INSTRUCTOR'S SOLUTIONS MANUAL

INTRODUCTORY STATISTICS: EXPLORING THE WORD THROUGH DATA

THIRD EDITION

Robert Gould

University of California, Los Angeles

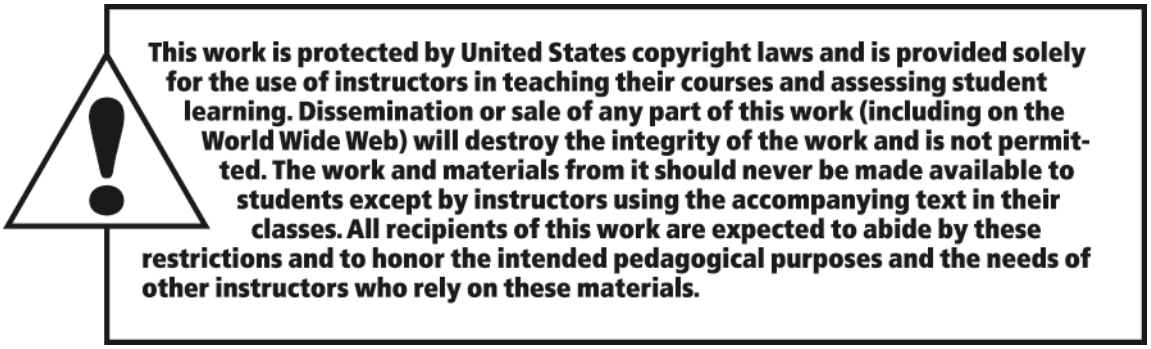
Rebecca Wong

West Valley College

Colleen Ryan

Moorpark Community College

Pearson



Reproduced by Pearson from electronic files supplied by the author.

ISBN-13: 978-0-13-519061-6
ISBN-10: 0-13-519061-4

CONTENTS

Chapter 1: Introduction to Data

Section 1.2: Classifying and Storing Data 1
Section 1.3: Investigating Data 3
Section 1.4: Organizing Categorical Data 3
Section 1.5: Collecting Data to Understand Causality 6
Chapter Review Exercises 7

Chapter 2: Picturing Variation with Graphs

Section 2.1: Visualizing Variation in Numerical Data
and Section 2.2: Summarizing Important Features of a Numerical Distribution 9
Section 2.3: Visualizing Variation in Categorical Variables
and Section 2.4: Summarizing Categorical Distributions 14
Section 2.5: Interpreting Graphs 15
Chapter Review Exercises 16

Chapter 3: Numerical Summaries of Center and Variation

Section 3.1: Summaries for Symmetric Distributions 19
Section 3.2: What's Unusual? The Empirical Rule and z-Scores 22
Section 3.3: Summaries for Skewed Distributions 23
Section 3.4: Comparing Measures of Center 24
Section 3.5: Using Boxplots for Displaying Summaries 27
Chapter Review Exercises 28

Chapter 4: Regression Analysis: Exploring Associations between Variables

Section 4.1: Visualizing Variability with a Scatterplot 31
Section 4.2: Measuring Strength of Association with Correlation 31
Section 4.3: Modeling Linear Trends 32
Section 4.4: Evaluating the Linear Model 37
Chapter Review Exercises 40

Chapter 5: Modeling Variation with Probability

Section 5.1: What Is Randomness? 49
Section 5.2: Finding Theoretical Probabilities 49
Section 5.3: Associations in Categorical Variables 54
Section 5.4: Finding Empirical and Simulated Probabilities 56
Chapter Review Exercises 58

Chapter 6: Modeling Random Events: The Normal and Binomial Models

Section 6.1: Probability Distributions Are Models of Random Experiments.....65
Section 6.2: The Normal Model.....67
Section 6.3: The Binomial Model (Optional).....79
Chapter Review Exercises.....81

Chapter 7: Survey Sampling and Inference

Section 7.1: Learning about the World through Surveys.....85
Section 7.2: Measuring the Quality of a Survey.....86
Section 7.3: The Central Limit Theorem for Sample Proportions.....88
Section 7.4: Estimating the Population Proportion with Confidence Intervals.....90
Section 7.5: Comparing Two Population Proportions with Confidence.....94
Chapter Review Exercises.....97

Chapter 8: Hypothesis Testing for Population Proportions

Section 8.1: The Essential Ingredients of Hypothesis Testing.....101
Section 8.2: Hypothesis Testing in Four Steps.....102
Section 8.3: Hypothesis Tests in Detail.....107
Section 8.4: Comparing Proportions from Two Populations.....108
Chapter Review Exercises.....112

Chapter 9: Inferring Population Means

Section 9.1: Sample Means of Random Samples.....121
Section 9.2: The Central Limit Theorem for Sample Means.....122
Section 9.3: Answering Questions about the Mean of a Population.....123
Section 9.4: Hypothesis Testing for Means.....125
Section 9.5: Comparing Two Population Means.....131
Chapter Review Exercises.....138

Chapter 10: Associations between Categorical Variables

Section 10.1: The Basic Ingredients for Testing with Categorical Variables.....147
Section 10.2: The Chi-Square Test for Goodness of Fit.....149
Section 10.3: Chi-Square Tests for Associations between Categorical Variables.....153
Section 10.4: Hypothesis Tests When Sample Sizes Are Small.....160
Chapter Review Exercises.....165

Chapter 11: Multiple Comparisons and Analysis of Variance

Section 11.1: Multiple Comparisons.....173
Section 11.2: The Analysis of Variance.....175
Section 11.3: The ANOVA Test.....176
Section 11.4: Post-Hoc Procedures.....180
Chapter Review Exercises.....184

Chapter 12: Experimental Design: Controlling Variation

Section 12.1: Variation Out of Control....187
Section 12.2: Controlling Variation in Surveys....192
Section 12.3: Reading Research Papers....192

Chapter 13: Inference without Normality

Section 13.1: Transforming Data....197
Section 13.2: The Sign Test for Paired Data....199
Section 13.3: Mann-Whitney Test for Two Independent Groups....201
Section 13.4: Randomization Tests....203
Chapter Review Exercises....204

Chapter 14: Inference for Regression

Section 14.1: The Linear Regression Model....209
Section 14.2: Using the Linear Model....210
Section 14.3: Predicting Values and Estimating Means....212
Chapter Review Exercises....213

Chapter 1: Introduction to Data

Section 1.2: Classifying and Storing Data

1.1 There are eight variables: “Female”, “Commute Distance”, “Hair Color”, “Ring Size”, “Height”, “Number of Aunts”, “College Units Acquired”, and “Living Situation”.

1.2 There are eleven observations.

1.3 a. Living situation is categorical.

b. Commute distance is numerical.

c. Number of aunts is numerical.

1.4 a. Ring size is numerical.

b. Hair color is categorical.

c. Height is numerical.

1.5 Answers will vary but could include such things as number of friends on Facebook or foot length. *Don’t copy these answers*.

1.6 Answers will vary but could include such things as class standing (“Freshman”, “Sophomore”, “Junior”, or “Senior”) or favorite color. *Don’t copy these answers*.

1.7 0 = male, 1 = female. The sum represents the total number females in the data set.

1.8 There would be seven 1’s and four 0’s.

1.9 Female is categorical with two categories. The 1’s represent females, and the 0’s represent males. If you added the numbers, you would get the number of females, so it makes sense here.

1.10 a. Freshman

0
1
1
0
1
1
0
1
1
0
0

b. numerical

c. categorical

1.11 a. The data is stacked.

b. 1 = male, 0 = female.

c.

Male	Female
1916	9802
183	153
836	1221
95	
512	

1.12 a. The data is unstacked.

b. Labels for columns will vary.

Gender	Age
1	29
1	23
1	30
1	32
1	25
0	24
0	24
0	32
0	35
0	23

c. Gender is categorical; Age is numerical

1.13 a. Stacked and coded:

Calories	Sweet
90	1
310	1
500	1
500	1
600	1
90	1
150	0
600	0
500	0
550	0

The second column could be labeled "Salty" with the 1's being 0's and the 0's being 1's.

b. Unstacked:

Sweet	Salty
90	150
310	600
500	500
500	550
600	
90	

1.14 a. Stacked and coded:

Cost	Male
10	1
15	1
15	1
25	1
12	1
8	0
30	0
15	0
15	0

The second column could be labeled "Female" with the 1's being 0's and the 0's being 1's.

b. Unstacked:

Male	Female
10	8
15	30
15	15
25	15
12	

Section 1.3: Investigating Data

1.15 Yes. Use College Units Acquired and Living Situation.

1.16 Yes. Use Female and Height.

1.17 No. Data on number of hours of study per week are not included in the table.

1.18 Yes. Use Ring Size and Height.

1.19 a. Yes. Use Date.

b. No. data on temperature are not included in the table.

c. Yes. Use Fatal and Species of Shark.

d. Yes. Use Location.

1.20 Use Time and Activity.

Section 1.4: Organizing Categorical Data

1.21 a. 33/40 = 82.5%

b. 32/45 = 71.1%

c. 33/65 = 50.8%

d. 82.5% of 250 = 206

1.22 a. 4/27 = 14.8%

b. 14/27 = 51.9%

c. 4/18 = 22.2%

d. 14.8% of 600 = 89 men

1.23 a. 15/38 = 39.5% of the class were male.

b. 0.64(234) = 149.994, so 150 men are in the class.

c. 0.40(x) = 20, so 20/0.40 = 50 total students in the class.

1.24 a. 0.35(346) = 121 male nurses.

b. 66/178 = 37.1% female engineers.

c. 0.65(x) = 169 so 169/0.65 = 260 lawyers in the firm.

1.25 The frequency of women 6, the proportion is 6/11, and the percentage is 54.5%.

1.26 The frequency is 8, the proportion is 8/11, and the percentage is 72.7%.

1.27 a. and b.

	Men	**Women**	**Total**
Dorm	3	4	7
Commuter	2	2	4
Total	5	6	11

c. 4/6 = 66.7%

d. 4/7 = 57.1%

e. 7/11 = 63.6%

f. 66.7% of 70 = 47

1.28 a. and b.

	Men	**Women**	**Total**
Brown	3	5	8
Black	2	0	2
Blonde	0	1	1
Total	5	6	11

c. 5/6 = 83.3%

d. 5/8 = 62.5%

e. 8/11 = 72.7%

f. 83.3% of 60 = 50

1.29 1.26(x) = 160328 so 160328/1.26 = 127,244 personal care aids in 2014

1.30 .1295(x) = 3480000 so 3480000/.1295 = $26,872,587.87 total candy sales

1.31

State	Prison	Rank Prison	Population	Population (thousands)	Prison per 1000	Rank Rate
California	136,088	1	39,144,818	39145	3.48	4
New York	52518	2	19,795,791	19796	2.65	5
Illinois	48278	3	12,859,995	12860	3.75	3
Louisiana	30030	4	4,670,724	4671	6.43	1
Mississippi	18793	5	2,992,333	29922	6.28	2

California has the highest prison population. Louisiana has the highest rate of imprisonment.

The two answers are different because the state populations are different.

1.32 a. Miami: 4,919,000/2891 = 1701 Detroit: 3,903,000/3267 = 1195

Atlanta: 3,500,000/5083 = 689 Seattle: 2,712,000/1768 = 1534

Baltimore: 2,076,000/1768 = 1174

Ranks: 1- Miami, 2- Seattle, 3- Detroit, 4- Baltimore, 5- Atlanta

b. Atlanta

c. Miami

1.33

Year	%Uncovered
1990	$\frac{34,719}{249,778} = 13.9\%$
2000	$\frac{36,586}{279,282} = 13.1\%$
2015	$\frac{29758}{316574} = 9.4\%$

The percentage of uninsured people have been declining.

1.34

Year	% Subscribers
2012	$\frac{103.6}{114.7} = 90.3\%$
2013	$\frac{103.3}{114.1} = 90.5\%$
2014	$\frac{103.7}{115.7} = 89.6\%$
2015	$\frac{100.2}{116.5} = 86.0\%$
2016	$\frac{97.8}{116.4} = 84.0\%$

The percentage of cable subscribers rose slightly between 2012 and 2013 but has declined each year since then.

1.35

Year	%Older Population
2020	$\frac{54.8}{334} = 16.4\%$
2030	$\frac{70.0}{358} = 19.6\%$
2040	$\frac{81.2}{380} = 21.4\%$
2050	$\frac{88.5}{400} = 22.1\%$

The percentage of older population is projected to increase.

1.36

Year	%Older Population
2000	$\frac{4.0}{8.2} = 48.8\%$
2005	$\frac{3.6}{7.6} = 47.4\%$
2010	$\frac{3.6}{6.8} = 52.9\%$
2014	$\frac{3.2}{6.9} = 46.4\%$

The rate has fluctuating over this period, decreasing, then increasing, and then decreasing again.

1.37 We don't know the percentage of female students in the two classes. The larger number of women at 8a.m. may just result from a larger number of students at 8 a.m., which may be because the class can accommodate more students because perhaps it is in a large lecture hall.

1.38 No, we need to know the population of each city so we can compare the rates.

Section 1.5 Collecting Data to Understand Causality

1.39 Observational study.

1.40 Controlled experiment.

1.41 Controlled experiment.

1.42 Controlled experiment.

1.43 Controlled experiment.

1.44 Observational study.

1.45 Anecdotal evidence are stories about individual cases. No cause-and effect conclusions can be drawn from anecdotal evidence.

1.46 These testimonials are anecdotal evidence. There is no control group and no comparison. No cause-and-effect conclusions can be drawn from anecdotal evidence.

1.47 This was an observational study, and from it you cannot conclude that the tutoring raises the grades. Possible confounders (answers may vary): 1. It may be the more highly motivated who attend the tutoring, and this motivation is what causes the grades to go up. 2. It could be that those with more time attend the tutoring, and it is the increased time studying that causes the grades to go up.

1.48 a. If the doctor decides on the treatment, you could have bias.

b. To remove this bias, randomly assign the patients to the different treatments.

c. If the doctor knows which treatment a patient had, that might influence his opinion about the effectiveness of the treatment.

d. To remove that bias, make the experiment double-blind. The talk-therapy-only patients should get a placebo, and no patients should know whether they have a placebo or antidepressant. In addition, the doctor should not know who took the antidepressants and who did not.

1.49 a. The sample size of this study is not large (40). The study was a controlled experiment and used random assignment. It was not double-blind since researchers new what group each participant was in.

b. The sample size of the study was small, so we should not conclude that physical activity while learning caused higher performance.

1.50 This is an observational study because researchers did not determine who received PCV7 and who did not. You cannot conclude causation from an observational study. We must assume that it is possible that there were confounding factors (such as other advances in medicine) that had a good effect on the rate of pneumonia.

1.51 a. Controlled experiment. Researchers used random assignment of subjects to treatment or control groups.

b. Yes. The experiment had a large sample size, was controlled, randomized, and double-blind; and used a placebo.

1.52 a. Observational study. There was no random assignment to treatment/control groups. The subjects kept a food diary and had their blood drawn.

b. We cannot make a cause-and-effect conclusion since this was an observational study.

1.53 No, this was not a controlled experiment. There was no random assignment to treatment/control groups and no use of a placebo.

1.54 No. There was no control group and no comparison. From observation of 12 children it is not possible to come to a conclusion that the vaccine causes autism. It may simply be that autism is usually noticed at the same age the vaccine is given.

