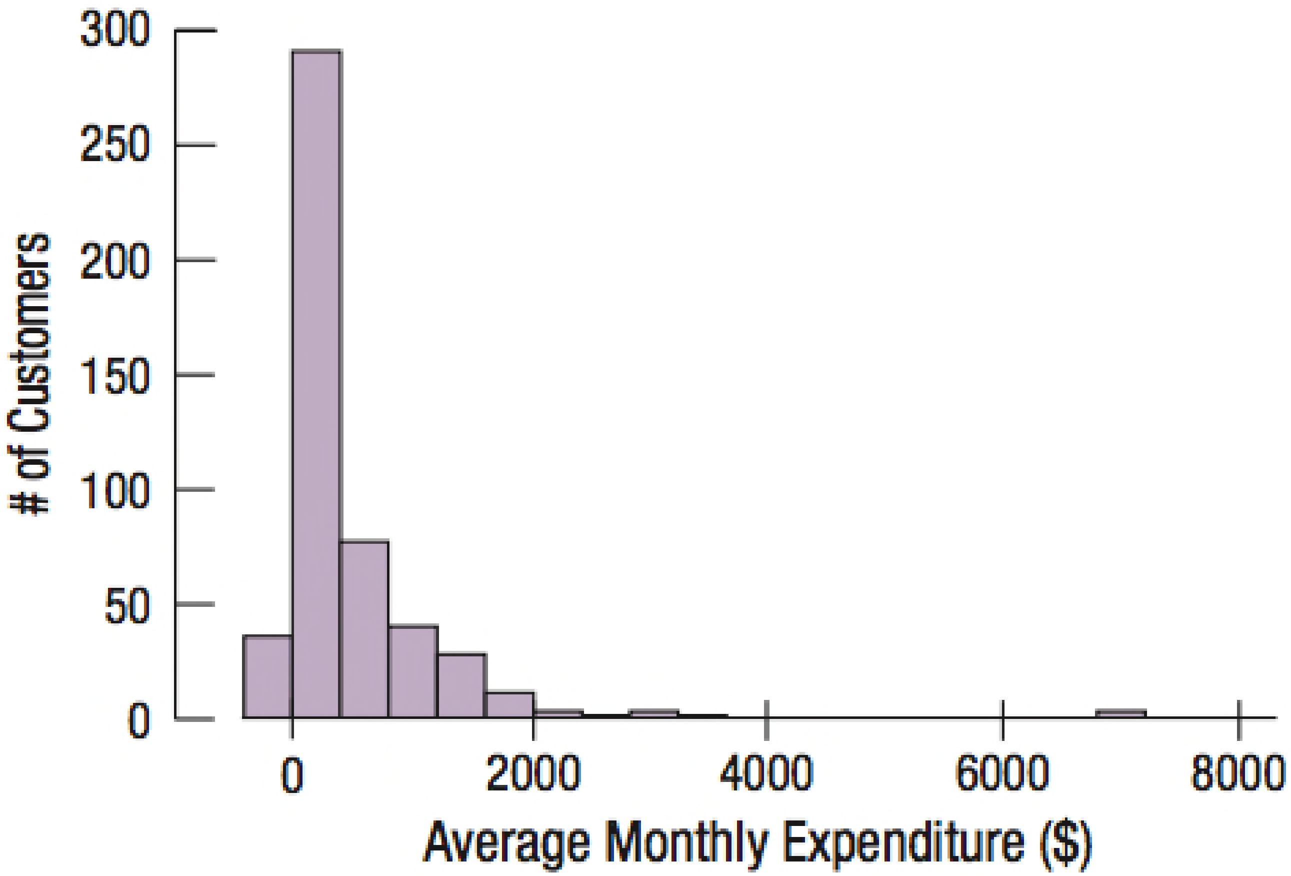
Outliers

- An Outlier is a data value that is far above or far below the rest of the data values.
- An outlier is sometimes just an error in the data collection.
- An outlier can also be the most important data value.
  - Income of a CEO
  - Temperature of a person with a high fever
  - Elevation at Death Valley

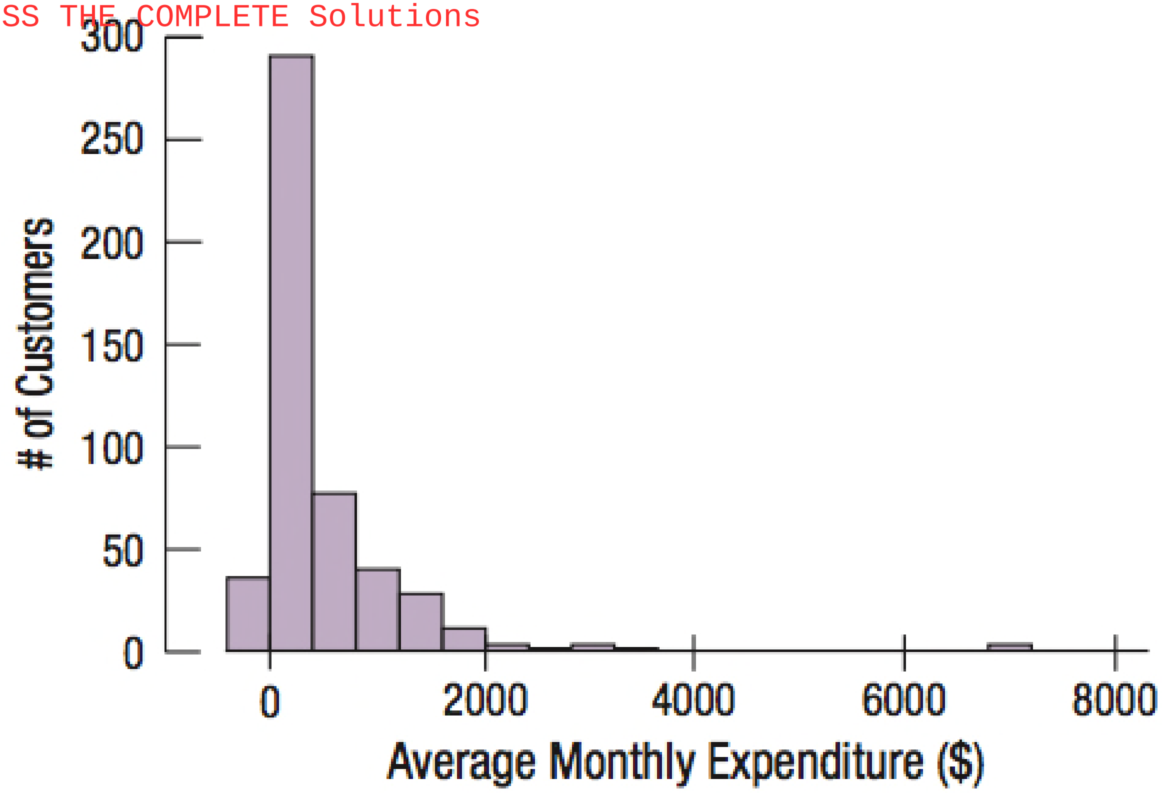


Example

The histogram shows the amount of money spent by a credit card company’s customers. Describe and interpret the distribution.



