INSTRUCTOR'S SOLUTIONS MANUAL

WILLIAM CRAINE III
Lansing High School

INTRO STATS

FIFTH EDITION

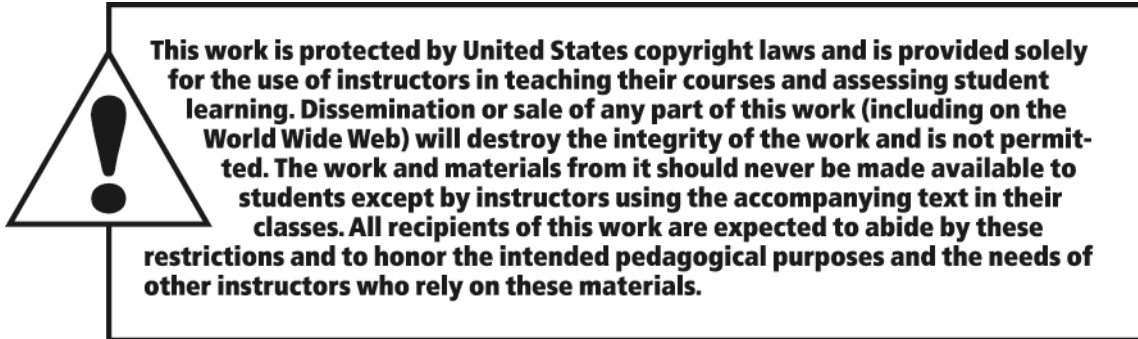## Richard De Veaux

Williams College

Paul Velleman

Cornell University

David Bock

Cornell University

P Pearson

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Reproduced by Pearson from electronic files supplied by the author.

Pearson

Contents

Chapter 1 – Stats Starts Here

Section 1.1

1. **Grocery shopping.** Discount cards at grocery stores allow the stores to collect information about the products that the customer purchases, what other products are purchased at the same time, whether or not the customer uses coupons, and the date and time that the products are purchased. This information can be linked to demographic information about the customer that was volunteered when applying for the card, such as the customer's name, address, sex, age, income level, and other variables. The grocery store chain will use that information to better market their products. This includes everything from printing out coupons at the checkout that are targeted to specific customers to deciding what television, print, or Internet advertisements to use.

2. **Online shopping.** Amazon hopes to gain all sorts of information about customer behavior, such as how long they spend looking at a page, whether or not they read reviews by other customers, what items they ultimately buy, and what items are bought together. They can then use this information to determine which other products to suggest to customers who buy similar items, to determine which advertisements to run in the margins, and to determine which items are the most popular so these items come up first in a search.

3. **Parking lots.** The owners of the parking garage can advertise about the availability of parking. They can also communicate with businesses about hours when more spots are available and when they should encourage more business.

4. **Satellites and global climate change.** This rise and fall of temperature and water levels can help in planning for future problems and guide public policy to protect our safety.

Section 1.2

5. **Super Bowl.** When collecting data about the Super Bowl, the games themselves are the *Who*.

6. **Nobel laureates.** Each year is a case, holding all of the information about that specific year. Therefore, the year is the *Who*.

7. **Health records.** The sample is about 5,000 people, and the population is all residents of the United States of America. The *Who* is the selected subjects and the *What* includes medical, dental, and physiological measurements and laboratory test results.

8. **Facebook.** The *Who* is the 350 million photos. The *What* might be information about the photos, for example: file format, file size, time and date when uploaded, people and places tagged, and GPS information.

Section 1.3

9. Grade level.

a) If we are, for example, comparing the percentage of first-graders who can tie their own shoes to the percentage of second-graders who can tie their own shoes, grade-level is treated as categorical. It is just a way to group the students. We would use the same methods if we were comparing boys to girls or brown-eyed kids to blue-eyed kids.

b) If we were studying the relationship between grade-level and height, we would be treating grade level as quantitative.

10. ZIP codes.

a) ZIP codes are categorical in the sense that they correspond to a location. The ZIP code 14850 is a standardized way of referring to Ithaca, NY.

b) ZIP codes generally increase as the location gets further from the east coast of the United States. For example, one of the ZIP codes for the city of Boston, MA is 02101. Kansas City, MO has a ZIP code of 64101, and Seattle, WA has a ZIP code of 98101.

11. Voters. The response is a categorical variable.

12. Job hunting. The answer is a categorical variable.

13. Medicine. The company is studying a quantitative variable.

14. Stress. The researcher is studying a quantitative variable.

Section 1.4

15. Voting and elections. Pollsters might consider whether a person voted previously or whether he or she could name the candidates. Voting previously and knowing the candidates may indicate a greater interest in the election.

16. Weather. Meteorologists can use the models to predict the average temperature ten days in advance and compare their predictions to the actual temperatures.

17. The News. Answers will vary.

18. The Internet. Answers will vary.

19. Gaydar. *Who* – 40 undergraduate women. *What* – Whether or not the women could identify the sexual orientation of men based on a picture. *Population of interest* – All women.

20. **Hula-hoops.** *Who* – An unknown number of participants. *What* – Heart rate, oxygen consumption, and rating of perceived exertion. *Population of interest* – All people.

21. **Bicycle Safety.** *Who* – 2,500 cars. *What* – Distance from the bicycle to the passing car (in inches). *Population of interest* – All cars passing bicyclists.

22. **Investments.** *Who* – 30 similar companies. *What* – 401(k) employee participation rates (in percent). *Population of interest* – All similar companies.

23. **Honesty.** *Who* – Workers who buy coffee in an office. *What* – amount of money contributed to the collection tray. *Population of interest* – All people in honor system payment situations.

24. **Blindness.** *Who* – 24 patients. *What* – Whether the patient had Stargardt's disease or dry age-related macular degeneration, and whether or not the stem cell therapy was effective in treating the condition. *Population of interest* – All people with these eye conditions.

25. **Not-so-diet soda.** *Who* – 474 participants. *What* – whether or not the participant drank two or more diet sodas per day, waist size at the beginning of the study, and waist size at the end of the study. *Population of interest* – All people.

26. **Molten iron.** *Who* – 10 crankshafts at Cleveland Casting. *What* – The pouring temperature (in degrees Fahrenheit) of molten iron. *Population of interest* – All crankshafts at Cleveland Casting.

27. **Weighing bears.** *Who* – 54 bears. *What* – Weight, neck size, length (no specified units), and sex. *When* – Not specified. *Where* – Not specified. *Why* - Since bears are difficult to weigh, the researchers hope to use the relationships between weight, neck size, length, and sex of bears to estimate the weight of bears, given the other, more observable features of the bear. *How* – Researchers collected data on 54 bears they were able to catch. *Variables* – There are 4 variables; weight, neck size, and length are quantitative variables, and sex is a categorical variable. No units are specified for the quantitative variables. *Concerns* – The researchers are (obviously!) only able to collect data from bears they were able to catch. This method is a good one, as long as the researchers believe the bears caught are representative of all bears, in regard to the relationships between weight, neck size, length, and sex.

28. **Schools.** *Who* – Students. *What* – Age (probably in years, though perhaps in years and months), race or ethnicity, number of absences, grade level, reading score, math score, and disabilities/special needs. *When* – This information must be kept current. *Where* – Not specified. *Why* – Keeping this information is a state requirement. *How* – The information is collected and stored as part of school records. *Variables* – There are seven variables. Race or ethnicity, grade level, and disabilities/special needs are categorical variables. Number of absences (days), age (years?), reading test score, and math test score are quantitative variables. *Concerns* – What tests are used to measure reading and math ability, and what are the units of measure for the tests?

