TT TEXTBOOK TESTBANKS

Solutions Manual for Statistics 5th Agresti

SOLUTIONS MANUAL SAMPLE PREVIEW

PUBLISHER
Pearson

COPYRIGHT
2021

RESOURCE TYPE
Solutions Manual

ISBN
9780136559504

INSTRUCTOR'S SOLUTIONS MANUAL

JAMES LAPP

STATISTICS: THE ART AND SCIENCE OF LEARNING FROM DATA

FIFTH EDITION

Alan Agresti

University of Florida

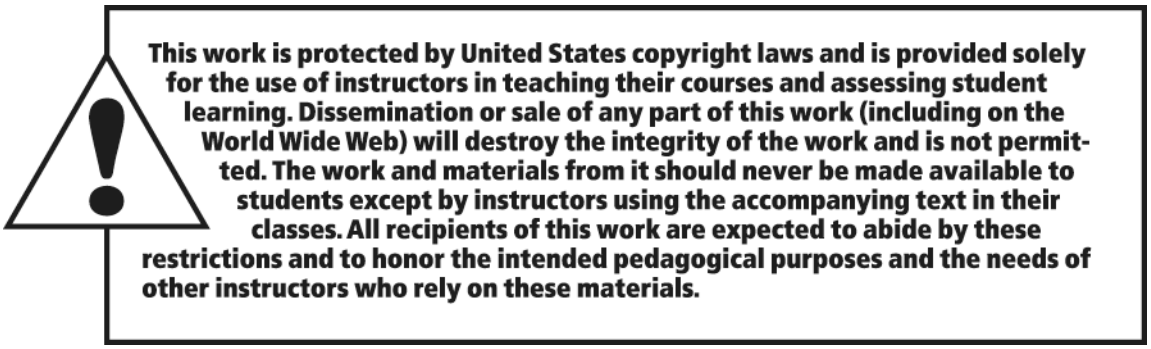
Christine Franklin

University of Georgia

Bernhard Klingenberg

Williams College and New College of Florida

Pearson



ISBN-13: 978-0-13-655984-9
ISBN-10: 0-13-655984-0

Table of Contents

Section 1.1: Using Data to Answer Statistical Questions

1.1 Aspirin and heart attacks

a. Aspects of the study that have to do with design include the sample of 22,000 physicians, the randomization of the halves of the sample to the two groups (aspirin and placebo), and the plan to obtain percentages of each group that have heart attacks.

b. Aspects having to do with description include the actual percentages of the people in the sample who have heart attacks (i.e., 0.9% for those taking aspirin and 1.7% for those taking placebo).

c. Aspects that have to do with inference include the use of statistical methods to conclude that taking aspirin reduces the risk of having a heart attack.

1.2 Poverty and race

a. The statistical question of interest is "What is the association between poverty and race for American households?". The data will come from the 75,000 households surveyed in the CPS.

b. Aspects having to do with description include the actual percentages of White, Black, and Asian households with income below the poverty line.

c. Aspects that have to do with inference include the prediction that among *all* Black households (not just the ones surveyed), it is plausible that between 27.5% and 29.1% have incomes below the poverty line.

1.3 GSS and heaven

a. 4,854 (as of 2020)

b. Yes, definitely: 63.8%, Yes, probably: 20.6%, No, probably not: 9.1%, No, definitely not: 6.6%

c. 2018: 61.2%, 2008: 64.3% (about 3 percentage points smaller)

1.4 GSS and heaven and hell

a. Yes, definitely: 52.3%; Yes, probably: 20.1%; No, probably not: 14.5%; No, definitely not: 13.1%

b. Yes, definitely: 51.6%; The percentage of "yes, definitely" responses was higher for belief in heaven in 2008 than in hell (61.2%)

1.5 GSS and work hours

a. 575 people responded to the question.

b. The most common response was 40 hours (36.9%).

Section 1.2: Sample Versus Population

1.6 Description and inference

a. With description, we are summarizing a group of numbers. We can use description with either samples or populations. With inference, we use data from samples to make conclusions or predictions about populations. For example, if we ask a sample of adults how many pets they own, and take the mean number of pets, that number is a description. If we use that number to predict the mean number of pets owned by the whole population, the predicted mean (or the predicted range for the mean) would be an inference.

b. Descriptive statistics would be useful to summarize data from a population. With a census, it would be unwieldy to examine everyone's ages, for example, but it would be useful to know a mean age. Inferential statistics are not needed, however, because we already have information about the population; we don't need to predict it.

1.7 Censorship

a. The sample is the 3,077 people who responded.

b. The population is all adults in the United States.

c. The statistic is 23% of respondents said antireligious books should be removed.

d. The population parameter is the proportion in the entire U.S. population who think antireligious books should be removed.

1.8 Concerned about global warming?

a. The sample is the set of polled Floridians. The population is the set of all adult Florida residents.

b. The percentages quoted are statistics since they are summaries of the sample.

1.9 Graduate school information

a. Each student in the program is a subject.

b. The sample is the students identified for an interview from the given program.

c. The population is all students in the program.

1.10 Is globalization good?

a. The samples are those people selected from each country to participate in the survey. The populations are all adults in Africa and all adults in North America.

b. These are statistics because they represent a summary of the sample data.

1.11 Graduating seniors' salaries

a. These are descriptive statistics. They are summarizing data from a population – all graduating seniors at a given school.

b. These analyses summarize data on a population – all graduating seniors at a given school; thus, the numerical summaries are best characterized as parameters.