1.55 a. Intervention remission: 11/33 = 33.3%; Control remission: 3 /34 = 8.8%

b. Controlled experiment. There was random assignment to treatment/control groups.

c. While this study did use random assignment to treatment/control groups, the sample size was fairly small (67 total) and there was no blinding in the experimental design. The difference in remission may indicate that the diet approach is promising and further research in this area is needed.

1.56 Ask whether there was random assignment to groups. Without random assignment there could be bias, and we cannot infer causation.

1.57 No. This is an observational study.

1.58 This is likely a conclusion from observational studies since it would not be ethical to randomly assign a subject to a group that drank large quantities of sugary drinks. Since this was likely based on observational studies, we cannot conclude drinking sugary beverages causes lower brain volume.

Chapter Review Exercises

1.59 a. 61/98 = 62.2%

b. 37/82 = 45.1%

c. Yes, this was a controlled experiment with random assignment. The difference in percentage of homes adopting smoking restrictions indicates the intervention may have been effective.

1.60 No. Cause-and-effect conclusions cannot be drawn from observational studies.

1.61 a. Gender (categorical) and whether students had received a speeding ticket (categorical)

b.

	Male	**Female**
Yes	6	5
No	4	10

c. Men: 6/10=60%; Women: 5/15 = 33.3%; a greater percentage of men reported receiving a speeding ticket.

1.62 a. Gender (categorical) and whether students had driven over 100 mph (categorical).

b.

	Male	**Female**
Yes	6	5
No	3	10

c. Men: 6/9 = 66.7%; Women: 5/15 = 33.3%; a greater percentage of men reported driving over 100 mph.

1.63 Answers will vary. *Students should not copy the words they see in these answers.* Randomly divide the group in half, using a coin flip for each woman: Heads she gets the vitamin D, and tails she gets the placebo (or vice versa). Make sure that neither the women themselves nor any of the people who come in contact with them know whether they got the treatment or the placebo ("double-blind"). Over a given length of time (such as three years), note which women had broken bones and which did not. Compare the percentage of women with broken bones in the vitamin D group with the percentage of women with broken bones in the placebo group.

1.64 Answers will vary. *Students should not copy the words they see here.* Randomly divide the group in half, using a coin flip for each person: Heads they get Coumadin, and tails they get aspirin (or vice versa). Make sure that neither the subjects nor any of the people who come in contact with them know which treatment they received ("double-blind"). Over a given length of time (such as three years), note which people had second strokes and which did not. Compare the percentage of people with second strokes in the Coumadin group with the percentage of people with second strokes in the aspirin group. There is no need for a placebo because we are comparing two treatments. However, it would be acceptable to have three groups, one of which received a placebo.

1.65 a. The treatment variable is mindful yoga participation. The response variable is alcohol use.

b. Controlled experiment (random assignment to treatment/control groups).

c. No, since the sample size was fairly small; however, the difference in outcomes for treatment/control groups may indicate that further research into the use of mindful yoga may be warranted.

1.66 a. The treatment variable was neurofeedback; the response variable is ADHD symptoms.

b. Controlled experiment (random assignment to treatment/control groups).

c. No because there were no significant differences in outcomes between any of the groups.

1.67 No. There was no control group and no random assignment to treatment or control groups.

1.68 a. Long course antibiotics: 39/238 = 16.4%; short course antibiotics: 77/229 = 33.6%.

The longer course recipients did better.

b.

	10 days	**5 days**
Failure	39	77
Success	199	152

c. Controlled experiment (random assignment to treatment/control groups).

d. Yes. This was a controlled, randomized experiment with a large sample size.

1.69 a. LD: 8% tumors; LL: 28% tumors A greater percentage of the 24 hours of light developed tumors.

b. A controlled experiment. You can tell by the random assignment.

c. Yes, we can conclude cause and effect because it was a controlled experiment, and random assignment will balance out potential confounding variables.

1.70 a. 43/53, or about 81.1%, of the males who were assigned to Scared Straight we rearrested. 37/55, or 67.3%, of those receiving no treatment were rearrested So the group from Scared Straight had a higher arrest rate.

b. No, Scared Straight does not cause a lower arrest rate because the arrest rate was higher.

Chapter 2: Picturing Variation with Graphs

Section 2.1: Visualizing Variation in Numerical Data and Section 2.2: Summarizing Important Features of a Numerical Distribution

2.1 a. 4 people had resting pulse rates more than 100.

b. $\frac{4}{125} = 3.2\%$ of the people had resting pulse rates of more than 100.

2.2 a. 8 people have glucose readings above 120 mg/dl.

b. $8 = 6.1\%$ of these people have glucose readings above 120 mg/dl.

2.3 New vertical axis labels: $\frac{1}{25} = 0.04$, $\frac{2}{25} = 0.08$, $\frac{3}{25} = 0.12$, $\frac{4}{25} = 0.16$, $\frac{5}{25} = 0.20$

2.4 a. The bin width is 100.

b. The histogram is bimodal because two bins have a much higher relative frequency than the others.

c. About 19% (combine 6% and 13%). Due to the scale on the graph, any answer between 18% to 20% is acceptable.

2.5 Yes, since only about 7% of the pulse rates were higher than 90 bpm. Conclusion might vary, but students must mention that 7% of pulse rates were higher than 90 bpm.

2.6 No, because on roughly half of the days the post office served more than 250 customers, so 250 would not be unusual.

2.7 a. Both cereals have similar center values (about 110 calories). The spread of the dotplots differ.

b. Cereal from manufacturer K tend to have more variation.

2.8 a. Both distributions have more than one mode. The center for the coins from the United States is much larger than the center for other countries. The spreads are similar.

b. Coins in the United States tend to weigh more, as we conclude because the center of the distribution is higher for the United States coins.

2.9 Roughly bell shaped. The lower bound is 0, the mean will be a number probably below 9, but a few students might have slept quite a bit (up to 12 hours?) which creates a right-skew.

2.10 Roughly right-skewed (most students with no tickets, very few with many tickets).

2.11 It would be bimodal because the men and women tend to have different heights and therefore different arm spans.

2.12 It might be bimodal because private colleges and public colleges tend to differ in amount of tuition.

2.13 About 75 beats per minute.

2.14 About 500 Calories.

2.15 The BMI for both groups are right skewed. For the men it is maybe bimodal (hard to tell). The typical values for the men and women are similar although the value for the men appears just a little bit larger than the typical value for the women. The women's values are more spread out.

2.16 a. Both distributions are right skewed. They have similar typical values.

b. The men's distribution is more spread out and has a greater percentage of values that are considered high. So, the women's levels are somewhat better.

2.17 a. The distribution is multimodal with modes at 12 years (high school), 14 years (junior college), 16 years (bachelor's degree), and 18 years (possible master's degree). It is also left-skewed with numbers as low as 0.

b. Estimate: 300 + 50 + 100 + 40 + 50, or about 500 to 600, had 16 or more years.

c. Between $\frac{500}{2018}$, or about 25%, and $\frac{600}{2018}$, or about 30%, have a bachelor's degree or higher. This is very similar to the 27% given.

2.18 a. The distribution is right-skewed.

b. About 2 or 3.

c. Between 80 and 100.

d. $\frac{80}{2000} = 4\%$ or $\frac{100}{2000} = 5\%$

2.19 Ford typically has higher monthly costs (the center is near 250 dollars compared with 225 for BMW) and more variation in monthly costs.

2.20 Both makes have similar typical mpg (around 23 mpg). BMW has more variation in mpg (more horizontal spread in the data).

2.21 1. The assessed values of homes would tend to be lower with a few higher values: This is histogram B.

2. The number of bedrooms in the houses would be slightly skewed right: This is histogram A.

3. The height of house (in stories) for a region would be that allows up to 3 stories would be histogram C.

2.22 1. The consumption of coffee by a person would be skewed right with many people who do not drink coffee and a few who drink a lot: This is histogram A.

2. The maximum speed driven in a car would be roughly symmetrical with a few students who drive very fast: This is histogram C.

3. The number of times a college student had breakfast would skew left with students who rarely eat breakfast: This is histogram B.

2.23 1. The heights of students would be bimodal and roughly symmetrical: This is histogram B.

2. The number of hours of sleep would be unimodal and roughly symmetrical, with any outliers more likely being fewer hours of sleep: This is histogram A.

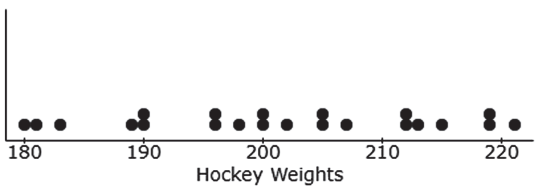
3. The number of accidents would be left skewed, with most student being involved in no or a few accidents: This is histogram C.

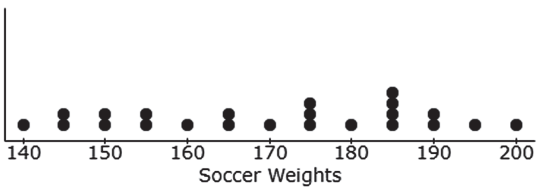
2.24 1. The SAT scores would be unimodal and roughly symmetrical: This is histogram C.

2. The weights of men and women would be bimodal and roughly symmetrical, but with more variation that SAT scores: This is histogram A.

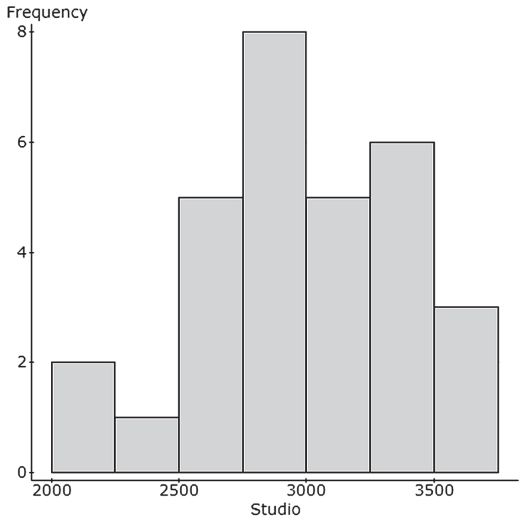
3. The ages of students would be left skewed, with most student being younger: This is histogram B.

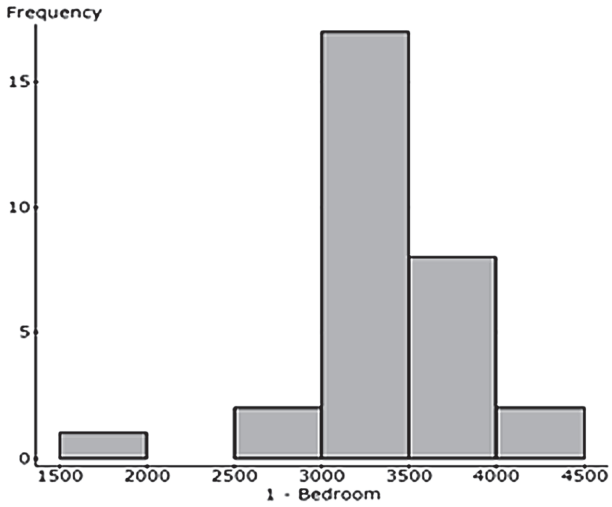
2.25 Students should display a pair of dotplots or histograms. One graph for Hockey and one for Soccer. The hockey team tends to be heavier than the soccer team (the typical hockey player weighs about 202 pounds while the typical soccer player weighs about 170 pounds). The soccer team has more variation in weights than the hockey team because there is more horizontal spread in the data. Statistical Question (answers may vary): Are hockey players heavier than soccer players? Which type of athlete has the most variability in weight?





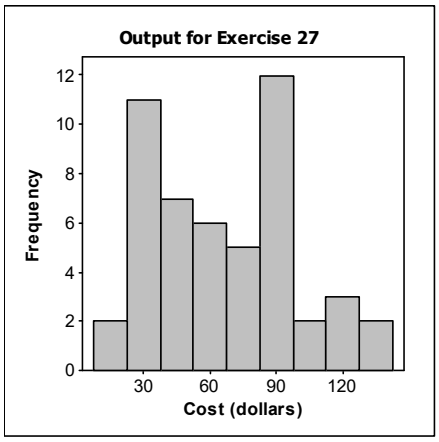
2.26 (Answers may vary). Which type of apartment tends to cost more? See histograms.



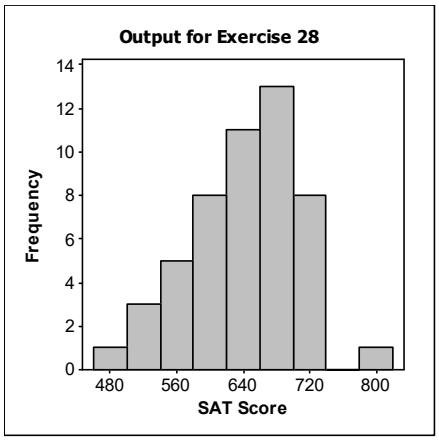


Studio apartments tend to be less expensive and have more variation in price than do one-bedrooms.

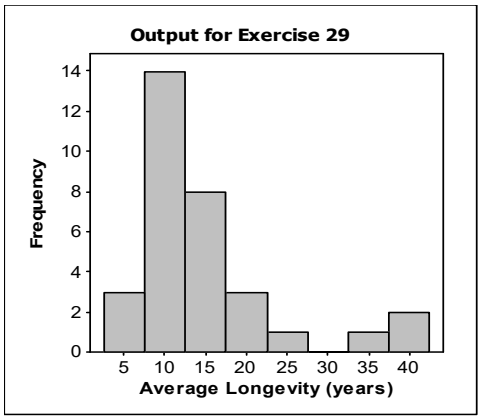
2.27 See histogram. The shape will depend on the binning used. The histogram is bimodal with modes at about \$30 and about \$90.



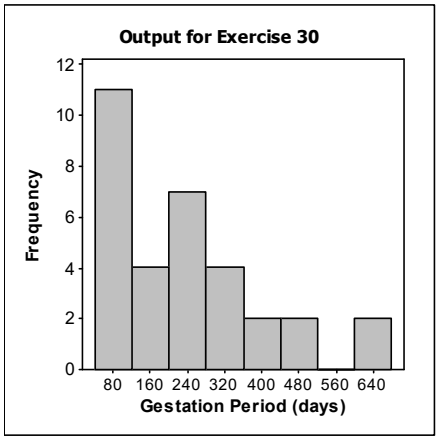
2.28 See histogram. The shape will depend on the binning used. The 800 score could be an outlier or not, and the graph could appear left-skewed or not.



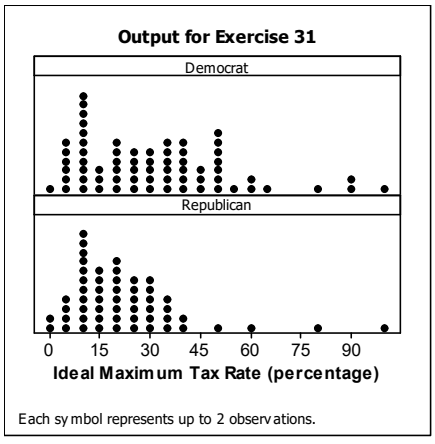
2.29 See histogram. The histogram is right-skewed. The typical value is around 12 (between 10 and 15) years, and there are three outliers: Asian elephant (40 years), African elephant (35 years), and hippo (41 years). Humans (75 years) would be way off to the right; they live much longer than other mammals.