29. Arby's menu. _Who_ – Arby's sandwiches. _What_ – type of meat, number of calories (in calories), and serving size (in ounces). _When_ – Not specified. _Where_ – Arby's restaurants. _Why_ – These data might be used to assess the nutritional value of the different sandwiches. _How_ – Information was gathered from each of the sandwiches on the menu at Arby's, resulting in a census. _Variables_ – There are three variables. Number of calories and serving size (ounces) are quantitative variables, and type of meat is a categorical variable.

30. Age and party. _Who_ – 1180 Americans. _What_ – Region, age (in years), political affiliation, and whether or not the person voted in the 2006 midterm Congressional election. _When_ – First quarter of 2007. _Where_ – United States. _Why_ – The information was gathered for presentation in a Gallup public opinion poll. _How_ – Phone Survey. _Variables_ – There are four variables. Region, political affiliation, and whether or not the person voted in 1998 are categorical variables, and age is a quantitative variable.

31. Babies. _Who_ – 882 births. _What_ – Mother's age (in years), length of pregnancy (in weeks), type of birth (caesarean, induced, or natural), level of prenatal care (none, minimal, or adequate), birth weight of baby (unit of measurement not specified, but probably pounds and ounces), gender of baby (male or female), and baby's health problems (none, minor, major). _When_ – 1998-2000. _Where_ – Large city hospital. _Why_ – Researchers were investigating the impact of prenatal care on newborn health. _How_ – It appears that they kept track of all births in the form of hospital records, although it is not specifically stated. _Variables_ – There are three quantitative variables: mother's age (years), length of pregnancy (, and birth weight of baby. There are four categorical variables: type of birth, level of prenatal care, gender of baby, and baby's health problems.

32. Flowers. _Who_ – 385 species of flowers. _What_ – Date of first flowering (in days). _When_ – Not specified. _Where_ – Southern England. _Why_ – The researchers believe that this indicates a warming of the overall climate. _How_ – Not specified. _Variables_ – Date of first flowering is a quantitative variable. _Concerns_ - Hopefully, date of first flowering was measured in days from January 1, or some other convention, to avoid problems with leap years.

33. Herbal medicine. _Who_ – experiment volunteers. _What_ – herbal cold remedy or sugar solution, and cold severity (0 to 5 scale). _When_ – Not specified. _Where_ – Major pharmaceutical firm. _Why_ – Scientists were testing the efficacy of an herbal compound on the severity of the common cold. _How_ – The scientists set up a controlled experiment. _Variables_ – There are two variables. Type of treatment (herbal or sugar solution) is categorical, and severity rating is quantitative. _Concerns_ – The severity of a cold seems subjective and difficult to quantify. Also, the scientists may feel pressure to report negative findings about the herbal product.

34. Vineyards. _Who_ – American Vineyards. _What_ – Size of vineyard (in acres), number of years in existence, state, varieties of grapes grown, average case price (in dollars), gross sales (probably in dollars), and percent profit. _When_ – Not specified. _Where_ – United States. _Why_ – Business analysts hoped to provide information that would be helpful to producers of American wines. _How_ – Not specified. _Variables_ – There are five quantitative variables and two categorical variables. Size of vineyard, number of years in existence, average case price, gross sales, and percent profit are quantitative variables. State and variety of grapes grown are categorical variables.

35. Streams. *Who* – Streams. *What* – Name of stream, substrate of the stream (limestone, shale, or mixed), acidity of the water (measured in pH), temperature (in degrees Celsius), and BCI (unknown units). *When* – Not specified. *Where* – Upstate New York. *Why* – Research is conducted for an Ecology class. *How* – Not specified. *Variables* – There are five variables. Name and substrate of the stream are categorical variables, and acidity, temperature, and BCI are quantitative variables.

36. Fuel economy. *Who* – Every model of automobile in the United States. *What* – Vehicle manufacturer, vehicle type, weight (probably in pounds), horsepower (in horsepower), and gas mileage (in miles per gallon) for city and highway driving. *When* – This information is collected currently. *Where* – United States. *Why* – The Environmental Protection Agency uses the information to track fuel economy of vehicles. *How* – The data is collected from the manufacturer of each model. *Variables* – There are six variables. City mileage, highway mileage, weight, and horsepower are quantitative variables. Manufacturer and type of car are categorical variables.

37. Refrigerators. *Who* – 353 refrigerator models. *What* – Brand, cost (probably in dollars), size (in cu. ft.), type, estimated annual energy cost (probably in dollars), overall rating, and repair history (in percent requiring repair over the past five years). *When* – 2013. *Where* – United States. *Why* – The information was compiled to provide information to the readers of *Consumer Reports*. *How* – Not specified. *Variables* – There are 7 variables. Brand, type, and overall rating are categorical variables. Cost, size, estimated energy cost, and repair history are quantitative variables.

38. Walking in circles. *Who* – 32 volunteers. *What* – Sex, height, handedness, the number of yards walked before going out of bounds, and the side of the field on which the person walked out of bounds. *When* – Not specified. *Where* – Not specified. *Why* – The researcher was interested in whether people walk in circles when lost. *How* – Data were collected by observing the people on the field, as well as by measuring and asking the participants. *Variables* – There are 5 variables. Sex, handedness, and side of the field are categorical variables. Height and number of yards walked are quantitative variables.

39. Kentucky Derby 2016. *Who* – Kentucky Derby races. *What* – Year, winner, jockey, trainer, owner, and time (in minutes, seconds, and hundredths of a second). *When* – 1875 – 2016. *Where* – Churchill Downs, Louisville, Kentucky. *Why* – Not specified. To examine the trends in the Kentucky Derby? *How* – Official statistics are kept for the race each year. *Variables* – There are 6 variables. Winner, jockey, trainer and owner are categorical variables. Date and duration are quantitative variables.

40. Indy 2016. *Who* – Indy 500 races. *What* – Year, driver, time (in minutes, seconds, and hundredths of a second), and speed (in miles per hour). *When* – 1911 – 2016. *Where* – Indianapolis, Indiana. *Why* – Not specified. To examine the trends in Indy 500 races? *How* – Official statistics are kept for the race every year. *Variables* – There are 4 variables. Driver is a categorical variable. Year, time, and speed are quantitative variables.

41. Kentucky Derby 2016 on the computer.

    a) Fonso was the winning horse in 1880.

    b) The length of the race changed in 1895, from 1.5 miles to 1.25 miles.

    c) The winning time in 1974 was 124 seconds.

    d) Secretariat ran the Derby in under 2 minutes in 1973, as did Monarchos in 2001.

42. Indy 500 2016 on the computer.

    a) The average speed of the winner in 1920 was 88.619 miles per hour.

    b) Bill Vukovich won the Indy 500 twice in the 1950s.

    c) There were only 6 Indy 500 races in the 1940s.

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.