1.12 At what age did women marry?

a. The mean age of 24.1 years for this sample is descriptive.

b. The historian estimates the age for the whole population of brides in early 19th century New England, estimating the average age to fall between 23.5 and 24.7. This is inferential.

c. The inference refers to the population of all New England brides between the years of 1800 and 1820.

d. The average of 24.1 years is based on a sample and is therefore a statistic.

1.13 Age pyramids as descriptive statistics

a. The bar graph for 1750 shows shorter and shorter bars as age increases indicating that there were few Swedish people who were old in 1750.

b. For every age range, the bars are much longer for both men and women in 2010 than in 1750.

c. The bars for women in their 70's and 80's in 2010 are longer than those for men of the same age in the same year.

d. The first manned space flight took place in 1961 so that people born during this era would fall in the 45–49 age group. This is the largest five-year group for both men and women.

1.14 Margin of Error in labor statistics

a. Yes; since $\$74,900 - \$680 = \$74,220$ and $\$74,900 + \$680 = \$75,580$, \$75,000 is well within the margin of error of \$680 for \$74,900.

b. No; we would expect that the national unemployment rate to be between $4.1\% - 0.1\% = 4.0\%$ and $4.1\% + 0.1\% = 4.2\%$. 3.5% is well below the lower value and, hence, not plausible.

1.15 National service

a. Yes, the populations are the same in the two studies. For both, it's all students at your school.

b. It is *very* unlikely that you will choose the same 20 students.

c. Although it is most likely that the sample proportions will not be the same, they should be close to each other.

1.16 Samples vary less with more data

a. It would be more surprising to flip a coin 500 times and observe all heads.

b. As the sample size increases, the amount by which sample proportions tend to vary decreases. The estimates from larger samples, therefore, tend to be more accurate than estimates from smaller samples. When the coin is flipped just 5 times, it's easy to see that we could get a sample with all heads. However, when the number of flips is increased to 500, it is much more likely that the sample proportion is near the population proportion of 0.5. It would be extremely unlikely to observe very few heads or almost all heads in 500 flips of a fair coin.

c. Answers will vary. For one set of 25 simulations flipping the coin five times each, the smallest proportion was 0.0 (i.e., 0% heads) and the largest 1 (i.e., 100%). 60% heads was the most common sample proportion.

d. Answers will vary. For one set of 25 simulations flipping the coin 500 times each, the smallest proportion was 0.44 and the largest was 0.54.

1.17 Comparing polls

a. The CNN poll predicts that the true proportion of Obama voters is within 3 percentage points of 49%.

b. The first four polls are all within the margin of error. Rand overestimated and Fox underestimated Obama's proportion. Generally, the polls are fairly accurate.

c. Answers will vary. For one set of 10 simulations, the smallest simulated proportion was 0.48 and the largest was 0.54.

1.18 Ebola outbreaks

The answer to this problem is based on a random process. This leads to potentially different answers each time it is performed. The binomial distribution (see Section 6.3) says that 14 or fewer people who died should occur in only about 1 in 100 simulations, so most students will likely not see any of these situations.

1.19 Statistical Significance

a. ii b. i c. i

1.20 Smoking cessation

The best option is iii.

Section 1.3: Organizing Data, Statistical Software, and the New Field of Data Science

1.21 Data file for friends

The results for this exercise will be different for each person who does it. The data files, however, should all look like this:

Friend	Characteristic 1	Characteristic 2
1		
2		
3		
4		

For each friend, you'll have a number or label under characteristics 1 and 2. For example, if you asked each friend for gender and hours of exercise per week, the first friend might have m (for male) under Characteristic 1, and 6 (for hours exercised per week) under Characteristic 2.

1.22 Shopping sales data file

Customer	Clothes	Sporting goods	Books	Food
1	$49	$0	$0	$16
2	$0	$150	$43	$25
3	$0	$0	$0	$0
4	$87	$15	$0	$12

1.23 Cleaning data and missing values

Traveler	Check in (min)	Luggage ($)	Airbnb
1	30	25	No
2	90	60	Yes
3	NA	NA	Yes

1.24 Wealth and health

Answers will vary. The following table reflects 2019 data.

Country	Per capita GDP (dollars/person)	Life Expectancy (years)
Afghanistan	502.10	64
Greece	19,582.50	82
Mexico	9,863.10	75
⋮	⋮	⋮

1.25 Create a data file with software

Your data file (from Exercise 1.22) will be in the following format.

Customer	Clothes	Sporting goods	Books	Food
1	$49	$0	$0	$16
2	$0	$150	$43	$25
3	$0	$0	$0	$0
4	$87	$15	$0	$12

1.26 Downloading a data file

a. Answers will vary.

b. Answers will vary.

1.27 Train and Test Data

The training data is the data set with labels. The test data is the data set without labels.

1.28 Cambridge Analytica

Cambridge Analytica allegedly obtained personal data from millions of Facebook users through personality quizzes and an app run on Facebook. They used the personality profiles for targeting Facebook users with specific political messages trying to influence their voting behavior in the 2016 U.S. presidential election.

1.29 Ethics in Data Analysis

For example, on Wikipedia, the number of articles mentioning males is much greater than the number of articles mentioning females. An algorithm trained on Wikipedia data for identifying the target of pronouns may do much better for male-related pronouns than female-related pronouns.

Chapter Exercises: Practicing the Basics

1.30 UW student survey

a. The population is the entire UW student body of 40,858. The sample is the 100 students who were asked to complete the questionnaire.

b. This value would not necessarily equal the value for the entire population of UW students. It is quite possible that the sample of 100 is not exactly representative of the whole student body. This percentage is only an estimate of the percentage of all students who would respond this way. It is unlikely that any single sample of 100 would have a percentage that was exactly the percentage of the entire population.

c. The numerical summary is a sample statistic because it only summarizes for a sample, not for a population.

