Appendix A: Answer Keys
Introduction to Data Analysis (2nd Edition)

Chapter 1 Answer Keys

Section 1A Answers

1. Bear Data

Age: Quantitative (units: months)
Month Bear Measured: Categorical
Gender: Categorical
Head Length: Quantitative (units: inches)
Head Width: Quantitative (units: inches)
Neck Circumference: Quantitative (units: inches)
Length: Quantitative (units: inches)
Chest: Quantitative (units: inches)
Weight: Quantitative (units: pounds)

2. Cereal Data

Name of Cereal: Categorical
Manufacturer: Categorical
Target: Categorical
Shelf Displayed: Categorical
Calories: Quantitative (units: number of calories per serving)
Carbs: Quantitative (units: grams per serving)
Fat: Quantitative (units: grams per serving)
Fiber: Quantitative (units: grams per serving)
Potassium: Quantitative (units: milligrams per serving)
Protein: Quantitative (units: grams per serving)
Sodium: Quantitative (units: milligrams per serving)
Sugar: Quantitative (units: grams per serving)
Vitamins: Quantitative (units: % of daily need per serving)
Consumer Report Magazine: Quantitative (units: Consumer Report Rating Points)
Serving Size: Quantitative (units: cups per serving)
Weight: Quantitative (units: ounces per serving)

3.

a) Milligrams of Aspirin: Quantitative
b) Types of Cars: Categorical
c) Smoke Marijuana or not: Categorical
d) Number of Bicycles: Quantitative
e) Types of Birds: Categorical
f) Grams of Gold: Quantitative
g) Types of Cardio Classes: Categorical
h) Number of Cardio Classes: Quantitative
i) City: Categorical
j) Money in Bank Accounts: Quantitative
k) Zip Codes: Categorical
l) Driver's License Numbers: Categorical
m) Number of Taxis: Quantitative

Section 1B Answers

1.

a) Population of Interest: All students at the college.
b) Method: Voluntary Response
c) Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

2.

a) Population of Interest: All students at the high school.
b) Method: Convenience
c) Will not represent the population very well. There is sampling bias, since the individuals were not chosen randomly.

3.

a) Population of Interest: All voters in Jamie's city.
b) Method: Simple Random Sample
c) Will represent the population well as long as there is no other types of bias present. No sampling bias.

4.

a) Population of Interest: All employees at the company.
b) Method: Census
c) Census is better than a random sample. Will represent the population very well as long as there is no other types of bias present. No sampling bias.

5.

a) Population of Interest: All people in Portland, Oregon.
b) Method: Convenience
c) Will not represent the population very well. There is sampling bias, since the individuals were not chosen randomly.

6.

a) Population of Interest: All people in Toronto.
b) Method: Simple Random Sample
c) Will represent the population well as long as there is no other types of bias present. No sampling bias.

7.

a) Population of Interest: All people that come to Hugo's library.
b) Method: Census
c) Census is better than a random sample. Will represent the population very well as long as there is no other ypes of bias present. No sampling bias.

8.

a) Population of Interest: All people that use smart phones.
b) Method: Voluntary Response
c) Will not represent the population very well. Sampling bias, since the individuals were not chosen randomly.

9.

a) Population of Interest: All students at that college.
b) Method: Simple Random Sample
c) Will represent the population well as long as there is no other types of bias present. No sampling bias.

Section 1C Answers

1.

a) Population: All people or objects to be studied. For example, all students at College of the Canyons.
b) Census: Collecting data from everyone in your population. For example, collecting data from all of the students at college of the canyons.
c) Sample: Collecting data from a subgroup of the population. For example, collecting data from fifty students at College of the Canyons.
d) Bias: When data does not reflect the population. For example, friends and family will not represent the population of all people in Los Angeles, CA.
e) Question Bias: Phrasing a question in order to force people to answer the way you want. For example, we want to collect data on smoking cigarettes, but give the person a lecture on how unhealthy cigarettes are before asking them.

f) Response Bias: When someone is likely to lie about the answer to a question. For example, asking people how much they weigh in pounds. They may not give you a truthful answer.
g) Sampling Bias: Not using randomization when collecting sample data. For example, collecting data from only your friends and family. This is not a random sample.
h) Deliberate Bias: Falsifying or changing your data or leaving out groups from your population of interest. For example, a person might remove all of the data from people that disagreed with their opinion.

i) Non-response Bias: When people are likely to not answer when asked to provide data. Randomly calling phone numbers to get data, but the person refuses to answer the phone.

2.

Population of interest: All people in the U.S.

Question Bias: The question was phrased to make people feel bad about answering no.
Response Bias: Vaccinations are a controversial issue and many people may feel scared to admit that they don't agree with vaccinations. There will be many people that lie.
Non-response Bias: There will be many people that randomly selected, but refuse to answer the question.

3.

Population of interest: All Americans.

Response Bias: Cocaine users would not feel comfortable answering the question honestly.
Non-response: Many people may be randomly selected, but will chose not to answer the question.

4.

Population of interest: All college students in Canada

Sampling Bias: The data was not collected randomly.
Deliberate Bias: Most of the colleges in Canada were left out since they only got data from a college near house.
Response Bias: Many people lie about their ages.
Non-response Bias: Many people will refuse to answer.

5.

Population of interest: All adults in Palmdale, CA.

Sampling Bias: The individuals were not selected randomly.
Deliberate Bias: Julie skipped streets that looked poor. These people are not being represented in the data.
Response Bias: People often lie about their income.
Non-response: Many people may not be home or refuse to answer the door.

6.

Population of interest: All students at the college

Response Bias: Many people may lie about their mental health status.
Non-response Bias: Many people may refuse to participate.
Question Bias: The question seems to make people feel bad about answering no.

7.

Population of interest: All pills made by the company.

Deliberate Bias: They deleted data that poorly reflected the pharmaceutical company.

8.

Population of interest: All cars made by this manufacturer

Deliberate Bias: They are leaving out all cars brought to private mechanics or other dealerships.
Non-response Bias: Many people may refuse to bring in the car for minor repairs.
Response Bias: Some people may lie about or forget to list problems with the car.

9.

Population of Interest: All people accused of a crime in the U.S.

Deliberate Bias: There is a conflict of interest. Northpointe should not be doing their own validation study. The validation study should be done by independent statisticians.

Section 1D Answers

1. Observational Study: Collecting data without trying to control confounding variables. Data collected by an observational study can show relationships but cannot prove cause and effect.

2. Experiment: A scientific method for controlling confounding variables and proving cause and effect.

3. Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable that causes the effect.

4. Response Variable: The dependent variable. In an experiment this the variable that measures the effect.

5. Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

6. Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

7. Placebo: A fake medicine or fake treatment used to control the placebo effect.

8. Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

9. Single Blind: When only the person receiving the treatment does not know if it is real or a placebo.

10. Double Blind: When both the person receiving the treatment and the person giving the treatment does not know if it is real or a placebo.

11.

a) They did use random assignment. The problem says they were randomly put into three groups.

b) Answers will vary. Confounding Variables: volume of the music, genetics, age, education level, etc.

c) Explanatory Variable: Type of Music
Response Variable: The amount of information they were able to memorize.

d) Control Group: The group that memorized information without music.
Treatment Groups: The groups that memorize information with their favorite music or with hated music.

e) Since the three groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, and genetics.) They must make the volume of the music the same for all participants. These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the no music group did significantly better, it proves that listening to music does not cause a person to memorize information better. It shows that memorizing information in silence is better.

12.

a) They did use random assignment. The problem says that participants were randomly put into the control and treatment groups.

b) Answers will vary. Confounding Variables: amount of motion, genetics, age, diet, pregnancy, etc.

c) Explanatory Variable: Taking Dramamine or not.
Response Variable: The amount of motion sickness.

d) Control Group: The group that took a placebo.
Treatment Group: The groups that took Dramamine.

e) Since the two groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, and genetics.) They must make the amount of motion the same for all participants. These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the treatment (Dramamine) group had significantly less motion sickness, it proves that Dramamine does decrease the amount of motion sickness.

13.

a) They did use random assignment. The problem says they were randomly put into two groups.

b) Answers will vary. Confounding Variables: age, education level, experience, type of job, where the applicant lives, poverty level, etc.

c) Explanatory Variable: Whether the applicant on the fake resume had a white or African American sounding name.
Response Variable: Whether or not the applicant received a call back for a job interview.

d) Control Group: The group that had a white sounding name.
Treatment Group: The group that had an African American sounding name.

e) Since the two groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, experience, type of job, where the applicant lives, poverty level, etc.) These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the applicants with African American names received significantly less call backs for job interviews than for applicants with white sounding names, it shows that whether the applicant had a white or African American sounding name did influence the chances of getting a call back for a job interview. Shows there is racial discrimination in the job market in Boston and Chicago.

Chapter 1 Review Sheet Answers

1.

a) Categorical since the data would consist of words.
b) Quantitative since it is numerical measurement data.
c) Categorical since the data would consist of words.
d) Categorical since the data would consist of words.
e) Quantitative since it is numerical measurement data.
f) Quantitative since it is numerical measurement data.

2.

a) Jim can ask every 5^{th} student that walks into the COC cafeteria about their salary. This would have a significant amount of sampling bias.

b) Jim can put a survey on Facebook asking how money COC students make. This would have a significant amount of sampling bias.

c) Jim can have a computer randomly select student ID numbers and then track down those students whose ID numbers were selected and ask them their salary. This would have no sampling bias.

d) Jim can ask other students in his COC classes about their salary. This would have a significant amount of sampling bias since it is not a random sample.

e) Jim can randomly select 10 section numbers at COC, and then go to those classes and get data from everyone in the class. Since he chose the groups randomly, this would not have much sampling bias.

f) Jim could walk around the COC campus asking female students about their salary. Later he could walk around asking male students about their salary. Later he could compare the female and male student salaries. Since this method was not randomly selected, there would be a lot of sampling bias.

3.

Population: The collection of all people or objects to be studied. For example, a marine biologist could study all dolphins in the world.

Census: Collecting data from everyone in a population. This is the best way to collect data and minimizes sampling bias. For example, suppose our population of interest was the students at Valencia high school. We could collect data from every student at Valencia high school.

Sample: Collecting data from a small subgroup of the population. For example, if our population was all people in Palmdale, CA, we might collect data from fifty people in Palmdale.

Random: When everyone in the population has a chance to be included in the sample. Suppose our population is all COC students. We could have a computer randomly select student ID numbers and then collect data from those students.

Bias: When data does not represent the population. Asking your friends and family will not represent the population of all people in the world.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. Sample mean averages, sample standard deviations, or sample percentages would all be examples of statistics.

4.

Sampling Bias: A type of bias that results from collecting sample data that is not random or representative of the population. For example, if our population was all adults in California, and our sample consists of asking our friends and family. To limit this bias, we could take a random sample instead.

Question Bias: A type of bias that results when someone phrases the question or gives extra information with the goal of swaying the person to answer a certain way. Instead of asking a person's opinion about raising taxes, the person first gives a speech about how they think raising taxes is terrible. To limit this bias we could simply ask if the person is for raising or lowering taxes and not give any extra information.

Response Bias: A type of bias that results when people do not answer truthfully or accurately. Asking people how much they weigh in pounds will result in many people lying about the answer. Instead of asking people, we could weigh them on a scale and assure them the data will not be released.

Deliberate Bias: A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. A common deliberate bias is to delete all of the data that makes your company look bad. We could avoid this bias by not deleting data or falsifying reports. Use the data to improve the company.

Non-response Bias: A type of bias that results when people refuse to participate or give data. When calling random phone numbers to collect data, many people will refuse to answer. To limit this bias, we may leave a message asking them to call us back and offering a gift card if they do.

5. Rachael will need a group of volunteers who want to participate in the experiment. She will need to randomly assign the volunteers into two groups. One group will be the treatment group and receive actual nicotine patches. The other group will be the control group and receive a fake patch (placebo). The placebo patch and the real patch should look identical. Patches should be given to patients using a double blind approach. No volunteer in the experiment will know if they are getting the real patch or a placebo. Also those directly giving the patch will not know either. This will control the placebo effect. Randomly assigning the groups will make them alike in many confounding variables. Rachael may also exercise direct control and manipulate the groups so that they are even more alike. There are many confounding variables including the level of addiction, the number of cigarettes smoked previously, genetics, age, gender, stress, job, etc. Answers may vary. Random assignment should control these confounding variables. If the experiment shows that those with the patch have a significantly higher percentage of quitting smoking, then it will prove that using the patch causes a person to quit smoking.

6.

An experiment creates two or more similar groups with either random assignment or using the same people twice. The similar groups control confounding variables and prove cause and effect. An observational study does not create similar groups and does not control confounding variables. An observational study just collects data and analyzes it, so it cannot prove cause and effect.

Experiment Example: Suppose we want to prove that drinking alcohol causes car accidents. We can have a group of volunteers that wish to participate. We create a driving course with cones. All of the volunteers drive the course sober and we keep track of the number of cones struck. All volunteers drive the same car, with no other distractions (no phones or radio). Then we allow the volunteers to drink alcohol until they all have similar blood alcohol content. Then they can re-drive the course and we keep track of the number of cones struck. If the number of cones is significantly more in the drunk drivers, we have proven that drinking alcohol causes car accidents.

Observational Study Example: Suppose we collect data on car accidents and how many of them involved drunk driving. There are many things that influence having a car accidents other than alcohol, so this data would not prove cause and effect.

Introduction to Data Analysis (2nd Edition)
Chapter 2 Answer Keys

Section 2A Answers

1. 3.9%
2. 88.3%
3. 0.61%
4. 9.2%
5. 21.7%
6. 0.38%
7. 65.1%
8. 7.05%
9. 0.014%
10. 70.05%
11. 0.58
12. 0.926
13. 0.08104
14. 0.00772
15. 0.0319
16. 0.08
17. 0.625
18. 0.0352
19. 0.00044
20. 0.03
21. 0.354
22. 0.026
23. 0.004
24. 0.026
25. 0.200
26. 5.7%
27. 12.3%
28. 74.0%
29. 2.7%
30. 0.3%

31. 6064 ÷ 10528 ≈ 0.576 = 57.6% of the LGBTQ students feel unsafe at school because of their sexual orientation.

32. 8970 ÷ 10528 ≈ 0.852 = 85.2% of the LGBTQ students have been verbally harassed.

33. 5117 ÷ 10528 ≈ 0.486 = 48.6% of the LGBTQ students experienced cyberbullying.

34. 1369 ÷ 10528 ≈ 0.130 = 13.0% of the LGBTQ students where physically assaulted.

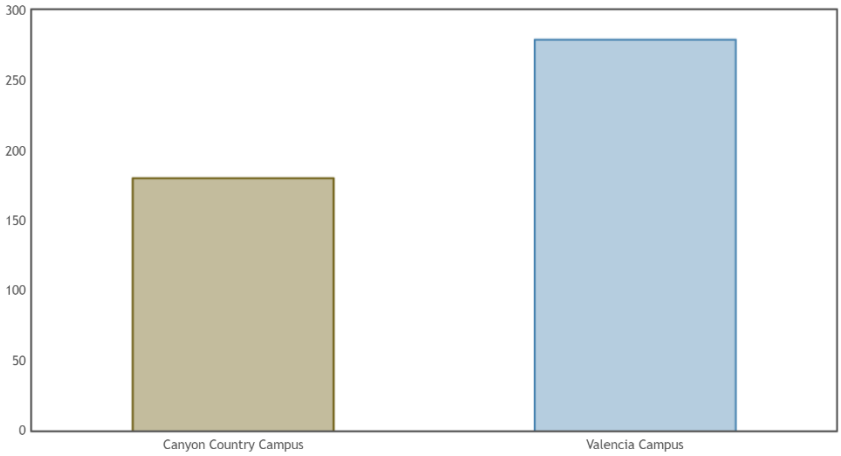
35. 57.6% = 0.576 of the LGBTQ students who were harassed or assaulted in school did not report the incident.

36. 63.5% = 0.635 of the LGBTQ students who reported an incident said that the school staff did not respond and told them to ignore it.

37. 45% = 0.45 of African American defendants were misclassified as high risk by the COMPASS program.

Section 2B Answers

1.

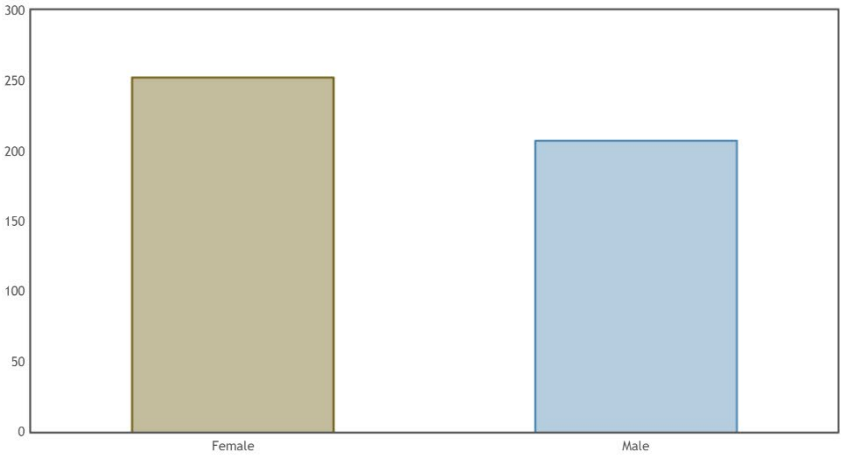


Summary Statistics

	Count	Proportion
Canyon Country Campus	180	0.392
Valencia Campus	279	0.608
Total	459	1.000

a) There are more students at Valencia.
b) There were 279 Math 075 students at the Valencia campus.
c) There were 180 Math 075 students at the Canyon Country campus.
d) 0.608 of the Math 075 students attend the Valencia campus.
e) 0.392 of the Math 075 students attend the Canyon Country campus.
f) 60.8% of the Math 075 students attend the Valencia campus.
g) 39.2% of the Math 075 students attend the Canyon Country campus.

2.

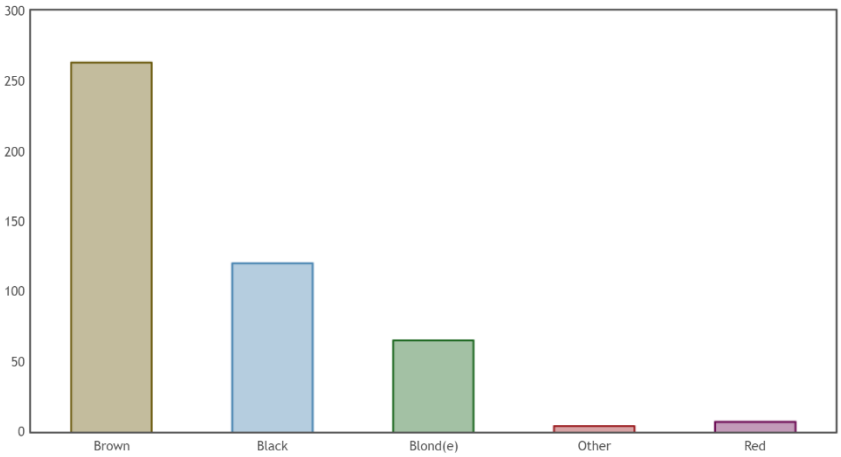


Summary Statistics

	Count	Proportion
Female	252	0.549
Male	207	0.451
Total	459	1.000

a) There were more Math 075 students that identified as female than male.
b) There were 252 Math 075 students that identified as female.
c) There were 207 Math 075 students that identified as male.
d) 0.549 of the Math 075 students identified as female.
e) 0.451 of the Math 075 students identified as male.
f) 54.9% of the Math 075 students identified as female.
g) 45.1% of the Math 075 students identified as male.