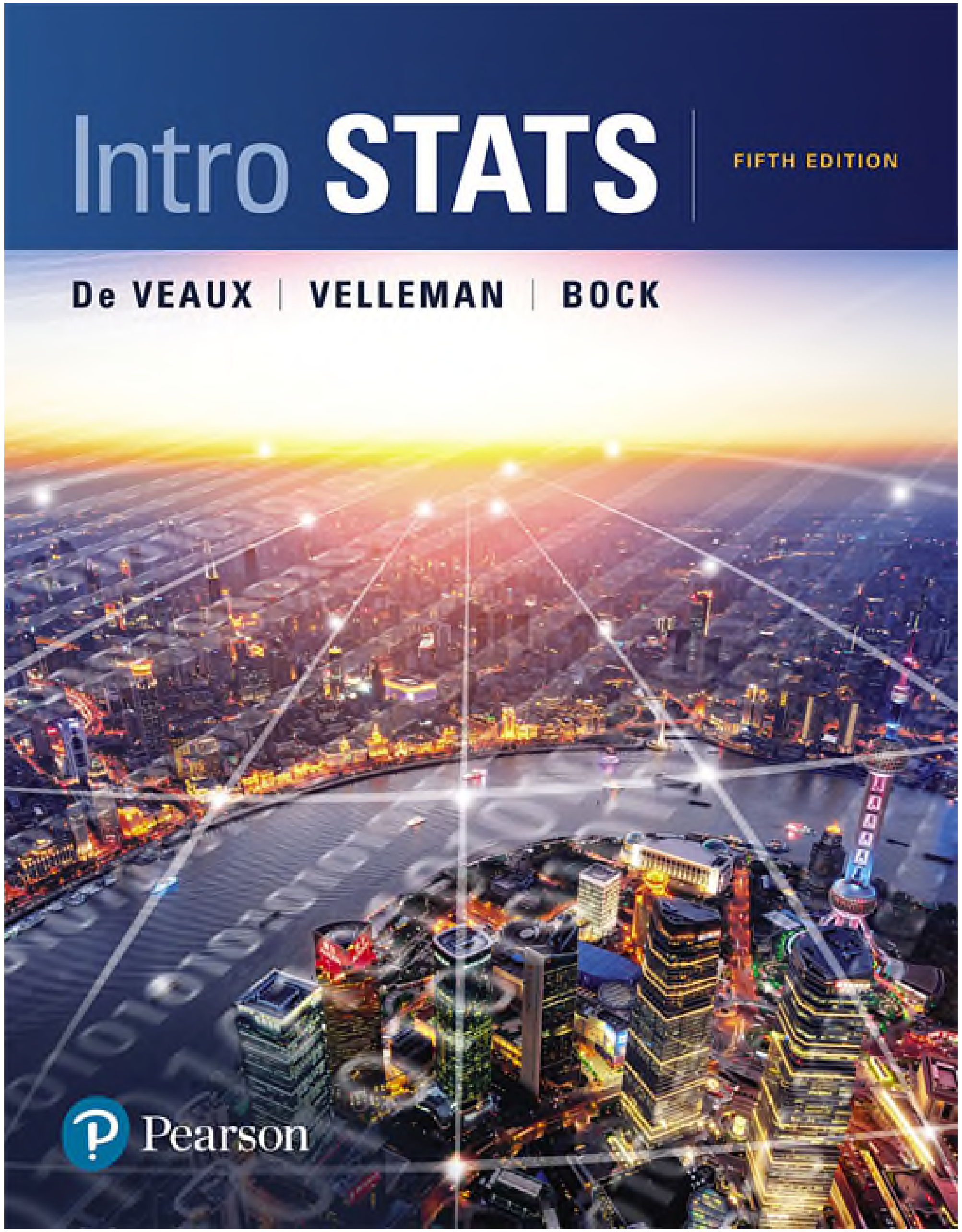
Example Continued



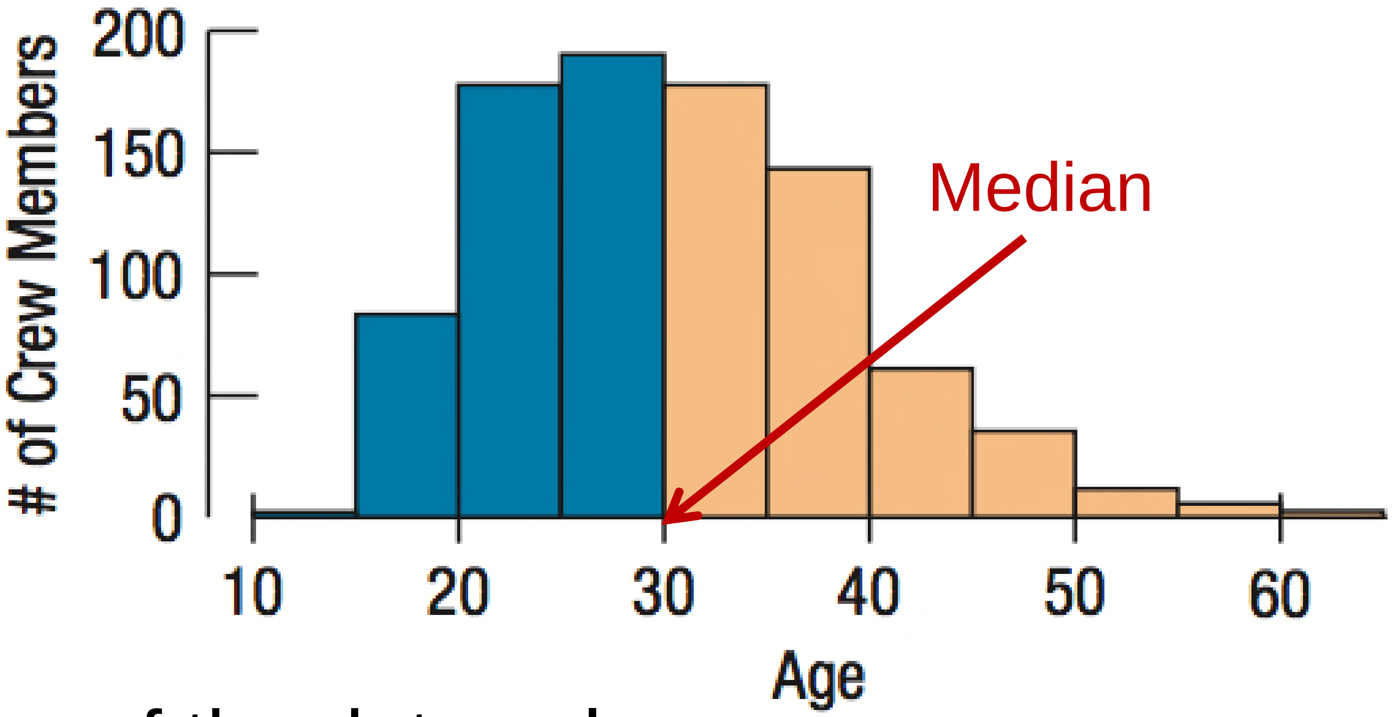
- The distribution is unimodal. Customers most commonly spent a small amount of money.
- The distribution is skewed right. Many customers spent only a small amount and a few were spread out at the high end.
- There is an outlier at around $7000. One customer spent much more than the rest of the customers.
- Some expenditures are negative.

2.4

Center



The Median



- Median: The center of the data values.

- Half of the data values are to the left of the median and half are to the right of the median.

- For symmetric distributions, the median is directly in the middle.

Calculating the Median: Odd Sample Size

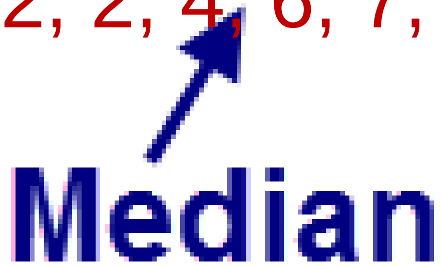
- First order the numbers.
- If there are an odd number of numbers, *n*, the median is at position $\frac{n+1}{2}$.
- Find the median of the numbers: 2, 4, 5, 6, 7, 9, 9.
- $\frac{n+1}{2}=\frac{7+1}{2}=4$ $\qquad \underline{2,4,5},\boxed{6},\underline{7,9,9}$
- The median is the fourth number: 6
- Note that there are 3 numbers to the left of 6 and 3 to the right.

Calculating the Median: Even Sample Size

- First order the numbers.