1.31 Euthanasia

a. The population is all American adults.

b. The sample data are summarized by a proportion, 0.598.

c. The population proportion who would commit suicide.

1.32 Sleep disorders among college students

It is very likely that between 25% and 29% of students are at risk for at least one sleep disorder.

1.33 Breaking down Brown versus Whitman

a. The results summarize sample data because not every voter in the 2010 California gubernatorial election was polled.

b. The percentages reported here are descriptive in that they describe the exact percentages of the sample polled who were Democrat and voted for Brown, who were Republican and voted for Brown and who were Independent and voted for Brown.

c. The inferential aspect of this analysis is that the exit poll results were used to predict what percentage of each of the three parties (Democrat, Republican and Independent) voted for Brown in the 2010 California gubernatorial election. The margins of error give a likely range for the population percentages for each of the three parties.

1.34 Online learning

a. The sample is the 100 students surveyed. The population is all students in this school.

b. (i) Descriptive statistics would give us information about the preferences of the 100 students in the sample.

(ii) Inferential statistics allow us to draw a conclusion about the preferences of the student body.

1.35 Website tracking

a. The sample is the 85 visitors to the site that day. The population of interest is all potential visitors to the site (in March).

b. The descriptive aspects are the average number of minutes the 85 visitors stayed on the site or the proportion of the 85 visitors who use a Chrome browser.
The inferential aspects are predicting the average number of minutes that visitors will stay on the site or checking if the proportion of all users of the site that use a Chrome browser is larger than 50%.

1.36 Support of labor unions

a. The range of likely values is $62\% - 4\% = 58\%$ to $62\% + 4\% = 56\%$.

b. i. descriptive statistics

c. ii. inferential statistics

Chapter Exercises: Concepts and Investigations

1.37 How sample means vary from one sample to the next

a. Answers will vary. For one set of simulations, the smallest sample proportion was 0.3 and the largest sample proportion was 1.0.

b. Answers will vary. For one set of simulations, the smallest sample proportion was 0.50 and the largest sample proportion was 0.71.

c. The sample size of 100 leads to a smaller margin of error.

d. Answers will vary. For one set of 100 simulations, the smallest sample proportion was 0.563 and the largest sample proportion was 0.638.

1.38 How sample proportions vary from one sample to the next

a. Answers will vary. For one set of simulations, the smallest sample average was 1.0 and the largest sample average was 5.5.

b. Answers will vary. For one set of simulations, the smallest sample average was 3.20 and the largest sample average was 4.35.

1.39 Statistics in the news

Answers will vary. If the article has numbers that summarize for a given group (sample or population), it's using descriptive statistics. If it uses numbers from a sample to predict something about a population, it's using inferential statistics.

1.40 What is statistics?

Answers will vary.

1.41 Surprising suicide data?

The likelihood of getting this result is extremely small.

1.42 Create a data file

See solution for Exercise 1.23 for format of data in MINITAB.

Chapter Exercises: Multiple Choice or True/False

1.43 Use of inferential statistics?

The best answer is (c).

1.44 NYC vs. Upstate significantly different?

The best answer is (c).

1.45 True or false?

False: We often want to describe the sample AND make inferences about the population.

1.46 True or false?

False.

Chapter Exercises: Student Activities

1.47 Margin of error

a. Answers will vary. One run of simulations yielded 0.491, 0.519, 0.496, 0.497, 0.492, 0.500, 0.518, 0.499, 0.503, and 0.497.

b. Answers will vary. All the intervals constructed using the values in (a) contained 0.50.

c. Answers will vary. About 95% of all intervals should contain 0.5.

1.48 Choosing a random sample

a. Answers will vary. With a class size of 30 and sampling 5 numbers, one simulation yielded 3, 4, 14, 28, and 22.

b. Answers will vary. With a class size of 30 and sampling 5 numbers, after 20 simulations, the number 2 was selected twice.

1.49 Gallup polls

Answers will vary. One possible answer is "70% of Americans say the U.S. should still play a strong role in world."

1.50 Getting to know the class

Answers will vary.

Section 2.1: Different Types of Data

2.1 Categorical/quantitative difference

a. Categorical variables are those in which observations belong to one of a set of categories, whereas quantitative variables are those on which observations are numerical.

b. An example of a categorical variable is religion. An example of a quantitative variable is temperature.

2.2 U.S. married-couple households

The variable summarized is categorical. The variable is type of U.S. married-couple households, and there are four types: traditional, dual-income with children, dual-income with no children, and other. These types are the categories.

2.3 Identify the variable type

a. quantitative

b. categorical

c. categorical

d. quantitative

2.4 Categorical or quantitative?

a. categorical

b. quantitative

c. categorical

d. quantitative

2.5 Discrete/continuous

a. A discrete variable is a quantitative variable for which the possible values are separate, such as 0, 1, 2, … . A continuous variable is a quantitative variable for which the possible values form an interval.

b. Example of a discrete variable: the number of children in a family (a given family can't have 2.43 children). Example of a continuous variable: temperature (we can have a temperature of 48.659).

2.6 Discrete or continuous?

a. continuous

b. discrete

c. continuous

d. discrete

2.7 Discrete or continuous 2

a. continuous

b. discrete

c. discrete

d. continuous

2.8 Nominal/ordinal

a. Categories for a nominal categorical variable are not ordered, while categories of an ordinal categorical variable can be ordered in some way, i.e., from worst to best.

b. Answers may vary. Nominal: Brand of Laptop (HP, Dell, Apple, Lenovo, …)
Ordinal: Rating of item on Amazon (1 star to 4 stars).