3.



Summary Statistics

	Count	Proportion
Brown	263	0.573
Black	120	0.261
Blond(e)	65	0.142
Other	4	0.0087
Red	7	0.015
Total	459	1.000

a) The hair color with the most Math 075 students was brown.
b) The hair color with the least Math 075 students was “other”.
c) 263 of the Math 075 students had brown hair.
d) 65 of the Math 075 students had blonde hair.
e) 0.015 of the Math 075 students had red hair.
f) 0.261 of the Math 075 students had black hair.
g) 1.5% of the Math 075 students had red hair.
h) 26.1% of the Math 075 students had black hair.

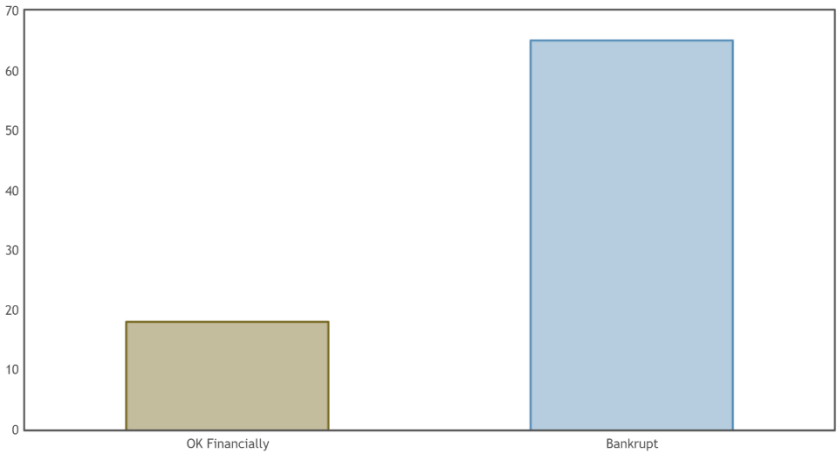
4.

a) Democratic party was most popular with Math 075 students.
b) Independent political party was least popular with Math 075 students.
c) 97 of the Math 075 students were republican.
d) 185 of the Math 075 students were democrat.
e) 19% of the Math 075 students identified as independent political party.
f) 20% of the Math 075 students identified as other political party.
g) 0.4 of the Math 075 students identified as democratic.
h) 0.21 of the Math 075 students identified as republican.

5.

a) In 2015, the most popular social media with Math 075 students was Instagram.
b) In 2015, the least popular social media with Math 075 students was “other” social media.
c) 94 of the Math 075 students prefer Snapchat.
d) 137 of the Math 075 students prefer Instagram.
e) 20% of the Math 075 students prefer Twitter.
f) 8% of the Math 075 students prefer “other” social media.
g) 0.3 of the Math 075 students prefer Instagram.
h) 0.2 of the Math 075 students prefer Snapchat.

6.

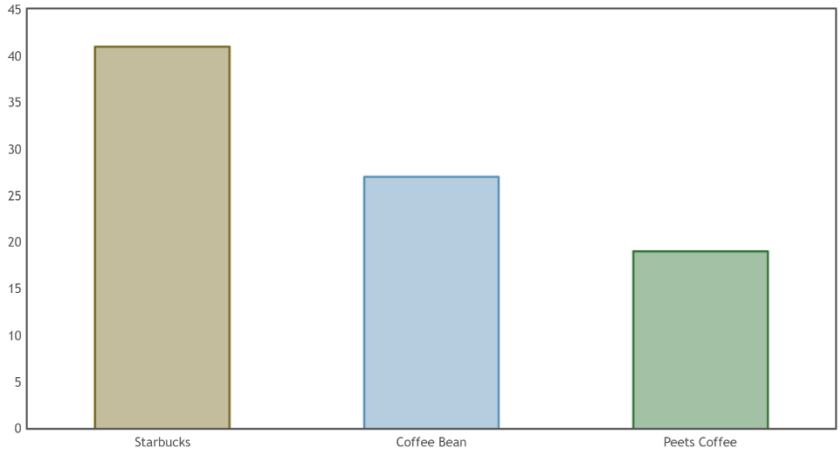


Summary Statistics

	Count	Proportion
OK Financially	18	0.217
Bankrupt	65	0.783
Total	83	1.000

a) 0.783 of the retired NFL players had gone bankrupt.
b) 78.3% of the retired NFL players had gone bankrupt.
c) 0.217 of the retired NFL players were doing ok financially.
d) 21.7% of the retired NFL players were doing ok financially.

7.



Summary Statistics

	Count	Proportion
Starbucks	41	0.471
Coffee Bean	27	0.31
Peets Coffee	19	0.218
Total	87	1.000

a) 0.471 of the Math 075 students prefer Starbucks.
b) 47.1% of the Math 075 students prefer Starbucks.
c) 0.31 of the Math 075 students prefer Coffee Bean.
d) 31% of the Math 075 students prefer Coffee Bean.
e) 0.218 of the Math 075 students prefer Peet's Coffee.
f) 21.8% of the Math 075 students prefer Peet's Coffee.

8.

a) 9.5% (0.095) of the LGBTQ students were not planning to continue their education due to high victimization against their sexual orientation.

b) 10.0% (0.100)of the LGBTQ students were not planning to continue their education due to high victimization against their gender expression.

c) 5.4% (0.054) of the LGBTQ students were not planning to continue their education due to lower level victimization against their sexual orientation.

d) 5.2% (0.052) of the LGBTQ students were not planning to continue their education due to lower level victimization against their gender expression.

9.

a) 66.3% (0.663) of the LGBTQ students attending a school without a Gay-Straight Alliance program feel unsafe because of sexual orientation.

b) 50.2% (0.502) of the LGBTQ students attending a school with a Gay-Straight Alliance program feel unsafe because of sexual orientation.

c) 48.2% (0.482) of the LGBTQ students attending a school without a Gay-Straight Alliance program feel unsafe because of gender expression.

d) 39.1% (0.391) of the LGBTQ students attending a school with a Gay-Straight Alliance program feel unsafe because of gender expression.

Section 2C Answers

1.

a) Ratio: $0.653 \div 0.347 \approx 1.88$

The percentage of the women that preferred athletic wear is 1.88 times larger than the percentage of women that preferred traditional jeans.

b) Percent of Increase: $\frac{(0.653-0.347)}{0.347} \times 100\% \approx 88.2\%$ (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of the women that prefer athletic wear is significantly higher than the percent of the women that prefer jeans.

d) We would advise the company to increase their supply of women's athletic wear and decrease the supply of women's jeans.

2.

a) Ratio: $0.163 \div 0.14 \approx 1.16$

The percentage of the patients on the medical/surgical ward is 1.16 times larger than the percentage of patients on the telemetry ward.

b) Percent of Increase: $\frac{(0.163-0.14)}{0.14} \times 100\% \approx 16.3\%$ (Small percent of increase)

c) The sample size was large enough, but the ratio and percent of increase were low. This data indicates that the percentage of patients on the medical/surgical floor and the telemetry floor are about the same.

d) We would advise the hospital to set aside similar amount resources for both floors.

3.

a) Ratio: $0.724 \div 0.276 \approx 2.62$

The percentage of employees without health insurance is 2.62 times larger than the percentage of employees with health insurance.

b) Percent of Increase: $\frac{(0.724-0.276)}{0.276} \times 100\% \approx 162.3\%$ (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of employees without health insurance is significantly larger than the percentage of employees with health insurance.

d) We would advise the company to increase access to their health insurance benefits.

4.

a) Ratio: $0.228 \div 0.18 \approx 1.27$

The percentage of the people that took the medicine and improved was only 1.27 times larger than the percentage in the placebo group.

b) Percent of Increase: $\frac{(0.228-0.18)}{0.18} \times 100\% \approx 26.7\%$ (Moderate percent of increase)

c) The sample size was large enough (barely), and the ratio and percent of increase were moderately high. This data indicates that the percentage of people that took the medicine and improved was not significantly larger than for the placebo group.

d) The experiment indicates that the depression medicine may not work. We recommend further testing.

5.

a) Ratio: $0.45 \div 0.23 \approx 1.96$

The percentage of African American defendants misclassified as high risk is 1.96 times larger than the percentage white defendants misclassified as high risk.

b) Percent of Increase: $\frac{(0.45-0.23)}{0.23} \times 100\% \approx 95.6\%$ (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of African American defendants misclassified as high risk is significantly higher than the percentage white defendants misclassified as high risk.

d) We would advise the court system to stop using the COMPAS program to determine if a defendant will repeat their crime. The system seems to be racially biased.

6.

a) Ratio: $0.18 \div 0.13 \approx 1.38$

The percentage of cars from Japan was 1.38 times larger than the percentage of cars from Germany.

b) Percent of Increase: $\frac{(0.18-0.13)}{0.13} \times 100\% \approx 38.5\%$ (Moderate percent of increase.)

c) The ratio and percent of increase were not significantly high which would usually indicate that the percentage of cars made in Japan was slightly higher than for Germany. However, the sample size was really too small to make a determination with this data.

7.

a) Ratio: $0.33 \div 0.29 \approx 1.14$

The percentage of cereals made by Kelloggs was only 1.14 times larger than the percentage of cereals made by General.

b) Percent of Increase: $\frac{(0.33-0.29)}{0.29} \times 100\% \approx 13.8\%$ (Low percent of increase)

c) The ratio and percent of increase were low which would indicate that the percentages were close. However, the sample size was really too small to make a determination with this data.

8.

a) Ratio: $0.67 \div 0.33 \approx 2.03$

The percentage of cereals made for adults was 2.03 times larger than the percentage of cereals made for children.

b) Percent of Increase: $\frac{(0.67-0.33)}{0.33} \times 100\% \approx 103.0\%$ (High percent of increase)

c) The ratio and percent of increase were high which would usually indicate that the percentage of cereals for adults is significantly higher than the percentage for children. However, the sample size was really too small to make a determination with this data.

9.

a) Ratio: $0.095 \div 0.054 \approx 1.76$

The percentage of highly victimized sexual orientation LGBTQ students not planning to continue school is 1.76 times larger than the percentage of lower level sexual orientation victimized LGBTQ students not planning to continue school.

b) Percent of Increase: $\frac{(0.095-0.054)}{0.054} \times 100\% \approx 75.9\%$ (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of highly victimized (sexual orientation) LGBTQ students not planning to continue school is significantly higher than the percentage of lower level victimized (sexual orientation) LGBTQ students not planning to continue school.

d) Ratio: $0.1 \div 0.052 \approx 1.92$

The percentage of highly victimized gender expression LGBTQ students not planning to continue school is 1.92 times larger than the percentage of lower level gender expression victimized LGBTQ students not planning to continue school.

e) Percent of Increase: $\frac{(0.1-0.052)}{0.052} \times 100\% \approx 92.3\%$ (Large percent of increase)

f) The ratio and percent of increase were significantly high. The sample size was large enough, so this data indicates that the percentage of highly victimized gender expression LGBTQ students not planning to continue school is significantly higher than the percentage of lower level victimized gender expression LGBTQ students not planning to continue school.

10.

a) Ratio: $0.663 \div 0.502 \approx 1.32$

The percentage of LGBTQ students from schools without a Gay-Straight Alliance that feel unsafe due to sexual orientation is 1.32 times larger than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to sexual orientation.

b) Percent of Increase: $\frac{(0.663-0.502)}{0.502} \times 100\% \approx 32.1\%$ (Moderate percent of increase)

c) The ratio and percent of increase were moderately high. Since the sample size was large enough, this data indicates that the percentage of LGBTQ students from schools without a Gay-Straight Alliance (GSA) that feel unsafe due to sexual orientation is moderately higher than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to sexual orientation. Schools without a GSA should consider instituting one.

d) Ratio: $0.482 \div 0.391 \approx 1.23$

The percentage of LGBTQ students from schools without a Gay-Straight Alliance that feel unsafe due to gender expression is 1.23 times larger than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to gender expression.

e) Percent of Increase: $\frac{(0.482-0.391)}{0.391} \times 100\% \approx 23.3\%$ (Moderate percent of increase)

f) The ratio and percent of increase were moderately high. Since the sample size was large enough, this data indicates that the percentage of LGBTQ students from schools without a Gay-Straight Alliance (GSA) that feel unsafe due to gender expression is moderately higher than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to gender expression. Schools without a GSA should consider instituting one.

11.

a) The percentage of call-backs for applicants with white sounding names was 1.5 times higher than for applicants with African American sounding names.

b) Percent of Increase: $\frac{(0.1006-0.0670)}{0.0670} \times 100\% \approx 50.1\%$ (High percent of increase)

c) The ratio and percent of increase were high. Since the sample size was large enough, this data indicates that the percentage of call-backs for applicants with white sounding names was significantly higher than for applicants with African American sounding names.

d) Since the experiment controlled confounding variables, this may indicate racial discrimination in the labor market in Boston and Chicago.

Section 2D Answers

1. 658 cars
2. 1472 people
3. 260 dogs
4. 77 cats
5. 315 bears
6. 20,246 car accidents
7. 10,800 cases of flu

8.

a) 0.15
b) 0.15 x 78300 ≈ 11,745
We estimate that there are 11,745 people in Chino Hills without insurance.

9.

a) 0.3
b) 0.3 x 305700 ≈ 91,710
We estimate that there are 91,710 people in Stockton which own guns.

10.

a) 0.093
b) 0.093 x 18400 ≈ 1,711
We estimate that there are 1,711 students at COC with diabetes.

11.

a) 0.159
b) 0.159 x 161000 ≈ 25,599
We estimate that there are 25,599 people in Lancaster struggling with hunger.

12.

a) 0.0147
b) 0.0147 x 136400 ≈ 2,005
We estimate that there are 2,005 people in Van Nuys with autism.

13.

a) 0.0051
b) 0.0051 x 1769000 ≈ 9,022
We estimate that there are 9,022 cars in San Francisco with defective air bags.

14.

a) 0.148
b) 0.148 x 305700 ≈ 45,244
We estimate that there are 45,244 people in Stockton, CA living in poverty.

15.

a) 0.33
b) 0.33 x 147 ≈ 49
We estimate that there are 49 doctors that have been sued for malpractice at that hospital.

16.

a) 0.78
b) 0.78 x 26682 ≈ 20,812
We estimate that there are 20,812 retired NFL players bankrupt or in financial stress.

17.

a) 0.6
b) 0.6 x 4374 ≈ 2,624
We estimate that there are 2,624 retired NBA players that have gone broke.

18.

a) 0.576
b) 0.576 x 244,000 ≈ 140,544
We estimate that there are 140,544 LGBTQ students between 13 and 17 years old in California that feel unsafe at their school due to sexual orientation.

19.

a) 0.852
b) 0.852 x 1994000 ≈ 1,698,888
We estimate that there are 1,698,888 LGBTQ students in the U.S. between 13 and 17 years old that have been verbally harassed at school.

20.

a) 0.486
b) 0.486 x 114000 ≈ 55,404
We estimate that there are 55,404 LGBTQ students between the ages of 13 and 17 in Florida that have experienced cyberbullying.

21.

a) 0.13
b) 0.13 x 113,000 ≈ 14,690
We estimate there are 14,690 LGBTQ students between in the ages of 13 and 17 in New York that have been physically assaulted.

Chapter 2 Review Sheet Answers

1. Quantitative
2. Categorical
3. Categorical
4. Quantitative

5. 0.0385
6. 0.926
7. 0.0051

8. 55.8%
9. 0.32%
10. 9.3%

11. 17/47 ≈ 0.362
12. 0.362 x 100% = 36.2%

13. 0.333
14. 0.333 x 58 ≈ 19 HIV deaths by Tuberculosis

15. 1478 Intel processors
16. 850 AMD processors
17. About 4% of processors are made by Mobile.
18. About 35% of processors are made by AMD.
19. About 61% of processors are made by Intel.

20. Intel had the most processors.
21. Mobile made the least processors.

22. Percent Ratio for Intel and AMD = 61/35 ≈ 1.7
The percentage of Intel is significantly higher as it is a larger sample size and the ratio is not close to 1.

Introduction to Data Analysis (2nd Edition)
Chapter 3 Answer Keys

Section 3A

1.

2 x 4 table

	A	AB	B	O	All
Rh+	3	1	2	9	15
Rh−	2	1	0	2	5
All	5	2	2	11	20

2.

2 x 4 table

	ER	ICU	Med/Surg	SDS	All
F	2	1	4	2	9
M	3	2	2	4	11
All	5	3	6	6	20

3.

2 x 4 table

	ER	ICU	Med/Surg	SDS	All
Rh+	4	2	5	4	15
Rh−	1	1	1	2	5
All	5	3	6	6	20

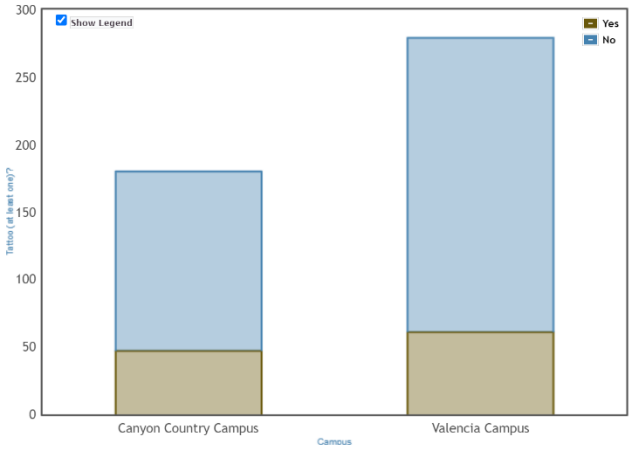
4.

4 x 4 table

	ER	ICU	Med/Surg	SDS	All
A	1	0	2	2	5
AB	1	0	1	0	2
B	0	1	0	1	2
O	3	2	3	3	11
All	5	3	6	6	20

5.

a-b)



Counts Table Switch Variables

Tattoo (at least one)? \ Campus	Canyon Country Campus	Valencia Campus	Total
Yes	47	61	108
No	133	218	351
Total	180	279	459

Proportions Row Column Overall

c) Grand Total = 459 students

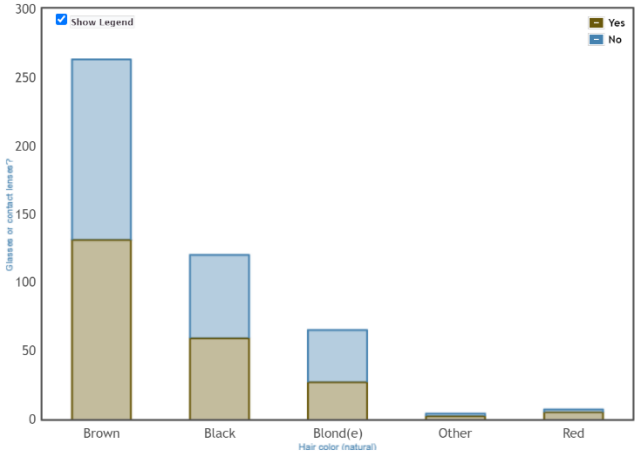
d) Total Valencia Campus = 279 students

e) Total with Tattoo = 108 students

f) 133 students both went to Canyon Country campus and did not have a tattoo.