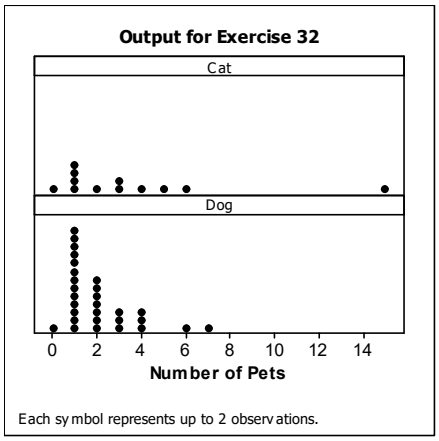
2.30 The histogram is right-skewed and also bimodal (at least with this grouping). The modes are at about 80 days and 240 days. The typical value is about 240 days (between 160 and 320 days). There are two outliers at more than 600 days, the Asian elephant and the African elephant. Humans (266 days) would be near the middle of the graph.



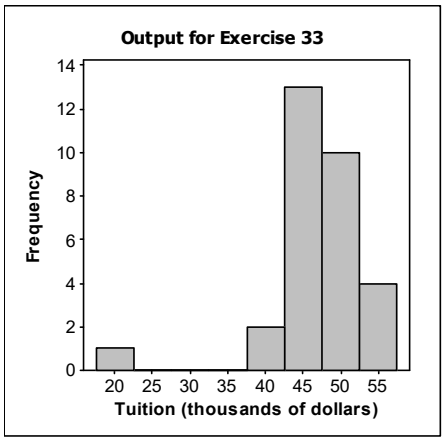
2.31 Both graphs are multimodal and right-skewed. The Democrats have a higher typical value, as shown by the fact that the center is roughly around 35 or 40%, while the center value for the Republicans is closer to 20 to 30%. Also note the much larger proportion of Democrats who think the rate should be 50% or higher. The distribution for the Democrats appears more spread out because the Democrats have more people responding with both lower and higher percentages.



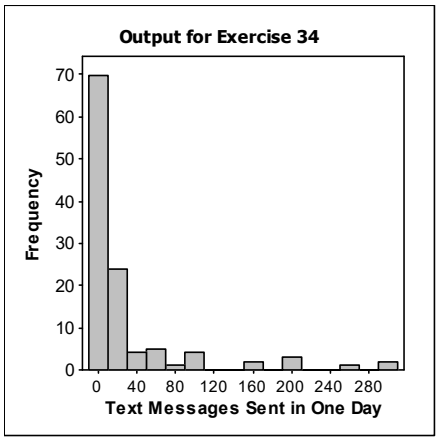
2.32 Both distributions are right-skewed. A large outlier did represent a cat lover, but typically, cat lovers and dog lovers both seem to have about 2 pets, although there are a whole lot of dog lovers with one dog.



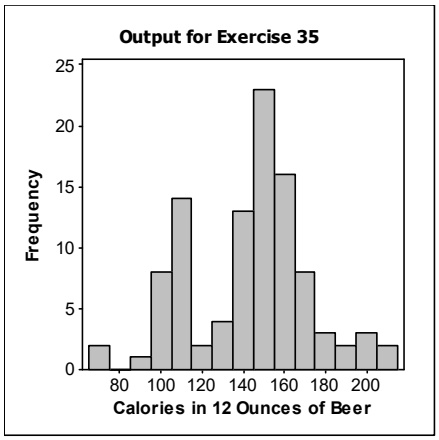
2.33 The distribution appears left-skewed because of the low-end outlier at about $20,000 (Brigham Young University).



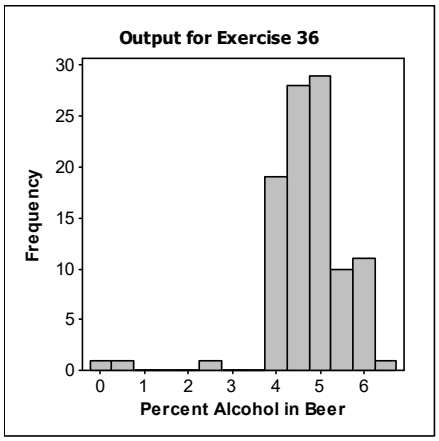
2.34 The histogram is strongly right-skewed, with outliers.



2.35 With this grouping the distribution appears bimodal with modes at about 110 and 150 calories. (With fewer—that is, wider—bins, it may not appear bimodal.) There is a low-end outlier at about 70 calories. There is a bit of left skew.



2.36 The distribution is left-skewed primarily because of the outliers at about 0% alcohol.



Section 2.3: Visualizing Variation in Categorical Variables and Section 2.4: Summarizing Categorical Distributions

2.37 No, the largest category is Wrong to Right, which suggests that changes tend to make the answers more likely to be right.

2.38 a. About 7.5 million.

b. About 5 million.

c. No, overweight and obesity do not result in the highest rate. That is from high blood pressure.

d. This is a Pareto chart.

2.39 A bar graph with the least variability would be one in which most children favored one particular flavor (like chocolate). A bar graph with most variability, would be one in which children were roughly equally split in their preference. with 1/3 choosing vanilla, 1/3 chocolate, 1/3 strawberry.

2.40 Least amount of variability would be one where most of the applicants had the same education goal (like transfer). Most variability would be one where applicants were roughly equally divided among the five choices.

2.41 a. "All of the Time" was the most frequent response.

b. The difference is about 10%. It is easier to see in the bar chart.

2.42 a. About 30%.

b. About 7%. It is easier to use the bar chart.

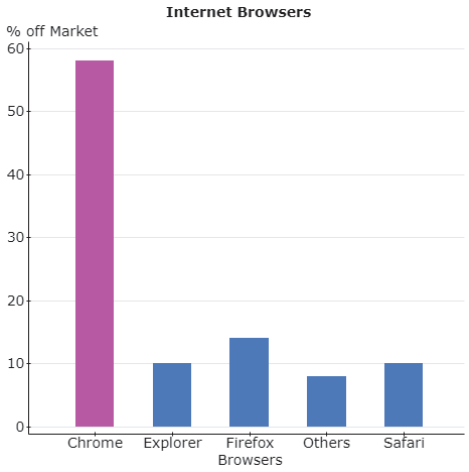
2.43 a. 40–59-year old's.

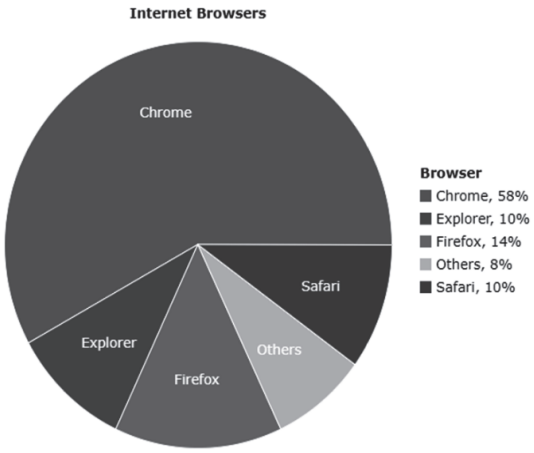
b. The obesity rates for women are slightly higher in the 20–39 and 60+ groups. The obesity rate for men is higher in the 40–59 age group.

2.44 a. Fitness rates are slightly higher in the West than in the other three regions.

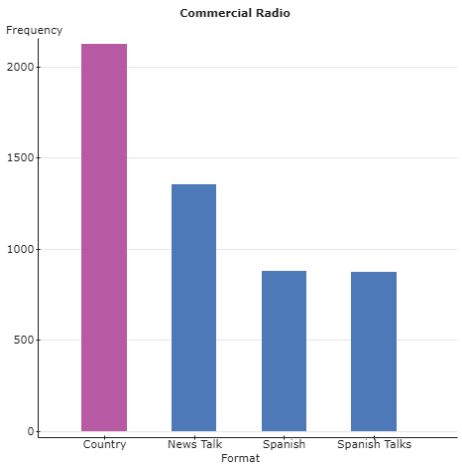
b. In all regions aerobic fitness rates were higher than muscle strengthening rates.

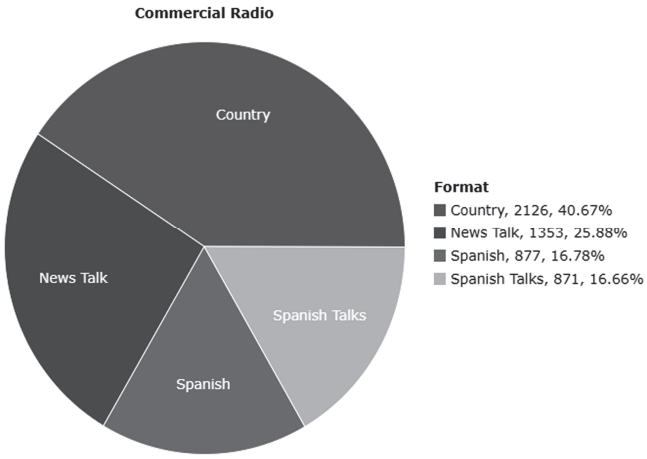
2.45 Bar graph or pie chart. Chrome controls the highest market share.





2.46 Bar graph or pie chart. Country is the most popular format.





Section 2.5: Interpreting Graphs

2.47 This is a histogram, which we can see because the bars touch. The software treated the values of the variable *Garage* as numbers. However, we wish them to be seen as categories. A bar graph or pie chart would be better for displaying the distribution.

2.48 The graph is a histogram (the bars touch), and histograms are used for numerical data. But this data set is categorical, and the numbers (1, 2, and 3) represent categories. A more appropriate graph would be a bar graph or pie graph.

2.49 Hours of sleep is a numerical variable. A histogram or dotplot would better enable us to see the distribution of values. Because there are so many possible numerical values, this pie chart has so many "slices" that it is difficult to tell which is which.

2.50 a. This is a bar chart (or bar graph), as you can see by the separation between bars.

b. These numerical data would be better shown as a pair of histograms (with a common horizontal axis) or a pair of dotplots. Bar graphs are for categorical data.

2.51 Those who still play tended to have practiced more as teenagers, which we can see because the center of the distribution for those who still play is about 2 or 2.5 hours, compared to only about 1 or 1.5 hours for those who do not. The distribution could be displayed as a pair of histograms or a pair of dotplots.

2.52 a. *Gender* is categorical and *Hours on Cell Phone* is also categorical.

b. Because in this data set both variables are categorical, the bar chart is appropriate.

c. You could make two histograms (or two dotplots) for the data because the time would be numerical. It would be ideal to use a common horizontal axis for easy comparison of the two graphs.

d. The distributions show that the women tend to talk more. (The mode for women is 4–8 hours, and the mode for the men is 0–4 hours.)

Chapter Review Exercises

2.53 Since the data are numerical, a pair of histograms or dotplots could be used, one for the males and one for the females. A statistical question is Who slept more on average, men or women?"

2.54 Since the data are categorical, a side-by-side bar chart could be used. Examples of statistical questions: Which major has the greatest percentage of men? Which major is most popular among women, and is this different than for men?

2.55 a. The diseases with higher rates for HRT were heart disease, stroke, pulmonary embolism, and breast cancer. The diseases with lower rates for HRT were endometrial cancer, colorectal cancer, and hip fracture.

b. Comparing the rates makes more sense than comparing just the numbers, in case there were more women in one group than in the other.

2.56 a. South Korea and the United States have the highest rate of access to the Internet.

b. China and Thailand have the highest percentage of music purchased over the Internet.

2.57 The vertical axis does not start at zero and exaggerates the differences. Make a graph for which the vertical axis starts at zero.

2.58 In histograms the bars should generally touch, and these don't touch. Also, we cannot see the top of the range because "More" is a poor label. Change the numbers on the horizontal axis and increase the width of the bins so as to make the bars touch.

2.59 The shapes are roughly bell-shaped and symmetric; the later period is warmer, but the spread is similar. This is consistent with theories on global warming. The difference is 57.9 – 56.7 = 1.2, so the difference is only a bit more than 1degree Fahrenheit.

2.60 The typical percentage of students with jobs at the top schools is higher than the percentage for the bottom 91 schools. In other words, you are more likely to find a job if you went to a law school in the top half of the rankings. Both histograms are left-skewed. Also, the range for the bottom schools is wider, because it goes down to lower employment rates.

2.61 a. A smaller percentage favor nuclear energy in 2016.

b. The Republicans show the most change.

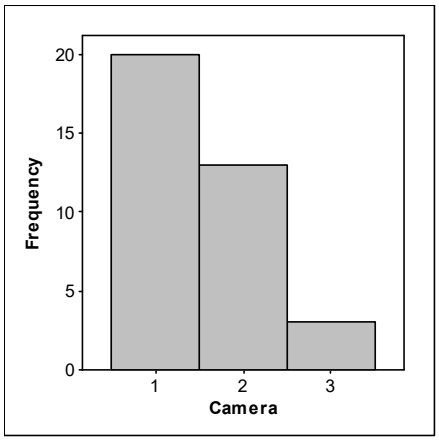
c. A side-by-side bar graph (Republican 2010 adjacent to Republican 2016) would make the comparison easier.

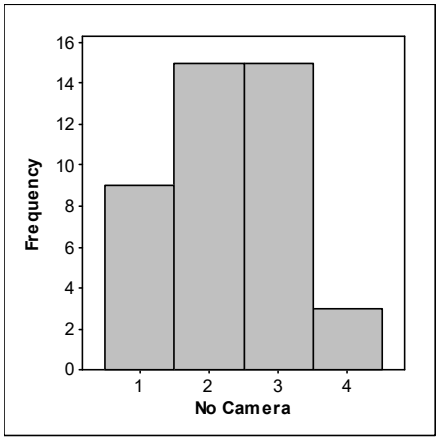
2.62 A greater percentage of people think stem cell research is morally acceptable in 2017 than in 2002.

2.63 The created 10-point dotplot will vary, but the dotplot for this exercise should be right-skewed.

2.64 The created 10-point dotplot will vary, but the dotplot for this exercise should be not be skewed.

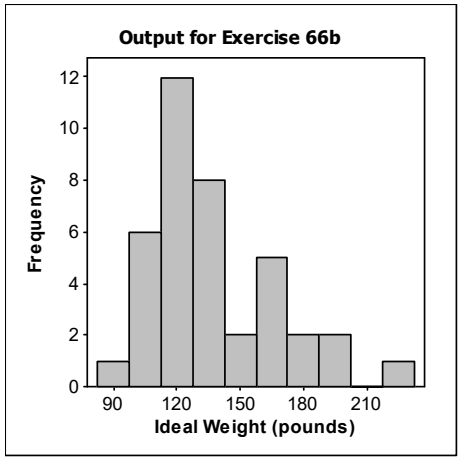
2.65 Graphs will vary. Histograms and dotplots are both appropriate. For the group without a camera the distribution is roughly symmetrical, and for the group with a camera it is right-skewed. Both are unimodal. The number of cars going through a yellow light tends to be less at intersections with cameras. Also, there is more variation in the intersections without cameras.