Cause of Death 2014
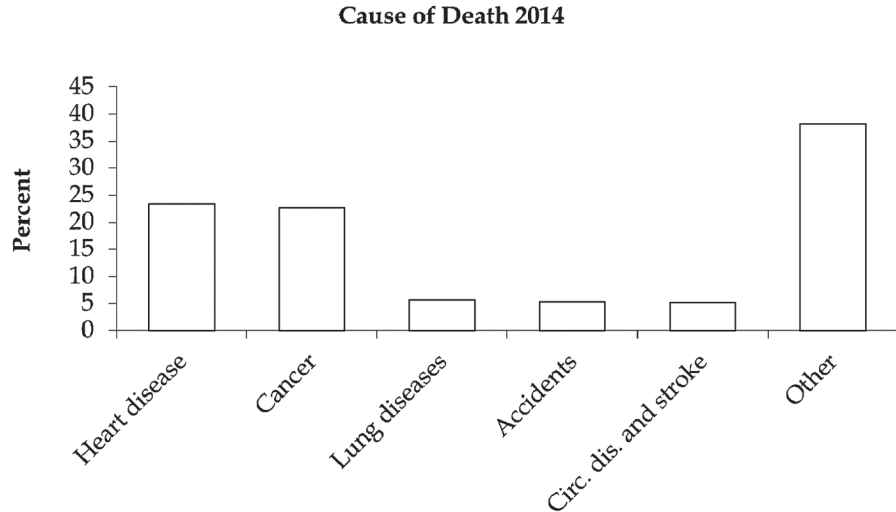
b) Since the percentages listed add up to 61.9%, other causes must account for 38.1% of US deaths.

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 'n' 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.

Total Points Scored (first 50 Super Bowls)

    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.
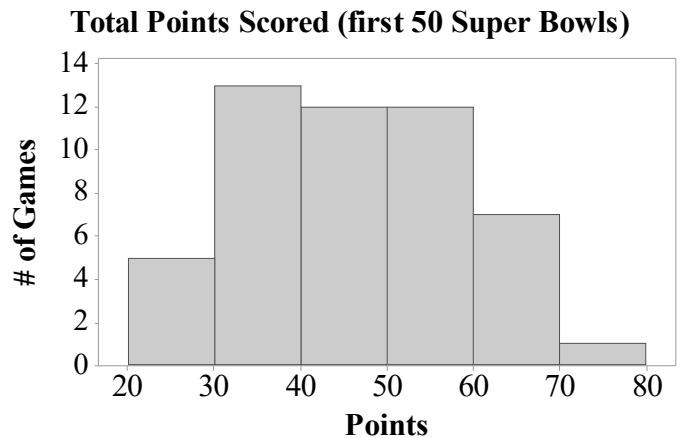
51. Test scores, large class.

    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.

Alex Rodriguez 1994 - 2016
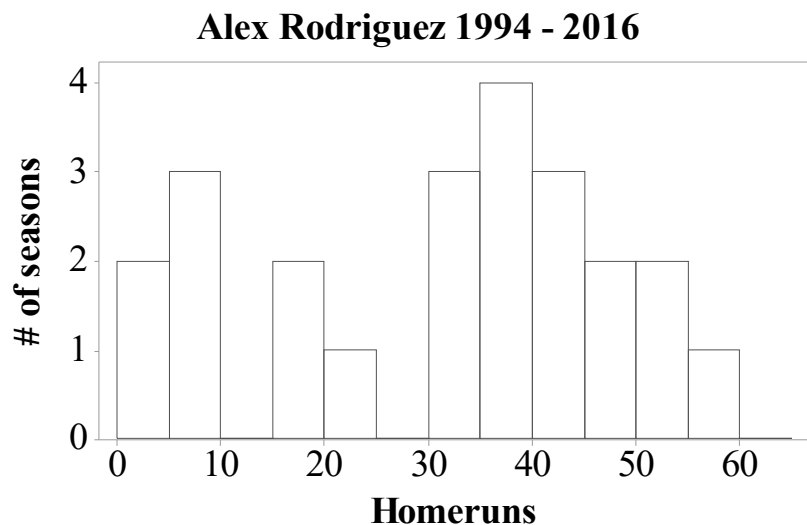
77. Acid rain.

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.

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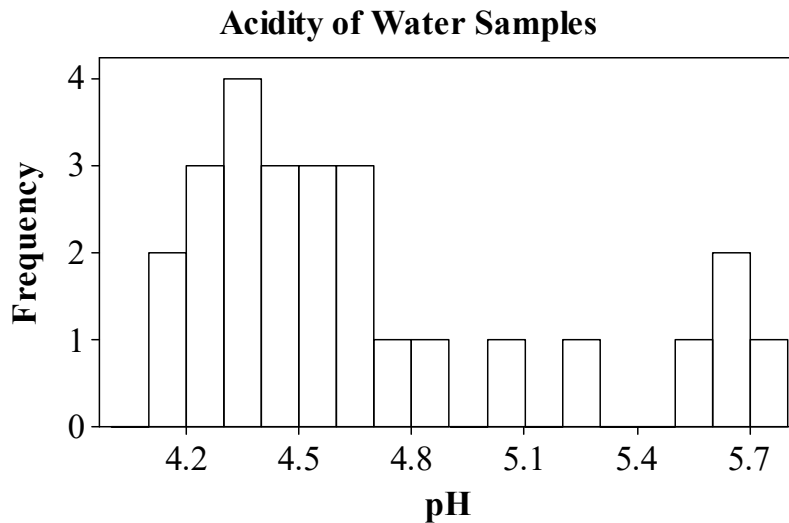
b) The cluster of high outliers contains many dates that were holidays in 1973. Traffic patterns would probably be different then, which might account for the difference.

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.

US Math Test Scores
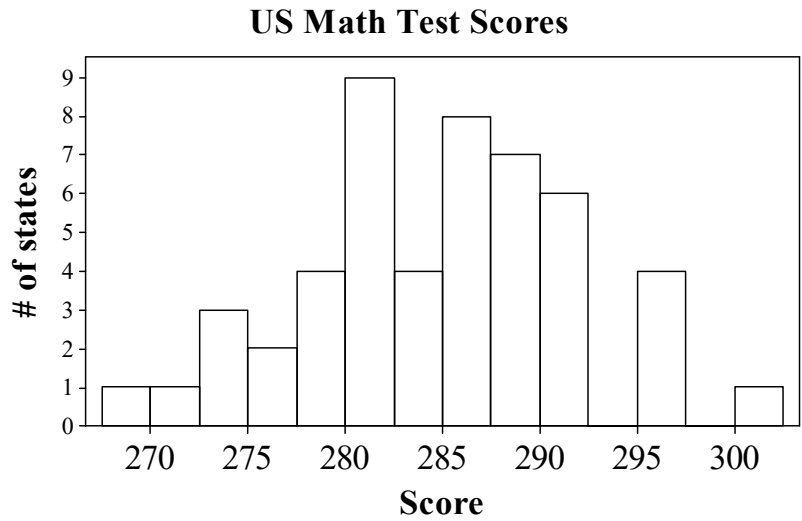
85. Population growth 2010.

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%.