6.

a-b)



Counts Table Switch Variables

Glasses or contact lenses? \ Hair color (natural)	Brown	Black	Blond(e)	Other	Red	Total
Yes	131	59	27	2	5	224
No	132	61	38	2	2	235
Total	263	120	65	4	7	459

Proportions Row Column Overall

c) Grand Total = 459 students

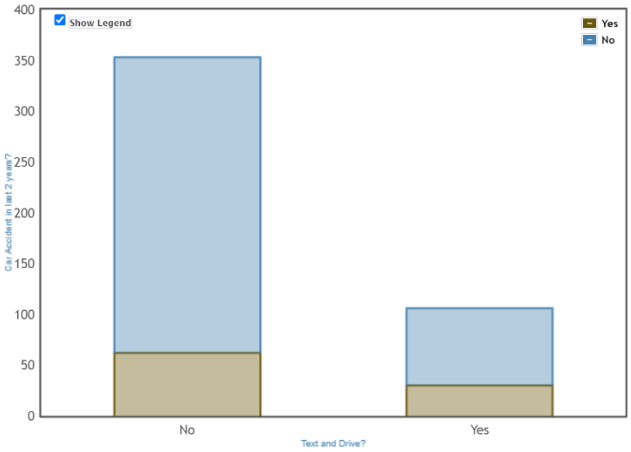
d) 224 students need contacts or glasses.

e) 263 students have brown hair.

f) 61 students both do not need contacts or glasses and have black hair.

7.

a-b)



Counts Table Switch Variables

Car Accident in last 2 years? \ Text and Drive?	No	Yes	Total
Yes	62	30	92
No	291	76	367
Total	353	106	459

Proportions Row Column Overall

c) Grand Total = 459 students

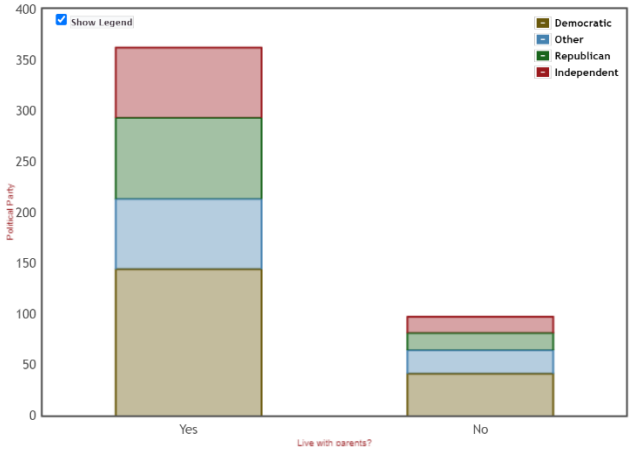
d) 353 students said they do not text and drive. I do not believe that all of these students told the truth. The actual count may be lower because of response bias.

e) 92 students have been in a car accident over the two year period.

f) 30 students both said they have been in a car accident and admitted to texting and driving.

8.

a-b)



Counts Table Switch Variables

Political Party \ Live with parents?	Yes	No	Total
Democratic	144	41	185
Other	69	23	92
Republican	80	17	97
Independent	69	16	85
Total	362	97	459

Proportions Row Column Overall

c) Grand Total = 459 students

d) 97 students said they do not live with their parents.

e) 85 students identify as “independent” political party.

f) 144 students both live with parents and identify as democrat.

Section 3B

1.

a) 108 students have at least one tattoo.
b) $108 \div 459 \approx 0.235$
c) $0.235 \times 100\% = 23.5\%$

2.

a) 99 students prefer Facebook.
b) $99 \div 459 \approx 0.216$
c) $0.216 \times 100\% = 21.6\%$

3.

a) 351 students do not have a tattoo.
b) $351 \div 459 \approx 0.765$
c) $0.765 \times 100\% = 76.5\%$

4.

a) 137 students prefer Instagram.
b) $137 \div 459 \approx 0.298$
c) $0.298 \times 100\% = 29.8\%$

5.

a) 33 students both have a tattoo and prefer Facebook.
b) $33 \div 459 \approx 0.072$
c) $0.072 \times 100\% = 7.2\%$

6.

a) 99 students both do not have a tattoo and prefer Instagram.
b) $99 \div 459 \approx 0.216$
c) $0.216 \times 100\% = 21.6\%$

7.

a) 79 students both do not have a tattoo and prefer Snapchat.
b) $79 \div 459 \approx 0.172$
c) $0.172 \times 100\% = 17.2\%$

8.

a) 9 students both have a tattoo and prefer “Other” social media.
b) $9 \div 459 \approx 0.020$
c) $0.020 \times 100\% = 2.0\%$

9.

a) $108 + 99 - 33 = 174$ students either have a tattoo or prefer Facebook.
b) $174 \div 459 \approx 0.379$
c) $0.379 \times 100\% = 37.9\%$

10.

a) $351 + 137 - 99 = 389$ students either do not have a tattoo or prefer Instagram.
b) $389 \div 459 \approx 0.847$
c) $0.847 \times 100\% = 84.7\%$

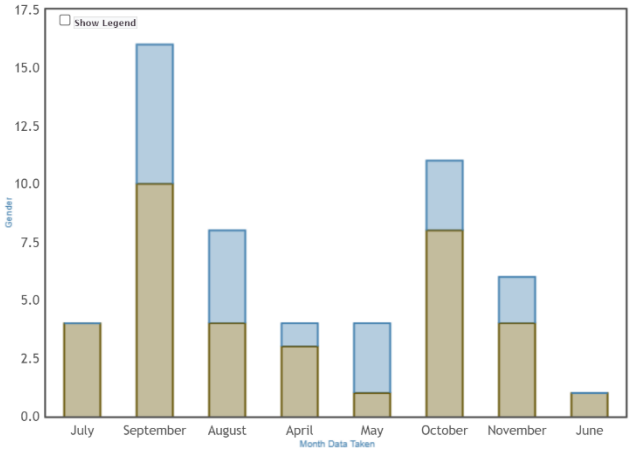
11.

a) 91 + 94 = 185 students prefer either Twitter or Snapchat. (Variables do not intersect.)
b) 185 ÷ 459 ≈ 0.403
c) 0.403 × 100% = 40.3%

12.

a) 108 + 38 – 9 = 137 students either have a tattoo or prefer “Other” social media.
b) 137 ÷ 459 ≈ 0.298
c) 0.298 × 100% = 29.8%

13.



Counts Table Switch Variables

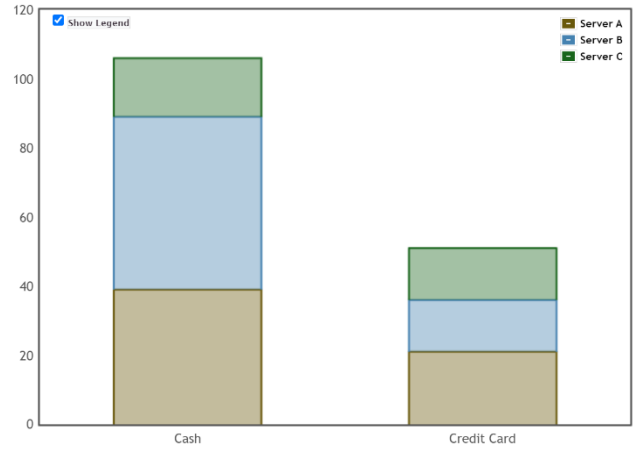
Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	4	10	4	3	1	8	4	1	35
female	0	6	4	1	3	3	2	0	19
Total	4	16	8	4	4	11	6	1	54

Proportions Row Column Overall

Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	0.074	0.185	0.074	0.056	0.019	0.148	0.074	0.019	0.648
female	0	0.111	0.074	0.019	0.056	0.056	0.037	0	0.352
Total	0.074	0.296	0.148	0.074	0.074	0.204	0.111	0.019	1

a) 0.296 (29.6%) of the bears were measured in September.
b) 0.352 (35.2%) of the bears were female.
c) 0.111 (11.1%) of the bears were both female and had data taken in September.
d) 0.296 + 0.352 – 0.111 = 0.537 (53.7%) of the bears were either female or had data taken in September.

14.



Counts Table Switch Variables

undefined \ undefined	Cash	Credit Card	Total
Server A	39	21	60
Server B	50	15	65
Server C	17	15	32
Total	106	51	157

Proportions Row Column Overall

undefined \ undefined	Cash	Credit Card	Total
Server A	0.248	0.134	0.382
Server B	0.318	0.096	0.414
Server C	0.108	0.096	0.204
Total	0.675	0.325	1

a) 0.675 (67.5%) of the bills were paid with cash.
b) 0.414 (41.4%) of the bills had server B as the server.
c) 0.318 (31.8%) of the bills were both served by server B and paid in cash.
d) 0.675 + 0.414 – 0.318 = 0.771 (77.1%) of the bills were either served by server B or paid in cash.

Section 3C

1.

a) 108 students have at least one tattoo.
b) 38 students both have a tattoo and prefer Instagram.
c) $38 \div 108 \approx 0.352$ of the tattoo students prefer Instagram.
d) $0.352 \times 100\% = 35.2\%$ of the tattoo students prefer Instagram.

2.

a) 91 students prefer Twitter?
b) 78 students both do not have a tattoo and prefer Twitter.
c) $78 \div 91 \approx 0.857$ of the Twitter students do not have a tattoo.
d) $0.857 \times 100\% = 85.7\%$ of the Twitter students do not have a tattoo.

3.

a) 351 students do not have a tattoo.
b) 66 students both do not have a tattoo and prefer Facebook.
c) $66 \div 351 \approx 0.188$ of the no tattoo students prefer Facebook.
d) $0.188 \times 100\% = 18.8\%$ of the no tattoo students prefer Facebook.

4.

a) 94 students prefer Snapchat.
b) 15 students both have a tattoo and prefer Snapchat.
c) $15 \div 94 \approx 0.160$ of the Snapchat students have a tattoo.
d) $0.160 \times 100\% = 16.0\%$ of the Snapchat students have a tattoo.

5.

a) 180 students went to the Canyon Country campus.
b) 138 students both drive alone and went to the Canyon Country campus.
c) $138 \div 180 \approx 0.767$ of the Canyon Country campus students drove alone to school.
d) $0.767 \times 100\% = 76.7\%$ of the Canyon Country campus students drove alone to school.

6.

a) 46 students were dropped off by someone.
b) 14 students were both dropped off and went to the Canyon Country campus.
c) $14 \div 46 \approx 0.304$ of the dropped off students went to the Canyon Country campus.
d) $0.304 \times 100\% = 30.4\%$ of the dropped off students went to the Canyon Country campus.

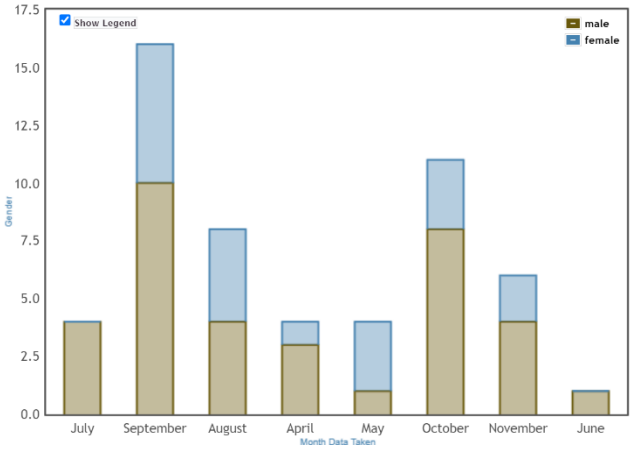
7.

a) 279 students went to the Valencia campus.
b) 22 students both carpool and went to the Valencia campus.
c) $22 \div 279 \approx 0.079$ of the Valencia campus students carpool to school.
d) $0.079 \times 100\% = 7.9\%$ of the Valencia campus students carpool to school.

8.

a) 24 students used public transportation to school.
b) 17 students both used public transportation and went to the Valencia campus.
c) $17 \div 24 \approx 0.708$ of the public transportation students went to the Valencia campus.
d) $0.708 \times 100\% = 70.8\%$ of the public transportation students went to the Valencia campus.

9.



Counts Table Switch Variables

Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	4	10	4	3	1	8	4	1	35
female	0	6	4	1	3	3	2	0	19
Total	4	16	8	4	4	11	6	1	54

Proportions Row | Column | Overall

Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	0.114	0.286	0.114	0.086	0.029	0.229	0.114	0.029	1
female	0	0.316	0.211	0.053	0.158	0.158	0.105	0	1
Total	0.074	0.296	0.148	0.074	0.074	0.204	0.111	0.019	1

a) 0.211 (21.1%) of the female bears were measured in August.

b) 0.114 (11.4%) of the male bears were measured in August.

c) The proportions in part (a) and (b) look significantly different. (85.1% increase)

d) 0.229 (22.9%) of the female bears were measured in October.

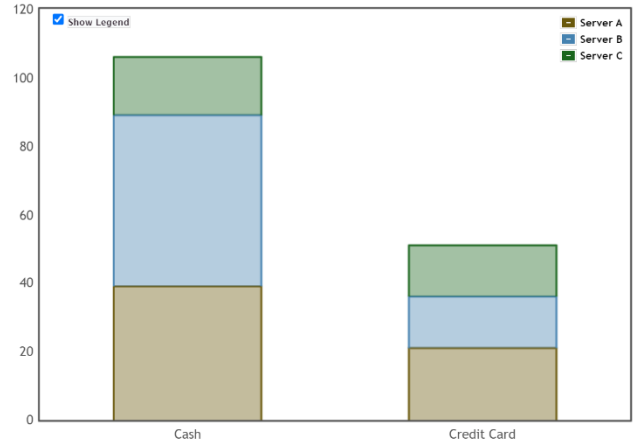
e) 0.158 (15.8%) of the male bears were measured in October

f) The proportions in part (d) and (e) have a 44.9% increase. They are different but may not be significant based on the small sample size.

g) Overall these two percentages indicate a difference between female and male bears in when they were measured. This difference in conditional probabilities may indicate that gender is related to the month the bear was measured.

h) No. Just because two variables are related, does not prove causation. This data was an observational study. It needed to use experimental design to prove cause and effect.

10.



Counts Table Switch Variables

undefined \ undefined	Cash	Credit Card	Total
Server A	39	21	60
Server B	50	15	65
Server C	17	15	32
Total	106	51	157

Proportions Row | Column | Overall

undefined \ undefined	Cash	Credit Card	Total
Server A	0.368	0.412	0.382
Server B	0.472	0.294	0.414
Server C	0.16	0.294	0.204
Total	1	1	1

a) 0.412 (41.2%) of the credit card customers were served by server A.
b) 0.294 (29.4%) of the credit card customers were served by server B.
c) 0.294 (29.4%) of the credit card customers were served by server C.
d) The conditional proportions in part (a), (b) and (c) do not look significantly different. Server A had a higher percentage, but servers B and C were the same.
e) Since the conditional proportions were not significantly different, this probably indicates that using a credit card is not related to the server.

Ch 3 Review Problem Answers

1.

	Everyone	Kids	Parents	All
Female Pet	2	0	2	4
Male Pet	4	1	4	9
All	6	1	6	13

2. 119/280 = 0.425 = 42.5%

3. 49/280 = 0.175 = 17.5%

4. 35/280 = 0.125 = 12.5%

5. 21/280 = 0.075 = 7.5%

6. 91/280 = 0.325 = 32.5%

7. 119/280 = 0.425 = 42.5%

8. 28/105 ≈ 0.267 = 26.7%

9. 21/49 ≈ 0.429 = 42.9%

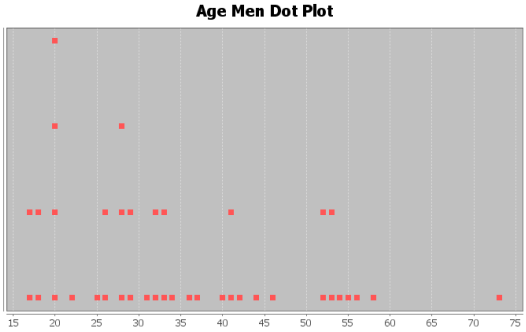
10. Significantly different

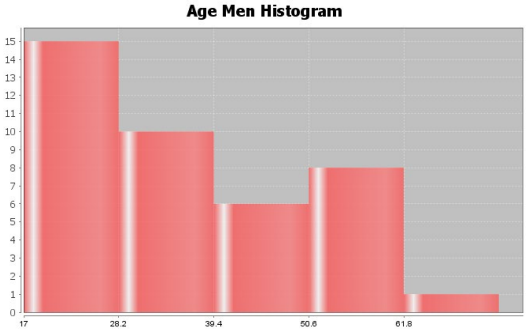
11. Data suggests that grade level is related to being democrat.