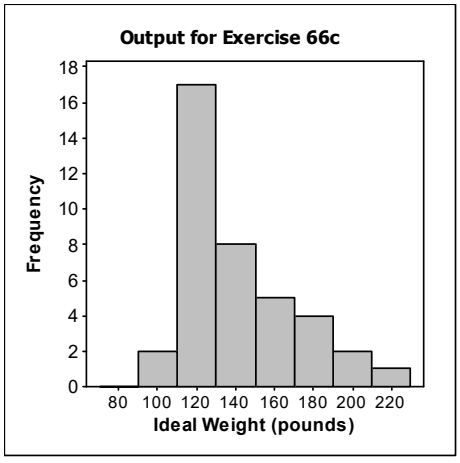




2.66 a. You might expect bimodality because men tend to have ideal weights that are larger than women's ideal weights.

b. and c.





Graphs may vary, depending on technology and the choice of bins for the second histogram. On the two graphs given here, the bin width for the first is 15 pounds and for the second is 20 pounds. The first distribution is bimodal and the second is not.

2.67 Both distributions are right-skewed. The typical speed for the men (a little above 100 mph) is a bit higher than the typical speed for the women (which appears to be closer to 90 mph). The spread for the men is larger primarily because of the outlier of 200 mph for the men.

2.68 Both graphs are relatively symmetric and unimodal. The center for the men is larger than the center for the women, showing that men tend to wear larger shoes than women. The spread is a bit more for women because their sizes range from about 5 to about 10 whereas the men's sizes range from about 8 to about 12. There are no outliers in either group.

2.69 The distribution should be right-skewed.

2.70 Since most of the physician's patients probably do not smoke and a few may be heavy smokers, the distribution should be right-skewed with lots of zeros and a few high numbers.

2.71 a. The tallest bar is Wrong to Right, which suggests that the instruction was correct.
b. For both instructors, the largest group is Wrong to Right, so it appears that changes made tend to raise the grades of the students.

2.72 a. The raw numbers would be affected by how many were in each group, and that might hide the rate. For example, because there are many more old women than old men, that information would hide the rates.

b. The males up to about 64 have a higher rate of visits to the ER. From 65 to 74 the rates are about the same, and for 75 and up the rates are higher for the women.

2.73 a. Facebook (only about 5% used it less often than weekly).

b. LinkedIn (only about 20% used it daily).

c. Facebook (around 75% were in one category—daily).

2.74 a. Facebook is used most frequently by men and women.

b. Facebook is used most frequently by both genders. Pinterest is used least by males, while Instagram is used least by females.

2.75 a. Histogram or dotplot

b. Side-by-side barplots showing gender frequencies separately for each zip code.

2.76 a. Stacked or side-by-side histograms or dotplots. Ideally histograms would show relative frequencies.

b. Bar chart by zip code frequency.

Chapter 3: Numerical Summaries of Center and Variation

Answers may vary slightly, especially for quartiles and interquartile ranges, due to type of technology used, or rounding.

Section 3.1: Summaries for Symmetric Distributions

3.1 c

3.2 b

3.3 The mean is between about 4 and 6 hours.

3.4 The mean is approximately 139 mEq/L.

3.5 a. The mean number of floors is 118.6.

b. The standard deviation of the number of floors is 26.0.

c. Dubai is farthest from the mean.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
# Floors	5	118.6	676.3	26.005769	11.630133	108	62	101	163	101	120

3.6 a. The typical height of the coasters is 385.2 feet.

b. The standard deviation of the heights is 64.0 feet.

c. Both the mean and the standard deviation would decrease. The new mean is 378 feet; the new standard deviation is 55.5 feet.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Height (in ft)	5	385.2	4097.7	64.01328	28.627609	415	146	310	456	325	420

3.7 a. The typical river is 2230.8 miles long.

b. The standard deviation is 957.4 miles. The Mississippi-Missouri-Red Rock River contributes most to the standard deviation because it is farthest from the mean.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Length (in miles)	5	2230.8	916540.7	957.36132	428.145	1900	2260	1450	3710	1459	2635

c. The mean would decrease, and the standard deviation will increase. New mean: 1992.3 miles, new standard deviation: 1036.5 miles.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Length (in miles)	5	2230.8	916540.7	957.36132	428.145	1900	2260	1450	3710	1459	2635

3.8 a. Typically, the first ladies had $\bar{x} = \frac{0+5+6+0+2+4}{6} = \frac{17}{6} = 2.8$ children. One could also say that the first ladies had about 2.8 children, on average.

b. 2.8 is a lot less than 7 or 8, so the first ladies tended to have fewer children than usual at that time in history.

c. Martha Jefferson, with 6, has the number farthest from the mean.

d. $s = \sqrt{\frac{32.84}{5}} = 2.6$

x	$x - \bar{x}$	$(x - \bar{x})^2$
0	–2.8	7.84
5	2.2	4.84
6	3.2	10.24
0	–2.8	7.84
2	–0.8	0.64
4	1.2	1.44
17	0	32.84

3.9 a. For the early 1900s the mean is 22.02 seconds; the standard deviation is 0.40 seconds.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Time (1900-1920)	5	22.02	0.162	0.40249224	0.18	22	1	21.6	22.6	21.7	22.2

b. For recent Olympics, the mean is 19.66 seconds; the standard deviation is 0.35 seconds.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Time (2000-2016)	5	19.66	0.123	0.35071356	0.15684387	19.8	0.8	19.3	20.1	19.3	19.8

c. Recent winners are faster and have less variation in their winning times.

3.10 a. Backstroke: Mean is 52.92 seconds standard deviation is 0.93 seconds.
Butterfly: Mean is 51 seconds; standard deviation is 0.76 seconds.

b. Butterfly tends to have a faster gold medal time and less variation in winning times.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
100-Meter Backstroke	5	52.92	0.867	0.93112835	0.41641326	52.6	2.1	52	54.1	52.2	53.7
100-Meter Butterfly	5	51	0.575	0.75828754	0.3391165	51.2	1.9	50.1	52	50.4	51.3

3.11 a. A total of 185 people were surveyed.

b. Comparing the means, men thought more should be spent on a wedding. Comparing the standard deviations, men had more variation in their responses.

3.12 Those who had never previously had a wedding thought more should be spent (greater mean). Those who had already had a wedding had more variation in their responses (greater standard deviation).

3.13 a. The mean for longboards is 12.4 days, which is more than the mean for shortboards, which is 9.8 days. So longboarders tend to get more surfing days.

b. The standard deviation of 5.2 days for the longboarders was larger than the standard deviation of 4.2 days for the shortboarders. So the longboarders have more variation in days.

Descriptive Statistics: Long, Short

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Long	30	12.367	5.249	4.000	8.000	11.500	16.750	22.000
Short	30	9.767	4.256	4.000	6.750	9.500	12.000	20.000

3.14 California: Mean is $6847.67; standard deviation is $429.04; Texas: Mean is $7220.17; standard deviation is $807.91. State college tuition costs tend to be higher and have more variation in Texas than in California.

3.15 San Jose tends to have a higher typical temperature; Denver has more variation in temperature.

3.16 New York City tends to have higher temperatures and more variation in temperature.

3.17 a. 20.3 to 40.1 pounds.

b. Less than one standard deviation from the mean.

3.18 a. Top: $52.2 + 2.5 = 54.7$; Bottom: $52.2 - 2.5 = 49.7$

b. No, 54 is not more than one standard deviation above the mean because it is not more than 54.7.

3.19 a. The mean is \$3.66 and represents the typical price of a container (59 to 64 ounces) of orange juice at this site.

b. The standard deviation is 0.51. Most orange juice of this size is within 51 cents of \$3.66.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
OJ Prices ($)	10	3.659	0.26183222	0.51169544	0.16181231	3.785	1.5	2.99	4.49	2.99	3.99

3.20 a. The mean is 6.8 years.

b. In 10 years, the mean is 16.8 years.

c. The standard deviation is 3.6 years.

d. In 10 years, standard deviation is 3.6 years.

e. The mean increased by 10 because the data values increased by 10. The standard deviation did not change because the variation of the values did not change.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Ages (years)	4	6.75	12.916667	3.5939764	1.7969882	7.5	8	2	10	4	9.5
Ages (Years)	4	16.75	12.916667	3.5939764	1.7969882	17.5	8	12	20	14	19.5

3.21 The standard deviation for the 100-meter event would be less. All the runners come to the finish line within a few seconds of each other. In the marathon, the runners can be quite widely spread out after running that long distance.

3.22 The standard deviation for the soccer team would probably be less because all the players might have healthful weights. The academic team might have members of widely different weights.

3.23 South Carolina: Mean is \$198.1 thousand; standard deviation is \$68.0 thousand. Tennessee: Mean is \$215.8 thousand; standard deviation is \$75.1 thousand. Houses in Tennessee are typically more expensive and have more variation in price than houses in South Carolina.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
SC (in $thousands)	15	198.13333	4625.5524	68.011414	17.560471	183	224	110	334	160	221
TN (in $thousands)	15	215.8	5640.0286	75.100124	19.390769	200	275	125	400	160	280

3.24 a. Florida has more expensive home prices. (Florida mean is \$244.4 thousand compared to Georgia mean of \$228.6 thousand)

b. Florida has more variation in home prices (Florida standard deviation is \$86.4 thousand compared to Georgia standard deviation of \$80.0 thousand)

c. The standard deviation would decrease. New standard deviation: \$68.3 thousand

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
FL ($ Thousands)	13	244.37692	7470.3736	86.431323	23.971736	229.9	292.5	139.9	432.4	187.9	312.3
GA ($ Thousands)	13	228.55385	6404.6127	80.028824	22.196002	220.5	261.1	123.9	385	170	265

3.25 a. The mean for the men was 10.5, and for the women it was 4.7, showing that the male drinkers typically drank more (on average, almost six drinks more) than the female drinkers.

b. The standard deviation for the men was 11.8, and the standard deviation for the women was 4.8, showing much more variation in the number of drinks for the men.

Descriptive Statistics: Drinks_Female, Drinks_Male

Variable	N	Mean	StDev
Drinks_Female	46	4.696	4.821
Drinks_Male	53	10.55	11.83

c. The mean for the men is now 8.6, and the mean for the women is still 4.7. Thus the mean for the men is still larger than the mean for the women, but not as much larger.

d. The standard deviation would be smaller without the two outliers. This is because the contribution to the standard deviation from these two outliers is large since they are farthest from the mean, and that contribution would be removed.

Descriptive Statistics: Drinks_Female, Drinks_Male w/o 48 & 70

Variable	N	Mean	StDev
Drinks_Female	46	4.696	4.821
Drinks_Male w/o 48 & 70	51	8.647	6.57

3.26 a. The mean weight for the babies born to the smoking mothers was 2863 grams, and the mean weight for the nonsmoking mothers was 3588 grams, so the typical baby of a smoking mother weighed less than the typical baby of a nonsmoking mother. The standard deviation for the smoking mothers was 957 g and for the nonsmoking mothers it was 597 g, showing that there was more variation with the smoking mothers.

b. The mean for the babies born to smoking mothers was 2957 grams without the outlier, and the mean for babies born to the nonsmoking mothers was still 3588 grams. The typical baby of a smoking mother weighed less than the typical baby of a nonsmoking mother, even without the outlier. The standard deviation for the smoking mothers was 871 g, and for the nonsmoking mothers it was 597 g, showing that there was more variation with the smoking mothers, even without the outlier. Removing the low outlier caused the mean to increase and the standard deviation to decrease.

Descriptive Statistics: NO, YES, YES_896 removed

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
NO	35	3588	597	2436	3080	3612	4060	4788
YES	22	2863	957	896	2230	2940	3367	5264
YES_896 removed	21	2957	871	1008	2492	2968	3430	5264

3.27 A standard deviation can be zero if all the values are the same (no variation between the data values).

3.28 The standard deviation cannot be negative because in the calculation all differences from the mean are squared, making all distances positive values.

Section 3.2: What's Unusual? The Empirical Rule and *z*-Scores

3.29 a. 68%

b. (16 + 34 + 27)/120 = 64%. The estimate is very close to 68%.

c. Between 555 and 819 runs.

3.30 a. Between 63.5 and 78.1 million BTU (70.8 ± 7.3)

b. Between 56.2 and 80.4 million BTU (70.8 ± 2(7.3))

c. Yes because 54 million BTU is more than 2 standard deviations from the mean

d. No, because 85.4 million BTU is not more than 2 standard deviations from the mean

3.31 a. Approximately 95%, using the empirical rule (35.9 ± 2(11.6))

b. Approximately 68% using the empirical rule (35.9 ± 11.6)

c. No, because it is not more than 2 standard deviations from the mean.

3.32 a. Approximately 68%, using the empirical rule (43.0 ± 11.3)

b. Approximately 95%, using the empirical rule (43.0 ± 2(11.3)).

c. No, because 51.9 is not more than 2 standard deviations from the mean.

3.33 a. $\frac{58-64}{3}=\frac{-6}{3}=-2$

b. $x=\bar{x}+zs=64+1(3)=64+3=67$ inches (or 5 feet 7 inches)

3.34 a. $x=\bar{x}+zs=64-1(3)=64-3=61$ inches (or 5 feet 1 inch)

b. $\frac{70-64}{3}=\frac{6}{3}=2$

3.35 An IQ below 80 is more unusual because 80 is 1.33 standard deviations from the mean while 110 is only 0.67 standard deviations from the mean.
$80-100/15=-1.33$ $\quad$ $110-100/15=.67$

3.36 They are equally unusual. The z-score for 9 days early is $\frac{263-272}{9}=\frac{-9}{9}=-1,$ and the z-score for 9 days late is $\frac{281-272}{9}=\frac{9}{9}=1,$ and those z-scores are equally distant from 0.

3.37 a. $z=\frac{2500-3462}{500}=\frac{-962}{500}=-1.92$

b. $z=\frac{2500-2622}{500}=\frac{-122}{500}=-0.24$

c. A birth rate of 2500 grams is more common (the z-score is closer to 0) for babies born one month early. In other words, there is a higher percentage of babies with low birth weight among those born one month early. This makes sense because babies gain weight during gestation, and babies born one month early have had less time to gain weight.

3.38 a. $z=\frac{45-52.2}{2.5}=\frac{-7.2}{2.5}=-2.88$

b. $z=\frac{45-47.4}{2.5}=\frac{-2.4}{2.5}=-0.96$

c. A birth length of 45 cm is more common for babies born one month early. That makes sense, because babies grow during gestation, and babies born one month early have had less time to grow.

3.39 a. $69=2(3)=75$ inches

b. $69-(1.5)3=64.5$ inches

3.40 a. $x=\bar{x}+zs=64-1(2.5)=61.5$ inches

b. Evelyn: $z=75-64/2.5=4.4$

Draymond: $z=79-69/3=3.33$

Evelyn is taller b/c she is further from the mean compared to Draymond.

Section 3.3: Summaries for Skewed Distributions

3.41 Two measures of the center of data are the mean and the median. The median is preferred for data that are strongly skewed or have outliers. If the data are relatively symmetric, the mean is preferred but the median is also okay.

3.42 Two measures of variation (spread) are the interquartile range (IQR) and the standard deviation. The IQR is preferred for data that are strongly skewed or have outliers. If the data are relatively symmetric, the standard deviation is preferred.