2.9 Nominal or ordinal?

a. The variable, gender, is nominal.

b. The variable, favorite color, is nominal.

c. The variable, pain as measured on the 11-point pain scale, is Ordinal.

d. The variable, Effectiveness of teacher, is Ordinal.

2.10 Number of children

a. The variable, number of children, is quantitative.

b. The variable, number of children, is discrete.

c.

No. children	0	1	2	3	4	5	6	7	8+
Frequency	663	349	623	383	166	74	47	23	16
Proportion	0.283	0.149	0.266	0.163	0.071	0.032	0.02	0.01	0.007
Percentage	28.3	14.9	26.6	16.3	7.1	3.2	2.0	1.0	0.7

d. 28.3% of respondents in the GSS had no children (the modal categary), 14.9% had exactly one child, 26.6% had two, and 16.3% had three children. The remaining 13.9% of respondents had four or more children.

2.11 Fatal Shark Attacks

a.

Location	Florida	Hawaii	California	Australia	South Africa
Frequency	2	2	4	15	13
Proportion	0.032	0.032	0.063	0.238	0.206
Percentage	3.2	3.2	6.3	23.8	20.6

Location	Réunion Island	Brazil	Bahamas	Other
Frequency	6	4	6	11
Proportion	0.095	0.063	0.095	0.175
Percentage	9.5	6.3	9.5	17.5

b. Australia is the modal category.

c. The regions with most frequent fatal shark attacks are Australia and South Africa.

2.12 Smartphone use in class

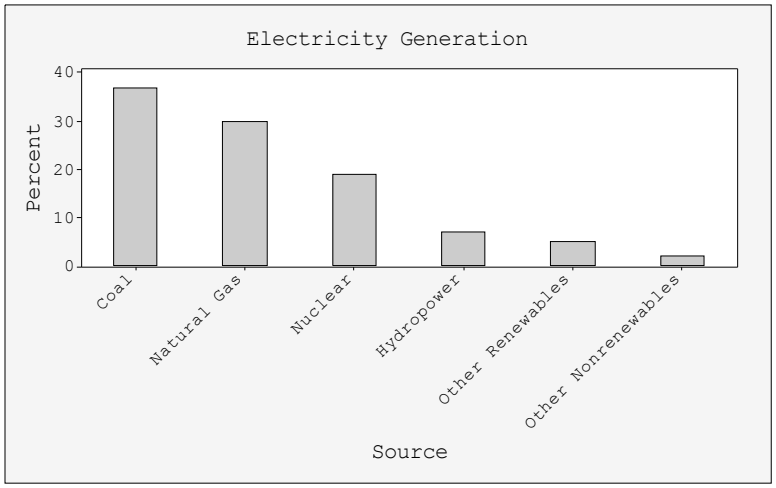
Of the 743 teenagers surveyed, 39.2% never lost focus, 28.8% rarely lost focus, 24.0% sometimes lost focus, and 8.1% often lost focus while checking their cell phone in class

Lost Focus?	Never	Rarely	Sometimes	Often
Frequency	291	214	178	60
Proportion	0.392	0.288	0.240	0.081
Percentage	39.2	28.8	24.0	8.1

Section 2.2: Graphical Summaries of Data

2.13 Generating Electricity

a.



b. Sketching a bar chart would be easier. Sketching the precise areas corresponding to the percentages is more challenging in a pie chart.

c. It is straightforward to judge the relative sizes when comparing the bars corresponding to the percentages.

d. Coal is the modal category.

2.14 What do alligators eat?

a. Primary food choice is categorical.

b. The modal category is "fish."

c. Approximately 43% of alligators ate fish as their primary food choice.

d. This is an example of a Pareto chart, a chart that is organized from most to least frequent choice.

2.15 Weather stations

a. The slices of the pie portray categories of a variable (i.e., regions).

b. The first number is the frequency, the number of weather stations in a given region. The second number is the percentage of all weather stations that are in this region.

c. It is easier to identify the modal category using a bar graph than using a pie chart because we can more easily compare the heights of bars than the slices of a piece of pie. For example, in this case, the slices for Midwest and West look very similar in size, but it would be clear from a bar graph that West was taller in height than Midwest.

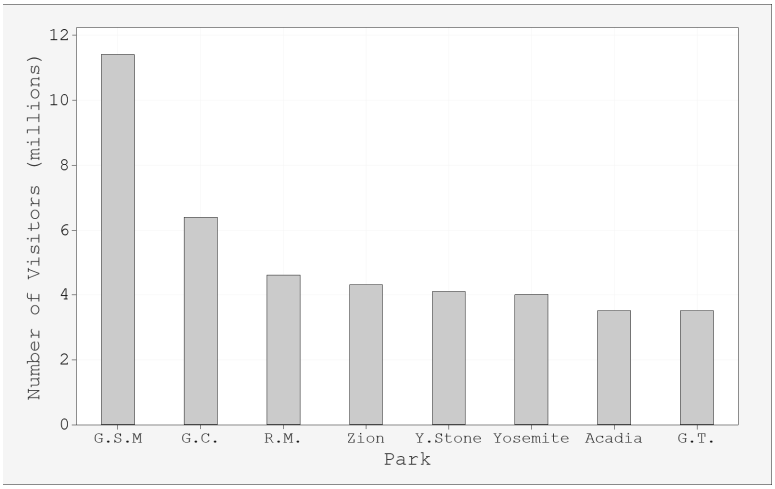
2.16 Most popular national park

a. National park visited is categorical.

b. A Pareto chart would make more sense because it allows the viewer to easily locate the categories with the highest and lowest frequencies.