Population Growth - 2000 to 2010
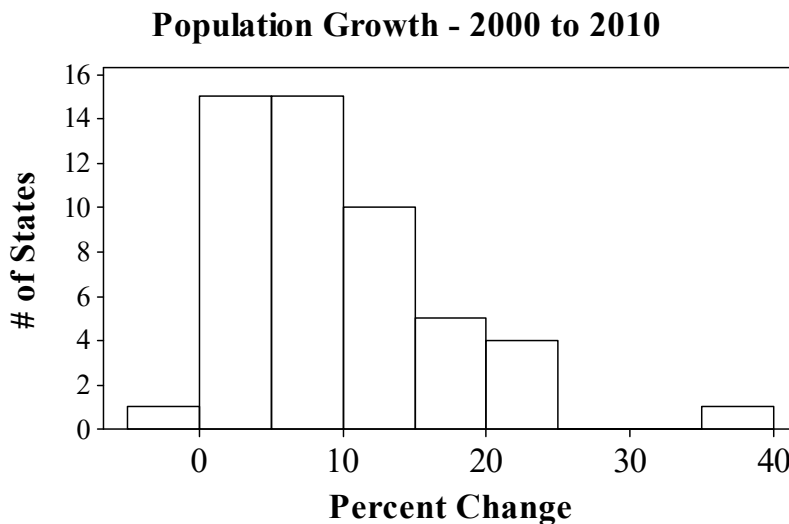
Chapter 3 – Relationships Between Categorical Variables – Contingency Tables

Section 3.1

1. **College value?**

	Poor	Only Fair	Good	Excellent	DK/NA	Total
US Adults	321	900	750	107	64	2142
Presidents	32	222	622	179	0	1055
Total	353	1122	1372	286	64	3197

   a) The percent of college presidents who think that higher education provides a poor value is $32/1055 \approx 3\%$.

   b) $(750 + 107)/2142 \approx 40\%$ of U.S. adults think that the value provided by the higher education system is either good or excellent.

   c) 15% of U.S. adults view college as a poor value, but only 3% of college presidents do. Similarly, U.S. adults are twice as likely to view college as an only fair value compared to the presidents (42% to 21%). Presidents are much more likely (76%) to rate colleges as a good or excellent value compared to U.S. adults (only 40%). So in short, college presidents have a much higher opinion of the value of college than U.S. adults do.

   d) In this random sample of U.S. adults, $107/2142 \approx 5.00\%$ responded that college provides an excellent value. This is only an estimate, however, of the percentage of all U.S. adults who feel the same way. The percentage of all U.S. adults is probably close to, but not exactly 5%.

2. **Cyber comparison shopping.**

	<$30K	$30K - $49.9K	$50K – $74.9K	>$75K	Total
Yes	207	115	134	204	660
No	625	406	260	417	1708
Total	832	521	394	621	2368

   a) $207/832 \approx 24.9\%$ of those earning less than $30K cyber comparison shop. $204/621 \approx 32.9\%$ of those earning more than $75K cyber comparison shop.

   b) $207/660 \approx 31.4\%$ of those who cyber comparison shop earn less than $30K.

   c) $660/2368 \approx 28\%$ of those in this random sample of people who owned cell phones use their phones to cyber comparison shop. First of all, not all shoppers have cell phones, so we wouldn't expect the percentage of all shoppers to be similar to the percentage found in this sample. Even if we were only concerned with all shoppers with phones, we would only expect the percentage of all shoppers to be reasonably close to, but not exactly, 28%.

Section 3.2

3. **College value again.**

   a) The conditional distribution of college presidents' opinions about the value of a college education is 3% poor; 21% only fair; 59% good; 17% excellent

   b) Omitting the 64 DK/NAs, the conditional distribution of the opinions U.S. adults about the value of a college education is 15% negative; 43% middle; 41% positive.

4. **Cyber comparison shopping again.**

   a) The conditional distribution of income for those who do not cyber compare prices is 37% <$30K; 24% $30K–$49.9K; 15% $50K–$74.9K; 24% >$75K.

   b) The conditional distribution of income for those who do cyber compare prices is 31% <$30K; 17% $30K–$49.9K; 20% $50K–$74.9K; 31% >$75K.

Section 3.3

5. **Diet and politics.**

   a) The distribution of political alignment among carnivores is about 25% conservative, 40% moderate, and 35% liberal. Omnivores were generally less conservative and more liberal, with about 10% conservative, 35% moderate, and 55% liberal. Vegetarians were even more liberal than the other groups, with 0% conservative, 30% moderate, and 70% liberal.

   b) The differences are quite large. There appears to be a strong association between diet preference and political alignment.

6. **Diet and politics revisited.**

   a) The distribution of diet preference among liberals is approximately 10% carnivore, 75% omnivore, and 15% vegetarian. Moderates tend to be slightly more carnivorous, with approximately 18% carnivores, 73% omnivores, and 9% vegetarians. Conservatives tend to be even more carnivorous, with 35% carnivores, 65% omnivores, and 0% vegetarians.

   b) The differences are quite large. There appears to be a strong association between political alignment and diet preference.

7. **Fish and prostate cancer revisited**

   a) Looking at the horizontal axis only, approximately 7% of the men had prostate cancer.

   b) There are more men who didn't have cancer and never or seldom ate fish. The rectangles are approximately the same height, but the bar for "no cancer" is much wider.

   c) The percentage of men who never/seldom ate fish is lower in the group with no cancer than in the group with cancer. Disregard the width, and look only at the height to compare the conditional distribution of fish consumption within each cancer group. The bar for "never/seldom" is slightly shorter within the "no cancer" group.

8. **College value? revisited.**

   a) Approximately 33% of the respondents were college presidents. Look only at the horizontal axis to estimate this percentage.

   b) This is difficult to compare on the mosaic plot. The rectangles appear to have similar areas, but the U.S. adults region is much wider and shorter, while the college presidents region is narrower, but taller. Looking back at the original data table, we can see that the college presidents have a greater number, 179 to 107.

c) Yes, it is clear that a greater percentage of college presidents said that college provides an excellent value, when compared to the percentage of U.S. adults who said the same thing. The bar for college presidents is much taller.

d) The differences seem quite large. It does appear there is a strong association between the responses and whether the respondent was a college president. College presidents were much more likely to say that college provides an excellent or good value, and less likely to say that college provides a poor value than U.S. adults in general.

Section 3.4

9. **Diet and politics III.**

   Men

	Carnivore	Omnivore	Vegetarian	Total
Liberal	9	74	5	88
Moderate	12	54	1	67
Conservative	9	14	0	23
Total	30	142	6	178

   Women

	Carnivore	Omnivore	Vegetarian	Total
Liberal	4	53	12	69
Moderate	4	27	6	37
Conservative	1	4	0	5
Total	9	84	18	111

   a) Men are more likely to be conservative carnivores. $9/178 \approx 5.1\%$ of the men are conservative carnivores, while only $1/111 \approx 0.9\%$ of the women are conservative carnivores.

   b) Liberal vegetarians are more likely to be women. Of the 17 liberal vegetarians, 12 of them are women. $12/17 \approx 70.6\%$ of liberal vegetarians are women.

10. **Being successful.**

    a) The table contains column percentages, so we can add percentages within each column. 18% + 48% = 66% of young women consider it very important or one of the most important things for them to be personally successful.

    b) The percentage is lower for men. 11% + 47% = 58% of young men consider it very important or one of the most important things for them to be personally successful.

    c) We cannot determine what percent of all women (from both age groups) feel this way, since we don't have the sample sizes from each age group, only the percentages.

    d) Young women are slightly more likely than young men (66% compared to 58%) to consider financial success important. Young men are slightly more likely to view this as somewhat important or not important.

Chapter Exercises

11. Movie genres and ratings.

a) $452/1529 \approx 29.6\%$ of the films were rated R.

b) $124/1529 \approx 8.1\%$ of the films were R-rated comedies.

c) $124/452 \approx 27.4\%$ of the R-rated films were comedies.

d) $124/312 \approx 39.7\%$ of the comedies were rated R.

12. Not the labor force.

a) $2207/12872 \approx 17.1\%$ of the unemployed were available to work now.

b) $1048/12872 \approx 8.1\%$ of the unemployed were available to work now and aged 25 to 54 years.

c) $208/4158 \approx 5.0\%$ of the unemployed 16- to 24-year-olds were in school or training.

d) $4158/12872 \approx 32.3\%$ of the unemployed were aged 16 to 24 years.