3.43

363	384	389	408	423	434	471	520	602	677
		Q1			Med		Q3		

a. Median: 428.5 million; about half the top 10 Marvel movies made more than $428.5 million.

b. Q1 = 389 million, Q3 = 520 million; IQR = 520 – 389 = 131 million. This is the range of the middle 50% of the sorted incomes in the top 10 Marvel movies.

c. Range = 677 – 363 = 314 million. The IQR is preferred over the range because the range depends on only two observations and because it very sensitive to any extreme values in the data.

3.44

346	366	407	487	543	547	643
	Q1		Med		Q3	

a. Median = 487 million; about half the top 7 DC movies made more than $487 million

b. IQR = 547 – 366 = 181 million; this is the range of the middle 50% of the sorted incomes in the top 7 DC movies.

c. Range = 643 – 346 = 297 million. The IQR is preferred over the range because the range depends on only two observations and because it very sensitive to any extreme values in the data.

3.45 Median = 471 million. About 50% of the top 7 Marvel movies made more than $471 million.

3.46 Median = 543 million. About 50% of the top 5 DC movies made more than $543 million.

3.47 a. 25%

b. 75%

c. 50%

d. IQR = Q3 – Q1 = 390 – 237.7 = 152.3. The range of the middle 50% of the sorted data is 152.3 million BTUs.

3.48 a. 25%

b. 75%

c. 91.2

d. more variability in Total Energy Consumption (IQR for Total Energy Consumption is greater than the IQR of Industrial Energy Consumption).

Section 3.4: Comparing Measures of Center

3.49 a. Outliers are observed values that are far from the main group of data. In a histogram they are separated from the others by space. If they are mistakes, they should be removed. If they are not mistakes do the analysis twice: once with and once without outliers.

b. The median is more resistant, which implies that it changes less than the mean (when the data with and without outliers are compared).

3.50 Because one of the sets is strongly skewed, compare the median and the IQR for both sets.

3.51 The corrected value will give a different mean but not a different median. Medians are not as affected by the size of extreme scores, but the mean is affected.

3.52 If you were representing the owners, you would use the mean, because it is higher, and you would want to convince the public that the ball players are already getting a high enough salary. If you were a player, you would use the median, because it is lower, and you would want to argue that salaries should go up because players are getting too little. The values are so different because of the extreme right skew and the large outliers.

3.53 a. The distribution is right-skewed.

b. The median and the IQR should be used to describe the distribution.

3.54 The distributions are roughly symmetric and unimodal, so the mean and standard deviation can be used to describe the center and the spread of the distribution.

3.55 a. The distributions are right-skewed.

b. The medians.

c. The interquartile ranges.

d. The typical Democratic senator has been in office 9 years, while the typical Republican senator has been in office 6 years. There is more variability in the experience of Democratic senators, with an IQR of 12 years compared to an IQR of 10 years for Republican senators.

3.56 a. Both distributions are slightly right-skewed.

b. The medians.

c. The interquartile ranges.

d. The typical player age is similar for both teams, with a median age of 28 years for the Cubs and 27 years for the A's. There is more variation among the ages of the Cubs. The IQR for the Cubs is 7 years while the IQR for the A's is 4 years.

3.57 a. The median is 48. 50% of the southern states have more than 48 capital prisoners.

b. Q1 = 32, Q3 = 152, IQR = 120

c. The mean is 90.8.

d. The mean is pulled up by the really large numbers, such as Texas (243) and Florida (374).

e. The median is unaffected by outliers.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Capital Prisoners	15	90.8	11147.6	105.5822	27.261206	48	374	0	374	32	152

3.58 a. The median is 8. 50% of the western states have fewer than 8 capital prisoners.

b. Q1 = 2, Q3 = 33, IQR: 33 – 3 = 31

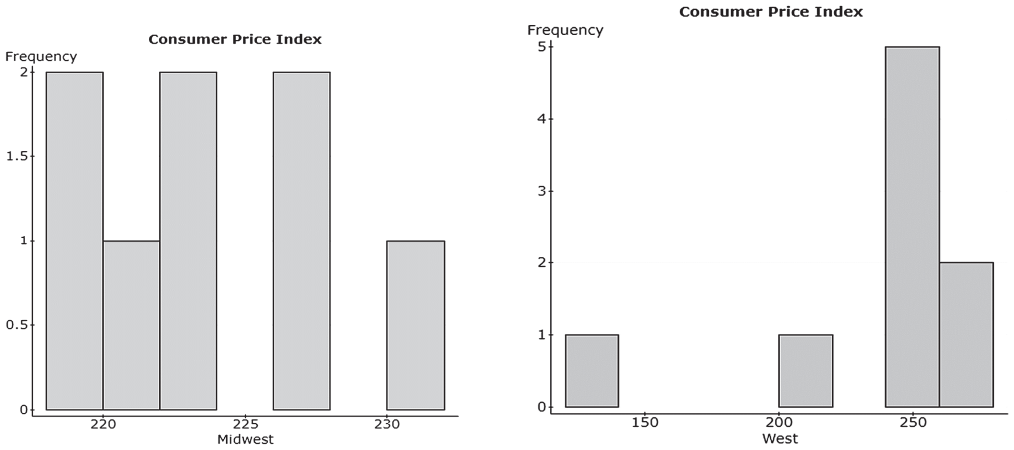
c. 78.38

d. The mean is pulled up by the really large numbers, such as California (746) and Arizona (125).

e. The median is unaffected by outliers.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Capital Prisoners	13	78.384615	41692.256	204.18682	56.631234	8	746	0	746	2	33

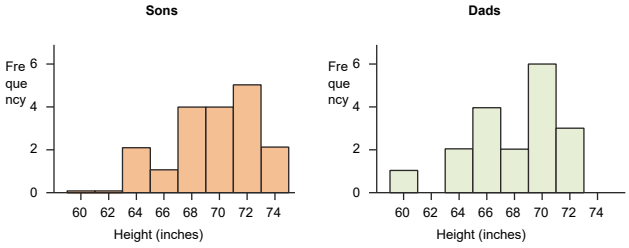
3.59



Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Midwest	8	223.6375	18.502679	4.301474	1.5208007	222.8	11.9	218.7	230.6	219.9	227.2
West	9	234.58889	1819.9511	42.660885	14.220295	244.6	141.4	128	269.4	240	258.6

Since the data for the West is left-skewed, use the median and IQR to compare the groups. The CPI for the West is higher than that of the Midwest (West median 244.6; Midwest median 222.8). There is more variability in the West CPI (West IQR 18.6; Midwest IQR 7.3). The West has one potentially low outlier.

3.60 a. Both data sets are slightly left-skewed.



b. The values below were obtained from Minitab.

Variable	Mean	StDev	Median	IQR
Sons	69.278	3.045	70.000	3.250
Dads	67.500	3.258	68.500	5.000

c. The mean for the sons is 69.3, which is larger than the mean for the dads, which is 67.5. Therefore, the sons tend to be taller than the dads. But the standard deviation for the dads (3.3) is larger than the standard deviation of the sons (3.0), showing that the dads' heights tend to be more spread out.

d. The median for the sons (70) is larger than the median for the dads (68.5), showing that the sons tend to be taller than the dads. The interquartile range for the dads (5) is larger than the interquartile range for the sons (3.25), showing that the dads' data tend to be more spread out. (The answer for the interquartile range may vary depending on the method used to find it.)

e. With strong skew in either group, you should compare medians and interquartile ranges. Because both data sets are a little skewed, you could argue either for comparing the means and standard deviations *or* for comparing the medians and interquartile ranges. Both comparisons show that the sons tend to be taller, but the data for the dads tend to be more spread out.

3.61 a. The balancing point of the histogram, the mean is approximately 80 millimeters.

b. $\bar{x} = 1(60) + 8(70) + 8(80) + 5(90) + 3(100) + 1(100) + 1(120)/27 = 83.0$ millimeters

c. It is an approximation because we used the left-hand side of the bin to estimate the data values contained in the bin.

3.62 a. Based on the histogram, the mean is approximately 80 millimeters.

b. $\bar{x} = 3(60) + 9(70) + 5(80) + 5(90) + 3(100) + 1(120) + 1(130)/27 = 81.9$ millimeters

c. It is an approximation because we used the left-hand side of the bin to estimate the data values contained in the bin.

Section 3.5: Using Boxplots for Displaying Summaries

3.63 a. South, West, Northeast, Midwest.

b. Northeast, South, Midwest, West.

c. The South and the Northeast each have one state with a high-priced low poverty rate.

d. The Northeast has the least amount of variability (smallest IQR).

e. The IQR is better because is it not influenced by unusually high or low values and because the data is not symmetric.

3.64 The data sets are right-skewed, as shown by the potential outliers and longer right whiskers, so the medians and interquartile ranges are appropriate.

Medians: W less than 50, MW between 50 and 100, S about 100, NE a bit less than 400.

Interquartile range (smallest to largest): W, S, MW, NE

3.65 a. The NFL has the highest ticket prices and the most variability in ticket prices (highest median and greatest IQR). The MLB has the lowest ticket prices.

b. Hockey tickets tend to be more expensive than basketball tickets (higher median). Both sports have some unusually high-priced tickets, and hockey has more variability in ticket prices (greater IQR).

3.66 a. The median for the western states is approximately 26%. The median for the eastern states is approximately 27%. A greater percentage of residents in eastern states have a BA.

b. The eastern states show more variability in the data with a wider range and one outlier.

3.67 Explanations of reasoning will vary.

a. Histogram 1 is left-skewed, histogram 2 is roughly bell-shaped (not very skewed), and histogram 3 is right-skewed.

b. Histogram 1 goes with Boxplot C, since the boxplot is left-skewed.
Histogram 2 goes with Boxplot B, since the boxplot is not very skewed.
Histogram 3 goes with Boxplot A, since the boxplot is right-skewed.

3.68 Explanations of reasoning will vary.

Histogram X goes with boxplot M; they both have a larger range with no outliers.
Histogram Y goes with boxplot C; they both have a larger range with two outliers
Histogram Z goes with boxplot P; they both show the smallest range.

3.69 The maximum value (756) is greater than the upper fence (237 + 1.5(237 - 63) = 498) and is considered a potential outlier. The minimum value (1) is not less than the lower fence.

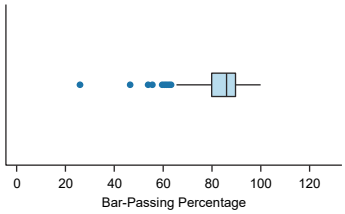
3.70 The maximum gas tax (68.7) is greater than the upper fence (51 + 1.5(51 - 40.2)) and may be considered a potential outlier. The minimum gas tax (30.7) is not less than the lower fence (40.2 - 1.5(51 - 40.2) = 24) and is not a potential outlier.

3.71 The whiskers are drawn to the upper and lower fences or to the maximum and minimum values if there are no potential outliers in the data set. There are no values less than the lower fence, so the left whisker is drawn to the minimum. Since the maximum value (128.016) is beyond the upper fence (33.223 + 1.5(33.223 - 8.526) = 70.3) the right whisker would be drawn to the upper fence (70.3).

3.72 Q1 - 1.5 * IQR = 80 - 1.5 * (90 - 80) = 80 – 1.5 * 10 = 80 - 15 = 65 percentage points.

The left-hand whisker will stop at 65, and there are quite a few outliers on the left side; these are schools with pass rates far lower than the typical school. On the right, Q3 + 1.5 * IQR = 90 + 15 = 105.

The highest school is at 100, so the right whisker stops at 100. There are no outliers on the right side.



3.73 The IQR is 90 - 78 = 12. 1.5 * 15 = 18, so any score below 78 - 18 = 60 is a potential outlier. We can see that there is at least one potential outlier (the minimum score of 40), but we don't know how many other potential outliers there are between 40 and 60. Therefore, we don't know which point to draw the left-side whisker to.

3.74 Yes, you can draw the boxplot. The IQR = 12, so 1.5 * 12 = 18. Thus, any point below 78 - 18 = 50 is a potential outlier. The smallest value is 60, so we know there are no potential outliers, and we should draw the left whisker to 60. On the right side, potential outliers are points greater than 90 + 18 = 108. Again, there are no potential outliers, and the whisker should stop at 100.

Chapter Review Exercises

3.75 a. The median is 41.05 cents/gallon. 50% of the southern states have gas taxes greater than 41.05 cents/gallon.

b. The middle 50% of the southern states have gas taxes in a range of 11.35 cents/gallon.

c. The mean is 43.6 cents/gallon.

d. The data may be right-skewed.

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Gas Tax (cents/gallons)	16	43.6	40.749333	6.3835204	1.5958801	41.05	19.8	35.2	55	38.85	50.2

3.76 a. 50% of the western states have gas taxes greater than 47 cents/gallon.

b. The middle 50% of western states have gas taxes in the range of 14.9 cents/gallon.

c. The mean is 46.7 cents/gallon.

d. The data may be fairly symmetric.

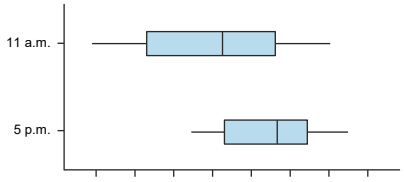
Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Gas Tax (cents/gallon)	14	46.571429	97.317582	9.8649674	2.6365234	47	32.2	30.7	62.9	37.4	52.3

3.77 The 5 p.m. class did better, typically; both the mean and the median are higher. Also, the spread (as reflected in both the standard deviation and the IQR) is larger for the 11 a.m. class, so the 5 p.m. class has less variation. The visual comparison is shown by the boxplots. Both distributions are slightly left-skewed. Therefore, you can compare the means and standard deviations or the medians and IQRs.

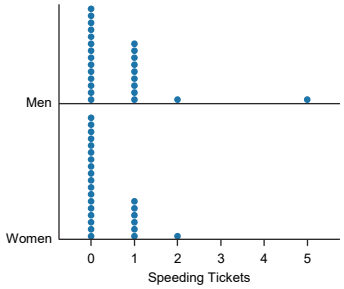
The visual comparison is shown by the boxplots. Both distributions are slightly left-skewed. Therefore, you can compare the means and the standard deviations *or* the medians and the IQRs.

Minitab Statistics

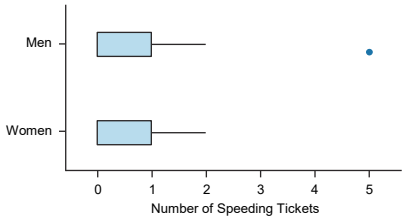
Variable	N	Mean	Median	StDev	Min	Max	Q1	Q3
11am	15	70.73	72.5	19.84	39	100	53	86
5pm	19	84.78	86.5	11.95	64.5	104.5	73	94



3.78



The dotplots show that both groups are unimodal. Both data sets are right-skewed. Refer to the boxplots.



Because of the outlier for the men, it is best to compare the medians and IQ ranges. They are the same: both medians are 0, and both IQ ranges are 1. The men's data include an outlier of 5 tickets.

Minitab Statistics: Men, Women

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Men	25	0.640	1.075	0.000	0.000	0.000	1.000	5.000
Women	25	0.320	0.557	0.000	0.000	0.000	1.000	2.000

3.79 The graph is bimodal, with modes around 65 inches (5 feet 5 inches) and around 69 inches (5 feet 9 inches). There are two modes because men tend to be taller than women.