- If there are an even number of numbers, *n*, the median is the average of the two middle numbers: $\frac{n}{2}, \frac{n}{2}+1$.

- Find the median of the numbers: 2, 2, 4, 6, 7, 8.



- $\frac{n}{2} = \frac{6}{2} = 3$

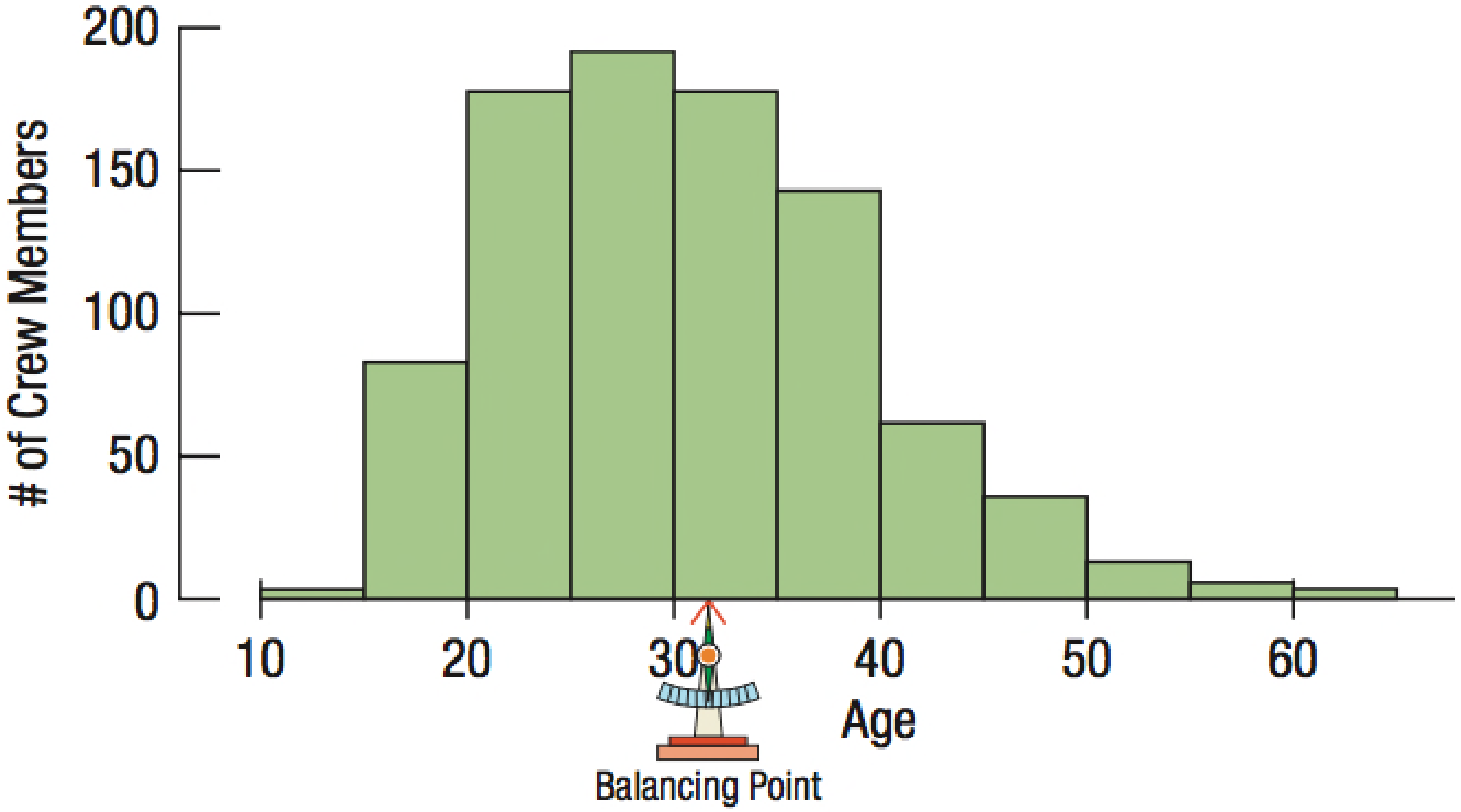
- The median is the average of the third and the fourth numbers: $\text{Median} = \frac{4+6}{2} = 5$

The Mean

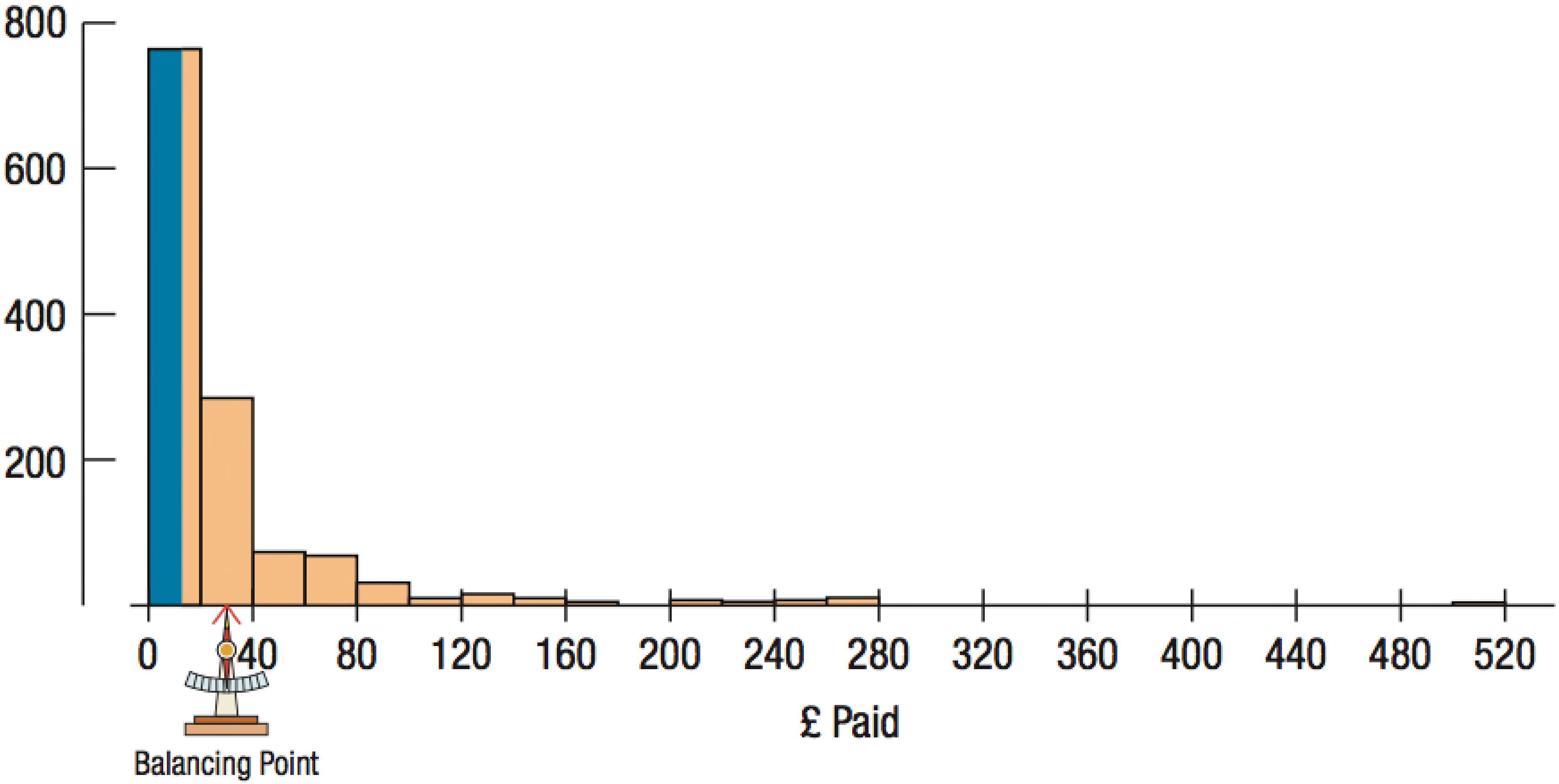
- The mean is the arithmetic average.
- Greek capital letter sigma $\sum$
- Add up all the values of the variable and divide by the number of data values.