13. Tables in the news. Answers will vary.

14. Graphs in the news. Answers will vary.

15. Poverty and region 2012.

The differences in poverty are not huge, but they may be real. The Northeast and Midwest have the lowest percentages of people living below the poverty level: 12.7% and 13.7%, respectively. In the West, 15.4% live below the poverty level, and the South has the highest rate at 16.8%.

16. Moviegoers and ethnicity.

a) Caucasians are 66% of the population but only 56% of ticket buyers. Hispanics are only 16% of the population, but they are 26% of ticket buyers. African Americans are more consistent: 12% of the population and 12% of the moviegoers.

b) There are ethnic differences among moviegoers, but they are not all that pronounced.

17. Death from the sky.

a) $100 - 60 - 30 - 1 - 0.2 - 0.17 = 8.63\%$ of estimated deaths are attributed to causes not listed here.

b) Regardless of the type of display chosen, it is difficult to display causes of death with percentages as low as 0.2% and 0.17%.

18. Cartoons.

a) Without a fun item: 61.4%, 59.1%, 61.4%, 17.0%, 22.7%

With fun item: 48.8%, 52.5%, 55.0%, 20.0%, 27.5%

Using relative frequencies makes these easier to compare, since the number of students in each group is different.

b) There is no evidence that cartoons help with understanding drawing a sample, the meaning of "random", or mean versus median, since the percentage of students answering the item correctly was higher in the group without the fun item. The percentage of students correctly answering the question about categorical versus quantitative variables was higher for the group with the fun item, but not significantly higher. There appears to be some evidence that students learn better using a fun item when studying the inadequacy of mean without standard deviation. The percentage of students answering the question correct in the fun item group was 4.8 percentage points higher than in the group with no fun item, 27.5% to 22.5%. However, with only 27.5% accuracy, it can't be said that they mastered the topic.

19. Smoking.

a) The smoking rate for 18-24-year-old men was 42.1% in 1974.

b) From 1974 to 2014, the smoking rate for 18-24-year-old men dropped from 42.1% to 18.5%

c) Men who were 18-24 years old in 1974 are in the 35-44 age group in 1994, the 45-54 age group in 2004, and the 2014. The smoking rate for this cohort has been decreasing through the years, from 42.1% to 33.2% to 26.7% to 18.8%. Although we don't have data on deaths in this table, it may very well be that the smokers have a higher death rate than the non-smokers, so this decrease doesn't necessarily mean that men in this cohort are quitting smoking.

20. Smoking women.

a) In each year, the women had a lower smoking rate than the men. Women had a 18-24-year-old smoking rate of 34.0% in 1974, compared to the 42.1% rate for men of the same age. The 18-24-year-old female smoking rates decline slowly over the years, and by 2014, the smoking rate for women is 16.5%, only 2 percentage points lower than the men.

b) Smokers are prone to lung cancer and other diseases. The smokers are less likely to live into old age and those left are less likely to be smokers.

21. Mothers and fathers 1965-2011.

a) Fathers spend the vast majority of their time on paid work, while mothers spend more time on child care and house work.

b) The time fathers spend on paid work has decreased, and the time they spend on child care and housework has increased. For mothers, the number of hours spent on paid work has significantly increased, and they have also increased their time spent on child care while reducing housework time.

c) Parents are spending more time on child care and paid work (13 hours to 21 hours and 50 hours to 58 hours). The time spent on housework has decreased from 36 hours to 28 hours.

d) Overall, parents in 2011 reported spending more time total on these tasks, a total of 107 hours in 2011 compared to 99 hours in 1965. Mothers increased their total working time by 3 hours, from 50 hours to 53 hours, while fathers increased their total working time by 5 hours, from 49 hours to 54 hours.

22. Mothers' and fathers' aspirations.

a) The actual percentages for working full-time differ the most, with 84% of fathers and 48% of mothers working full-time. This is a difference of 36 percentage points.

b) 48% of mothers work full-time, while only 27% desire to work full-time.

c) Only 7% of fathers stay at home, compared to 30% of mothers.

23. Teen smokers.

According to the Monitoring the Future study, teen smoking brand preferences differ somewhat by region. Although Marlboro is the most popular brand in each region, with about 58% of teen smokers preferring this brand in each region, teen smokers from the South prefer Newports at a higher percentage than teen smokers from the West, 22.5% to approximately 10%, respectively. Camels are more popular in the West, with 9.5% of teen smokers preferring this brand, compared to only 3.3% in the South. Teen smokers in the West are also more likely to have no particular brand than teen smokers in the South. 12.9% of teen smokers in the West have no particular brand, compared to only 6.7% in the South. Both regions have about 9% of teen smokers that prefer one of over 20 other brands.

24. Being successful revisited.

a) The four groups are approximately the same size, since the columns are approximately the same width.

b) Age is the more important factor. The male/female distributions are more similar to each another within each age group than the male/male and female/female distributions are across age groups. The older group is more likely to respond Not important; the younger more likely to respond Very important but not the most.

c) Women aged 18-34 are most likely to respond that being successful is the most important.

25. Diet and politics IV.

a) There are more men in the survey. The male columns are generally wider than the female columns.

b) We can't compare the genders within each category of political ideology since the sample sizes differ. We can, however, note that the male bars are narrowest in the Liberal category, and widest in the Conservative category, indicating that a small share of liberals and a large share of conservatives among the men. The women show the opposite association. The widest female bar is Liberal, while the narrowest is Conservative. In other words, there is an association between gender and political ideology. Males tend to be more conservative and females tend to be more liberal.

c) There is an association between politics and diet. Conservatives are more likely to be carnivores, while liberals are more likely to be vegetarians.

d) The association between politics and diet seems to differ between men and women. Differences in vegetarianism across political ideology is more pronounces in females than

in males. Differences in carnivorous eating habits across political ideology is more pronounced in males than females.

26. Handguns.

76.4% of handguns involved in Milwaukee buyback programs are small caliber, while only 20.3% of homicides are committed with small caliber handguns. Along the same lines, only 19.3% of buyback handguns are of medium caliber, while 54.7% of homicides involve medium caliber handguns. A similar disparity is seen in large caliber handguns. Only 2.1% of buyback handguns are large caliber, but this caliber is used in 10.8% of homicides. Finally, 2.2% of buyback handguns are of other calibers, while 14.2% of homicides are committed with handguns of other calibers. Generally, the handguns that are involved in buyback programs are not the same caliber as handguns used in homicides in Milwaukee.

27. Job satisfaction.

a) This is a table of column percents. The columns add up to 100%, while the rows do not.

b) i) This can't be found from the table. We don't know what the percent of respondents who are very satisfied.

ii) This can't be found from the table. We don't know what the percent of respondents who are dissatisfied.

iii) 39% of respondents who are dissatisfied with their current job are actually better off than their parents were at the same age.

iv) This can't be found from the table. We don't know what the percent of respondents who are very satisfied.

28. A sense of identity.

a) These are row percents. The rows add up to nearly 100%, while the columns do not. The differences from 100% are too large to be attributed to rounding error, so it is likely that some respondents replied something else other than the provided responses.

b) i) This can't be found from the table. We don't know what the percent of respondents who are self-employed.

ii) 65% of respondents who are employed by non-profits get a sense of identity from their job.

iii) This can't be found from the table. We don't know what the percent of respondents who think of their job as just something to do for a living.

iv) This can't be found from the table. We don't know what the percent of respondents who get a sense of identity from their job.

29. Seniors.

a) A table with marginal totals is to the right. There are 268 White graduates and 325 total graduates. $268/325 \approx 82.5\%$ of the graduates are white.