Introduction to Data Analysis
Chapter 4 Answer Key

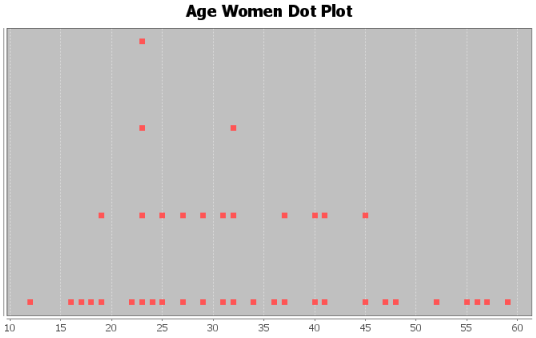
Section 4A Answers

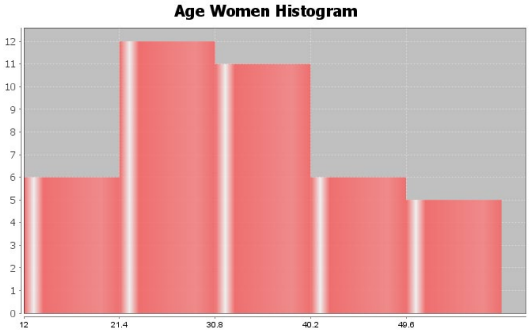
1. Men's Age (years): Skewed Right



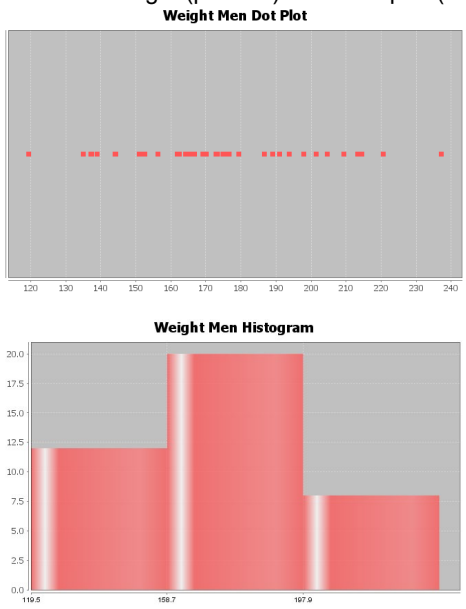


2. Women's Age (years): Almost Bell Shaped (Slightly Skewed Right)

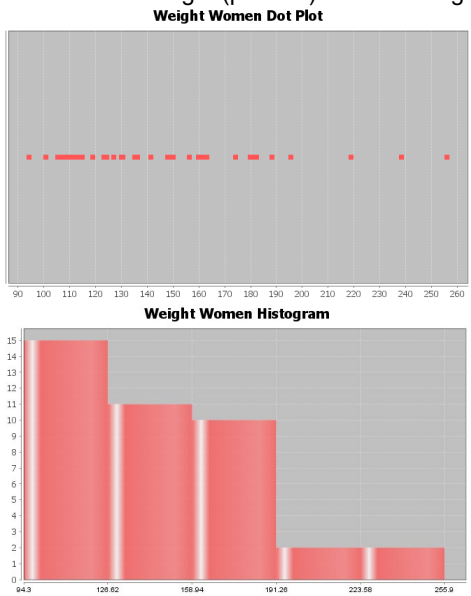




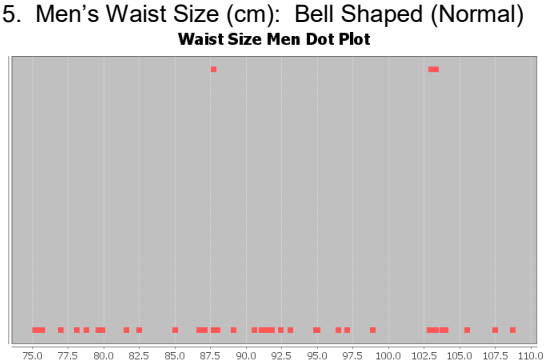
3. Men's Weight (pounds): Bell Shaped (Normal)

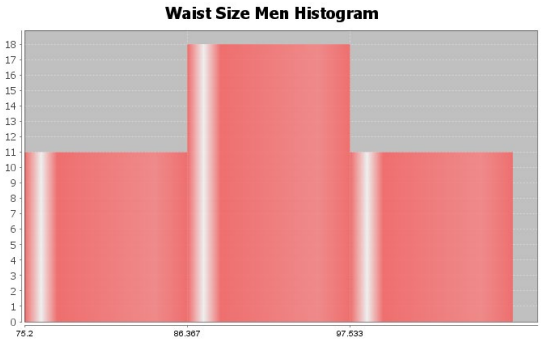


4. Women's Weight (pounds): Skewed Right

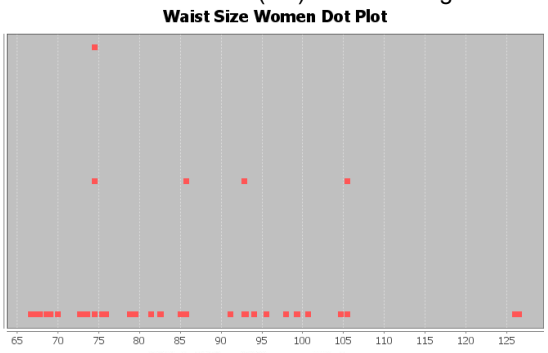


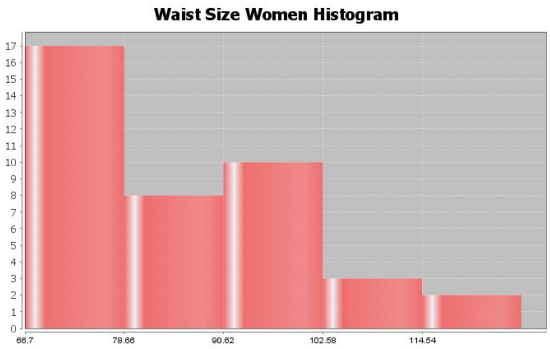
5. Men's Waist Size (cm): Bell Shaped (Normal)



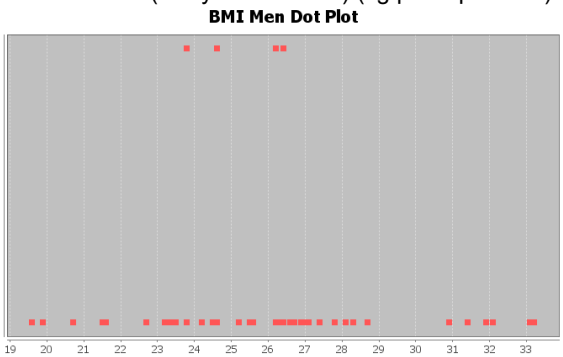


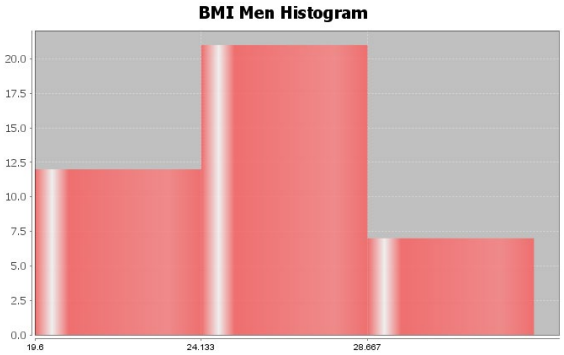
6. Women's Waist Size (cm): Skewed Right



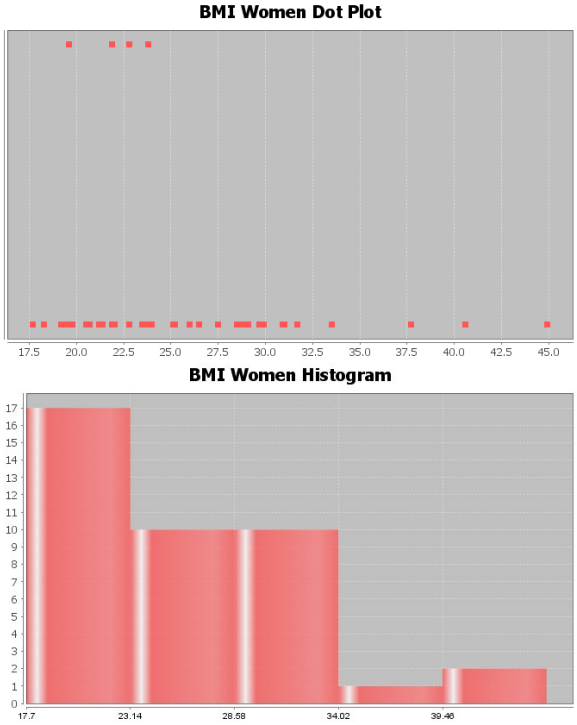


7. Men's BMI (Body Mass Index) (kg per sq meters): Bell Shaped (Normal)





8. Women's BMI (Body Mass Index) (kg per sq meters): Skewed Right



Section 4B Answers

1. Men's Age (years): Skewed Right

Best Measure of Center: Median

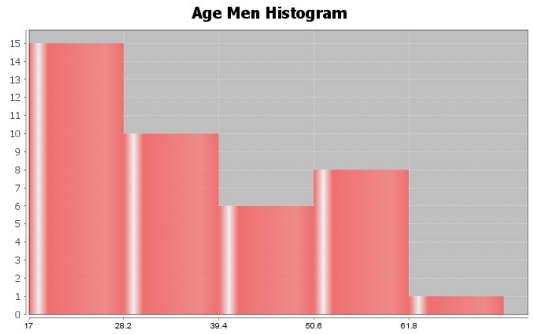
Descriptive Statistics

Variable	Mean
C15 Men Age (years)	35.475

Variable	Median	Mode	N for mode
C15 Men Age (years)	32.5	20.0	4

Variable	Min	Max
C15 Men Age (years)	17.0	73.0

Midrange = (17 + 73) /2 = 45



2. Women's Age (years): Almost Bell Shaped (Slightly Skewed Right)

Best Measure of Center: Median (If said skewed)
Best Measure of Center: Mean (If said almost Bell Shaped)

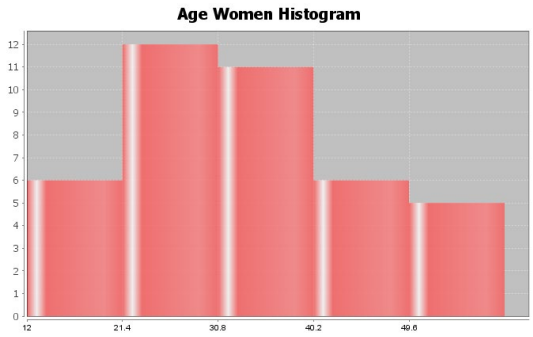
Descriptive Statistics

Variable	Mean
C1 Women Age (years)	33.225

Variable	Median	Mode	N for mode
C1 Women Age (years)	31.5	23.0	4

Variable	Min	Max
C1 Women Age (years)	12.0	59.0

Midrange = (12 + 59) / 2 = 35.5



3. Men's Weight (pounds): Bell Shaped (Normal)

Best Measure of Center: Mean

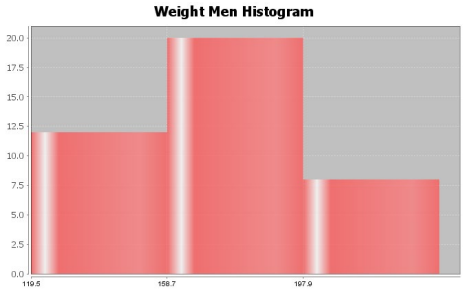
Descriptive Statistics

Variable	Mean
C17 Men Wt (Lbs)	172.55

Variable	Median	Mode	N for mode
C17 Men Wt (Lbs)	169.95	*	0

Variable	Min	Max
C17 Men Wt (Lbs)	119.5	237.1

MIdrange = (119.5 + 237.1) / 2 = 178.3



4. Women's Weight (pounds): Skewed Right

Best Measure of Center: Median

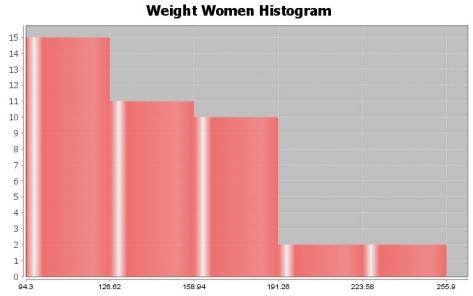
Descriptive Statistics

Variable	Mean
C3 Women Wt (Lbs)	146.220

Variable	Median	Mode	N for mode
C3 Women Wt (Lbs)	135.8	*	0

Variable	Min	Max
C3 Women Wt (Lbs)	94.3	255.9

Midrange = (94.3 + 255.9) / 2 = 175.1



5. Men's Waist Size (cm): Bell Shaped (Normal)

Best Measure of Center: Mean

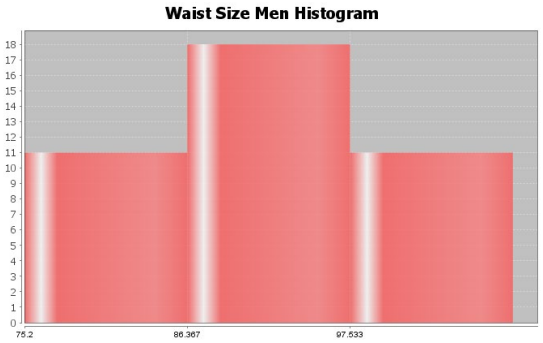
Descriptive Statistics

Variable	Mean
C18 Men Waist (cm)	91.285

Variable	Median	Mode	N for mode
C18 Men Waist (cm)	91.200	87.7, 103.0, 103.3	2

Variable	Min	Max
C18 Men Waist (cm)	75.2	108.7

Midrange = (75.2 + 108.7) / 2 = 91.95



6. Women's Waist Size (cm): Skewed Right

Best Measure of Center: Median

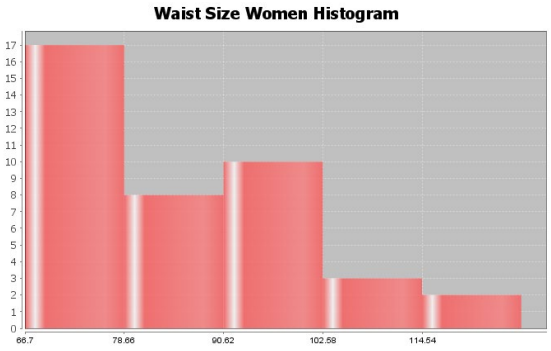
Descriptive Statistics

Variable	Mean
C4 Women Waist (cm)	85.033

Variable	Median	Mode	N for mode
C4 Women Waist (cm)	81.95	74.5	3

Variable	Min	Max
C4 Women Waist (cm)	66.7	126.5

Midrange = (66.7 + 126.5) / 2 = 96.6



7. Men's BMI (Body Mass Index) (kg per sq meters): Bell Shaped (Normal)

Best Measure of Center: Mean

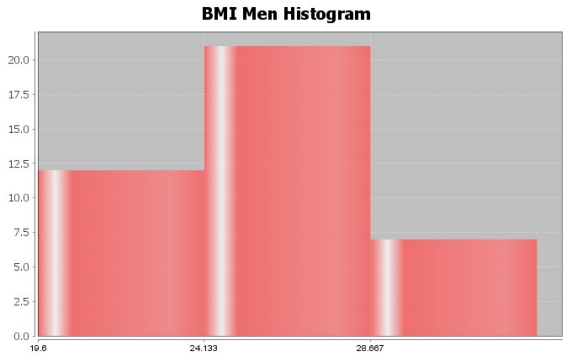
Descriptive Statistics

Variable	Mean
C23 Men BMI	25.998

Variable	Median	Mode	N for mode
C23 Men BMI	26.2	26.4, 24.6, 23.8, 26.2	2

Variable	Min	Max
C23 Men BMI	19.6	33.2

Midrange = (19.6 + 33.2) / 2 = 26.4



8. Women's BMI (Body Mass Index) (kg per sq meters): Skewed Right

Best Measure of Center: Median

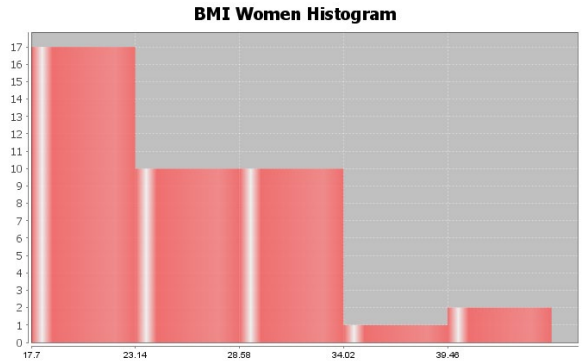
Descriptive Statistics

Variable	Mean
C9 Women BMI	25.74

Variable	Median	Mode	N for mode
C9 Women BMI	23.9	22.8, 19.6, 23.8, 21.9	2

Variable	Min	Max
C9 Women BMI	17.7	44.9

Midrange = (17.7 + 44.9) / 2 = 31.3



Section 4C Answers

1. Mean = 84/18 ≈ 4.7
2. Mean = 326/12 ≈ 27.2
3. Mean = 68/7 ≈ 9.71
4. Mean = 53.8/12 ≈ 4.48
5. Mean = 33.21/11 ≈ 3.019

6. Answers may vary (10,11,12,14,15,16)
7. Answers may vary (9,10,11,12,14,15,16,17)
8. Answers may vary (18,19,20,21,21.5,22,23,24,25)
9. Answers may vary (17,18,19,20,21,21.5,22,23,24,25,26)

10. The numbers are balanced around 10. 5 and 15 are 5 places from 10. 6 and 14 are four places from 10. 7 and 13 are both three places from 10. 8 and 12 are two places from 10. 9 and 11 are both one place from 10. The total distance from 10 for numbers above = total distance from 10 for numbers below.

Section 4D Answers

1.
Mean = 7
Sum of Squares = 154
Sample Size (total frequency) = 6
Degrees of Freedom = n-1 = 6-1 = 5
Standard Deviation = square root (154/5) = square root (30.8) ≈ 5.5

2.
Mean = 8
Sum of Squares = 100
Sample Size (total frequency) = 8
Degrees of Freedom = n-1 = 8-1 = 7
Standard Deviation = square root (100/7) = square root (14.28571429) ≈ 3.8

3. Bear Ages (Months): Shape Skewed Right

Mean and Standard Deviation are NOT accurate (not bell shaped)

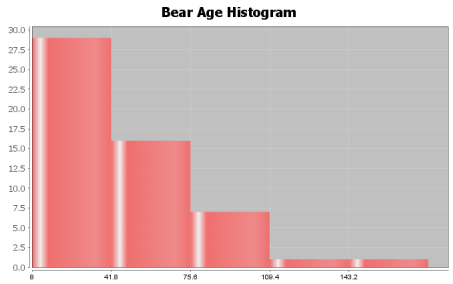
Standard Deviation Sentence: Typical bear ages were 33.7 months from the mean of 43.5 months.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C1 AGE (months)	43.519	33.721	1137.085

Variable	IQR
C1 AGE (months)	41.0

Variable	Range
C1 AGE (months)	169.0



4. Bear Neck Circumference (Inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

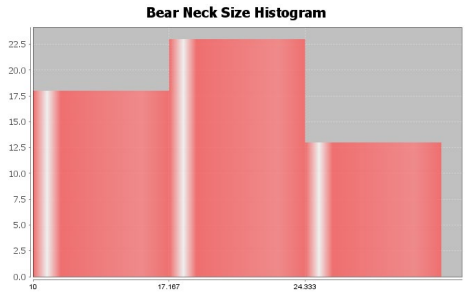
Standard Deviation Sentence: Typical bear neck sizes were 5.64 inches from the mean of 20.56 inches.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C6 Neck Circum (in)	20.556	5.641	31.818

Variable	IQR
C6 Neck Circum (in)	8.125

Variable	Range
C6 Neck Circum (in)	21.5



5. Bear Length (Inches): Skewed Left

Mean and Standard Deviation are NOT accurate (not bell shaped)

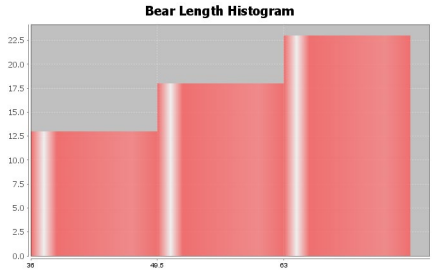
Standard Deviation Sentence: Typical bear lengths were 10.70 inches from the mean of 58.62 inches.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C7 Length (in)	58.617	10.701	114.509

Variable	IQR
C7 Length (in)	16.875

Variable	Range
C7 Length (in)	40.5



6. Bear Chest Size (Inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

Standard Deviation Sentence: Typical bear chest sizes were 9.4 inches from the mean of 35.7 inches.