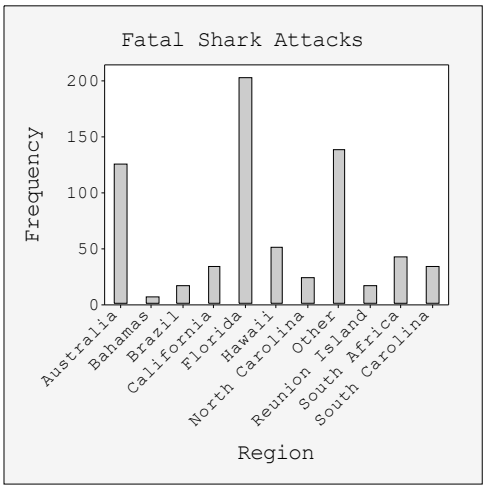
c. A dot plot or stem-and-leaf plot do not make sense because the data are categorical; these two types of plots are used with quantitative data (and also with data that have relatively few observations).

d.

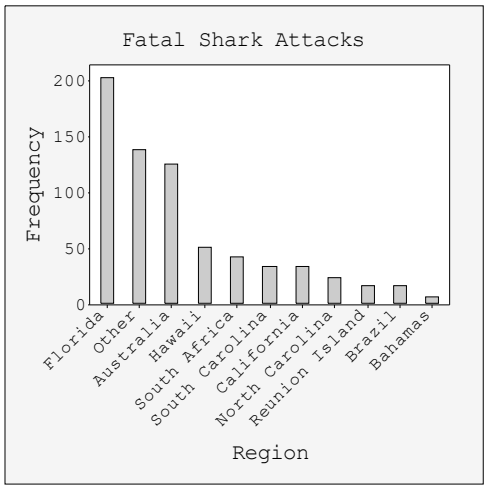


2.17 Pareto chart for fatal shark attacks

(i) Alphabetically



(ii) Pareto chart



With a Pareto chart, it is straightforward to identify the few regions with the largest number of fatal shark attacks.

2.18 Sugar dot plot

a. The minimum sugar value is zero grams, and the maximum is 18 grams.

b. The sugar outcome that occurs most frequently is called the mode. For this data set there are five modes: three, four, eleven, twelve and fourteen grams, each occurring twice.

c. Answers will vary.

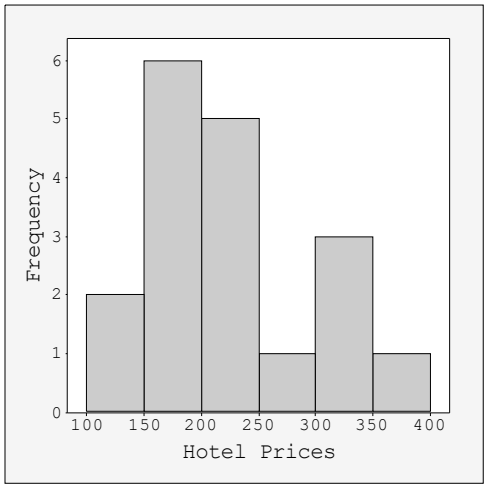
2.19 Spring break hotel prices

a.
1	24677999
2	133445
3	1338

b.
1	24
1	677999
2	13344
2	5
3	133
3	8

The plot with split stems gives a clearer picture of the shape of the distribution.

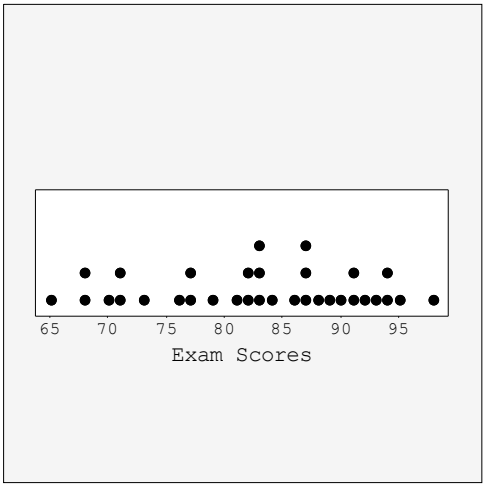
c. Most hotels charge between $150 and $250 per night, with a few charging more. The distribution of prices is right-skewed.



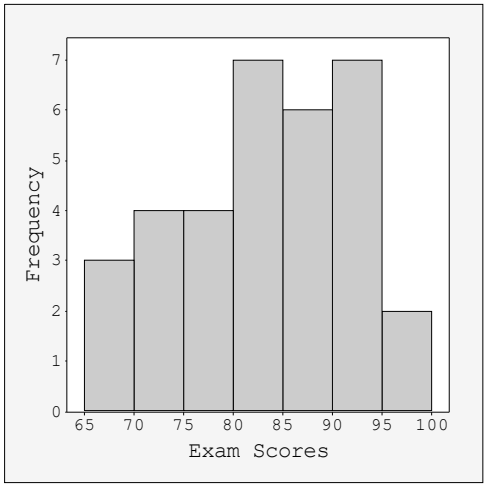
2.20 Graphing exam scores

a. There are 33 students in the class; the minimum score is 65 and the maximum is 98.

b.



c.

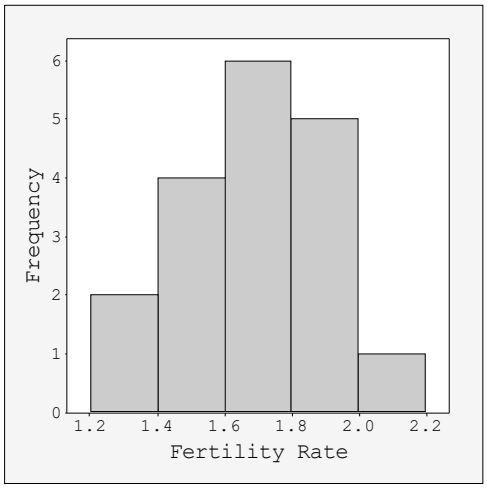


2.21 Fertility rates

a. 1 | 33455566677788889
   2 | 2

   A disadvantage of this plot is that it is too compact making it difficult to visualize where the data fall.

b. 1 | 33
   1 | 4555
   1 | 666777
   1 | 88889
   2 |
   2 | 2

c. 