Plans	White	Minority	TOTAL
4-year college	198	44	242
2-year college	36	6	42
Military	4	1	5
Employment	14	3	17
Other	16	3	19
TOTAL	268	57	325

b) There are 42 graduates planning to attend 2-year colleges. $42/325 \approx 12.9\%$

c) 36 white graduates are planning to attend 2-year colleges. $36/325 \approx 11.1\%$

d) 36 white graduates are planning to attend 2-year colleges and there are 268 whites graduates. $36/268 \approx 13.4\%$

e) There are 42 graduates planning to attend 2-year colleges, and 36 of them are white. $36/42 \approx 85.7\%$

30. Politics.

a) There are 192 students taking Intro Stats. Of those, 115, or about 59.9%, are male.

b) There are 192 students taking Intro Stats. Of those, 27, or about 14.1%, consider themselves to be "Conservative".

c) There are 115 males taking Intro Stats. Of those, 21, or about 18.3%, consider themselves to be "Conservative".

d) There are 192 students taking Intro Stats. Of those, 21, or about 10.9%, are males who consider themselves to be "Conservative".

31. Movies 06-15.

a) This is a table of column percents. The columns add up to 100%, while the rows do not.
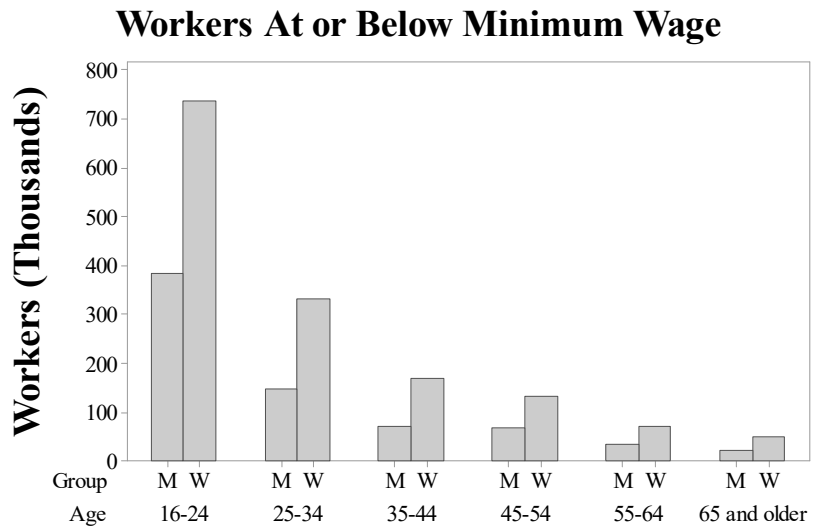
b) Movies rated G and PG have become slightly less common, while movies rated PG-13 and R have become slightly more common.

c) For Dramas, the percentages of PG and R have decreased while the percentage of PG-13 dramas has increased significantly. For Comedies, there has been a large increase in the percentage of R-rated films.

32. Minimum wage workers.

a) $7701/37972 \approx 20.3\%$ of the woman are aged 16-24.

b) The number of workers at or below minimum wage is highest for each sex at 16-24 years, with 384,000 male workers and 738,000 female workers. The number of workers at or below minimum wage decreases as the age increases for each sex, with only 22,000 male and 50,000 female workers. At each age category, the number of male workers at or below minimum wage is roughly half the number of female workers.


Workers At or Below Minimum Wage

33. More about seniors.

a) For white students, 73.9% plan to attend a 4-year college, 13.4% plan to attend a 2-year college, 1.5% plan on the military, 5.2% plan to be employed, and 6.0% have other plans.

b) For minority students, 77.2% plan to attend a 4-year college, 10.5% plan to attend a 2-year college, 1.8% plan on the military, 5.3% plan to be employed, and 5.3% have other plans.


Post High School Plans
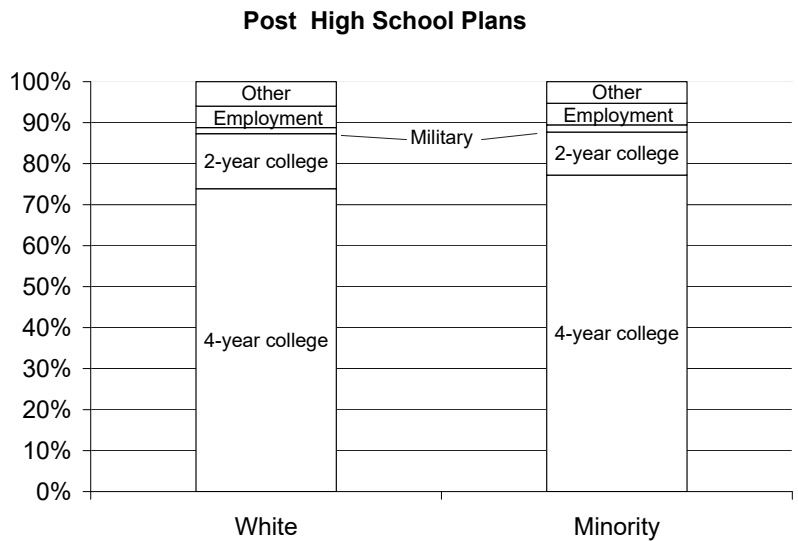
c) A segmented bar chart is a good display of these data.

d) The conditional distributions of plans for Whites and Minorities are similar: White – 74% 4-year college, 13% 2-year college, 2% military, 5% employment, 6% other. Minority – 77% 4-year college, 11% 2-year college, 2% military, 5% employment, 5% other. Caution should be used with the percentages for Minority graduates, because the total is so small. Each graduate is almost 2%. Still, the conditional distributions of plans are essentially the same for the two groups. There is little evidence of an association between race and plans for after graduation.

34. Politics revisited.

Politics of an Intro Stats Course

a) The females in this course were 45.5% Liberal, 46.8% Moderate, and 7.8% Conservative.

b) The males in this course were 43.5% Liberal, 38.3% Moderate, and 18.3% Conservative.

c) A segmented bar chart comparing the distributions is at the right.

d) Politics and sex do not appear to be independent in this course. Although the percentage of liberals was roughly the same for each sex, females had a greater percentage of moderates and a lower percentage of conservatives than males.
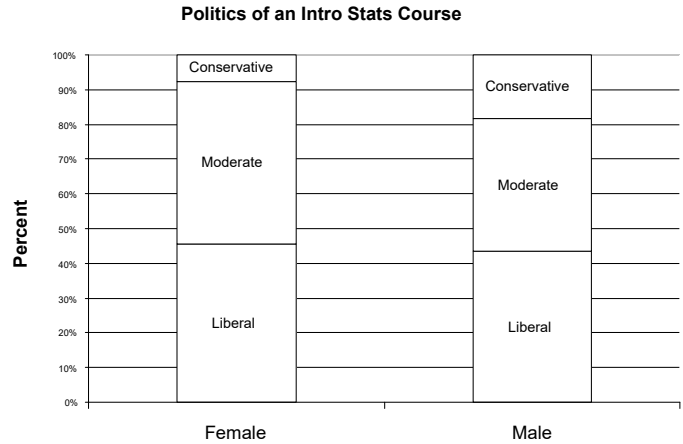
35. Magnet schools revisited.

a) There were 1755 qualified applicants to the Houston Independent School District's magnet schools program. Of those, 292, or about 16.6% were Asian.

b) There were 931 students accepted to the magnet schools program. Of those, 110, or about 11.8% were Asian.

c) There were 292 Asian applicants. Of those, 110, or about 37.7%, were accepted.

d) There were 1755 total applicants. Of those, 931, or about 53%, were accepted.