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

The Mean


# of Crew Members
200
150
100
50
0
10
20
30
40
50
60
Age
Balancing Point

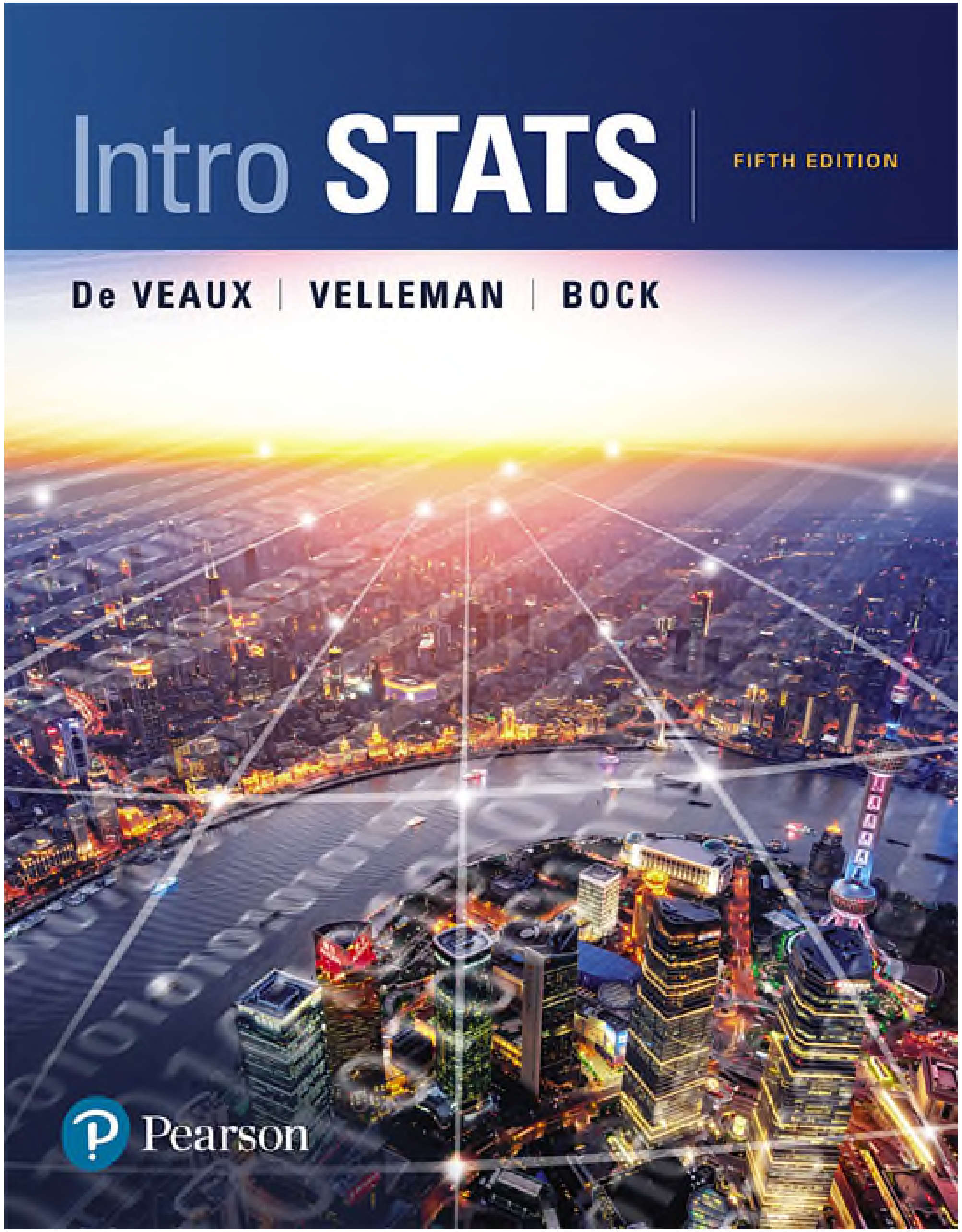
Mean or Median?



Mean or Median?

- Calculate both.
- Investigate outliers: correct them, use them, or set them aside.
- Consider what you want to know about the data.
- Decide whether to report just one of them, or both.
- Some fields have a standard practice: Economics – median for income distribution.

2.5



Spread

Spread

- Locating the center is only part of the story.
- Are the data all near the center or are they spread out?
- Is the highest value much higher than the lowest value?
- To describe data, we must discuss both the center and the spread.

Range

- The range is the difference between the maximum and minimum values.

  Range = Maximum – Minimum

- The ages of the guests at your dinner party are:

  16, 18, 23, 23, 27, 35, 74

- The *range* is: 74 – 16 = 58

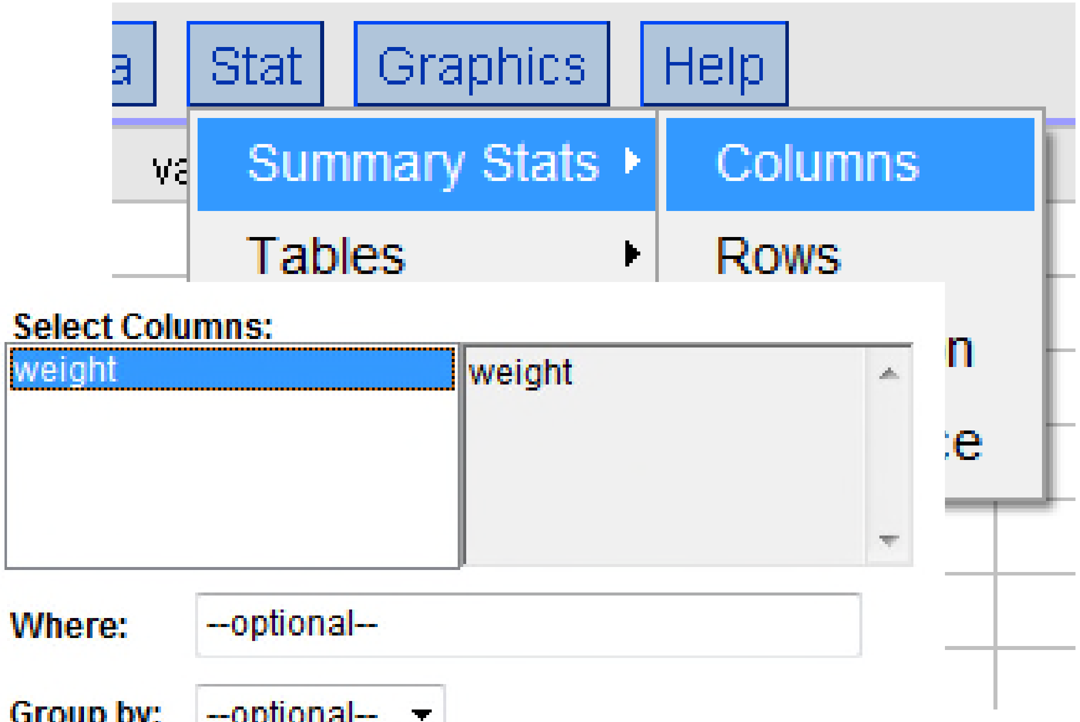
- The range is sensitive to outliers. A single high or low value will affect the range significantly.