2.22 Histogram of fertility rates

a. 13 bins were formed, each of length 0.5, starting at 1 and ending at 7.5.

b. Mexico would fall in the third bin, 2.0–2.5.

c. The *y*-axis for this histogram shows the number of countries in each bin, not the percent. Since the histogram depicts the distribution for 200 countries, and about 60 are in the bin from 1.5 to 2, around $60 / 200 = 0.3$, or 30% of countries, have fertility rate between 1.5 and 2.

2.23 Histogram for sugar

a. –1 to 1, 1 to 3, 3 to 5, 5 to 7, 7 to 9, 9 to 11, 11 to 13, 13 to 15, 15 to 17, and 17 to 19

b. The distribution is bimodal; child cereals, on average, have more sugar than adult cereals have.

c. The dot and stem-and-leaf plots allow us to see all the individual data points.

d. The relative differences among bars would remain the same.

2.24 Shape of the histogram

a. Assessed value of houses in a large city – skewed to the right (a long right tail) because of some very expensive homes.

b. Number of times checking account overdrawn in the past year for the faculty at the local university – skewed to the right because of the few faculty who overdraw frequently.

c. IQ for the general population – symmetric because most would be in the middle, with some higher and some lower; there is no reason to expect more to be higher or lower (particularly because IQ is constructed as a comparison to the general population's "norms").

d. The height of female college students – symmetric because most would fall in the middle, going down to a few short students and up to a few tall students.

2.25 More shapes of histograms

a. The scores of students (out of 100 points) on a very easy exam in which most score perfectly or nearly so, but a few score very poorly – skewed to the left because of the few who score poorly.

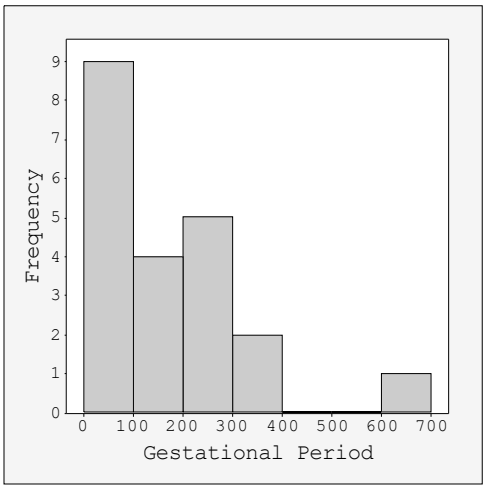
b. The weekly church contribution for all members of a congregation, in which the three wealthiest members contribute generously each week – skewed to the right because of the few wealthy members' contributions.

c. Time needed to complete a difficult exam (maximum time is 1 hour) – skewed to the left because most take almost or all of the whole time, whereas a few finish very quickly.

d. Number of music CDs (compact discs) owned, for each student in your school – skewed to the right because of a few students' huge CD collections.

2.26 Gestational Period

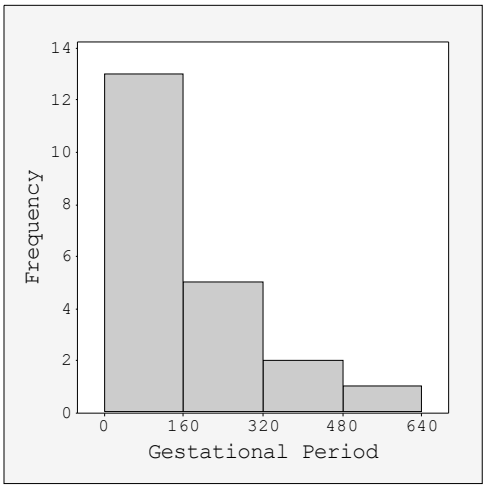
a.

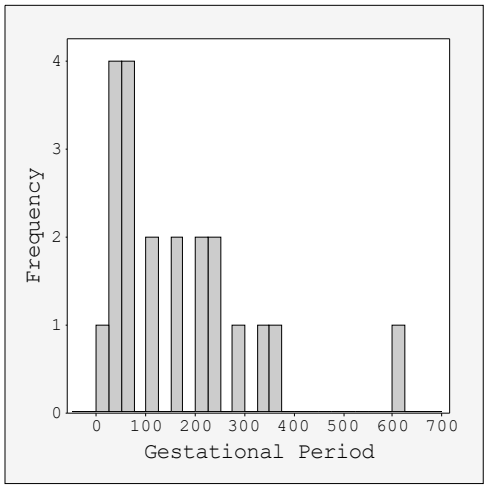


b. The elephant, with a gestational period of 624 days, is unusual.

c. The distribution is right-skewed.

d. Neither of the two histograms accurately summarizes the distribution. The one with 4 intervals is too coarse, the one with 30 intervals too fine.





2.27 How often do students read the newspaper?

a. This is a discrete variable because the value for each person would be a whole number. One could not read a newspaper 5.76 times per week, for example.

b. (i) The minimum response is zero.

(ii) The maximum response is nine.

(iii) Two students did not read the newspaper at all.

(iv) The mode is three.

c. This distribution is unimodal and somewhat skewed to the right.