36. More politics.

a)

Distribution of Sex Across Political Categories
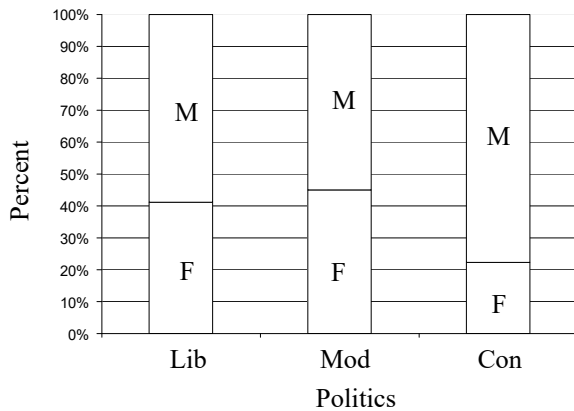
b) The percentage of males and females varies across political categories. The percentage of self-identified Liberals and Moderates who are female is about twice the percentage of Conservatives who are female. This suggests that *sex* and *politics* are not independent.

37. Back to school.

There were 1,755 qualified applicants for admission to the magnet schools program. 53% were accepted, 17% were wait-listed, and the other 30% were turned away. While the overall acceptance rate was 53%, 93.8% of Blacks and Hispanics were accepted, compared to only 37.7% of Asians, and 35.5% of whites. Overall, 29.5% of applicants were Black or Hispanics, but only 6% of those turned away were Black or Hispanic. Asians accounted for 16.6% of applicants, but 25.3% of those turned away. It appears that the admissions decisions were not independent of the applicant's ethnicity.

38. Parking lots.

a) In order to get percentages, first we need totals. Here is the same table, with row and column totals. Foreign cars are defined as non-American. There are 45+102=147 non-American cars or 147/359 ≈ 40.95%.

	Driver		
Origin	Student	Staff	Total
American	107	105	212
European	33	12	45
Asian	55	47	102
Total	195	164	359

b) There are 212 American cars of which 107 or 107/212 ≈ 50.47% were owned by students.

c) There are 195 students of whom 107 or 107/195 ≈ 54.87% owned American cars.

d) The marginal distribution of Origin is displayed in the third column of the table at the right: 59% American, 13% European, and 28% Asian.

Origin	Totals
American	212 (59%)
European	45 (13%)
Asian	102 (28%)
Total	359

e) The conditional distribution of Origin for Students is: 55% (107 of 195) American, 17% (33 of 195) European, and 28% (55 of 195) Asian.
The conditional distribution of Origin for Staff is:
64.0% (105 of 164) American, 7.3% (12 of 164) European, and 28.7% (47 of 164) Asian.

f) The percentages in the conditional distributions of Origin by Driver (students and staff) seem slightly different. Let's look at a segmented bar chart of Origin by Driver, to compare the conditional distributions graphically.

The conditional distributions of Origin by Driver have similarities and differences. Although students appear to own a higher percentage of European cars and a smaller percentage of American cars than the staff, the two groups own nearly the same percentage of Asian cars. However, because of the differences, there is evidence of an association between Driver and Origin of the car.

Conditional Distribution of Origin by Driver
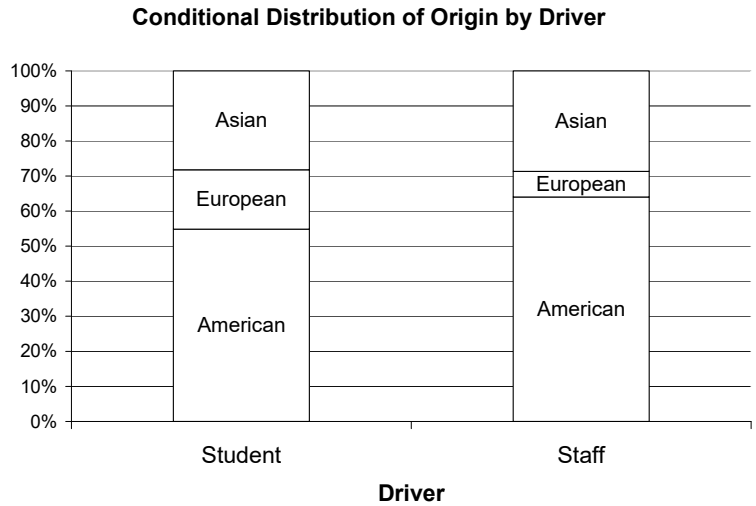
39. Weather forecasts.

a) The table shows the marginal totals. It rained on 34 of 365 days, or 9.3% of the days.

b) Rain was predicted on 90 of 365 days. $90/365 \approx 24.7\%$ of the days.

		Actual Weather		
		Rain	No Rain	**Total**
Forecast	Rain	27	63	90
	No Rain	7	268	275
	Total	34	331	365

c) The forecast of Rain was correct on 27 of the days it actually rained and the forecast of No Rain was correct on 268 of the days it didn't rain. So, the forecast was correct a total of 295 times. $295/365 \approx 80.8\%$ of the days.

d) On rainy days, rain had been predicted 27 out of 34 times (79.4%). On days when it did not rain, forecasters were correct in their predictions 268 out of 331 times (81.0%). These two percentages are very close. There is no evidence of an association between the type of weather and the ability of the forecasters to make an accurate prediction.

Weather Forecast Accuracy
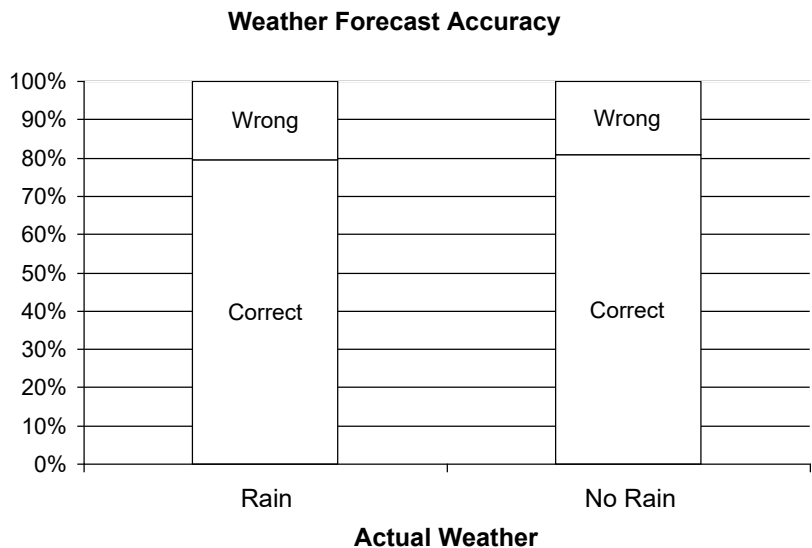
40. Twin births.

a) Of the 278,000 mothers who had twins in 1995-1997, 63,000 had inadequate health care during their pregnancies. 63,000/278,000 = 22.7%