Percentiles and Quartiles

- Percentiles divide the data in one hundred groups.
- The *n*th percentile is the data value such that *n* percent of the data lies below that value.
- For large data sets, the median is the 50th percentile.
- The median of the lower half of the data is the 25th percentile and is called the first quartile (Q1).
- The median of the upper half of the data is the 75th percentile and is called the third quartile (Q3).

StatCrunch, Q1, Median, and Q3

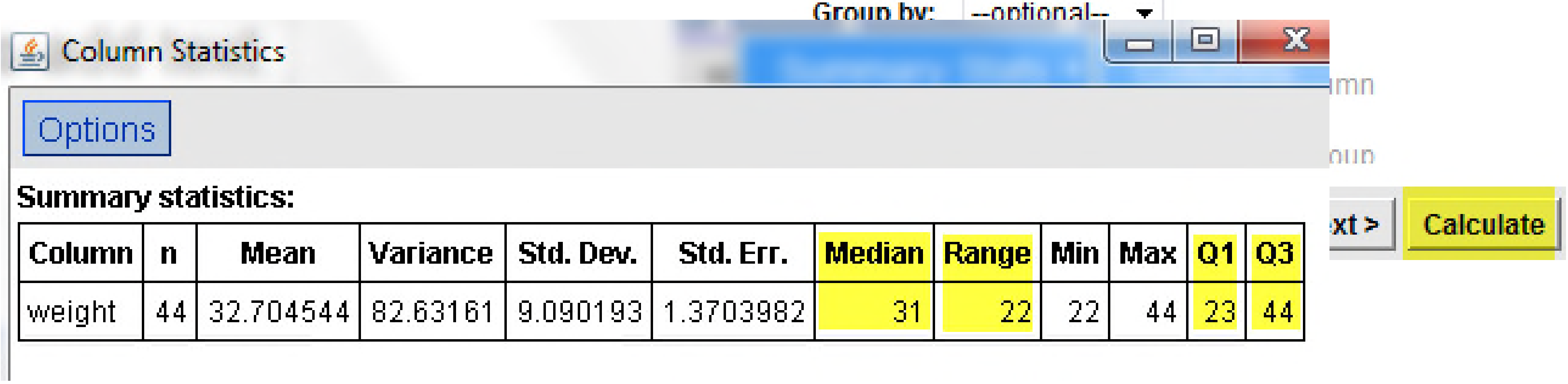
- Enter the data.
- Stat → Summary Stats → Columns
- Click on the variable and then Calculate.



Select Columns:
weight

Where: --optional--

Group by: --optional--



Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
weight	44	32.704544	82.63161	9.090193	1.3703982	31	22	22	44	23	44

The Interquartile Range

- The Interquartile Range (IQR) is the difference between the upper quartile and the lower quartile

  IQR = Q3 – Q1

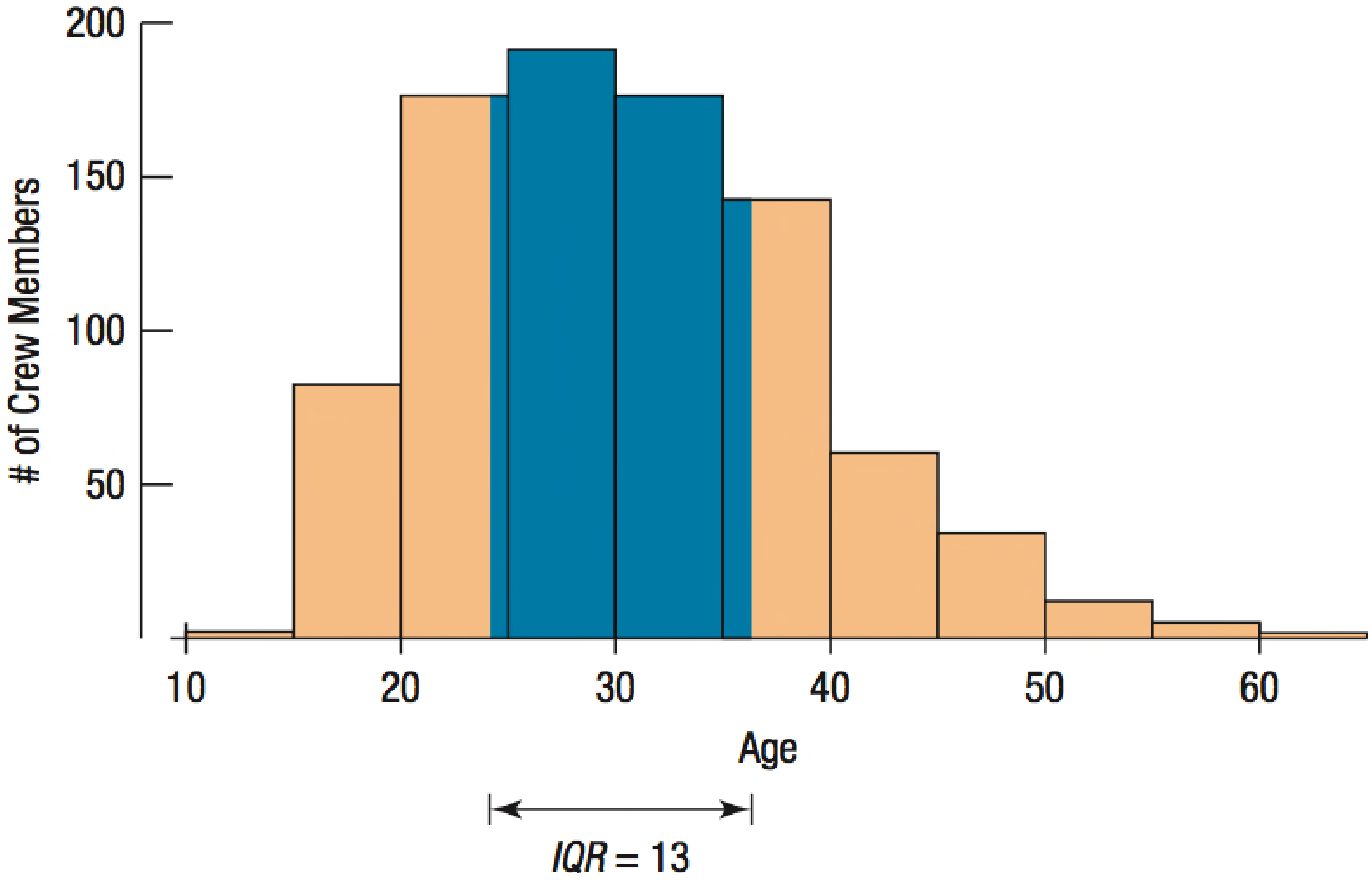
- The IQR measures the range of the middle half of the data.

- Example: If Q1 = 23 and Q3 = 44 then

  IQR = 44 – 23 = 21.

The Interquartile Range

- *IQR* = 37 – 24 = 13 years



Benefits and Drawbacks of the IQR

- The Interquartile Range is not sensitive to outliers.
- The IQR provides a reasonable summary of the spread of the distribution.
- The IQR shows where typical values are, except for the case of a bimodal distribution.
- The IQR is not great for a general audience since most people do not know what it is.

The Standard Deviation

- IQR always reasonable, but ignores most of data
- Usually use the mean as central value
- Least squares property – sum up squares of all the differences of data values from mean
- Residuals – difference of the values from the mean

The Variance

$$s^2 = \frac{\sum \left(y - \bar{y}\right)^2}{n - 1}$$

- The variance is a measure of how far the data is spread out from the mean.
- The difference from the mean is: $y - \bar{y}$.
- To make it positive, square it.
- Then find the average of all of these distances, except instead of dividing by n, divide by $n - 1$.
- Use s^2 to represent the variance.
- The variance's units are the square of the original units.