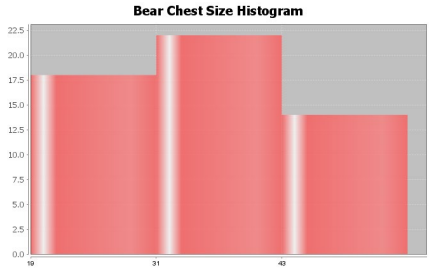
Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C8 Chest (in)	35.663	9.352	87.455

Variable	IQR
C8 Chest (in)	15.25

Variable	Range

C8 Chest (in)	36.0



7. Bear Weight (pounds): Skewed Right

Mean and Standard Deviation are NOT accurate (not bell shaped)

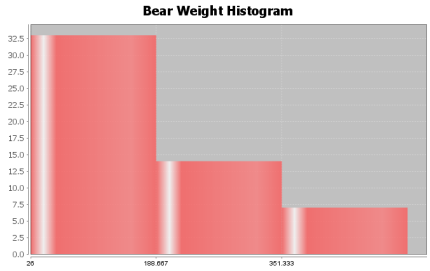
Standard Deviation Sentence: Typical bear lengths were 121.8 pounds from the mean of 182.9 pounds.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C9 Weight (Lbs)	182.889	121.801	14835.535

Variable	IQR
C9 Weight (Lbs)	158.0

Variable	Range
C9 Weight (Lbs)	488.0



8. Bear Head Length (inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

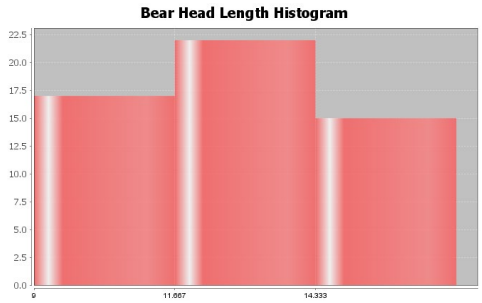
Standard Deviation Sentence: Typical bear head lengths were 2.14 inches from the mean of 12.95 inches.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C4 Head Length (In)	12.954	2.144	4.597

Variable	IQR
C4 Head Length (In)	3.0

Variable	Range
C4 Head Length (In)	8.0



9. Answers may vary (16,17,18,19,21,22,23,24) (This example checked with statistics software, see below)

Descriptive Statistics

Variable	Mean	Standard Deviation
3D#9	20.0	2.928

Variable	N total
3D#9	8

10. Answers may vary (9,10,11,12,13,27,28,29,30,31) (This example checked with statistics software, see below)

Descriptive Statistics

Variable	Mean	Standard Deviation
3D#10	20.0	9.603

Variable	N total
3D#10	10

Section 4E Answers

1a. Mean Average
1b. Mean Average
1c. Standard Deviation
1d. One Standard Deviation is Typical
1e. 68%
1f. Two Standard Deviations (or more) is considered unusual
1g. 2.5%
1h. 2.5%
1i. First calculate the unusual high cutoff by adding two standard deviations to the mean. Then look on the dotplot and see if any dots are higher than the cutoff.

1j. First calculate the unusual low cutoff by subtracting two standard deviations from the mean. Then look on the dotplot and see if any dots are lower than the cutoff.

2.
This data measured the lengths of the head of 54 bears in inches.
The data was bell shaped (normal).
The best measure of center was the mean of 12.95 inches. So the average length of the bear heads was 12.95 inches.
The best measure of spread was the standard deviation of 2.14 inches. This implies that typical bear head lengths were 2.14 inches from the mean. In fact, typical bear heads were between 10.81 inches and 15.10 inches.
There were no unusual values in the data set. The smallest bear head was 9 inches and the largest was 17 inches. Neither was unusual.

Unusual high cutoff = 17.24 (no values above 17.24)
Unusual low cutoff = 8.67 (no values below 8.67)

3.
This data measured the neck circumference of 54 bears in inches.
The data was bell shaped (normal).
The best measure of center was the mean of 20.56 inches. So the average bear neck circumference was 20.56 inches.
The best measure of spread was the standard deviation of 5.64 inches. This implies that typical bear neck sizes were 5.64 inches from the mean. In fact, typical bear neck circumferences were between 14.92 inches and 26.2 inches.
There were no unusual values in the data set. The smallest bear neck circumference was 10 inches and the largest was 31.5 inches. Neither was unusual.

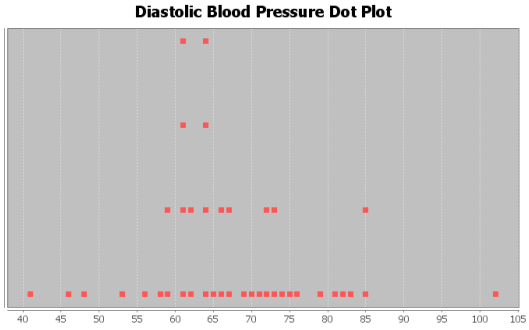
Unusual high cutoff = 31.84 (no values above 31.84)
Unusual low cutoff = 9.28 (no values below 9.28)

4.
This data measured the chest size of 54 bears in inches.
The data was bell shaped (normal).
The best measure of center was the mean of 35.66 inches. So the average chest size of the bears was 35.66 inches.
The best measure of spread was the standard deviation of 9.35 inches. This implies that typical bear chest sizes were 9.35 inches from the mean. In fact, typical bear chest sizes were between 26.31 inches and 45.01 inches.
The smallest bear chest size was 19 inches. This was not unusual. The largest bear chest size was 55 inches. This was unusually high. This was the only unusual value in the data set.

Unusual high cutoff = 54.36 (there was one value above 54.36)
Unusual low cutoff = 16.96 (no values below 16.96)

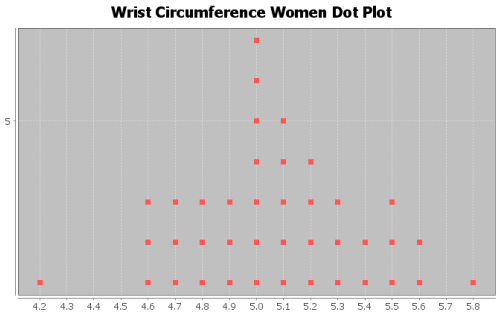
5.
This data measured the diastolic blood pressure of 40 women in (mm of mercury).
The data was bell shaped (normal).
The best measure of center was the mean of 67.4 mm of mercury. So the average diastolic blood pressure of these women was 67.4 mm of mercury.
The best measure of spread was the standard deviation of 11.6 mm of mercury. This implies that typical diastolic blood pressures were 11.6 mm of mercury from the mean. In fact, typical diastolic blood pressures for these women were between 55.8 and 79.0 mm of mercury.
The lowest diastolic blood pressure for these women was 41 mm of mercury. This was unusually low. The highest diastolic blood pressure was 102 mm of mercury. This was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 90.6 (there was one value (102) that was above 90.6)
Unusual low cutoff = 44.2 (there was one value (41) that was below 44.2)



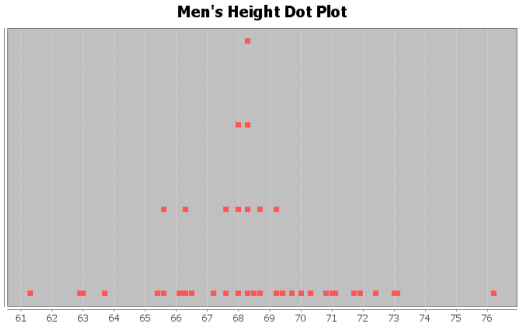
6.
This data measured the wrist circumference of 40 women in inches.
The data was bell shaped (normal).
The best measure of center was the mean of 5.07 inches. So the average wrist size of these women was 5.07 inches.
The best measure of spread was the standard deviation of 0.33 inches. This implies that typical wrist sizes for these women were 0.33 inches from the mean. In fact, typical wrist sizes for these women were between 4.74 and 5.40 inches.
The smallest wrist circumference for these women was 4.2 inches. This was unusually low. The largest wrist circumference was 5.8 inches. This was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 5.73 (there was one value (5.8) that was above 5.73)
Unusual low cutoff = 4.41 (there was one value (4.2) that was below 4.41)



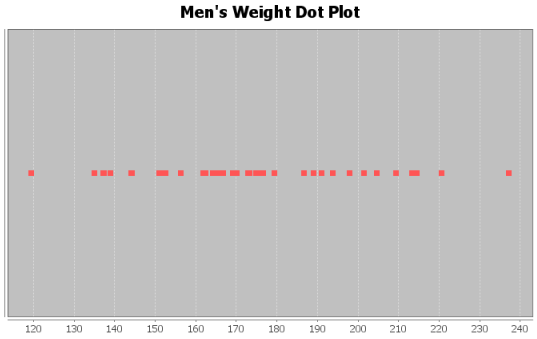
7.
This data measured the height of 40 men in inches.
The data was bell shaped (normal).
The best measure of center was the mean of 68.34 inches. So the average height of these men was 68.34 inches.
The best measure of spread was the standard deviation of 3.02 inches. This implies that typical heights of these men were 3.02 inches from the mean. In fact, typical heights for these men were between 65.32 inches and 71.36 inches.
The shortest man in the data was 61.3 inches. This height was unusually low. The tallest man in the data was 76.2 inches. This height was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 74.38 (there was only one value (76.2) that was above 74.38)
Unusual low cutoff = 62.30 (there was only one value (61.3) that was below 62.30)



8.
This data measured the weight of 40 men in pounds.
The data was bell shaped (normal).
The best measure of center was the mean of 172.55 pounds. So the average weight of these men was 172.55 pounds.
The best measure of spread was the standard deviation of 26.33 pounds. This implies that typical weights of these men were 26.33 pounds from the mean. In fact, typical weights for these men were between 146.22 pounds and 198.88 pounds.
The lightest man in the data was 119.5 pounds. This weight was unusually low. The heaviest man in the data was 237.1 pounds. This weight was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 225.21 (there was only one value (237.1) that was above 225.21)
Unusual low cutoff = 119.89 (there was only one value (119.5) that was below 119.89)



Answers for Chapter 4 Review Problems

1. Men's Diastolic BP
Shape: Skewed Left
Best Measure of Center (Best Average): Median

2. Men's Heights (inches)
Shape: Bell Shaped (Normal)
Best Measure of Center (Best Average): Mean

3. Men's Pulse Rates
Shape: Skewed Right
Best Measure of Center (Best Average): Median

4.
Mean = 216.1 / 13 ≈ 16.62

5.
Standard Deviation: How far typical values are from the mean in a bell shaped (normal) data set.

6.
The mean and standard deviation are only accurate if the data is bell shaped (normal).

7. Middle 68%
8. Top 2.5%
9. Bottom 2.5%

10. Bell Shaped (Normal)
11. 478 total students

12. Yes. The mean and standard deviation are accurate representations of center and spread because the data set is bell shaped (normal).

13. Average Math Intimidation score = 6.159 (mean)

14. Average Distance from the mean = 2.418 (standard deviation)

15. 3.741 ≤ typical math intimidation scores ≤ 8.577

16. Unusual High Cutoff = 10.995

17. Unusual Low Cutoff = 1.323

18. No. There are no unusually high values. (No values above the unusual high cutoff of 10.995)

19. None

20. Yes. There was one unusually low math intimidation score.

21. There are many people that answered 1. This was an unusually low value since it was below the unusually low cutoff of 1.323.

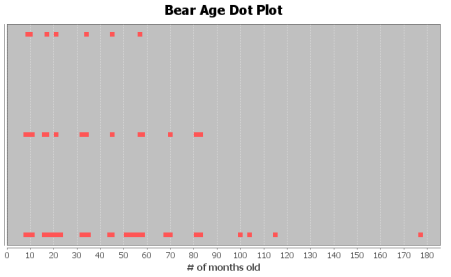
Introduction to Data Analysis (2nd edition)
Chapter 5 Answer Key

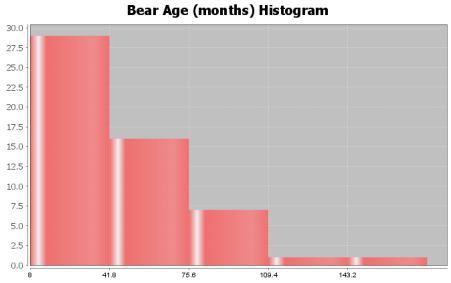
Section 5A Answers

1. Bear Ages (Months)

Shape: Skewed Right

Best Measure of Center: Median

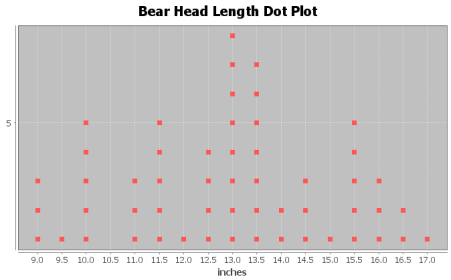


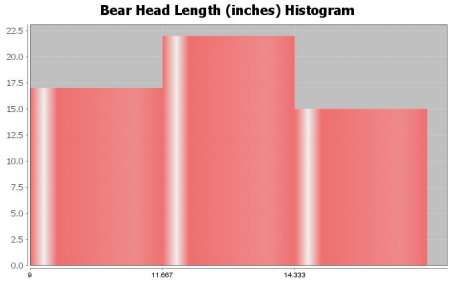


2. Head Length (inches)

Shape: Bell Shaped

Best Measure of Center: Mean

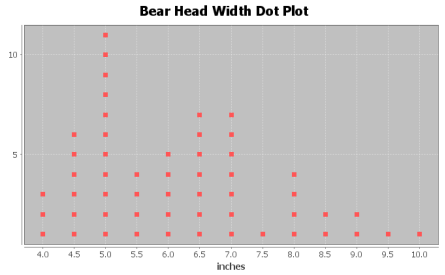


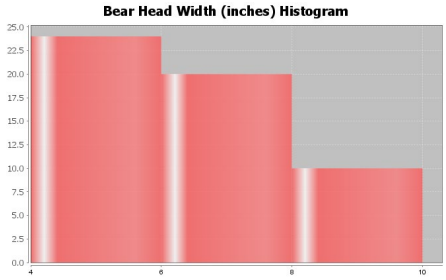


3. Bear Head Width (Inches)

Shape: Skewed Right

Best Measure of Center: Median

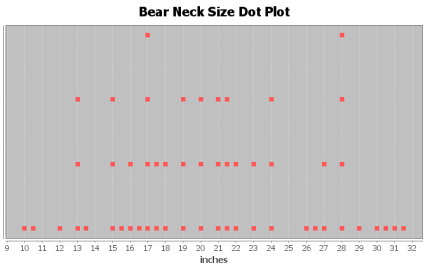


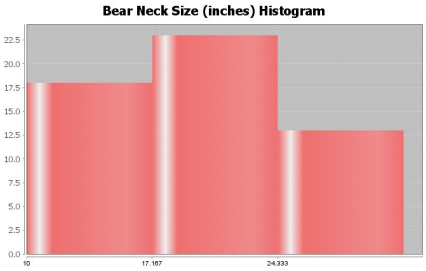


4. Bear Neck Size (inches)

Shape: Bell Shaped

Best Measure of Center: Mean

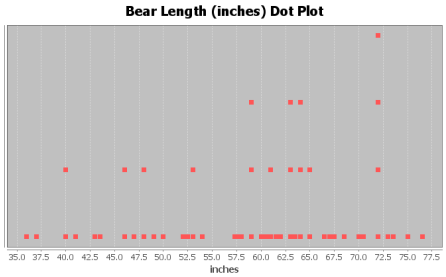


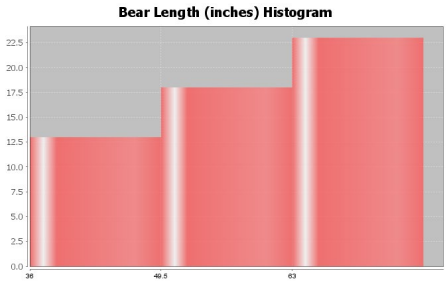


5. Bear Length (inches)

Shape: Skewed Left

Best Measure of Center: Median

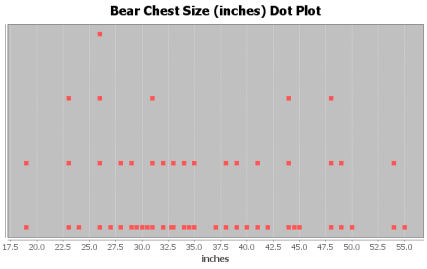


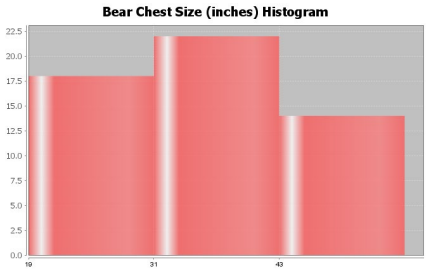


6. Bear Chest Size (inches)

Shape: Bell Shaped

Best Measure of Center: Mean

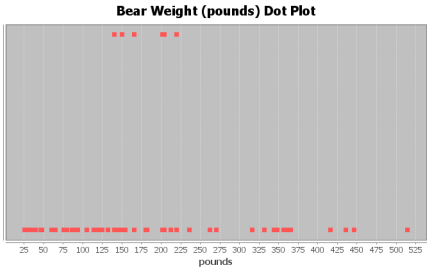


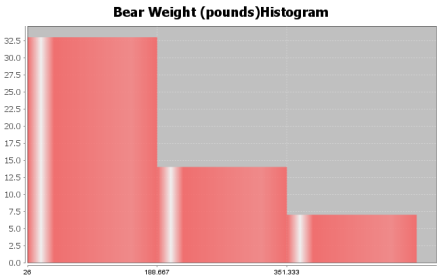


7. Bear Weight (pounds)

Shape: Skewed Right

Best Measure of Center: Median





Section 5B Answers

1a. Median = 17 (one number in middle)

1b. Median = (7.2 + 10.4)/2 = 8.8 (two numbers in middle)

1c. Median = 71 (one number in middle)

1d. Median = 157 (one number in middle)

1e. In order first: 1.9, 2.3, 2.8, 4.6, 6.1, 7.5, 8.3, 9.4
Median = (4.6 + 6.1)/2 = 5.35 (two numbers in middle)

1f. In order first: 18, 19, 20, 21, 23, 25, 26, 28, 29, 31, 32
Median = 25 (one number in middle)

2. Bear Data Medians

Descriptive Statistics

Variable	Median
C1 AGE (months)	34.0
C4 Head Length (In)	13.0
C5 Head Width (In)	6.0
C6 Neck Circum (in)	20.0
C7 Length (in)	60.75
C8 Chest (in)	34.0
C9 Weight (Lbs)	150.0

Section 5C Answers

1a. (answers may vary, these do not include the median in Q1 and Q3 calculation)
Median = 17
Q1 = (8+9)/2 = 8.5
Q3 = (26+29)/2 = 27.5
IQR = 27.5-8.5 = 19
Five Number Summary: 5, 8.5, 17, 27.5, 36

1b. (answers may vary, these do not include the median in Q1 and Q3 calculation)
Median = (7.2 + 10.4)/2 = 8.8
Q1 = 5.1
Q3 = 14.7
IQR = 14.7 – 5.1 = 9.6
Five Number Summary = 2.1 , 5.1 , 8.8 , 14.7 , 16.0

1c. (answers may vary, these do not include the median in Q1 and Q3 calculation)
Median = 71
Q1 = 41
Q3 = 88
IQR = 88 – 41 = 47
Five Number Summary = 31 , 41 , 71 , 88 , 103

1d. (answers may vary, these do not include the median in Q1 and Q3 calculation)
Median = 157
Q1 = (152+154)/2 = 153
Q3 = (163+164)/2 = 163.5
IQR = 163.5 – 153 = 10.5
Five Number Summary = 150 , 153 , 157 , 163.5 , 165

1e. (answers may vary, these do not include the median in Q1 and Q3 calculation)
In order first: 1.9, 2.3, 2.8, 4.6, 6.1, 7.5, 8.3, 9.4
Median = (4.6 + 6.1)/2 = 5.35
Q1 = (2.3 + 2.8)/2 = 2.55
Q3 = (7.5 + 8.3)/2 = 7.9
IQR = 7.9 – 2.55 = 5.35
Five Number Summary = 1.9 , 2.55 , 5.35 , 7.9 , 9.4

1f. (answers may vary, these do not include the median in Q1 and Q3 calculation)
In order first: 18, 19, 20, 21, 23, 25, 26, 28, 29, 31, 32
Median = 25
Q1 = 20
Q3 = 29
IQR = 29 – 20 = 9
Five Number Summary = 18 , 20 , 25 , 29 , 32

2. (Answers my vary depending on computer program used. These are from Statcato)