2.28 Blossom widths

a. The distribution is slightly right-skewed (or roughly symmetric). Most blossoms have a width between 3.2 and 3.6 in. There is one blossom with an unusual small width for that species of less than 2.4 in.

b. The distribution is left-skewed. Most blossoms have a width between 2.8 and 3.2 in.

c. $\frac{6+15+24}{50} = 0.90$ or 90%

2.28 (continued)

d. No. We don't know how many blossoms in the interval from 2.8 to 3.2 in. are actually wider than 3 in.

2.29 Central Park temperatures

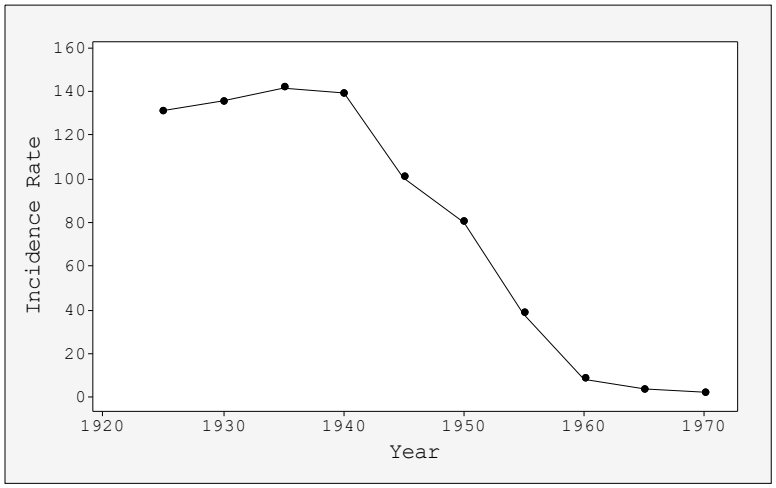
a. The distribution is somewhat skewed to the left.

b. A time plot connects the data points over time to show time trends.

c. A histogram shows the number of observations at each level more easily than does the time plot. We also can see the shape of the distribution from the histogram but not from the time plot.

2.30 Is whooping cough close to being eradicated?

a. One can see in the time plot below that after an initial slight increase, there was a sharp and steady decrease in incidence of whooping cough starting around 1940. The decrease leveled off starting around 1960. These data suggest that the whooping cough vaccination was proving effective in reducing the incidence of whooping cough.

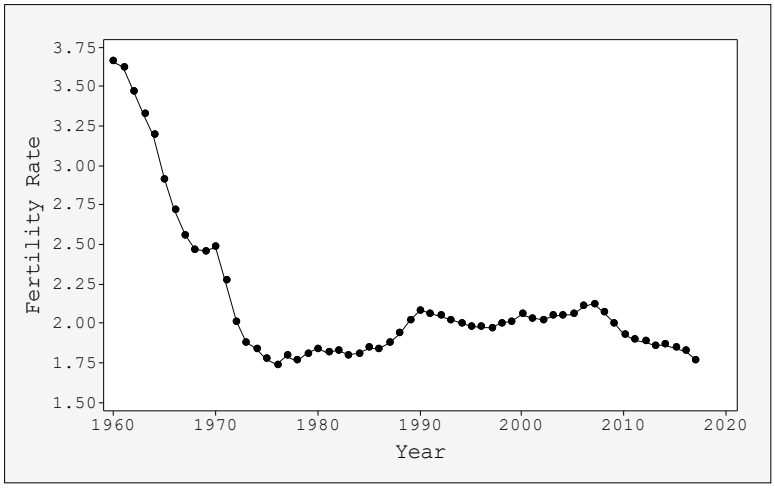


b. The incidence rate stayed low until about 2000, after which a sharp increase can be observed. No, the United States is not close to eradicating whooping cough. Potential reasons for this include fewer people deciding to get vaccinated and less efficient vaccinations.

c. A histogram would not address this question because it does not show the rates for each year; we would not be able to see changes over time.

2.31 Fertility rate over time

The fertility rate dropped sharply from about 3.6 in 1960 to about 1.8 in 1980, rebounded a bit to 2.1 in 2000, and has been dropping ever since to a current value of 1.8.



Section 2.3: Measuring the Center of Quantitative Data

2.32 Median versus mean

a. Median (The distribution would be right-skewed.)

b. Median (The distribution would be left-skewed.)

c. Mean (The distribution would be symmetric.)

2.33 More median versus mean

a. Median (The distribution would be right-skewed.)

b. Mean (The distribution would be symmetric.)

c. Median (The distribution would be left-skewed.)

2.34 More on CO_2 emissions

a. Mean: $\bar{x} = \frac{\sum x}{n} = \frac{8.0+5.3+1.8+1.7+1.2+0.8+0.6+0.5+0.4+0.4}{10} = \frac{20.7}{10} = 2.07$

Median: Find the middle value: $\frac{n+1}{2} = \frac{10+1}{2} = 5\frac{1}{2}$ th position

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

The median is $\frac{0.8+1.2}{2} = 1$.

b. Comparing absolute emission values for nations with different population sizes might be misleading because nations with larger populations tend to have larger total emissions. When viewed per capita, a different picture might emerge.

2.35 Resistance to an outlier

a. The median for all three data sets is ten. The values for all three sets of observations are already arranged in numerical order, and the middle number for each is 10.

b. Set 1: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+12}{5} = \frac{50}{5} = 10$

Set 2: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+100}{5} = \frac{138}{5} = 27.6$

Set 3: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+1000}{5} = \frac{1038}{5} = 207.6$

c. As the highest value becomes more and more of an extreme outlier, the median is unaffected, whereas the mean increases as the outlier becomes more extreme.

2.36 Income and health insurance

The distributions for both will be skewed to the right because the mean is much larger than median.