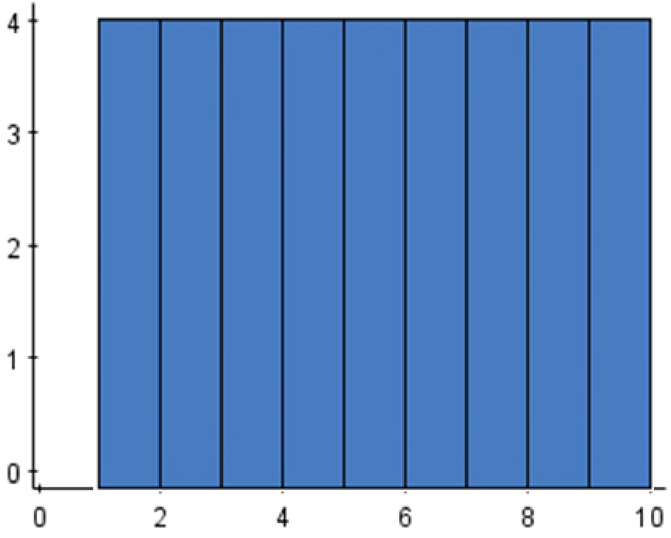
Standard Deviation

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

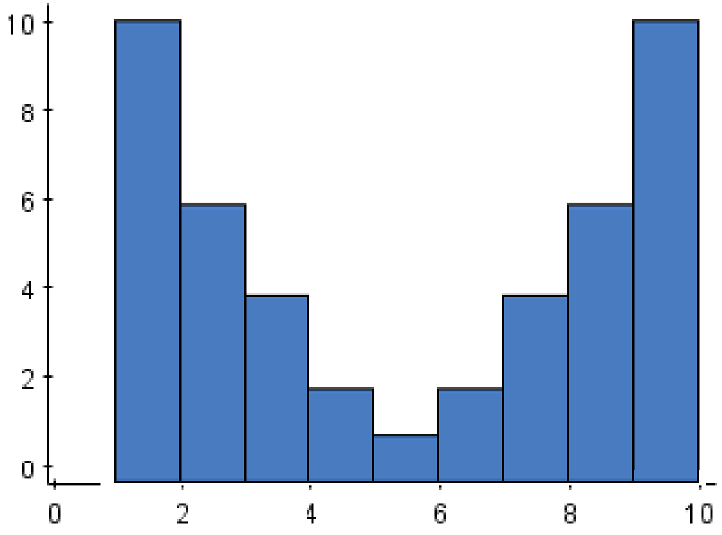
- Taking the square root of the variance gives the standard deviation, which will have the same units as *y*.
- The standard deviation is a number that is close to the average distances that the *y*-values are from the mean.
- If data values are close to the mean (less spread out), then the standard deviation will be small.
- If data values are far from the mean (more spread out), then the standard deviation will be large.

The Standard Deviation and Histograms

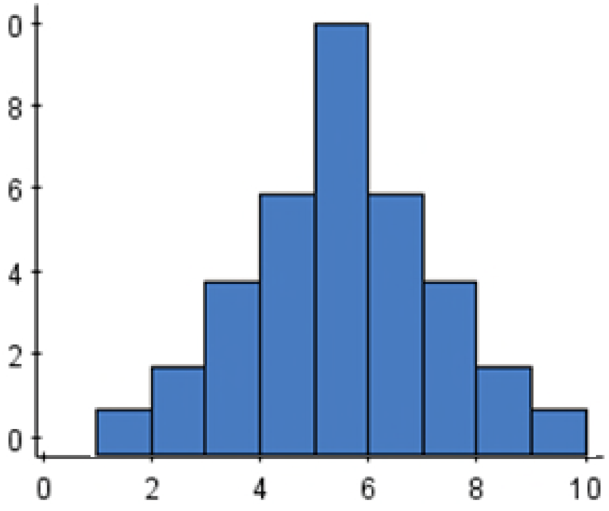
Order the histograms below from smallest standard deviation to largest standard deviation.



A



B



C

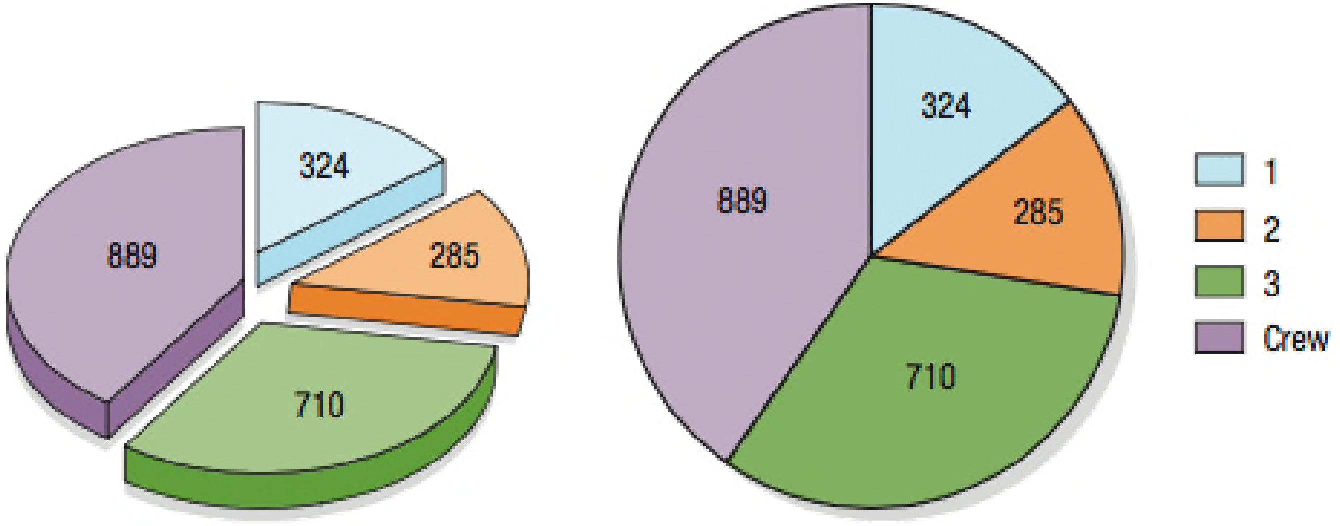
Answer: C, A, B

What to Tell About a Quantitative Variable

- Make a histogram or stem-and-leaf display
- Discuss shape: unimodal, symmetric, free of outliers
- Center and Spread
  - Median with IQR, Mean with Standard Deviation
  - Skewed shape – report median and IQR
  - Symmetric shape – report mean and standard deviation
- Discuss unusual features
  - Multiple modes
  - Outliers

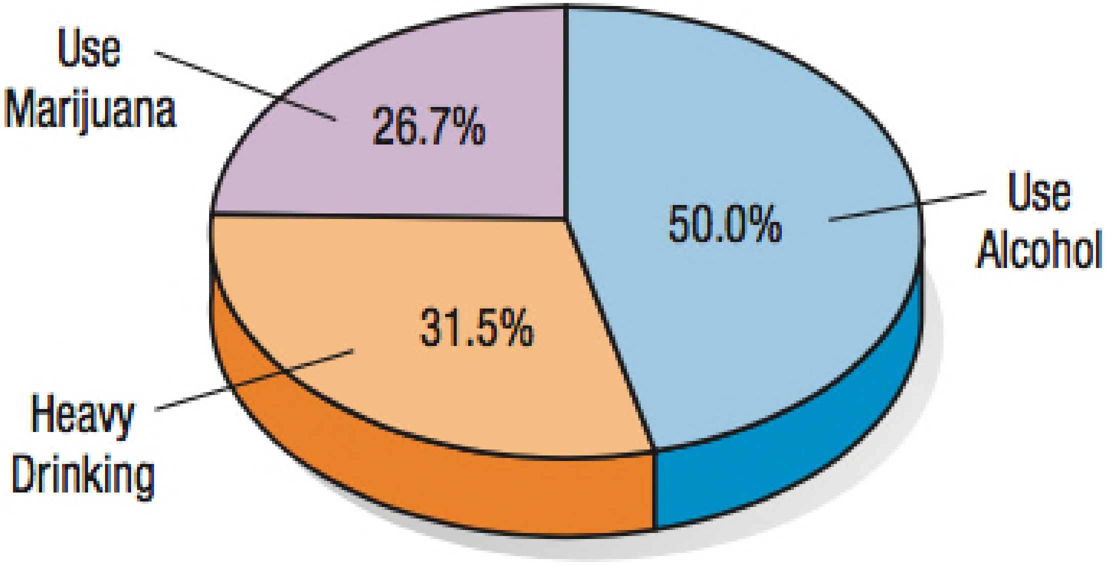
What Can Go Wrong?