3.80 a. Bimodal.

b. It could be from gender (men are faster than women) or from event (Olympic runners are faster than amateurs).

c. The bigger group (the mound is bigger) had the bigger (slower) times, so the mounds are probably due to the event (Olympic vs. amateur), not to gender. It would be unlikely for the larger mound for the slower group to represent women because there are fewer women who run in marathons.

3.81 Students should provide histograms for males and females; because of lack of symmetry, they should compare medians and IQRs. Both groups typically watch about the same amount and have similar variability of lack of symmetry they should compare medians and IQRs.

3.82 Students should provide histograms for both males and female because both groups had similar ideal retirement ages and similar variability in responses.

3.83 a. The mean is approximately 1000 calories.

b. An estimate of the standard deviation is (2200 – 100)/6 = 350 calories.

3.84 a. The mean is 1(3) + 4(4) + 27(5) + 65(6) + 62(7) + 51(8) + 11(9) + 3(10)/224 = 6.8 hours.

b. Since the data are fairly symmetric, we expect the mean and the median to be about the same.

3.85 Answers will vary.

3.86 Answers will vary.

3.87 Answers will vary.

3.88 Answers will vary.

3.89 Students should make a histogram for western states and southern states. Because the distributions are not symmetric, students should compare the medians and the IQRs. The western states tend to have a higher percentage of the population with a bachelor's degree. The southern states tend to have more variability.

3.90 Students should make a histogram for the northeastern states and for the midwestern states. Because the distributions are not symmetric, students should compare the medians and the IQRs. The unemployment rate for northeastern states is slightly higher than that of midwestern states. There is more variation in unemployment rates among midwestern states. (The median unemployment rate: Northeast 5.05, Midwest 4.4; IQR: Northeast 1.1, Midwest 1.55.)

3.91 a. Since the distributions are slightly right-skewed, compare the medians.

b. Football ticket prices tend to be higher than hockey tickets. Football tickets also show more variation in price. (IQR: Football $84, Hockey $67).

3.92 a. Since the distributions are not symmetric, compare the medians.

b. NBA tickets are typically more expensive than MLB tickets ($73.5 compared to $55.5). There is more variation in NBA ticket prices (comparing IQRs, $41 compared to $9.50). Both sports have potential high outliers in the data, since both maximum values are beyond the upper fence for each data set.

3.93 a. 80 + 2(4) = 88

b. 80 – 1.5(4) = 74

3.94 38 – 2 = 36 inches

3.95 The *z*-score for the SAT of 750 is (750 – 500)/ 100 = 2.5, and the *z*-score for the ACT of 28 is (28-21)/5 = 1.4. The score of 750 is more unusual because its *z*-score is farther from 0.

3.96 The boy's *z*-score is (43 – 38)/2 = 2.5 and the girl's *z*-score is (57 – 54.5)/2.5 = 1, so the boy is more unusually tall for his age and gender.

3.97 a. The distribution is right-skewed.

b. Because the data are right skewed, the mean would be greater than the median.

c. The majority of ticket prices would be less than the mean price.

3.98 a. The shape is right-skewed, the median is 20, and the interquartile range is 35 - 19 = 16. There is an outlier at 66.

b. The mean is 27.3, and the median is 20. The mean is much larger because of the outlier, 66. The mean and the median should be marked on the histogram of the data.

3.99 Answers will vary. "Who ran faster, grade 11 or grade 12 boys?" "Which group had the most consistent times, grade 11 or grade 12 boys?"

3.100 Answers vary. Possible answers include "Players in which position typically weigh more: shooting guards or centers?" "Which position shows more variability in player weights?"

3.101 Answers vary. Possible answers include "Which buildings are typically taller: those made of concrete or those made of steel?" "Is there more variability in the heights of building constructed in the 2000s or before 2000?"

3.102 Answers vary. Possible answers include "Do cereals on the top shelf typically contain more calories than those on the bottom shelf?" "Is there more variability in sugar content among cold cereals or hot cereals?"

Chapter 4: Regression Analysis: Exploring Associations between Variables

Section 4.1: Visualizing Variability with a Scatterplot

4.1 The critical reading score is a somewhat better predictor because the vertical spread is less, suggesting a more accurate prediction of GPA.

4.2 The relationship is stronger for years employed by the company because the vertical spread is less, suggesting a stronger relationship.

4.3 The trend appears roughly linear and positive up to about age 24, and then it starts to curve.

4.4 The trend seems near zero. This means that age does not seem to be related to GPA for these students.

4.5 There is very little trend. It appears that number of credits acquired is not associated with GPA.

4.6 The trend tends to be positive but shows some curvature (nonlinear).

4.7 The trend is positive. Students with more sisters tended to have more brothers. This trend makes sense, because large families are likely to have a large number of sons and a large number of daughters.

4.8 The trend tends to be positive. The 2000-square-foot house with a price of about $2500 thousand appears to be an outlier, and the 5000-square-foot home is another potential outlier.

4.9 There is a slightly negative trend. The negative trend suggests that the more hours of work a student has, the fewer hours of TV the student tends to watch. The person who works 70 hours appears to be an outlier because that point is separated from the other points by a large amount.

4.10 Very little trend. The number of hours of work does not seem to be related to the number of hours of sleep for these students.

4.11 There is a slight negative trend that suggests that older adults tend to sleep a bit less than younger adults. Some may say there is no trend.

4.12 Positive: taller women tend to weigh more.

Section 4.2: Measuring Strength of Association with Correlation

4.13 a. You should not find the correlation because the trend is not linear.

b. You may find the correlation because the trend is linear.

4.14 Linear regression should not be used because the trend is not linear. The highest fertility rate occurs as about 29 years of age.

4.15 The correlation coefficient is positive since the graph shows an upward trend.

4.16 The correlation coefficient is negative since the graph shows a downward trend.

4.17 Since it has a stronger positive association, 0.767 goes with graph A.
Since it has a weaker positive association, 0.299 goes with graph B.
Since it is the only graph with a negative association, –0.980 goes with graph C.

4.18 Since it is the only graph with a negative association, –0.903 goes with graph B.
Since it has a weaker positive association, 0.374 goes with graph A.
Since it has a stronger positive association, 0.777 goes with graph C.

4.19 Since it has a stronger positive association, 0.87 goes with graph A.
Since it is the only graph with a negative association, –0.47 goes with graph B.
Since it has a weaker positive association, 0.67 goes with graph C.

4.20 Since it has a stronger positive association, 0.98 goes with graph A.
Since it is the only graph with a negative association, –0.51 goes with graph B.
Since it has a weaker positive association, 0.18 goes with graph C.

4.21 R (correlation coefficient) = 0.68518783

a. $r = 0.69$

b. $r = 0.69$; the correlation coefficient stays the same.

c. $r = 0.69$. Adding a constant to all y-values does not change the value of r.

d. $r = 0.69$; the correlation coefficient stays the same.

4.22 a. $r = 0.86$

b. $r = 0.86$. Multiplying by a constant does not change the value of r.

c. $r = 0.86$. Adding a constant does not change the value of r because the strength of the association is not affected.

4.23 The correlation would not change. The correlation does not depend on which variable is the predictor and which is the response.

4.24 The new correlation would also be 0.91. The correlation remains unchanged if you change units or multiply the numbers for a variable by a positive constant. (Correlations are determined by z-scores, and the z-scores would not change because any change in a number is accompanied by a similar change in the mean.)

4.25 The correlation is 0.904. The professors that have high overall quality scores tend to also have high easiness scores.

4.26 The correlation is 0.891. This means that people with a large number of female cousins tend also to have a large number of male cousins, and people with a small number of female cousins tend also to have a small number of male cousins.

4.27 Higher gym usage is associated with higher GPA.

4.28 We expect the correlation is positive. Higher levels of education are associated with longer life expectancy.

Section 4.3: Modeling Linear Trends

4.29 a. The independent variable is median starting salary, and the dependent variable is median mid-career salary.

b. Salary distributions are usually skewed. Medians are therefore a more meaningful measure of center.

c. Between $110,000 and $120,000.

d. $\text{Mid-Career} = -7699 + 1.989\ \text{Starting} = -7699 + 1.989(60,000) = \$111,641$

e. Answers will vary. The number of hours worked per week, the amount of additional education required, gender, and the type of career are all factors that might influence mid-career salary.

4.30 a. The mother's height is the independent variable, and the daughter's height is the dependent variable.

b. About 62 or 63 inches.

c. $\text{Daughter} = 29.92 + 0.5417\ \text{Mother} = 29.92 + 0.5417(60) = 62.4$ inches

d. For each additional inch in the mother's height, the average daughter's height increases by about 0.54 inch.

e. Answers will vary. The father's height and the health of the daughter are two of many factors.

4.31 a. The median pay for women is about \$690 when pay for men is \$850.

b. predicted women = –62.69 + 0.887(850) = 691.26.

c. The slope is 0.887. Each additional dollar in men's' pay is associated with an average increase of \$0.887 in the median womens' pay.

d. The *y*-intercept is -62.69. It is not appropriate to interpret it in this context because the median income for men (*x*) cannot be zero.

4.32 a. The selling price for a 2000-square-foot home would be about \$1,400,000.

b. Predicted selling price = 756,789 + 327.29(2000) = \$1,411,369.

c. The slope of the regression equation is 327.29. For every additional square foot in size, selling price tends to increase by \$327.29.

d. The *y*-intercept is 756,789. It is not appropriate to interpret the *y*-intercept in this context because it is not possible to sell a home that is 0 square feet.

4.33 a. Predicted Armspan = 16.8 + 2.25 Height

b. $b = r\frac{s_y}{s_x} = 0.948\left(\frac{8.10}{3.41}\right) = 2.25$

c. $a = \overline{y} - b\overline{x} = 159.86 - 2.25\left(63.59\right) = 16.8$

d. Predicted Armspan = 16.8 + 2.25 Height = 16.8 + 2.25(64) 160.8, or about 161 cm

4.34 a. Predicted Foot = 5.67 + 0.998 Hand

b. $b = r\frac{s_y}{s_x} = 0.948\left(\frac{1.230}{1.168}\right) = 0.998$

c. $a = \overline{y} - b\overline{x} = 23.318 - 0.998\left(17.682\right) = 5.67$

d. Predicted Foot = 5.67 + 0.998 Hand = 5.67 + 0.998(18) = 23.634, or about 23.6 cm

4.35 a. Predicted Armspan = 6.24 + 2.515 Height (Rounding may vary.)

b. Minitab: slope = 2.51, intercept = 6.2

StatCrunch: slope = 2.514674, intercept = 6.2408333

Excel: slope = 2.514674, intercept = 6.240833

TI-84: slope = 2.514673913, intercept = 6.240833333

4.36 a. Predicted FootL = 15.81 + 0.5627 HandL (Rounding may vary.)

b. Minitab: slope = 0.563, intercept = 15.8

StatCrunch: slope = 0.5626551, intercept = 15.807631

Excel: slope = 0.562655, intercept = 15.80763

TI-84: slope = 0.5626550868, intercept = 15.80763027

4.37 The association for the women is stronger because the correlation (*r*-value) is closer to 1.

4.38 a. The men's line is higher, which means that the men tend to weigh more than the women at all ages shown.

b. The men's line is steeper. This tells us that older men tend to differ more in weight from younger men than older women tend to differ from younger women.

4.39 a. Based on the scatterplot there is not a strong association between these two variables.

b. The numerical value of the correlation would be close to zero because there is not an association between these variables.

c. Since there is not an association between these variables, we cannot use singles percentage to predict doubles percentage.

4.40 The absolute value of r is 1 because the points perfectly follow a linear trend. The sign is negative because the trend goes downward to the right. The points follow a downward straight-line pattern, so $r = -1$.

4.41 Explanations will vary.

	x	***y***
a.	odometer reading	price
b.	household size	water bill
c.	time spent in gym	weight

4.42 Explanations will vary.

	x	***y***
a.	stress level	blood pressure
b.	length of commute	mo. gas expense
c.	height	maximum speed

4.43 a. The higher the percentage of smoke-free homes in a state, the lower the percentage of high school students who smoke tends to be.

b. Predicted Pct. Smokers = 56.32 – 0.464 (Pct. Smoke Free) = 56.32 – 0.464(70) = 23.84, or about 24%

4.44 a. The higher the percentage of adults who smoke, the higher the percentage of smoking high school students tends to be.

b. Predicted Pct. Smokers = –0.838 + 1.124 (Pct. Adult Smoke) = –0.838 + 1.124(25) = 27.262, or about 27%

4.45 a. As driver age increases, insurance prices decrease but then begin to increase again at around 65 years of age. Younger drivers and older drivers tend to have more accidents so they are charged more for insurance.

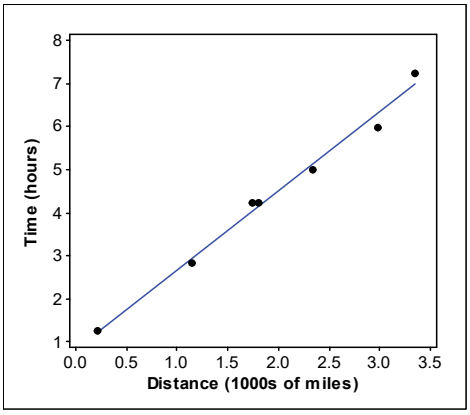
b. It would not be appropriate to do a linear regression analysis on these data because the data do not follow a linear trend.

4.46 a. Life insurance is very inexpensive for young people, but premiums begin to increase faster and faster at around 60 years of age.

b. It would not be appropriate to do a linear regression analysis on these data because the data do not follow a linear trend.

4.47 The answers follow the guidance on page 209.

1:



2: The linear model is appropriate. The points suggest a straight line.

3: Predicted Time = 0.8394 + 1.838 Distance

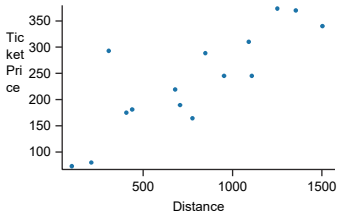
```
Regression Analysis: Time (hours) versus Distance (1000s of miles)
The regression equation is
Time (hours) = 0.8394 + 1.838 Distance (1000s of miles)
```

4: Each additional thousand miles takes, on average, about 1.84 more hours (or 110 minutes) to arrive.

5: The additional time shown by the intercept might be due to the time it takes for the plane to taxi to the runway, delays, the slower initial speed, and similar delays in the landing as well. The time for this appears to be about 0.84 hours (or 50 minutes).

6: Predicted Time = 0.8394 + 1.838(3) = 0.8394 + 5.514 = 6.35 hours. The predicted time for a flight from Boston to Seattle is about 6.35 hours.