2.37 Labor dispute

Management would want to use the mean because it would be skewed right by the outliers – the few members of management who make a whole lot of money. The mean income would be higher because of the outliers. The workers would prefer the median because it is not affected by the large outliers. It is a more accurate measure of the actual typical income.

2.38 Cereal sodium

The moderate skewness to the left causes the mean to be lower than the median.

2.39 Center of plots

a. The mean and median would be the same for the dot plots to the middle and to the right because the distributions are symmetric.

b. The distribution to the left is skewed to the right, and the mean would be higher than the median would. The mean would be pulled toward the higher, atypical values.

2.40 Public transportation – center

a. The mean is 2, the median is 0, and the mode is 0. Thus, the average score is 2, the middle score is 0 (indicating that the mean is skewed by outliers), and the most common score also is 0.

Mean: $\bar{x} = \frac{\sum x}{n} = \frac{0+0+4+0+0+0+10+0+6+0}{10} = \frac{20}{10} = 2$

Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, 10

Mode: the most common score is zero.

b. Now the mean is 10, but the median is still 0.

Mean: $\bar{x} = \frac{\sum x}{n} = \frac{0+0+4+0+0+0+10+0+6+90}{10} = \frac{110}{10} = 10$

Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, 10, 90

The median is not affected by the magnitude of the highest score, the outlier. Because there are so many zeros, even though we've added one score, the median remains zero. The mean, however, is affected by the magnitude of this new score, an extreme outlier.

2.41 Public transportation – outlier

a. The mean versus median applet confirms that the median is not affected by the magnitude of the highest score. Because there are so many zeros, even though we've added one score, the median remains zero. The mean, however, is affected by the magnitude of this new score, an extreme outlier.

b. The applet demonstrates that the outlier has a weaker effect when there are more scores near the original mean.

2.42 Baseball salaries

There are a few valuable players who receive exorbitant salaries, whereas the typical player is paid much less (although still a lot by most people's standards!). The very high salaries of the few affect the mean, but not the median.

2.43 More baseball salaries

Answers will vary.

2.44 Fertility statistics

a. The median fertility rate among the 18 countries was 1.65. Thus, about half of the countries listed have mean fertility rates at or below 1.65 with the remaining countries having mean fertility rates above 1.65.

b. The mean fertility rate among the 18 countries was 1.65.

c. Since the population of adult women varies greatly among the countries, it is necessary to calculate an overall fertility rate for each. This rate is the mean number of children per adult woman. The mean for a variable need not be one of the possible values for the variable. If we used the median number of children for each country, almost all countries would have a median of 1 or maybe 2 children per adult woman, and comparing countries is less informative.

2.45 Sex partners

a.

Number of partners	Number of respondents
0	102
1	233
2	18
3	9
4	2
5	1
Total	**365**

If the data is sorted from smallest to largest, the median is the number in the $(365+1)/2 = 183$rd position.

Since 102 respondents answered 0 and 233 answered 1, the median is 1.

2.45 (continued)

b. Mean: $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{102(0)+233(1)+18(2)+9(3)+2(4)+1(5)}{365} = \dfrac{309}{365} = 0.85$

c. Since the total number of respondents is 365, the median is still the value in the 183rd place when the data are sorted from smallest to largest. Since 233 respondents gave an answer of 1, the median is still 1. However, the value of the mean changes:

Mean: $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{0(0)+233(1)+18(2)+9(3)+2(4)+103(5)}{365} = \dfrac{819}{365} = 2.2$

2.46 Marriage statistics for 20 to 24-year-olds

a. Women: The mean is 0.274, the median is 0.

$\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{7350(0)+2587(1)+80(2)}{10,017} = \dfrac{2747}{10,017} = 0.274$; The median is the middle score. With 10,017 scores, the median is the score in the 5,009th position. Thus, the median is 0.

Men: The mean is 0.161, the median is 0.

$\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{8,418(0)+1,594(1)+10(2)}{10,022} = \dfrac{1,614}{10,022} = 0.161$; The median is the middle score. With 10,022 scores, the median is the score between the 5,011th and 5,012th positions. Thus, the median is 0.

b. Using the medians, it seems that there is no difference. Using the mean, in this age group, women have, on average, been married more often.

2.47 Accidents

In this case, the mean is likely to be more useful because it uses the numerical values of all of the observations, not just the ordering. Because so many people would report 0 motor accidents, the median is not very useful. It ignores too much of the data.

Section 2.4: Measuring the Variability of Quantitative Data

2.48 Sick leave

a. The range is 6; this is the distance from the smallest to the largest observation. In this case, there are six days separating the fewest and most sick days taken $(6-0=6)$.

b. The standard deviation is the typical distance of an observation from the mean (which is 1.25).

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{(0-1.25)^2+\cdots+(0-1.25)^2+(4-1.25)^2+(6-1.25)^2}{7} = \frac{39.5}{7} = 5.643$$

$$s = \sqrt{s^2} = \sqrt{5.643} = 2.38$$

The standard deviation of 2.38 indicates a typical number of sick days taken is 2.38 days from the mean of 1.25.

c. Redo (a) and (b).

a. The range is 60; this is the distance from the smallest to the largest observation. In this case, there are sixty days separating the fewest and most sick days taken $(60-0=60)$.

b. The standard deviation is the typical distance of an observation from the mean (which is 8).

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{(0-8)^2+\cdots+(0-8)^2+(4-8)^2+(60-8)^2}{7} = \frac{3104}{7} = 443.43$$

$$s = \sqrt{s^2} = \sqrt{443.43} = 21.06$$

The standard deviation of 21.06 indicates a typical number of sick days taken is 21.06 days from the mean of 8. The range and mean both increase when an outlier is added.