Twin Births 1995-97 (in thousands)				
Level of Prenatal Care	**Preterm (Induced or Caesarean)**	**Preterm (without procedures)**	**Term or Postterm**	**Total**
Intensive	18	15	28	61
Adequate	46	43	65	154
Inadequate	12	13	38	63
Total	76	71	131	278

b) There were 76,000 induced or Caesarean births and 71,000 preterm births without these procedures. (76,000 + 71,000)/278,000 = 52.9%

c) Among the mothers who did not receive adequate medical care, there were 12,000 induced or Caesarean births and 13,000 preterm births without these procedures. 63,000 mothers of twins did not receive adequate medical care. (12,000 + 13,000)/63,000 = 39.7%

d) A segmented bar chart is at the right.
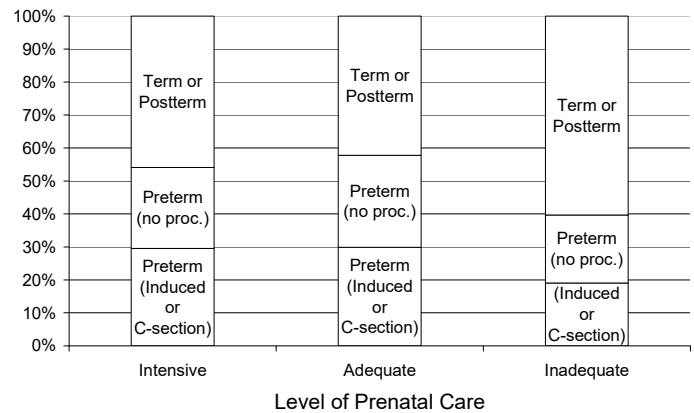
e) 52.9% of all twin births were preterm, while only 39.7% of births in which inadequate medical care was received were preterm. This is evidence of an association between level of prenatal care and twin birth outcome. If these variables were independent, we would expect the percentages to be roughly the same. Generally, those mothers who received adequate medical care were more likely to have preterm births than mothers who received intensive medical care, who were in turn more likely to have preterm births than mothers who received inadequate health care. This does *not* imply that mothers should receive inadequate health care do decrease their chances of having a preterm birth, since it is likely that women that have some complication *during* their pregnancy (that might lead to a preterm birth), would seek intensive or adequate prenatal care.
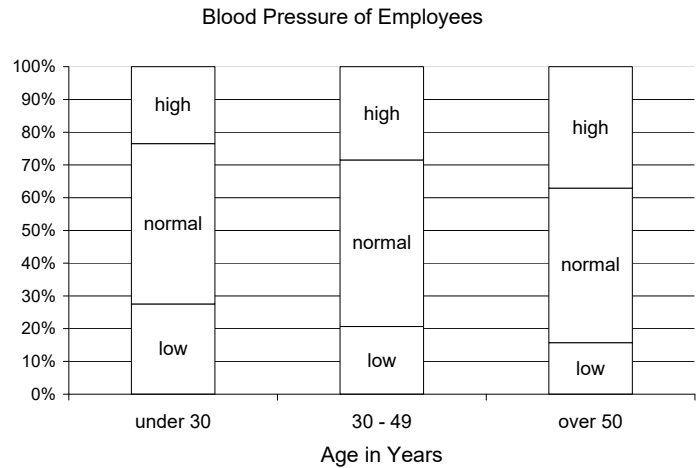
41. Blood pressure.

a) The marginal distribution of blood pressure for the employees of the company is the total column of the table, converted to percentages. 20% low, 49% normal and 31% high blood pressure.

Blood pressure	under 30	30 - 49	over 50	Total
low	27	37	31	95
normal	48	91	93	232
high	23	51	73	147
Total	98	179	197	474

b) The conditional distribution of blood pressure within each age category is:
Under 30 : 28% low, 49% normal, 23% high
30 – 49 : 21% low, 51% normal, 28% high
Over 50 : 16% low, 47% normal, 37% high

c) A segmented bar chart of the conditional distributions of blood pressure by age category is at the right.

Blood Pressure of Employees

d) In this company, as age increases, the percentage of employees with low blood pressure decreases, and the percentage of employees with high blood pressure increases.

e) No, this does not prove that people's blood pressure increases as they age. Generally, an association between two variables does not imply a cause-and-effect relationship. Specifically, these data come from only one company and cannot be applied to all people. Furthermore, there may be some other variable that is linked to both age and blood pressure. Only a controlled experiment can isolate the relationship between age and blood pressure.

42. Obesity and exercise.

a) Participants were categorized as Normal, Overweight or Obese, according to their Body Mass Index. Within each classification of BMI (column), participants self reported exercise levels. Therefore, these are column percentages. The percentages sum to 100% in each column, *not* across each row.

b) A segmented bar chart of the conditional distributions of level of physical activity by Body Mass Index category is at the right.

Body Mass Index and Level of Physical Activity
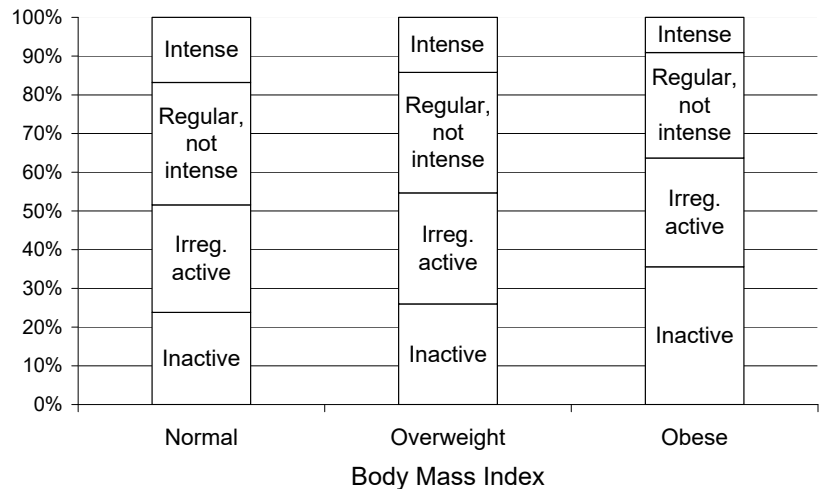
c) No, even though the graphical displays provide strong evidence that lack of exercise and BMI are not independent. All three BMI categories have nearly the same percentage of subjects who report "Regular, not intense" or "Irregularly active", but as we move from Normal to Overweight to Obese we see a decrease in the percentage of subjects who report "Regular, intense" physical activity (16.8% to 14.2% to

9.1%), while the percentage of subjects who report themselves as "Inactive" increases. While it may seem logical that lack of exercise causes obesity, association between variables does not imply a cause-and-effect relationship. A lurking variable (for example, overall health) might influence both BMI and level of physical activity, or perhaps lack of exercise is *caused by* obesity. Only a controlled experiment could isolate the relationship between BMI and level of physically activity.

43. Anorexia.

These data provide no evidence that Prozac might be helpful in treating anorexia. About 71% of the patients who took Prozac were diagnosed as "Healthy", while about 73% of the patients who took a placebo were diagnosed as "Healthy". Even though the percentage was higher for the placebo patients, this does not mean that Prozac is hurting patients. The difference between 71% and 73% is not likely to be statistically significant.

44. Antidepressants and bone fractures.

These data provide evidence that taking a certain class of antidepressants (SSRI) might be associated with a greater risk of bone fractures. Approximately 10% of the patients taking this class of antidepressants experience bone fractures. This is compared to only approximately 5% in the group that were not taking the antidepressants.

45. Driver's licenses 2014.

a) There are 8.5 million drivers under 20 and a total of 214.1 million drivers in the U.S. That's about 4.0% of U.S. drivers under 20.

b) There are 105.9 million males out of 214.1 million total U.S. drivers, or about 49.5%.

c) Each age category appears to have about 50% male and 50% female drivers. At younger ages, males form the slight majority of drivers. The percentage of male drivers shrinks until female drivers hold a slight majority, among older drivers.

d) There appears to be a slight association between age and gender of U.S. drivers. Younger drivers are slightly more likely to be male, and older drivers are slightly more likely to be female.