4.48 a. From the graph it looks like there is a strong linear association between the variables.



b. $r = 0.882$; ticket price = 94.52 + 0.18 distance

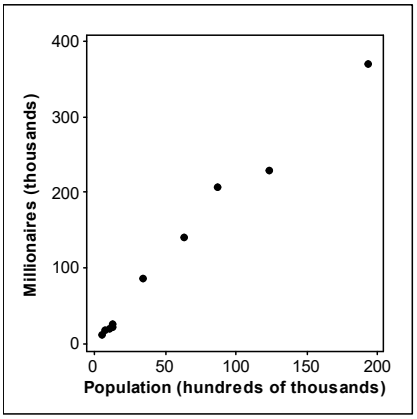
c. Each additional mile is associated with an average increase of 0.18 in ticket price.

d. It would be inappropriate to interpret the y-intercept (94.52) because it would not be possible to have a trip with a distance of 0 miles.

e. Ticket price = 94.52 + 0.18(572) = $197.48

4.49 a. The slope and correlation will be positive: The more population, the more millionaires there tend to be.

b.



c. $r = 0.992$

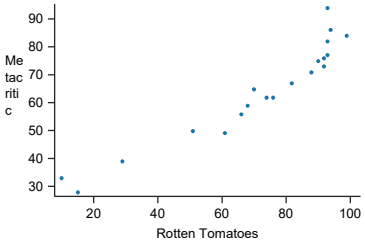
```
Correlations: Millionaires (thousands), Population (hundreds of thousands)
Pearson correlation of Millionaires (thousands) and Population (hundreds of
     Thousands) = 0.992
```

d. The slope is 1.9. For each additional hundred thousand in the population, there is an additional 1.9 thousand millionaires.

```
Regression Analysis: Millionaires (thousands) versus Population (hundreds of thousands)
The regression equation is
Millionaires (thousands) = 6.30 + 1.92 Population (hundreds of thousands)
```

e. Do not focus on the intercept, because it does not make sense to look for millionaires in states with no people.

4.50 a. From the graph it looks like there is a strong linear association between the variables.

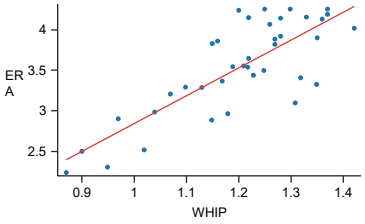


b. $r = 0.95$. Metacritic = 18.83 + 0.63 Rotten Tomatoes

c. Each additional Rotten Tomatoes point is associated with an average increase of 0.63 Metacritic points.

d. It would be inappropriate to interpret the y-intercept (18.83) because there are no movies with a 0 Rotten Tomatoes rating.

4.51 a.

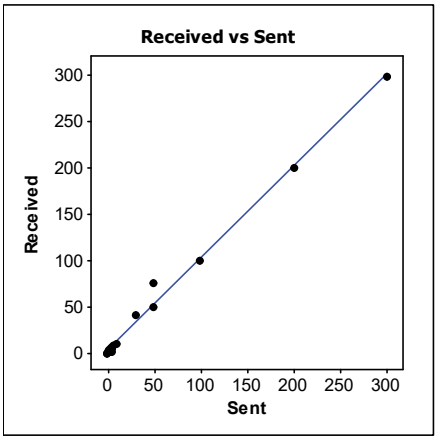


b. ERA = -0.578 + 3.436 WHIP

c. Each additional point in WHIP rating is associated with an average increase of 3.436 in ERA.

d. It would be inappropriate to interpret the y-intercept because there are no pitchers with a 0 WHIP rating.

4.52 a.



b. Predicted Texts Sent = –1.63 + 0.993(Texts Received); See scatterplot in part a for graph of line of best fit.

Regression Analysis: Received versus Sent

```
The regression equation is
Received = 1.87 + 1.001 Sent
```

c. For each additional text sent, there is an average of 1 more text received.

d. People who do not send texts still receive, on average, 2 texts.

Section 4.4: Evaluating the Linear Model

4.53 a. Influential points are outliers in the data that can have large effects on the regression line. When influential points are present in the data, do the regression and correlation with and without these points and comment on the difference.

b. The coefficient of determination is the square of the correlation coefficient; it measures the percentage of variation in the *y*-variable that is explained by the regression line.

c. Extrapolation means using the regression equation to make predictions beyond the range of the data. Extrapolation should not be used.

4.54 a. We expect monthly personal savings to decrease because the correlation is negative.

b. No. We cannot conclude a causal relationship because the data were not from a controlled, randomized experiment.

4.55 Older children have larger shoes and have studied math longer. Large shoes do not cause higher grades. Both are affected by age.

4.56 Growth rates slow as people get older. You should not extrapolate. (That is, you should not predict outside the range of the data.)

4.57 $(0.67)^2 = 0.4489$, so the coefficient of determination is about 45%. Therefore, 45% of the variation in weight can be explained by the regression line.

4.58 $(-0.70)^2 = 0.49$, so –0.70 has a coefficient of determination of 49%, and $(0.50)^2 = 0.25$, so 0.50 has a coefficient of determination of 25%. Thus the correlation of –0.70 is stronger than a correlation of 0.50.

4.59 Part of the poor historical performance could be due to chance, and if so, regression toward the mean predicts that stocks turning in a lower than average performance should tend to perform closer to the mean in the future. In other words, they might tend to increase.

4.60 Their new blood pressures will tend to be closer to normal, on average. Part of the high reading might be due to chance, and regression toward the mean predicts that a repeated measurement will be closer to the typical value. In other words, the new blood pressures might tend to be lower.

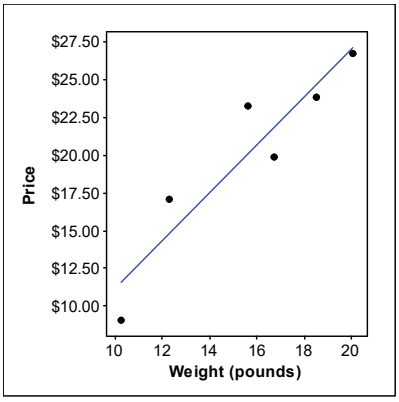
4.61 a. The slope of –2099 means that the salary is $2099 less for each year later that the person was hired, or $2099 more for each year earlier.

b. The intercept ($4,255,000) would be the salary for a person who started in the year 0, which is inappropriate (and ridiculous).

4.62 a. The slope of 0.95 means that for each additional city mpg, the highway value goes up by about 0.95 mpg, or nearly 1 mpg.

b. The intercept is 7.79. If a car were to get 0 mpg in the city, it would get 7.79 mpg on the highway. That conclusion is inappropriate because no cars get 0 mpg in the city.

4.63 a.



b. $r = 0.933$; A positive correlation suggests that larger turkeys tend to have a higher prices.

Correlations: Weight (pounds), Price
Pearson correlation of Weight (pounds) and Price = 0.933

c. Predicted Price = –4.49 + 1.57 Weight

Regression Analysis: Price versus Weight (pounds)
The regression equation is
Price = - 4.49 + 1.57 Weight (pounds)

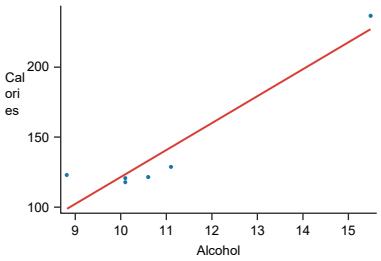
d. The slope of 1.57 means that for each additional pound, the price goes up by $1.57. The interpretation of the intercept is inappropriate, because it is not possible to have a turkey that weighs 0 pounds.

e. $r = –0.375$; Predicted Price = 26.87 – 0.533 Weight. The negative correlation and slope imply that the bigger the turkey, the less it costs! The 30-pound free turkey was an influential point, which really changed the results.

Correlations: Weight (pounds), Price
Pearson correlation of Weight (pounds) and Price = -0.375
Regression Analysis: Price versus Weight (pounds)
The regression equation is
Price = 26.87 - 0.553 Weight (pounds)

f. $r^2 = 0.87.87\%$ of the variation in turkey price is explained by weight.

4.64 a. It looks like there is a strong linear relationship between the variables.



b. $r = 0.95$; The sign is positive. As % alcohol increases, calories also increases.

c. Calories = -68.20 + 19.02% alcohol;

d. $r^2 = 0.91$; 91% of the variation in calories can be explained by the % alcohol.

e. $r = -0.35$; slope = -6.03. This influential point has made the correlation weaker and negative.

4.65 a. Positive.

b. Slope = 0.327; each additional $1 in teacher pay is associated with an increase in per pupil spending of 0.327.

c. The y-intercept is -5922; it would not be appropriate to interpret the y-intercept because there is no state an average teacher salary of $0.

d. –5922 + 0.327(60000) = $13,698.

4.66 a. The correlation is near 0, as is the slope.

b. The average teacher pay is not associated with the percentage graduating from high school so the regression equation should be used to make the prediction.

4.67 a. Since the y-values decrease as the x-values increase, there is a negative correlation.

b. For each additional hour of work, the score tended to go down by 0.48 point.

c. A student who did not work would expect to get about 87 on average.

4.68 a. Since the *y*-values increase as the *x*-values increase, the data suggest a positive trend. This means that there tends to be more trash if there are more people living in the house.

b. $r = \sqrt{0.76} = 0.877$

c. The slope is 11.30. For each additional person, on the average there are an additional 11.3 pounds of trash.

d. This suggests that with 0 people there should be 2.34 pounds of trash. But we should not draw that conclusion, because it is not meaningful to think of a house producing trash all by itself (without people present).

4.69 There is a stronger association between home runs and strikeouts ($r = 0.64$ compared to $r = -0.09$).

4.70 There is a stronger association between 3-point percentage and free-throw percentage ($r = 0.57$ compared to $r = -0.05$).

4.71 a.

```
Dependent Variable: 4th Grade Math
Independent Variable: 4th Grade Reading
4th Grade Math = 33.027886 + 0.6961738 4th Grade Reading
Sample size: 20
R (correlation coefficient) = 0.84877382
```

$r = 0.85$;
Predicted Math Score = 33.03 + 0.70 (Reading score)
Predicted Math Score = 33.03 + 0.70(70) = 82.03. The predicted math score is an 82.

b.

```
Dependent Variable: 4th Grade Reading
Independent Variable: 4th Grade Math
4th Grade Reading = -15.334137 + 1.0348235 4th Grade Math
Sample size: 20
R (correlation coefficient) = 0.84877382
```

$r = 0.85$;
Predicted Reading Score = –15.33 + 1.03(70) = 57. The predicted reading score is a 57.

c. Changing the choice of the dependent and independent variables does not change *r* but does change the regression equation.

4.72 a.

```
Dependent Variable: SAT Critical Reading Score
Independent Variable: SAT Math Score
SAT Critical Reading Score = -12.314155 + 1.0236995 SAT Math Score
Sample size: 29
R (correlation coefficient) = 0.98027436
```

$r = .98$; Math = 32.37 + 0.94 Reading; 32.37 + 0.94(600) = 596; Predicted critical reading score is 596.
Predicted Math Score = 32.37 + 0.94 (Reading score)
Predicted Math Score = 32.37 + 0.94(600) = 596 The predicted math score is an 596.

b.

```
Dependent Variable: SAT Math Score
Independent Variable: SAT Critical Reading Score
SAT Math Score = 32.367205 + 0.93869128 SAT Critical Reading Score
Sample size: 29
R (correlation coefficient) = 0.98027436
```

Predicted Reading Score = –12.31 + 1.02 (Math score)
Predicted Reading Score = –12.31 + 1.02(600) = 600. The predicted reading score is an 600.

c. Changing the choice of the dependent and independent variables does not change *r* but does change the regression equation.

4.73 The answers follow the guided steps.

1: a. $b = r \cdot \dfrac{s_{\text{final}}}{s_{\text{midterm}}} = 0.7 \cdot \dfrac{10}{10} = 0.7$

b. $a = \bar{y} - b\bar{x} = 75 - 0.7(75) = 75 - 52.5 = 22.5$

c. Predicted Final = 22.5 + 0.7 Midterm

2: Predicted Final = 22.5 + 0.7 Midterm = 22.5 + 0.7(95) = 22.5 + 66.5 = 89

3: The score of 89 is lower than 95 because of regression toward the mean.

4.74 a. Predicted Final = 18 + 0.75 Midterm; $b = r \cdot \frac{s_y}{s_x} = 0.75 \cdot \frac{8}{8} = 0.75$;

$a = \bar{y} - b\bar{x} = 72 - 0.75(72) = 72 - 54 = 18$

b. Predicted Final = 18 + 0.75(55) = 59.25

c. It is higher than 55 because of regression toward the mean.

d. It should be lower than 100 because of regression toward the mean.

Chapter Review Exercises

4.75 a. $r = 0.941$; Predicted Weight = –245 + 5.80 Height

Regression Analysis: Weight (pounds) versus Height (inches)
The regression equation is
Weight (pounds) = - 245 + 5.80 Height (inches)
Correlations: Height (inches), Weight (pounds)
Pearson correlation of Height (inches) and Weight (pounds) = 0.941

b.

Height (cm)	Weight (kg)
$60(2.54) = 152.40$	$\frac{105}{2.205} = 47.6191$
$66(2.54) = 167.64$	$\frac{140}{2.205} = 63.4921$
$72(2.54) = 182.88$	$\frac{185}{2.205} = 83.9002$
$70(2.54) = 177.80$	$\frac{145}{2.205} = 65.7596$
$63(2.54) = 160.02$	$\frac{120}{2.205} = 54.4218$

c. The correlation between height and weight is 0.941. It does not matter whether you use inches and pounds or centimeters and kilograms. A change of units does not affect the correlation because it has no units.

Correlations: Height (cm), Weight (kg)
Pearson correlation of Height (cm) and Weight (kg) = 0.941

d. The equations are different, using different units will result in different results.

Predicted Weight Pounds = –245 + 5.80 Height (inches)

Predicted Weight Kilograms = –111 + 1.03 Height (centimeters)

Regression Analysis: Weight (lb) versus Height (in)
The regression equation is
Weight (pounds) = - 245 + 5.80 Height (inches)
Regression Analysis: Weight (kg) versus Height (cm)
The regression equation is
Weight (kg) = - 111 + 1.03 Height (cm)

4.76 a. Predicted Weight = –169.5 + 5.174 Height (inches): The slope of 5.174 means that for each additional inch in height, a man weighs about 5 more pounds on average. It is not appropriate to interpret the intercept, because no one has a height near 0.

Regression Analysis: Weight (pounds) versus Height (inches)
The regression equation is
Weight (pounds) = - 169.5 + 5.174 Height (inches)

b. $r = 0.665$

Correlations: Height (inches), Weight (pounds)
Pearson correlation of Height (inches) and Weight (pounds) = 0.665

c. $(0.665)^2 = 44\%$ of the variation in weight can be explained by the regression line.

d. The correlation is the same.

Correlations: Height (centimeters), Weight (pounds)
Pearson correlation of Height (centimeters) and Weight (pounds) = 0.665

e. The new equation is Predicted Weight = –169.5 + 2.04 Height (cm): The slope of 2.04 means that for each additional centimeter of height, a man weighs about 2 more pounds on average.

Regression Analysis: Weight (pounds) versus Height (centimeters)
The regression equation is
Weight (pounds) = - 169.5 + 2.0369 Height (centimeters)

f. Changing units changes the equation but does not change the correlation. Note that the intercept is the same in both equations since they both predict the weight of a man with no height, but it has the same value only because the units of weight are the same in both equations. (Compare with Exercise 4.75 where the units of both variables are changed.)

4.77 a.