- Don’t violate the area principle.



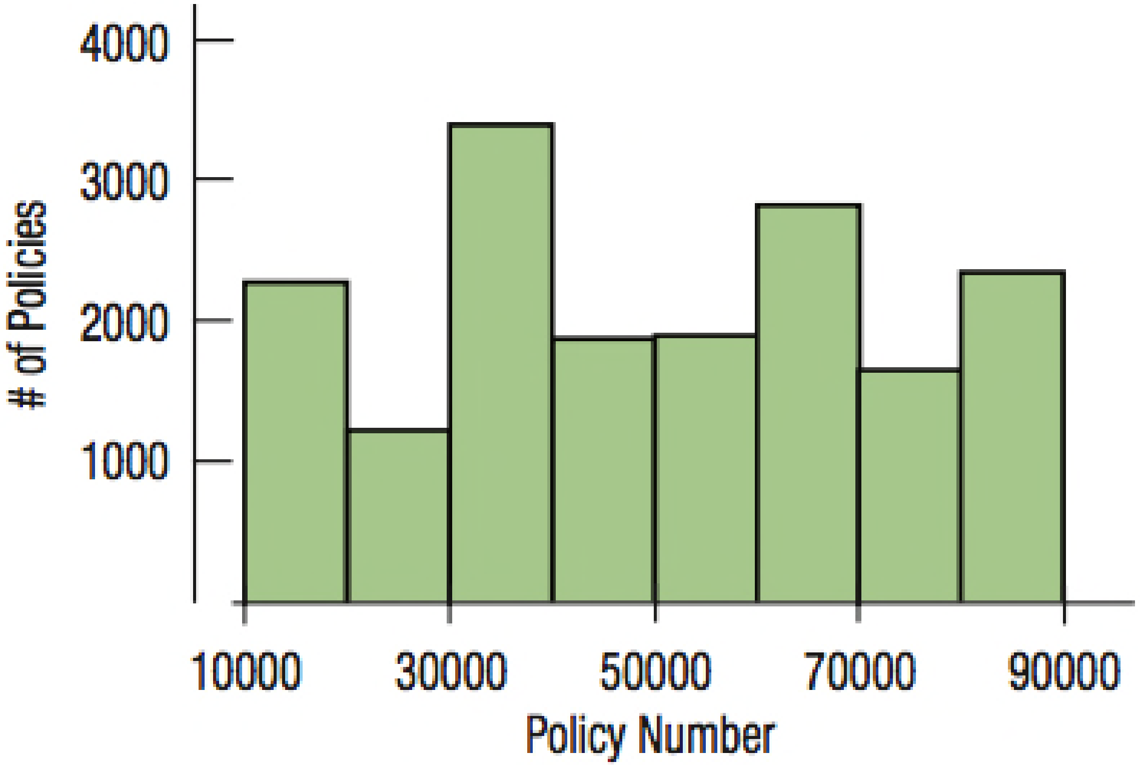
What Can Go Wrong?

- Keep it honest.



What Can Go Wrong?

- Don’t make a histogram of a categorical variable.

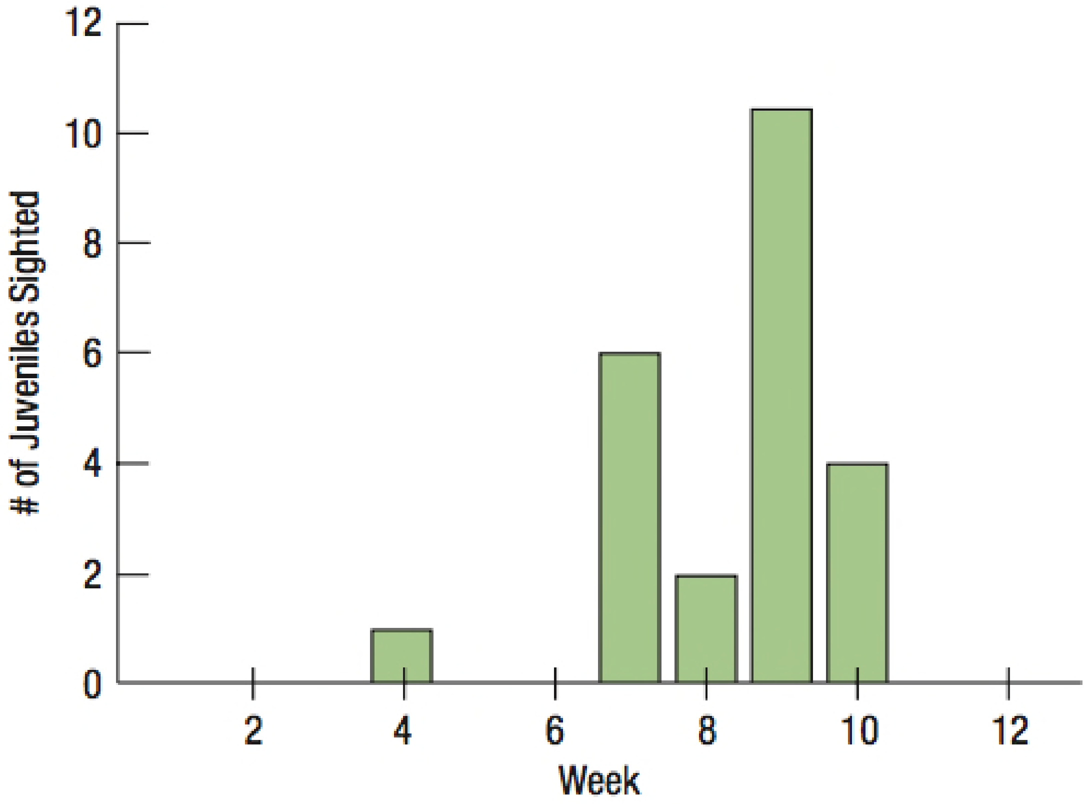


What Can Go Wrong?