Descriptive Statistics

Variable	Q1	Median	Q3	IQR
C1 AGE (months)	17.0	34.0	58.0	41.0
C4 Head Length (In)	11.5	13.0	14.5	3.0
C5 Head Width (In)	5.0	6.0	7.0	2.0
C6 Neck Circum (in)	16.375	20.0	24.5	8.125
C7 Length (in)	49.75	60.75	66.625	16.875
C8 Chest (in)	28.75	34.0	44.0	15.25
C9 Weight (Lbs)	84.5	150.0	242.5	158.0

Variable	Min	Max
C1 AGE (months)	8.0	177.0
C4 Head Length (In)	9.0	17.0
C5 Head Width (In)	4.0	10.0
C6 Neck Circum (in)	10.0	31.5
C7 Length (in)	36.0	76.5
C8 Chest (in)	19.0	55.0
C9 Weight (Lbs)	26.0	514.0

2a. Bear Age (months) Five Number Summary: 8 , 17 , 34 , 58 , 177

2b. Bear Head Length (inches) Five Number Summary: 9 , 11.5 , 13 , 14.5 , 17

2c. Bear Head Width (inches) Five Number Summary: 4 , 5 , 6 , 7 , 10

2d. Bear Neck Size (inches) Five Number Summary: 10 , 16.375 , 20 , 24.5 , 31.5

2e. Bear Length (inches) Five Number Summary: 36 , 49.75 , 60.75 , 66.625 , 76.5

2f. Bear Chest Size (inches) Five Number Summary: 19 , 28.75 , 34 , 44 , 55

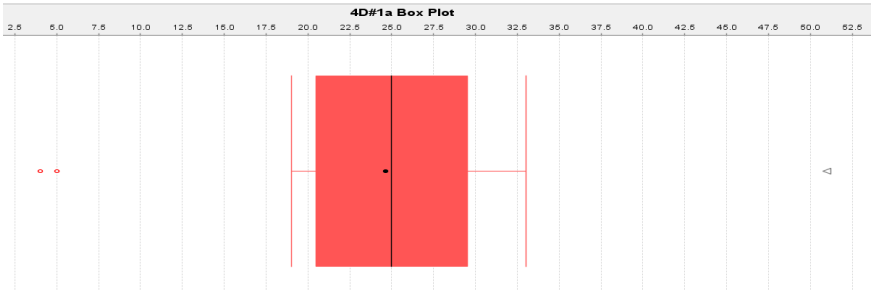
2g. Bear Weight (pounds) Five Number Summary: 26 , 84.5 , 150 , 242.5 , 514

Section 5D Answers

1a. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 25
Q1 = 20.5
Q3 = 29.5
IQR = 29.5 – 20.5 = 9
Unusual High Cutoff (for Skewed Data) = Q3 + (1.5xIQR) = 29.5 + (1.5x9) = 43
Unusual Low Cutoff (for Skewed Data) = Q1 – (1.5xIQR) = 20.5 – (1.5x9) = 7
Unusually High values: 51
Unusually Low values: 4 and 5

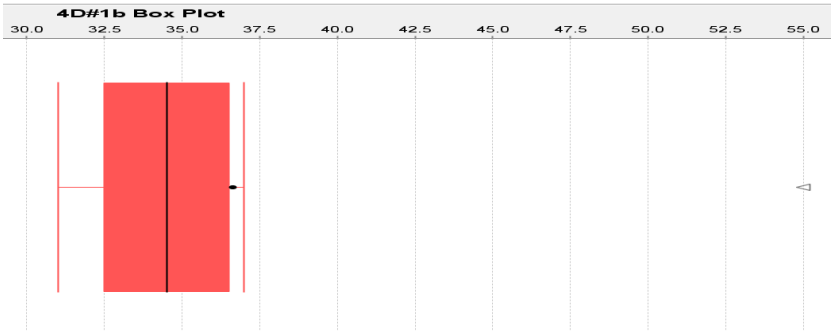
High Whisker (largest # that is not unusual): 33
Low Whisker (smallest # that is not unusual): 19



1b. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 34.5
Q1 = 32.5
Q3 = 36.5
IQR = 4
Unusual High Cutoff (for Skewed Data) = Q3 + (1.5xIQR) = 36.5 + (1.5x4) = 42.5
Unusual Low Cutoff (for Skewed Data) = Q1 – (1.5xIQR) = 32.5 – (1.5x4) = 26.5
Unusually High values: 55
Unusually Low values: none

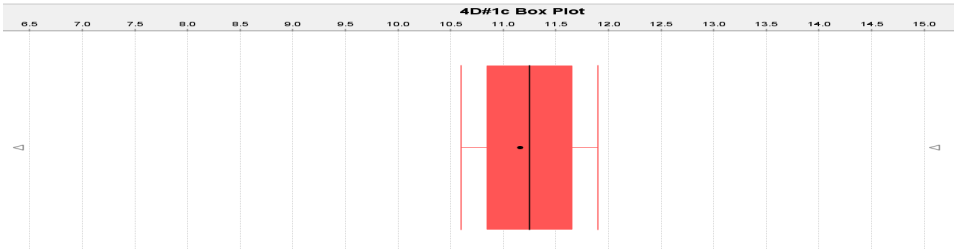
High Whisker (largest # that is not unusual): 37
Low Whisker (smallest # that is not unusual): 31



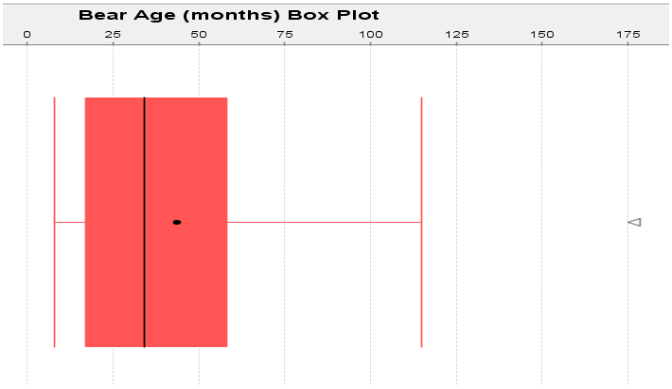
1c. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 11.25
Q1 = 10.85
Q3 = 11.65
IQR = 0.8
Unusual High Cutoff (for Skewed Data) = Q3 + (1.5xIQR) = 11.65 + (1.5x0.8) = 12.85
Unusual Low Cutoff (for Skewed Data) = Q1 – (1.5xIQR) = 10.85 – (1.5x0.8) = 9.65
Unusually High values: 15.1
Unusually Low values: 6.4

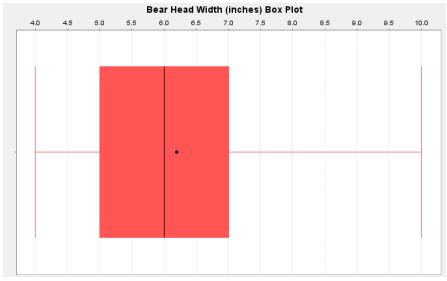
High Whisker (largest # that is not unusual): 11.9
Low Whisker (smallest # that is not unusual): 10.6



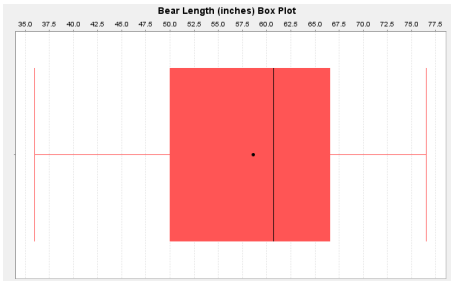
2a.


Bear Age (months) Box Plot
0
25
50
75
100
125
150
175

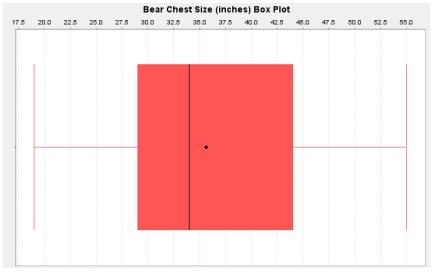
2b.


Bear Head Width (inches) Box Plot
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0

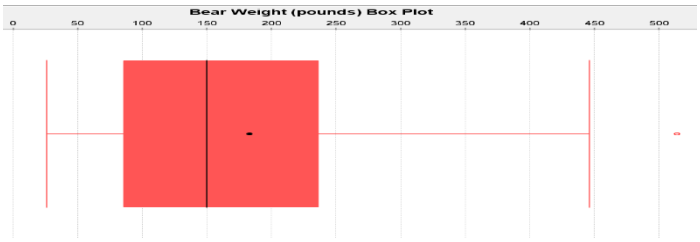
2c.


Bear Length (inches) Box Plot
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5

2d.


Bear Chest Size (inches) Box Plot
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0

2e.



3. Spring Bears *(These are approximate values from the graph.)*
Average Weight = Median ≈ 153 pounds

4. Spring Bears *(These are approximate values from the graph.)*
Typical Spread = IQR ≈ 28 pounds
137 pounds (Q1) ≤ Typical Spring Bear Weights ≤ 165 pounds (Q3)

5. Yes. There was one unusual value in the spring bear data.
Unusual value ≈ 197 pounds *(This is an approximate value from the graph.)*

Section 5E Answers

1a. Q1 is a measure of position.
1b. Mean is a measure of center.
1c. Variance is a measure of spread.
1d. Midrange is a measure of center.
1e. Standard Deviation is a measure of spread.
1f. Minimum value is a measure of position.
1g. Q3 is a measure of position.
1h. Mode is a measure of center.
1i. IQR is a measure of spread.
1j. Median is a measure of center.
1k. Range is a measure of spread.
1l. Maximum value is a measure of position.

2.

Mean: The mean is the center or average for bell shaped data sets that balances the distances. If this data was bell shaped we would use the mean of $1149.05 as the average.

Standard Deviation: The standard deviation measures how far typical values are from the mean in a bell shaped data set. If this data was bell shaped, then the typical values would be $516 from the mean.

Variance: Variance is a measure of spread used in ANOVA testing that is equal to the standard deviation squared.

Q1: The first quartile tells us that approximately 25% of the values in the data set are lower than $703.45.

Median: The median is a center or average when the data is in order. If this data set was skewed, we would use the median as our average. So if the data was skewed the average salary would be $1015.74.

Q3: The third quartile tells us that approximately 75% of the values in the data set are lower than $1496.11.

IQR: The interquartile range is the most accurate measure of spread for skewed data sets. It measures the spread for the middle 50% of the data. If this data was skewed, we would say that typical salaries are $792.66 from each other.

Mode: The mode is a measure of center that gives the number or numbers in the data that appear most often. There was no mode in the salary data since all the salaries appeared only once. "N for mode" tells us how many times the mode appears.

Min: The minimum is the smallest value in the data set. The lowest salary in the data was $371.57 and all other salaries in the data set are greater than $371.57.

Max: The maximum is the largest value in the data set. The highest salary in the data was $2396.28 and all other salaries in the data set are lower than $2396.28.

Range: The overall range of a data set is a quick measure of spread that is not very accurate because it does not measure typical values and may be influenced by unusual values. It is calculated by the subtracting the max and the min. The overall range of this data set was $2024.71. So all values in the data were within $2024.71 from each other.

N Total: The sample size or total frequency tells you how many numbers are in the data set. In this case there were 35 salaries in this data set.

Answers to Chapter 5 Review Sheet Problems

1. Shape = Skewed Right
Use the Median & IQR for center and spread.

2. Shape = Skewed Left
Use the Median & IQR for center and spread.

3. Shape = Bell Shaped (Normal)
Use the mean and standard deviation for center and spread.

4. (Answers may vary)
Median = (37 + 41)/2 = 78/2 = 39
Q1 = (26+28)/2 = 27
Q3 = (48+51)/2 = 49.5
IQR = Q3 – Q1 = 49.5 – 27 = 22.5

5. Interquartile Range (IQR): The interquartile range or IQR measures how far typical values are from each other in skewed data sets. IQR measure the spread for the middle 50% of the data values. To calculate IQR you subtract Q3 – Q1.

6. We should use the median as our center (average) and the IQR as our spread when the data is skewed (or not bell shaped).

7. Women's Cholesterol

8. Milligrams per deciliter (mg per dL)

9. Skewed Right

10. 38 numbers in data set

11. Yes. The median and IQR are accurate measures of center and spread because the data is skewed.

12. Average = 215 mg per dL (median)

13. Typical Distance from each other = 186.75 mg per dL (IQR)

14. 124.5 mg per dL (Q1) ≤ typical values ≤ 311.25 mg per dL (Q3)

15. No. No unusual low values on the boxplot.

16. Yes. The boxplot shows three unusually high values.

17. Unusual Values in the Data: 596 mg per dL, 600 mg per dL, and 920 mg per dL

18. About 75%

19. About 25%

20. About 50%

21. 531 mg per dL

22. False. There were the same amount of numbers.

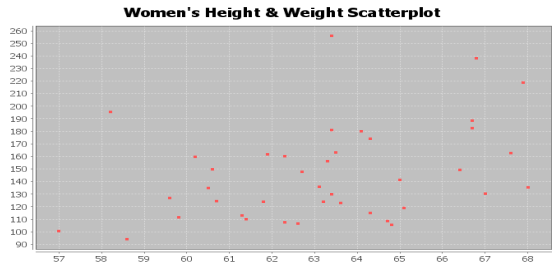
Introduction to Data Analysis
Chapter 6 Answer Key

Section 6A Answers

1.

Explanatory Variable (X): Height of woman
Response Variable (Y): Weight of woman

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a woman grows taller, her weight may increase. Weight changes may not indicate that a woman's height is changing.



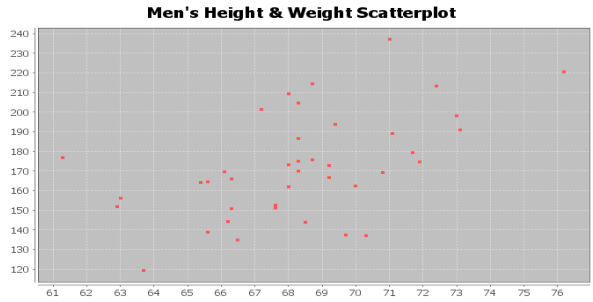
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There are a couple points at (58.2 in, 196 Lb) and (63.4 in, 256 Lb) that don't seem to fit the pattern and could be unusual points (outliers).

2.

Explanatory Variable (X): Height of man
Response Variable (Y): Weight of man

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a man grows taller, his weight may increase. Weight changes may not indicate that a man's height is changing.



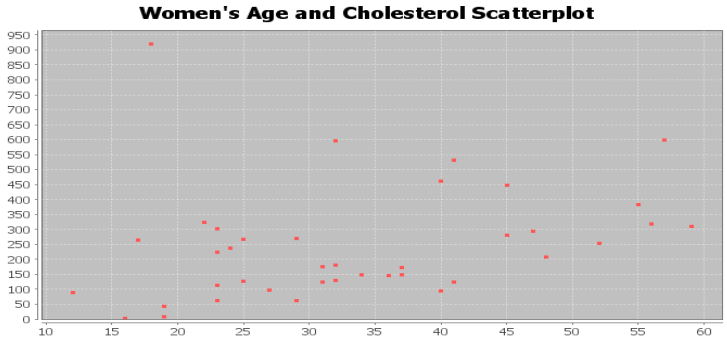
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There does not seem to be any points that are not following the linear pattern. No outliers.

3.

Explanatory Variable (X): Age of woman
Response Variable (Y): Cholesterol of woman

A woman's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



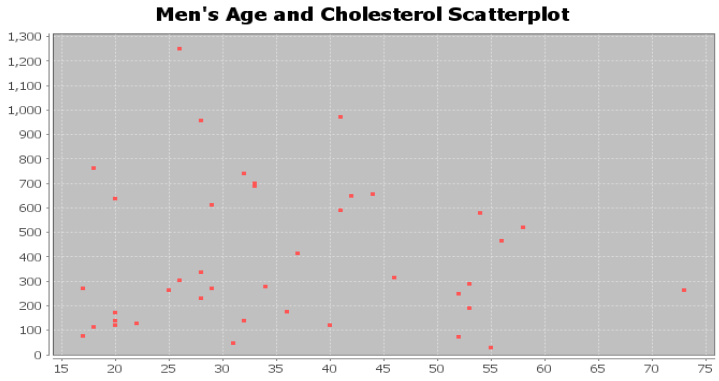
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There is one point at (18 years old, 920 mg per dL) that doesn't seem to fit the pattern and is unusual (outlier).

4.

Explanatory Variable (X): Age of man
Response Variable (Y): Cholesterol of man

A man's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



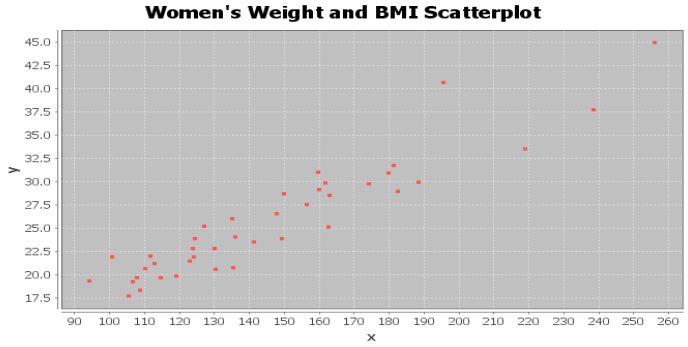
The scatterplot does not show any linear or curved trend. The dots seem to be all over without a distinguishable pattern. This indicates there is probably no relationship between men's age and cholesterol.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual.

5.

Explanatory Variable (X): Weight of woman
Response Variable (Y): Body Mass Index (BMI) of woman

Weight and Body Mass Index respond to each other, so we can chose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



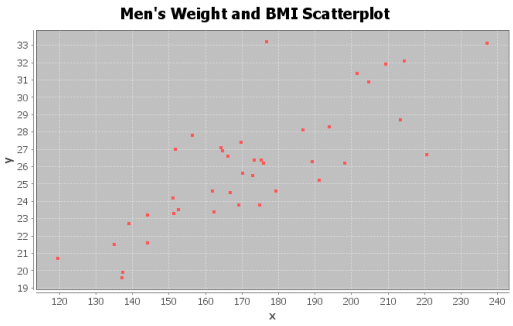
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There do not appear to be any outliers. All the points appear close to a line.

6.

Explanatory Variable (X): Weight of man
Response Variable (Y): Body Mass Index (BMI) of man

Weight and Body Mass Index respond to each other, so we can chose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



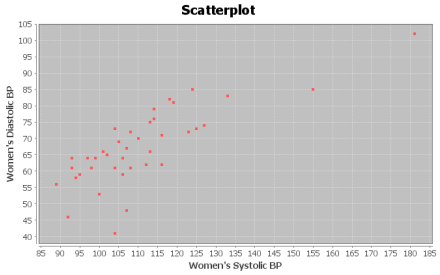
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There do not appear to be any outliers. All the points appear close to a line. The most unusual point was (178 Lbs ,33.2 kg/m^2), but this does not seem to be very far from the linear pattern.

7.

Explanatory Variable (X): Systolic Blood Pressure woman
Response Variable (Y): Diastolic Blood Pressure woman

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



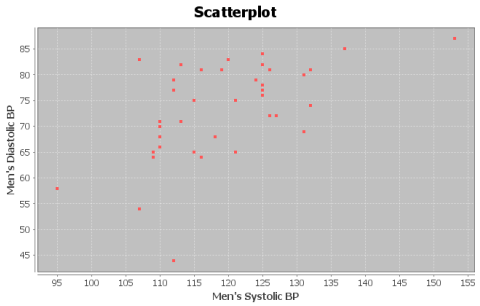
The scatterplot seems to show a positive linear trend. The dots to tend to increase from left to right and are close to a line.

