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

Pearson

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 330 Hudson Street, NY NY 10013

ISBN-13: 978-0-13-426536-0
ISBN-10: 0-13-426536-X

Contents

Chapter 1	Stats Starts Here	*1*
Chapter 2	Displaying and Describing Data	*7*
Chapter 3	Relationships Between Categorical Variables -Contingency Tables	*17*
Chapter 4	Understanding and Comparing Distributions	*35*
Chapter 5	The Standard Deviation as Ruler and the Normal Model	*53*
Review of Part I	Exploring and Understanding Data	*78*
Chapter 6	Scatterplots, Association, and Correlation	*97*
Chapter 7	Linear Regression	*112*
Chapter 8	Regression Wisdom	*145*
Chapter 9	Multiple Regression	*182*
Review of Part II	Exploring Relationships Between Variables	*190*
Chapter 10	Sample Surveys	*216*
Chapter 11	Experiments and Observational Studies	*227*
Review of Part III	Gathering Data	*245*
Chapter 12	From Randomness to Probability	*257*
Chapter 13	Sampling Distribution Models and Confidence Intervals for Proportions	*283*
Chapter 14	Inferences About Means	*301*
Chapter 15	Testing Hypotheses	*327*
Chapter 16	More About Tests and Intervals	*359*
Review of Part IV	From the Data at Hand to the World at Large	*375*
Chapter 17	Comparing Groups	*403*
Chapter 18	Paired Samples and Blocks	*451*
Chapter 19	Comparing Counts	*472*
Chapter 20	Inferences for Regression	*497*
Review of Part V	Inference for Relationships	*524*
Parts I – V	*Cumulative Review Exercises*	*566*

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.

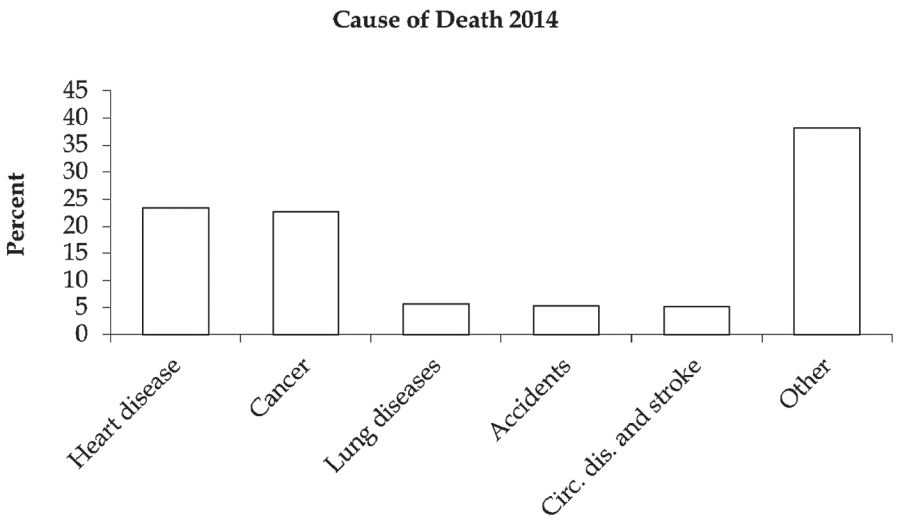
b) It is easy to tell from either chart; sometimes differences are easier to see on the bar chart because slices of the pie chart look too similar in size.

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.



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.

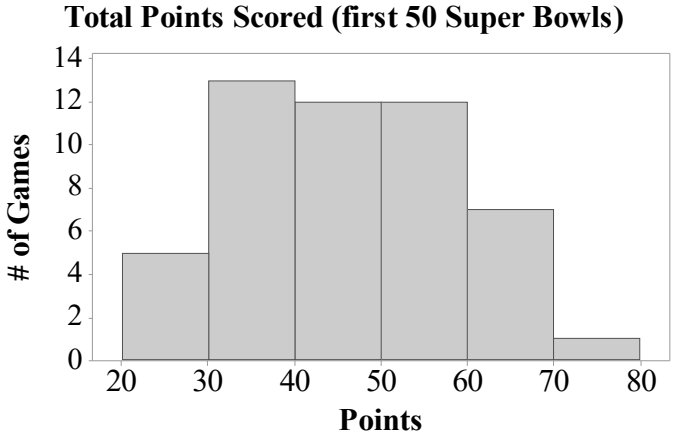
47. Heart attack stays.

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.

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.