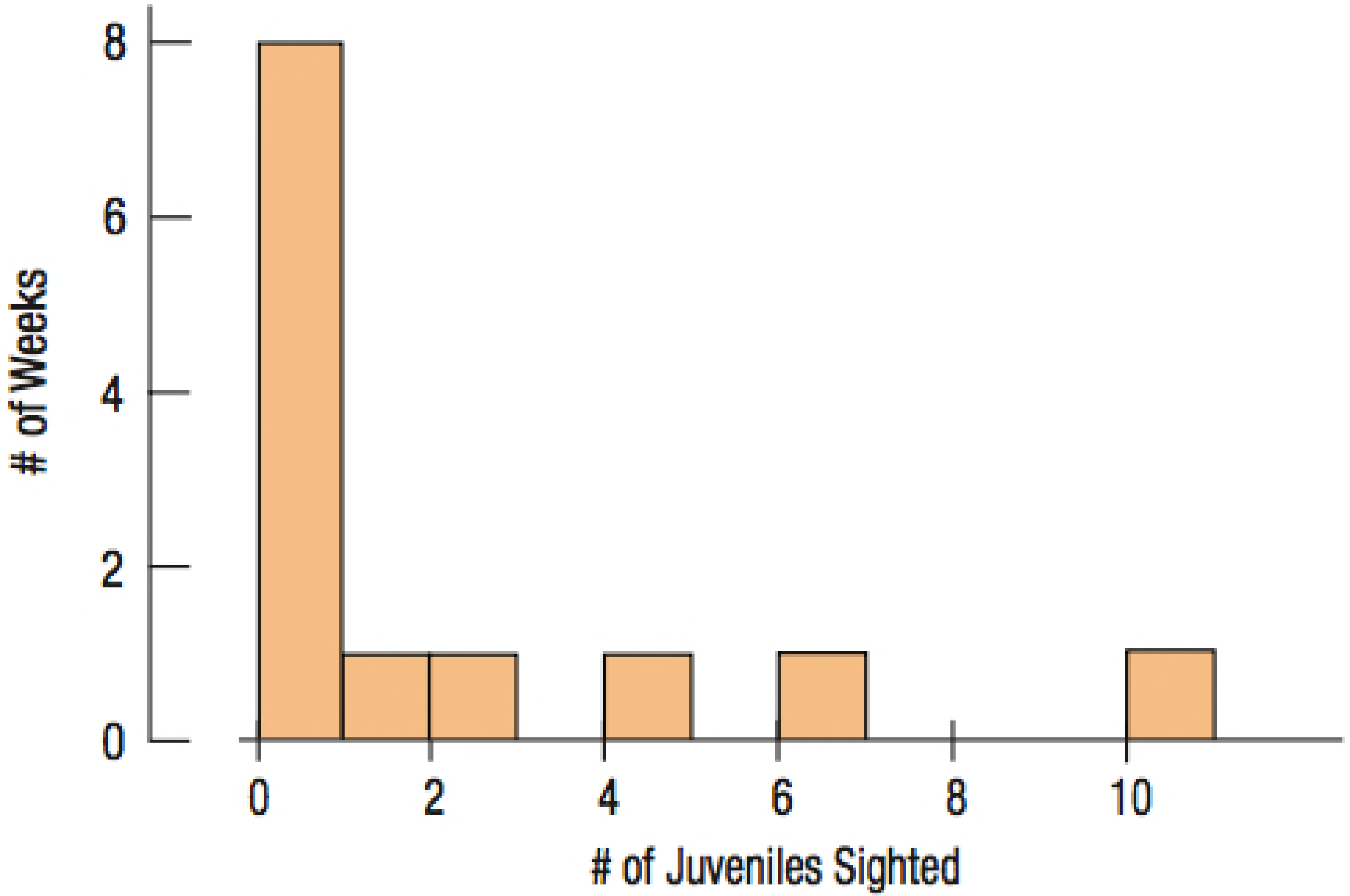
- Don’t look for the shape, center, and spread of a bar chart.
- Don’t compute numerical summaries of a categorical variable.

What Can Go Wrong?

- Don’t use bars in every display—save them for histograms and bar charts.

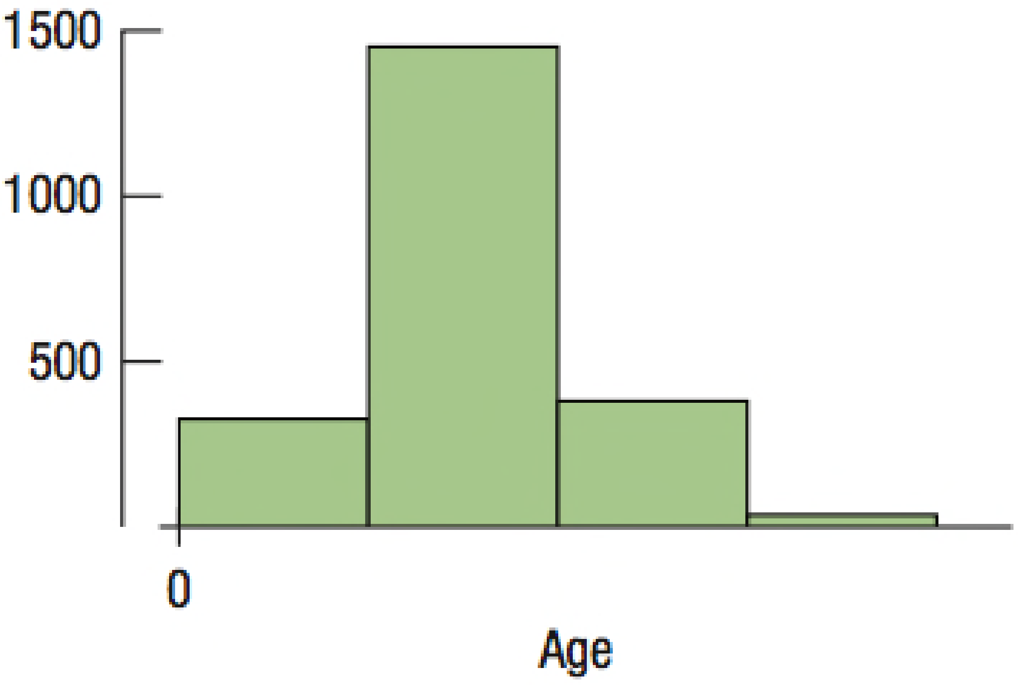


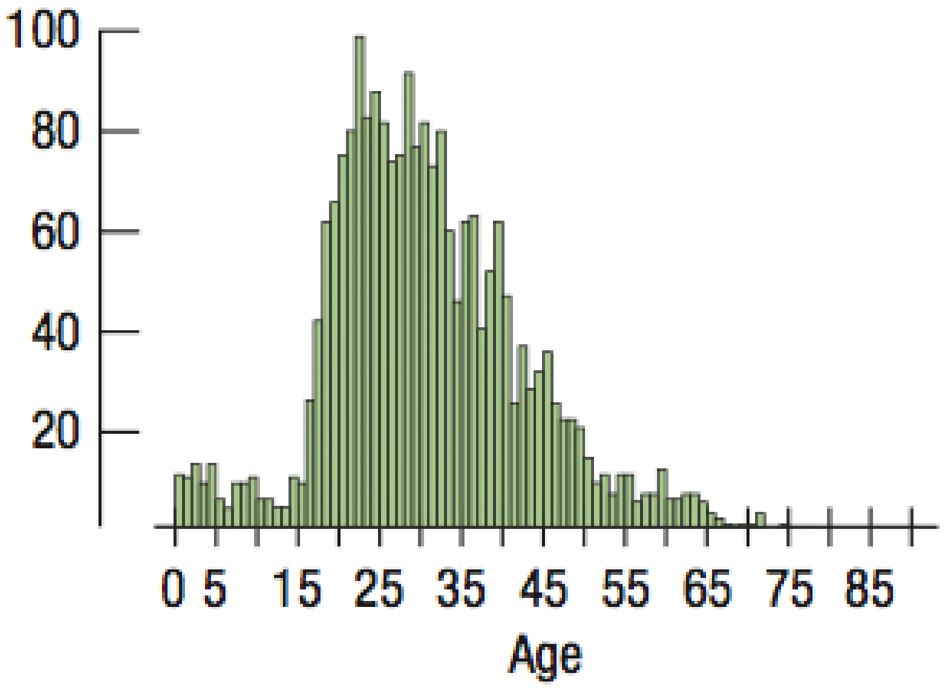
What Can Go Wrong?



What Can Go Wrong?

- Choose a bin width appropriate to the data.





What Can Go Wrong?

- Do a reality check.
- Don’t forget to sort the values before finding the median or percentiles.
- Don’t worry about small differences due to different methods or rounding.
- Don’t report too many decimal places.

What Can Go Wrong?

- Don’t round in the middle of a calculation.
- Watch out for multiple modes.
- Beware of outliers.
- Beware of inappropriate summaries.