Dependent Variable: Calories
Independent Variable: Carbs (grams)
Calories = 3.2622227 + 10.806543 Carbs (grams)
Sample size: 48
R (correlation coefficient) = 0.85821218

$r = 0.86$; Predicted Calories = 3.26 + 10.81(Carbs); Slope = 10.81; Each additional gram of carbohydrates is associated with an increase of 10.81 calories
Predicted Calories = 3.26 + 10.81(55) = 597.81 calories

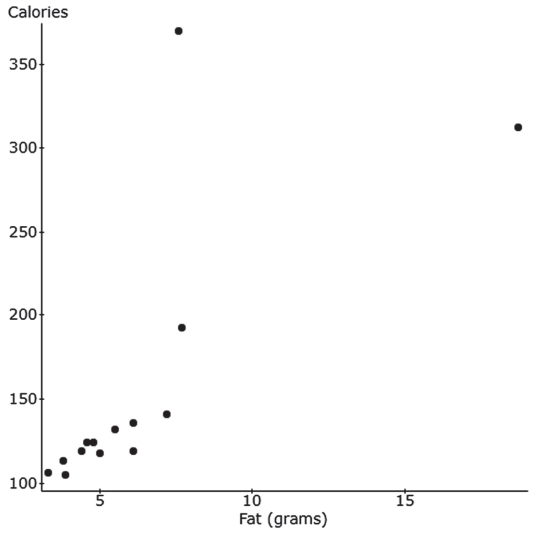
b.

Dependent Variable: Calories
Independent Variable: Sugars (grams)
Calories = 198.73601 + 32.745312 Sugars (grams)
Sample size: 48
R (correlation coefficient) = 0.79253604

$r = 0.79$; Predicted Calories = 198.74 + 32.75(sugars)
Predicted Calories = 198.74 + 32.75(10) = 526.24 calories

c. While both are fairly good, the number of carbs is a better predictor ($r = 0.86$ compared to $r = 0.79$).

4.78 a. Yes, there seems to be a linear trend.



b.

Dependent Variable: Calories
Independent Variable: Fat (grams)
Calories = 61.996495 + 15.153879 Fat (grams)
Sample size: 14
R (correlation coefficient) = 0.713427
R-sq = 0.50897808

$r = 0.713$

c. The slope of the regression equation is 15.15. Each additional gram of fat is associated with an increase of about 15 calories.

d. The y-intercept of the regression equation is 62.0. It represents the number of calories in a granola bar containing 0 grams of fat. Since no bars in our sample contained 0 grams of fat, it is not appropriate to interpret the y-intercept in this problem.

e. The coefficient of determination is 50.9%. About 51% of the variation in calories is explained by fat content.

f. Predicted Calories = 62.0 + 15.15(7) = 168 calories
A granola bar containing 7 grams of fat will have about 168 calories.

g. Since there are no granola bars in the data set with fat content near 25 grams, it would be inappropriate to use the regression equation to predict the fat content of a granola bar with 25 grams of fat.

h. Remove the bar with 7.6 grams of fat and 370 calories from the data set. Recalculating r, $r = 0.98$, indicating a stronger linear correlation. The slope and the y-intercept of the regression equation change after removing this data point. The new regression equation is calories = 55.8 + 13.8 fat.

4.79 a. There were no women taller than 69 inches, so the line should stop at 69 inches to avoid extrapolating.

b. Men who are the same height as women wear shoes that are, on average, larger sizes.

c. The mean increase in shoe size based on height is the same for men and women.

4.80 a. The older people tend to get somewhat less sleep for both genders.

b. The average decrease in sleep for each additional year of age is the same for men and women.

c. The relationship between sleep and age is nearly the same for men and women.

d. There were no men above about 47 years of age. Extending the line past 47 years would be extrapolating.

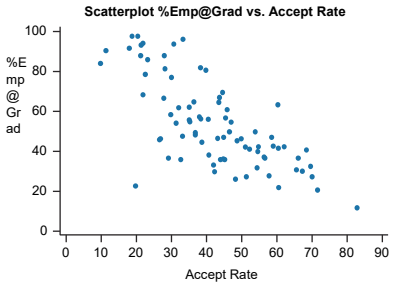
4.81 Among those who exercise, the effect of age on weight is less. An additional year of age does not lead to as great an increase in the average weight for exercisers as it does for non-exercisers.

4.82 a. The correlation is positive.

b. The correlation is not positive for every grade. It might be near zero or slightly negative for some grades.

c. Test scores are confounded with age. Older students are taller and also score higher. Thus the principal is not correct.

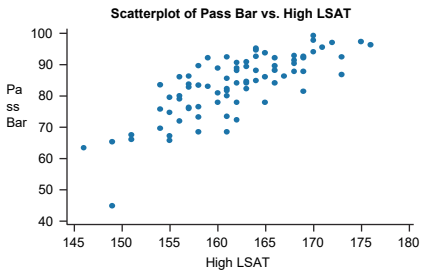
4.83 a.



There seems to be a linear trend. The less-selective law schools also tend to have a lower employment rate at graduation.

b. I. % Employed = 97.59 - 1.03 Acceptance Rate
II. Each additional percentage point in acceptance rate is associated with an average decrease of 1.03 percentage points in the rate of employment at graduation.
III. It would be inappropriate to interpret the *y*-intercept because there is no school with an acceptance rate of 0%.
IV. $r^2 = 52\%$. 52% of the variation in employment rate at graduation can be explained by the acceptance rate.
V. The regression equation predicts an employment rate of about 32.6% for a school with an acceptance rate of 50%.

4.84 a.



There seems to be a linear trend to the data. Schools with high average LSAT scores tend to have higher percentages of students passing the bar exam.

b. I. Pass Bar = -105.8 + 1.117 High LSAT.
II. Each additional point on the LSAT is associated with a mean increase of 1.1 in the bar-pass rate.
III. It would be inappropriate to interpret the *y*-intercept since no school admits students with an LSAT score of 0.
IV. $r^2 = 55.6\%$. About 55.6% of the variation in bar-pass rates can be explained by LSAT score.
V. The regression equation predicts a bar-pass rate of 61.75 for schools with an LSAT score of 150.

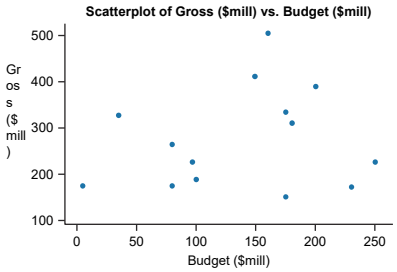
4.85 A comparison of scatterplots shows a stronger association between calories and fat compared with calories and carbohydrates. The correlation coefficient for fat and calories is $r = 0.82$ while the correlation coefficient for carbohydrates and calories is 0.39. Fat is a better predictor of the number of calories in these snack foods.

4.86 Education is a bit more closely correlated with salary than is years of employment. Yes, education pays off, but there is not much difference between the correlations.

Correlations: Salary, YrsEm, Educ

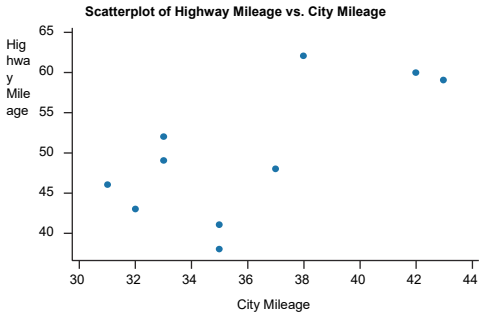
	Salary	YrsEm
YrsEm	0.765	
Educ	0.777	0.607

4.87



It is not appropriate to fit a linear regression model, because the trend is not linear. However, we can see that some big-budget films didn't do as well compared to some lower budget films. The one point at the top, Wonder Women, had the highest gross but did not have the highest budget.

4.88



Regression equation: Highway Mileage = –0.35 + 1.40 City Mileage. A car that gets 40 miles per gallon in the city is predicted to get –0.35 + 1.40(40) = 55.7 miles per gallon on the highway.

It would not be appropriate to use the regression equation to predict the highway mileage for a car with a city mileage of 60 miles pe gallon because the data do not include values greater than 43, and so we would be extrapolating.

4.89 a. The positive trend shows that the more stories there are, the taller the building tends to be.

b. Predicted Height = 115.4 + 12.85(100) = 115.4 + 1285 = 1400.4, or about 1400 feet

c. The slope of 12.85 means that buildings with one additional story tend to have an average of 12.35 feet of additional height.

d. Because there are no building with 0 stories, the interpretation of the intercept is not appropriate.

e. About 71% of the variation in height can be explained by the regression, and about 29% is not explained.

4.90 a. The trend is negative. An increase in high school graduation rates is associated with a decrease in poverty rate.

b. Each additional 1% in high school graduation rate is associated with a decrease of 0.70% in poverty rate.

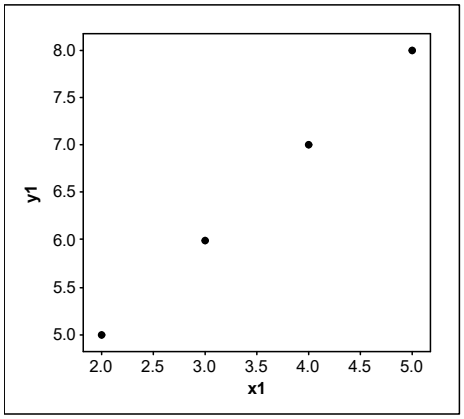
c. About 61% of the variation in poverty rates can be explained by graduation rates.

d. High school. Compare values of the correlation coefficients: High school and poverty: r = -0.78; BA and poverty: r = -0.40; advanced degree and poverty: r = -0.15.

4.91 Answers will vary.

$x1$	$y1$
2	5
3	6
4	7
5	8

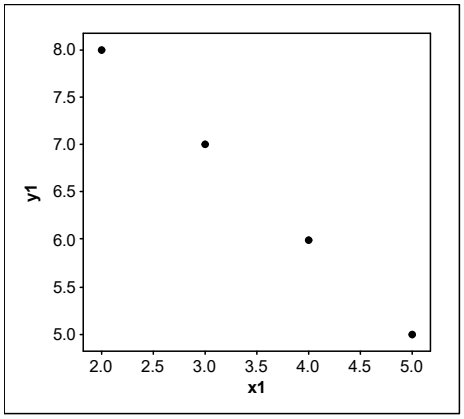
Correlations: x1, y1
Pearson correlation of x1 and y1 = 1.000



4.92 Answers will vary.

$x1$	$y1$
2	8
3	7
4	6
5	5

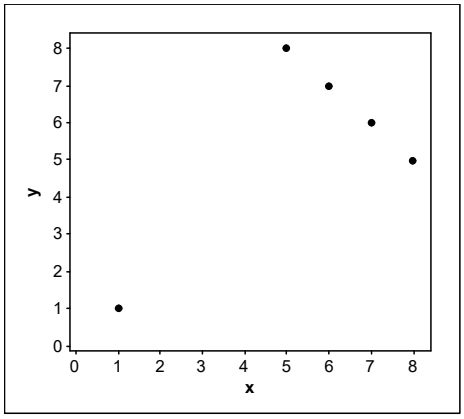
Correlations: x1, y1
Pearson correlation of x1 and y1 = -1.000



4.93 Answers will vary.

Dataset 1		Dataset 2	
*x*1	*y*1	*x*2	*y*2
5	8	5	8
6	7	6	7
7	6	7	6
8	5	8	5
		1	1

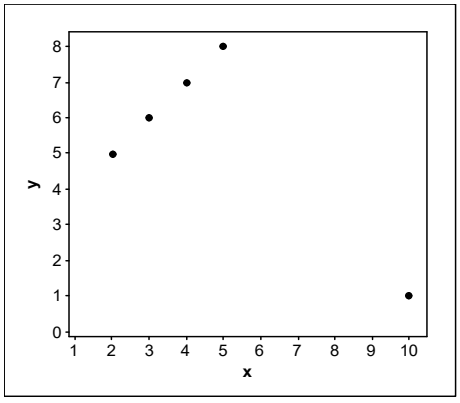
Correlations: x1, y1
Pearson correlation of x1 and y1 = -1.000
Correlations: x2, y2
Pearson correlation of x2 and y2 = 0.658



4.94 Answers will vary.

Dataset 1		Dataset 2	
*x*1	*y*1	*x*2	*y*2
2	5	2	5
3	6	3	6
4	7	4	7
5	8	5	8
		10	1

Correlations: x1, y1
Pearson correlation of x1 and y1 = 1.000
Correlations: x2, y2
Pearson correlation of x2 and y2 = -0.701



4.95 The trend is positive. In general, if one twin has a higher-than-average level of education, so does the other twin. The point that shows one twin with 1 year of education and the other twin with 12 years is an outlier. (Another point showing one twin with 15 years and the other with 8 years is a bit unusual, as well.)

4.96 Answers may vary. The trend is very weak and potentially positive. The trend is weak enough that it is difficult to tell whether it is linear or not.

4.97 There appears to be a positive trend. It appears that the number of hours of homework tends to increase slightly with enrollment in more units.

4.98 The number of hours of exercise and the number of hours of homework do not appear to be related. The correlation should be near zero.

4.99 Linear regression is not appropriate because the trend is not linear, it is curved.

4.100 Finding the correlation is not appropriate because the trend is not linear; it is curved.

4.101 The cholesterol going down might be partly caused by regression toward the mean.

4.102 Regression toward the mean might contribute to raising the scores of the students who scored low on the first test.

Chapter 5: Modeling Variation with Probability

Section 5.1: What Is Randomness?

5.1 Answers may vary with different random numbers.

a. 5 5 1 8 5 7 4 8 3 4

b. T T H T T T H T H H

c. 4/10 = 40% were heads.

5.2 a. 2 6 4 2 7 4 0 6 5 0 7 0 2 5 1 8 4 4 1 3

C C C C P C C C P C P C C P P C C C P P

b. 13/20 = 65% were assigned to the computer group.

c. 0, 1, 2, 3, 4 could have been assigned to the computer group and 5, 6, 7, 8, 9 could have been assigned to the pen and paper group.

d. The digits 0, 1, 2, 3, 4 could have assigned participants to Computer, and the digits 5, 6, 7, 8, 9 could have assigned participants to Paper.

5.3 This is a theoretical probability, because it is not based on an experiment.

5.4 This is an empirical probability, because it is based on an experiment.

5.5 This is an empirical probability, because it is based on an experiment.

5.6 This is a theoretical probability, because it is not based on an experiment.

Section 5.2: Finding Theoretical Probabilities

5.7 a. The seven equally likely outcomes are Town, Wu, Hein, Lee, Marland, Penner, and Holmes.

b. The probability that the new patient will be assigned to a female doctor is $4/7$, or about 57.1%.

c. The probability that new patient will be assigned to a male doctor is $3/7$, or about 42.9%.

d. Yes, the event "male doctor" is equivalent to the event "not a female doctor".

5.8 a. The eight equally likely outcomes are Nagle, Crouse, Warren, Tejada, Tran, Cochran, Perry, Rivas.

b. The probability of a student being assigned to an experienced teacher is $5/8$, or 62.5%.

c. The complement of event in part (b) is Cochran, Perry, and Rivas: These are the inexperienced teachers. The probability is $3/8 = 37.5\%$.

5.9 a. 0.26 can be the probability of an event.

b. -0.26 cannot be a probability because it is negative.

c. 2.6 cannot be the probability of an event since 2.6 is greater than 1.

d. 2.6% can be the probability of an event.

e. 26 cannot be the probability of an event since 26 is greater than 1.

5.10 a, b, and d, can be the probability of an event.

c. 9.9 cannot be a probability because it greater than 1.

e. -0.90 cannot be a probability because it is negative.

5.11 a. $P(\text{a heart}) = 13/52$, or $1/4$ or 25%