There do not appear to be any outliers. All the points appear close to a line.

8.

Explanatory Variable (X): Systolic Blood Pressure man
Response Variable (Y): Diastolic Blood Pressure man

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



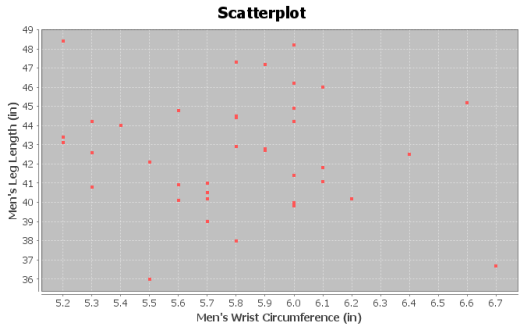
The scatterplot seems to show some positive linear trend. The dots to tend to increase from left to right and could be close to a line.

There seems to be one unusual point at (112 mg/dL , 44 mg/dL). This may be an outlier. Another point to consider is (107 mg/dL , 84 mg/dL) but this does not seem to be very far from the linear pattern.

9.

Explanatory Variable (X): Wrist Circumference man
Response Variable (Y): Leg Length man

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose leg length as the response variable (y) because I was interested in seeing if we can predict leg length from the wrist size.



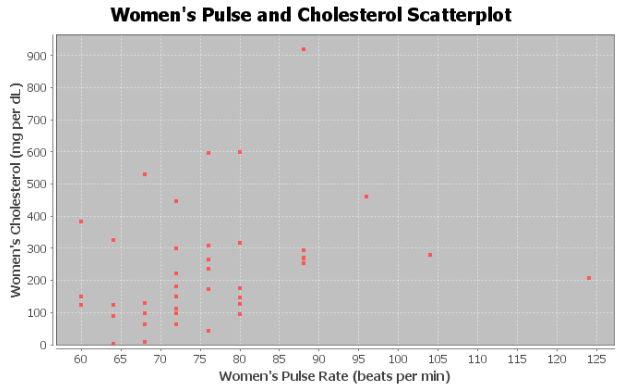
The scatterplot does not show any linear or curved trend. The dots seem to be all over without a distinguishable pattern. This indicates there is probably no relationship between men's wrist circumference and leg length.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual.

10.

Explanatory Variable (X): Pulse Rate Woman (beats per min)
Response Variable (Y): Cholesterol of Woman (mg per dL)

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. Checking cholesterol is more difficult as it requires a blood test, while pulse is relatively easy to check. I chose cholesterol as the response variable (y) because that is more difficult to measure and I was interested in seeing if we can predict cholesterol from pulse.



There does seem to be some positive linear trend, though the points are not as close to a line as I would like. The points show a slight upward trend from left to right.

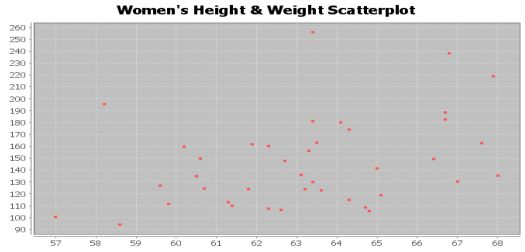
There seems to be one unusual point (outlier) at (88 BPM , 920 mg/dL). Another point to consider is (124 BPM, 201 mg/dL). This may also be unusual.

Section 6B Answers

1.

Explanatory Variable (X): Height of woman
Response Variable (Y): Weight of woman

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a woman grows taller, her weight may increase. Weight changes may not indicate that a woman's height is changing.



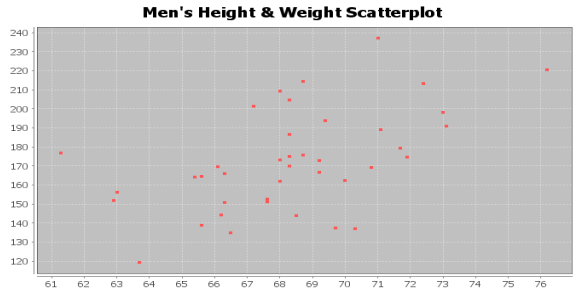
Correlation Coefficient r = +0.3644
This means that there is a weak positive linear correlation between the height and weight of the women in the data set.

There are a couple points at (58.2 in, 196 Lb) and (63.4 in, 256 Lb) that don't seem to fit the pattern and could be unusual points (outliers). The correlation coefficient r is not very strong, indicating that these outliers are influential.

2.

Explanatory Variable (X): Height of man
Response Variable (Y): Weight of man

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a man grows taller, his weight may increase. Weight changes may not indicate that a man's height is changing.



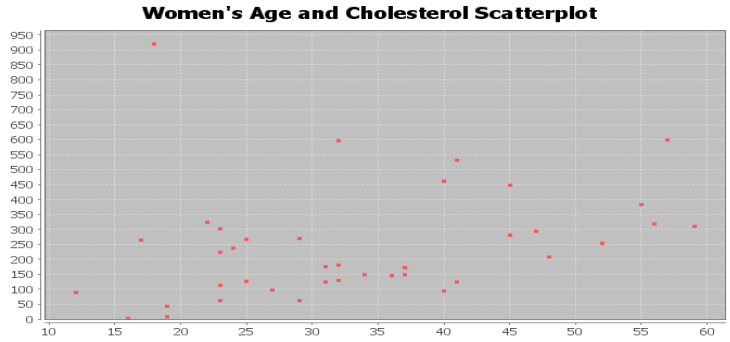
Correlation Coefficient r = +0.5222
This tells us that there is a moderate positive correlation between the height and weight of the men in the data set.

There does not seem to be any points that are not following the linear pattern. No outliers. The correlation coefficient r being moderately high confirms that there are probably no influential outliers.

3.

Explanatory Variable (X): Age of woman
Response Variable (Y): Cholesterol of woman

A woman's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



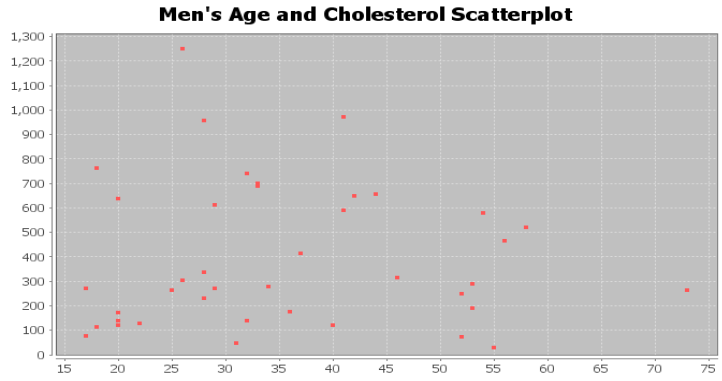
Correlation Coefficient r = +0.3022
This means that there is a weak positive linear correlation between the age and cholesterol of the women in the data set.

There is one point at (18 years old, 920 mg per dL) that doesn't seem to fit the pattern and is unusual (outlier). The correlation coefficient r is not very strong, indicating that this outlier is influential.

4.

Explanatory Variable (X): Age of man
Response Variable (Y): Cholesterol of man

A man's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



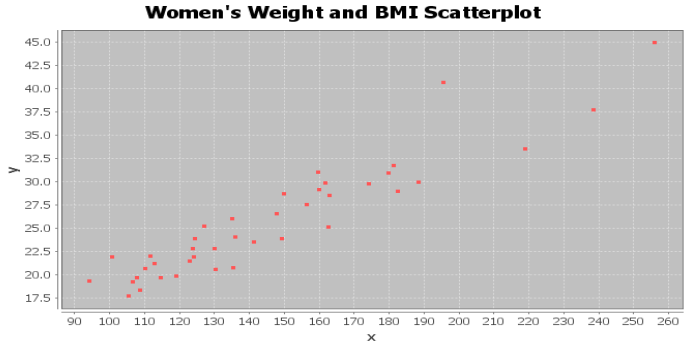
Correlation Coefficient r = -0.0154
The correlation coefficient is close to zero, so there is no correlation between men's age and cholesterol.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual. The correlation coefficient confirms this. There is no correlation.

5.

Explanatory Variable (X): Weight of woman
Response Variable (Y): Body Mass Index (BMI) of woman

Weight and Body Mass Index respond to each other, so we can chose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



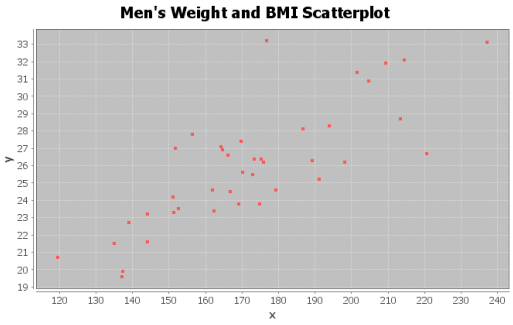
Correlation Coefficient r = +0.9361
There is a very strong positive correlation between the weight and BMI of the women in the data set.

There do not appear to be any outliers. All the points appear close to a line. The correlations coefficient being so strong indicates there are no outliers.

6.

Explanatory Variable (X): Weight of man
Response Variable (Y): Body Mass Index (BMI) of man

Weight and Body Mass Index respond to each other, so we can chose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



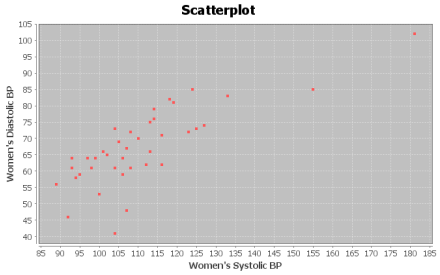
Correlation Coefficient r = +0.7997
There is a strong positive correlation between the weight and BMI of the men in the data set.

There do not appear to be any outliers. All the points appear close to a line. The most unusual point was (178 Lbs ,33.2 kg/m^2), but this does not seem to be very far from the linear pattern. The correlation coefficient confirms this since it is strong. So this possible outlier is not influential.

7.

Explanatory Variable (X): Systolic Blood Pressure woman
Response Variable (Y): Diastolic Blood Pressure woman

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



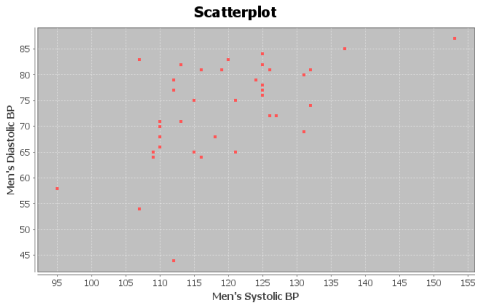
Correlation Coefficient r = +0.7854
This indicates that there is a strong positive correlation between the systolic and diastolic blood pressure of the women in the data set.

There do not appear to be any outliers. All the points appear close to a line. The correlations coefficient r confirms this since it is strong. There are no influential outliers.

8.

Explanatory Variable (X): Systolic Blood Pressure man
Response Variable (Y): Diastolic Blood Pressure man

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



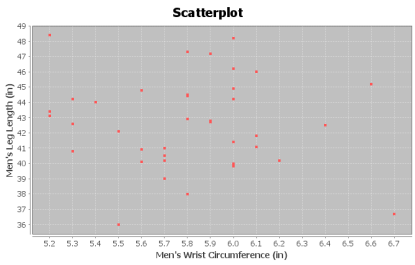
Correlation Coefficient r = +0.5517
This tells us there is a moderate positive correlation between the systolic and diastolic blood pressure of the men in this data set.

There seems to be one unusual point at (112 mg/dL , 44 mg/dL). This may be an outlier. Another point to consider is (107 mg/dL , 84 mg/dL) but this does not seem to be very far from the linear pattern. The correlation coefficient is only moderate and not strong. The (112 , 44) may be having a small influence on the correlation.

9.

Explanatory Variable (X): Wrist Circumference man
Response Variable (Y): Leg Length man

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose leg length as the response variable (y) because I was interested in seeing if we can predict leg length from the wrist size.



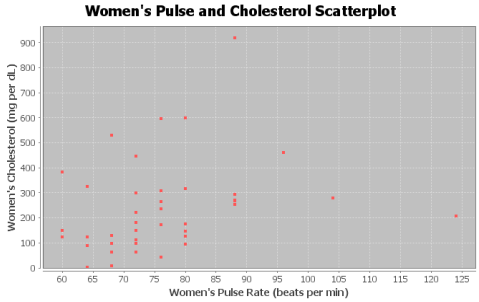
Correlation Coefficient r = -0.0789
The correlation coefficient is close to zero, so there is no correlation between the wrist circumference and leg length of the men in the data set.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual. The correlation coefficient confirms this. There is no correlation.

10.

Explanatory Variable (X): Pulse Rate Woman (beats per min)
Response Variable (Y): Cholesterol of Woman (mg per dL)

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. Checking cholesterol is more difficult as it requires a blood test, while pulse is relatively easy to check. I chose cholesterol as the response variable (y) because that is more difficult to measure and I was interested in seeing if we can predict cholesterol from pulse.



Correlation Coefficient r = +0.2659
This tells us that there is weak positive correlation between the pulse and cholesterol of the women in the data set.

There seems to be one unusual point (outlier) at (88 BPM , 920 mg/dL). Another point to consider is (124 BPM, 201 mg/dL). This may also be unusual. The correlation coefficient is very weak indicating these are both influential outliers.

Section 6C Answers

1a. 79.0% of the variability in the men's weight can be explained by the linear relationship with the waist size. This tells us there is a very strong relationship between the variables.

1b. 0.31% of the variability in the men's weight can be explained by the linear relationship with the pulse rate. This tells us there is no relationship between these variables.

1c. 12.4% of the variability in the men's weight can be explained by the linear relationship with the systolic blood pressure. This tells us there is a weak relationship between these variables.

1d. 15.0% of the variability in the men's weight can be explained by the linear relationship with the diastolic blood pressure. This tells us there is a weak relationship between these variables.

1e. 0.07% of the variability in the men's weight can be explained by the linear relationship with the cholesterol. This tells us there is no relationship between these variables.

1f. 64% of the variability in the men's weight can be explained by the linear relationship with the body mass index. This tells us there is a very strong relationship between the variables.

1g. 13.8% of the variability in the men's weight can be explained by the linear relationship with the leg length. This tells us there is a weak relationship between these variables.

1h. 40.3% of the variability in the men's weight can be explained by the linear relationship with the elbow circumference. This tells us there is a moderate relationship between these variables.

1i. 27.0% of the variability in the men's weight can be explained by the linear relationship with the wrist circumference. This tells us there is a moderate relationship between these variables.

1j. 67.5% of the variability in the men's weight can be explained by the linear relationship with the arm length. This tells us there is a very strong relationship between the variables.

2.

r^2 = (0.7287)^2 = 0.531 = 53.1%

53.1% of the variability in total trash can be explained by the linear relationship with paper trash.

Confounding Variables (answers may vary): metal trash, food trash, plastic trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does <u>not</u> imply that one variable causes another. There are many factors involved.

3.

r^2 = (0.5862)^2 = 0.344 = 34.4%

34.4% of the variability in metal trash can be explained by the linear relationship with plastic trash.

Confounding Variables (answers may vary): paper trash, food trash, total trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does <u>not</u> imply that one variable causes another. There are many factors involved.

4.

r^2 = (0.5833)^2 = 0.340 = 34.0%

34.0% of the variability in total trash can be explained by the linear relationship with food trash.

Confounding Variables (answers may vary): metal trash, paper trash, plastic trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

5.

r^2 = (-0.8713)^2 = 0.759 = 75.9%

75.9% of the variability in miles per gallon can be explained by the linear relationship with horsepower.

Confounding Variables (answers may vary): weight of car, type of gas, type of engine, type of carborator, freeway or road driving, wind resistance

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

6.

r^2 = (0.9404)^2 = 0.884 = 88.4%

88.4% of the variability in profit can be explained by the linear relationship with the number of cars sold.

Confounding Variables (answers may vary): costs of company, number of cars available, type of cars, area, talent of the sales employees, number of employees

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

7.

r^2 = (0.6727)^2 = 0.453 = 45.3%

45.3% of the variability in number of flowers can be explained by the linear relationship with amount of fertilizer.

Confounding Variables (answers may vary): type of flowers, weather, climate, temperature, area, amount of carbon dioxide, quality of soil before fertilizer was added

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

8.

r^2 = (-0.9429)^2 = 0.889 = 88.9%

88.9% of the variability in this stock price can be explained by the linear relationship with the number of weeks (time).

Confounding Variables (answers may vary): type of stock, overall stock market trends, national debt, unemployment rates

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

9.

Men's Body Mass Index multivariable study

Age/BMI: $r^2 = 0.071 = 7.1\%$

Height/BMI: $r^2 = 0.008 = 0.8\%$

Weight/BMI: $r^2 = 0.640 = 64.0\%$

Waist/BMI: $r^2 = 0.731 = 73.1\%$

Cholesterol/BMI: $r^2 = 0.012 = 1.2\%$

Follow up: Waist size had the strongest relationship with body mass index. Weight also had a strong relationship. A study of body mass index should focus on waist size and weight. Age had a weak relationship. Height and Cholesterol had virtually no relationship with BMI.

Surprises (answers may vary): Height is used in the calculation of body mass index, yet the correlation study indicated no relationship. This was surprising.

10:

Bear Weight multivariable study

Bear Age / Bear Weight: $r^2 = 0.561 = 56.1\%$

Bear Head Length / Bear Weight: $r^2 = 0.696 = 69.6\%$

Bear Head Width / Bear Weight: $r^2 = 0.614 = 61.4\%$

Bear Neck Size / Bear Weight: $r^2 = 0.873 = 87.3\%$

Bear Length / Bear Weight: $r^2 = 0.747 = 74.7\%$

Bear Chest Size / Bear Weight: $r^2 = 0.928 = 92.8\%$

Follow up: Chest Size of the bear had the strongest relationship with weight. Neck size and overall length had very strong relationships with weight also. Age, Head width and head length also had strong relationships with weight.

Surprises (answers may vary): All of the variables were related to weight. There were not any variables that were not related to the weight and they were all pretty strong relationships.

Section 6D Answers

1.

Slope = r times Sy / Sx = 0.7287 x 12.46 / 4.268 = 2.1784

Slope Sentence: For every one ton increase in paper trash, the total trash increases about 2.178 tons.

Y int = y mean – (slope)(x mean) = 27.44 – (2.1784 x 9.428) = 27.44 – 20.0538 = 6.902

Y int Sentence: If there was zero tons of paper trash, there would still be about 6.902 tons of total trash.

Equation of Regression Line: Y = 6.902 + 2.178 X

2.

Slope = r times Sy / Sx = 0.5862 x 1.091 / 1.065 = 0.6005 = 0.601

Slope Sentence: For every one ton increase in plastic trash, the metal trash increases about 0.6 tons.

Y int = y mean – (slope)(x mean) = 2.218 – (0.6005x1.911) = 2.218 – 1.14755 = 1.07

Y int Sentence: If there was zero tons of plastic trash, there would still be about 1.07 tons of metal trash.

Equation of Regression Line: Y = 1.07 + 0.601 X

3.

Slope = r times Sy / Sx = 0.5833 x 12.46 / 3.297 = 2.204

Slope Sentence: For every one ton increase in food trash, the total trash increases about 2.204 tons.