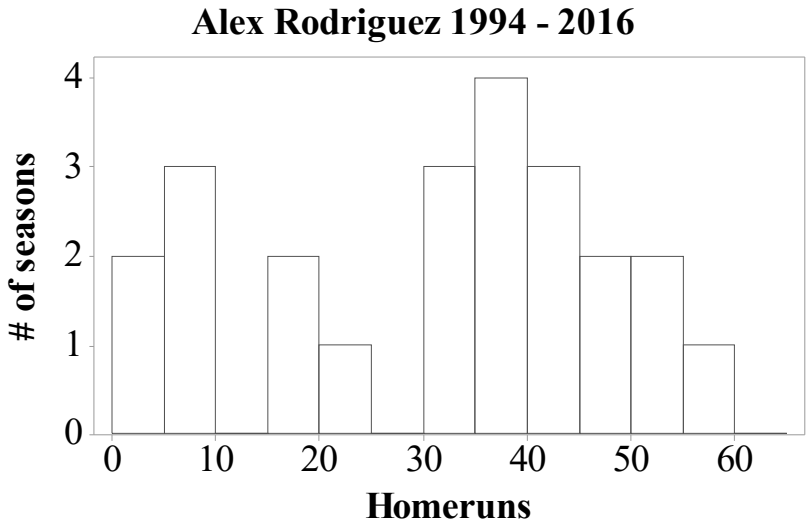
73. A-Rod 2016.

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.

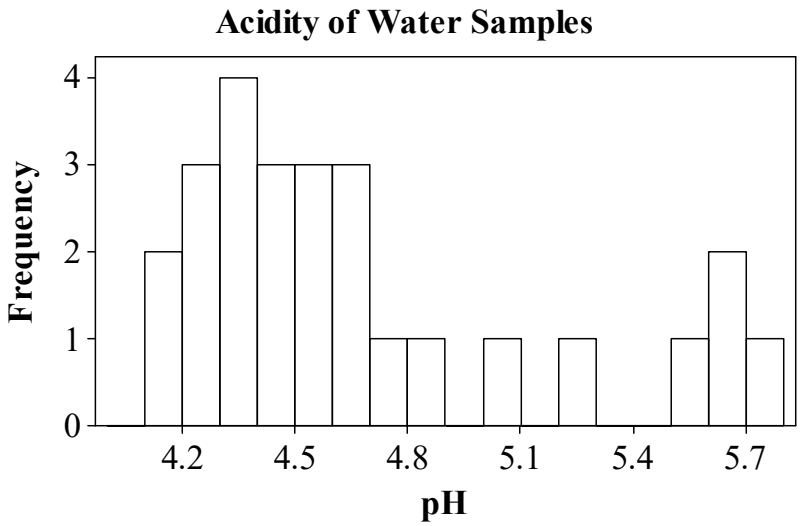
a) This is not a histogram. The horizontal axis should contain the number of home runs per year, split into bins of a convenient width. The vertical axis should show the frequency; that is, the number of years in which A-Rod hit a number of home runs within the interval of each bin. The display shown 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.

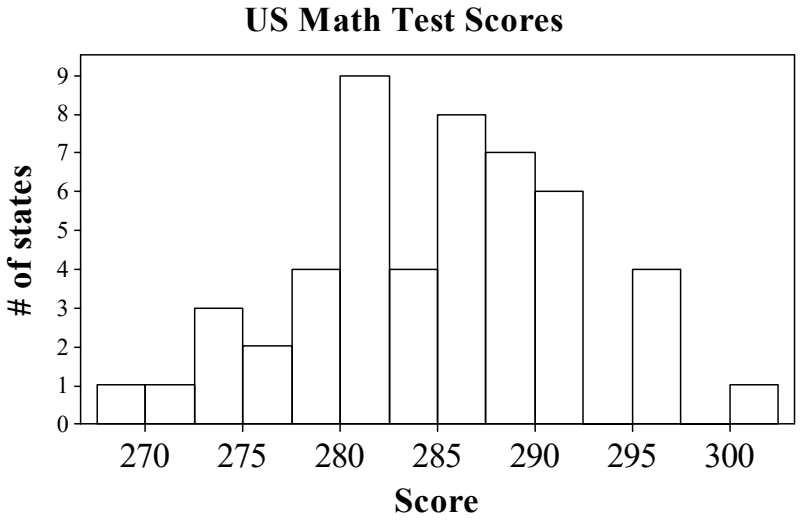
79. Final grades.

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%.