Y int = y mean – (slope)(x mean) = 27.44 – (2.204x4.816) = 27.44 – 10.614 = 16.826

Y int Sentence: If there was zero tons of food trash, there would still be about 16.826 tons of total trash.

Equation of Regression Line: Y = 16.826 + 2.204 X

4.

Slope = r times Sy / Sx = 0.9404 x 175.615 / 7.512 = 21.985

Slope Sentence: For every one car sold, the profits increases about 21.985 thousand dollars ($21985).

Y int = y mean – (slope)(x mean) = 420.25 – (21.985x21.667) = 420.25 – 476.349 = -56.099

Y int Sentence: If there was zero cars sold, the profits would be about -56.099 thousand dollars (loss of $56099).

Equation of Regression Line: Y = -56.099 + 21.985 X

5.

Slope = r times Sy / Sx = 0.6727 x 1.356 / 1.165 = 0.783

Slope Sentence: For every one pound of fertilizer added, the number of flowers per square foot increase about 0.783.

Y int = y mean – (slope)(x mean) = 13.867 – (0.783x3.387) = 13.867 – 2.652 = 11.215

Y int Sentence: If there was zero fertilizer, there would still be about 11.215 flowers per square foot.

Equation of Regression Line: Y = 11.215 + 0.783 X

6.

Slope = r times Sy / Sx = -0.9429 x 17.031 / 5.916 = -2.714

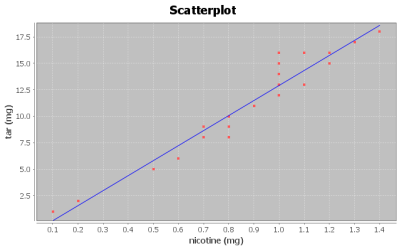
Slope Sentence: For every one week that goes by, the stock price decreases about $2.71 on average.

Y int = y mean – (slope)(x mean) = 270.6 – (-2.714x10.5) = 270.6 – (-28.497) = 270.6 + 28.497 = 299.097

Y int Sentence: At week zero, the stock price was $299.10 per share.

Equation of Regression Line: Y = 299.097 – 2.714 X

7.



Correlation Coefficient r = 0.9614
There is a very strong positive correlation between the amount of nicotine and tar. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting tar.

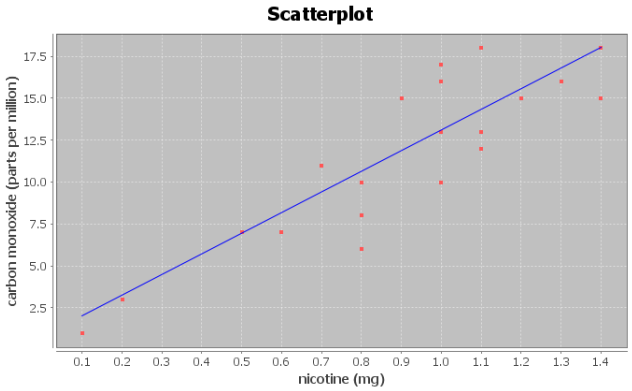
Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = -1.2713$
$b_1 = 14.2076$

Y-intercept = -1.2713
If there was zero mg of nicotine, then the amount of tar would be about -1.2713 mg.
Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict tar when nicotine is zero.

Slope = 14.2076
For every 1 mg of nicotine added to a cigarette, they add 14.2076 mg of tar.

Regression Line Equation: Y = -1.2713 + 14.2076 X

8.



Correlation Coefficient r = 0.8633
There is a strong positive correlation between the amount of nicotine and carbon monoxide. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting carbon monoxide.

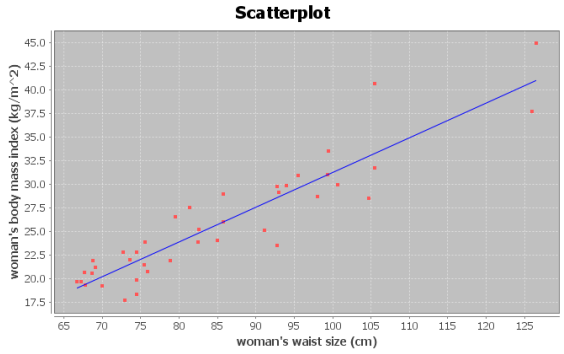
Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 0.7950$
$b_1 = 12.3057$

Y-intercept = 0.7950
If a cigarette with zero mg of nicotine is lighted, it will still release 0.7950 ppm of carbon monoxide.

Slope = 12.3057
For every 1 mg of nicotine added to a cigarette, the amount of carbon monoxide increases 12.3057 ppm.

Regression Line Equation: $Y = 0.7950 + 12.3057 X$

9.



Correlation Coefficient r = 0.9181
There is a very strong positive correlation between the women's waist size and body mass index in the data set. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting body mass index from waist size.

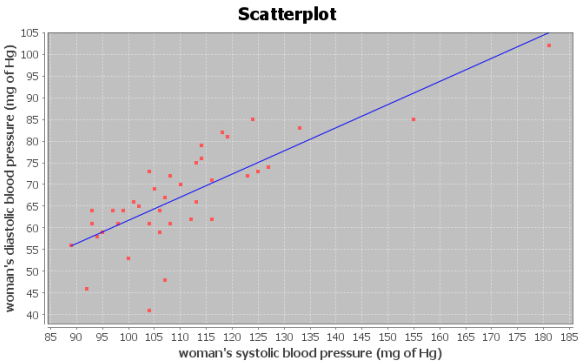
Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = -5.5117$
$b_1 = 0.3675$

Y-intercept = -5.5117
If a woman had a waist size of zero centimeters, she would have a predicted body mass index of -5.5117.
Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict BMI for a waist size of zero. A waist size of zero and a body mass index of negative 5.5 are both impossible.

Slope = 0.3675
For every 1 cm increase in waist size, the women's body mass index increases 0.3675.

Regression Line Equation: $Y = -5.5117 + 0.3675 X$

10.



Correlation Coefficient r = 0.7854
There is a strong positive correlation between the women's systolic and diastolic blood pressure in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting diastolic BP from systolic BP.

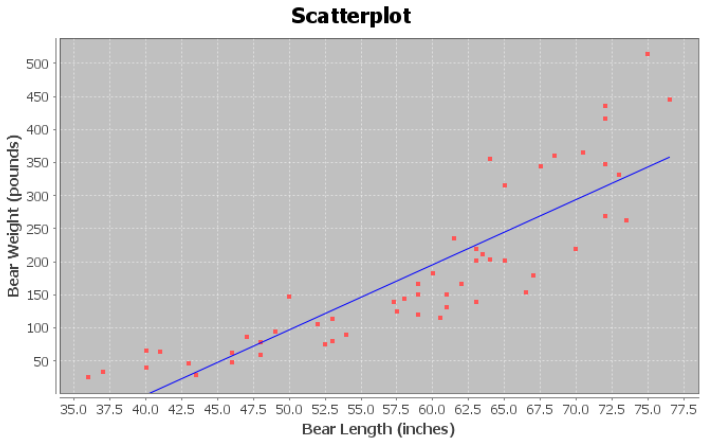
Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 8.3079$
$b_1 = 0.5335$

Y-intercept = 8.3079
If a woman had a systolic blood pressure of zero mm of Hg, she would have a predicted diastolic blood pressure of 8.3079 mm of Hg.
Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict diastolic blood pressure for a systolic blood pressure of zero. A living person cannot have a blood pressure of zero.

Slope = 0.5335
For every 1 mm of Hg increase in systolic blood pressure, the women's diastolic blood pressure increases 0.5335 mm of Hg.

Regression Line Equation: $Y = 8.3079 + 0.5335 X$

11.



Correlation Coefficient r = 0.8644
There is a strong positive correlation between the length and weight of the bears. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting bear weights.

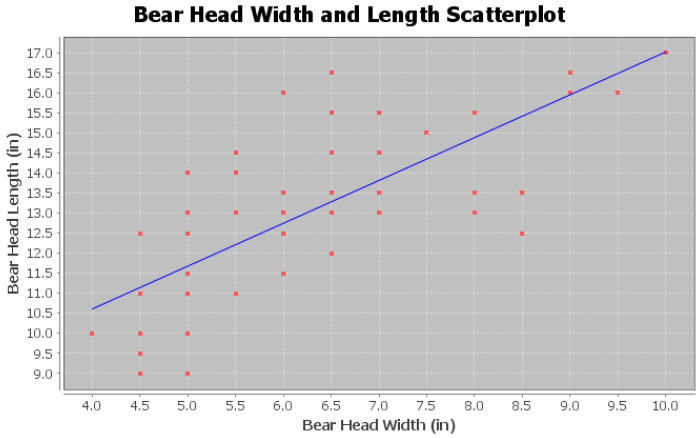
Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = -393.8391$
$b_1 = 9.8390$

Y-intercept = -393.8391
If a bear has a length of zero inches, then the bear would have a predicted weight of -393.8391 pounds.
Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict a bear weight from a length of zero. A bear length of zero and a weight of -393.8391 pounds are both impossible.

Slope = 9.8390
For every 1 inch longer a bear gets, the weight of the bear increases about 9.8390 pounds.

Regression Line Equation: $Y = -393.8391 + 9.8390 X$

12.



Correlation Coefficient r = 0.7535
There is a strong positive correlation between the head width and head length of the bears. The regression line fits the data very well with no influential outliers. The regression line will be accurate for predicting bear head lengths from the bear head width.

Regression:
Regression equation $Y = b_0 + b_1x$
$b_0 = 6.3362$
$b_1 = 1.0683$

Y-intercept = 6.3362
If a bear has a head width of zero inches, then the bear would have a predicted head length of 6.3362 inches.
Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict a bear head length from a head width of zero. A bear head width of zero is impossible.

Slope = 1.0683
For every 1 inch wider a bear's head gets, the head length of the bear increases about 1.0683 inches.

Regression Line Equation: $Y = 6.3362 + 1.0683 X$

Section 6E Answers

1.

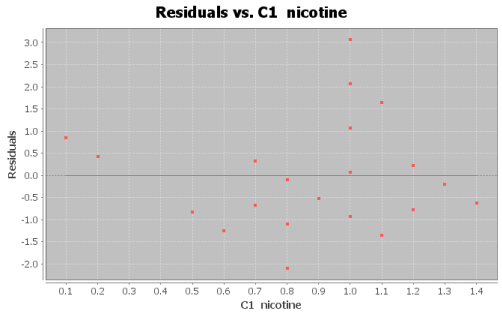
Se = 1.2984 mg of tar

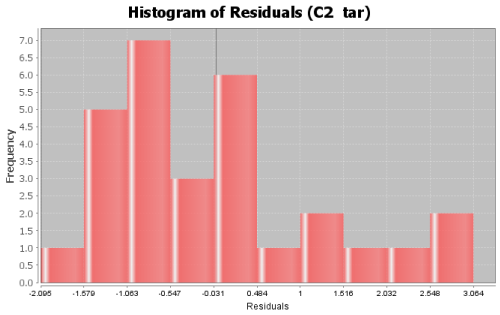
The points in the scatterplot are 1.2984 mg of tar from the regression line on average.

If we use a nicotine value in the scope of the x values and regression line to predict the amount of tar a cigarette has, our prediction could have an average error of 1.2984 mg of tar.

The residual plot does not look evenly spread out. It looks "V" shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram is not bell shaped (skewed right). It is also not centered at zero.





2.

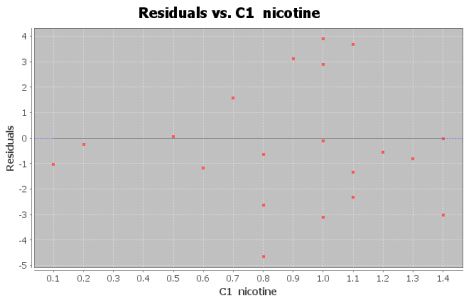
Se = 2.2961 PPM

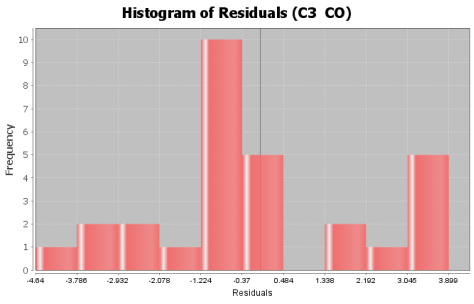
The points in the scatterplot are 2.2961 parts per million (ppm) from the regression line on average.

If we use a nicotine value in the scope of the x values and regression line to predict the amount of carbon monoxide a cigarette releases when burned, our prediction could have an average error of 2.2961 ppm.

The residual plot does not look evenly spread out. It looks "V" shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram does look relatively bell shaped. However it is not centered at zero.





3.

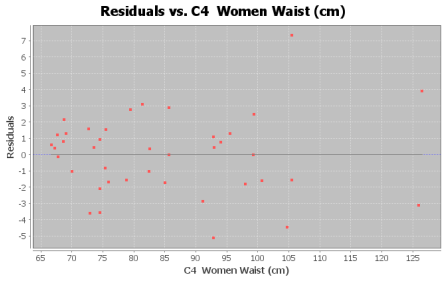
Se = 2.4761 kg/m^2

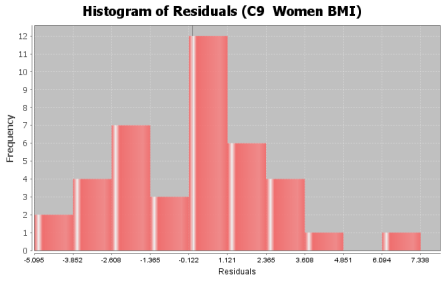
The points in the scatterplot are 2.4761 kg/m^2 from the regression line on average.

If we use a woman's waist size in the scope of the x values and regression line to predict the woman's body mass index, our prediction could have an average error of 2.4761 kg/m^2.

Overall the residual plot is pretty evenly spread out. There is one point that is far away from the regression line in the top right section of the residual plot that does make the graph look slightly "V" shaped (fan shaped).

The histogram does look relatively bell shaped. It is also centered at zero.





4.

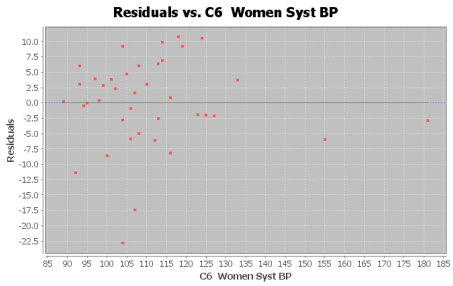
Se = 7.2912 mm of Hg

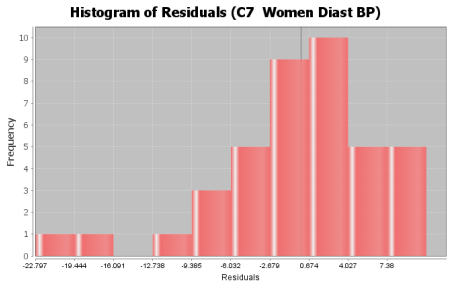
The points in the scatterplot are 7.2912 mm of Hg from the regression line on average.

If we use a woman's systolic blood pressure in the scope of the x values and regression line to predict the woman's diastolic blood pressure, our prediction could have an average error of 7.2912 mm of Hg.

The residual plot does not look evenly spread out. It looks "V" shaped (fan shaped). It is more spread out on the left side of the graph and less spread out on the right side of the graph.

The histogram does not look bell shaped (skewed left). The center is a little off from zero.





5.

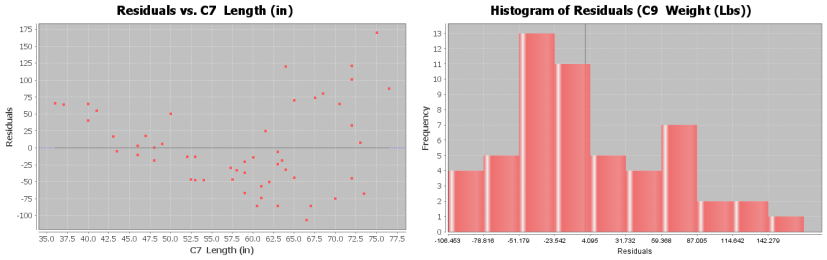
Se = 61.8272 pounds

The points in the scatterplot are 61.8272 pounds from the regression line on average.

If we use a bears length in the scope of the x values and regression line to predict the bears weight, our prediction could have an average error of 61.8272 pounds.

The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram does not look bell shaped (skewed right). The center is a little off from zero.



6.

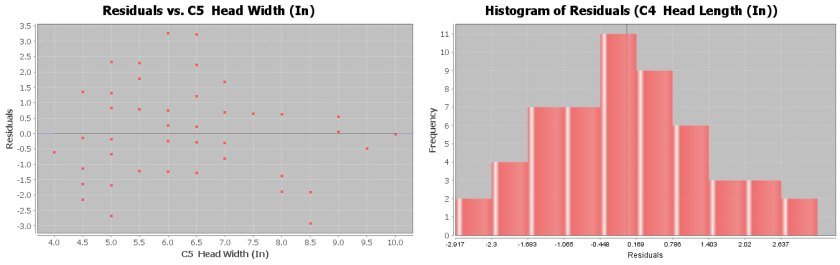
Se = 1.4231 inches

The points in the scatterplot are 1.4231 inches from the regression line on average.

If we use a bears head width in the scope of the x values and regression line to predict the bears head length, our prediction could have an average error of 1.4231 inches.

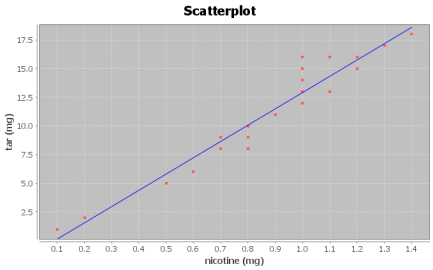
The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the left side of the graph and less spread out on the right side of the graph.

The histogram does look bell shaped. The histogram also looks centered at zero.



Section 6F Answers

1a.



Correlation Coefficient r = 0.9614
There is a very strong positive correlation between the amount of nicotine and tar. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting tar.

1b.

Regression Line Equation: Y = -1.2713 + 14.2076 X

1c.

0.1 mg nicotine ≤ scope of X values ≤ 1.4 mg of nicotine

Zero is an extrapolation since it does not fall in the scope of the X values. This is why the Y intercept -1.2713 does not make sense. You cannot have a negative amount of tar. The formula is not designed to plug in zero for x. Not surprising the predicted Y value from an extrapolation can be dramatically wrong.

1d.

Prediction for X = 0.8 mg nicotine:
Y = -1.2713 + 14.2076 X

Y = -1.2713 + 14.2076 (0.8)

Y = -1.2713 + 11.36608

Y = 10.09478 ≈ 10.1 mg of tar

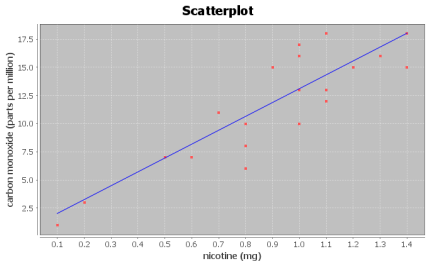
A cigarette with 0.8 mg of nicotine would be predicted to have 10.1 mg of tar. This prediction could have an average error of 1.2984 mg of tar (Se). So our prediction of 10.1 mg of tar could be about 1.3 mg too high or too low.

1e.

Prediction for X = 4.75 mg nicotine:
We should not use this formula to predict the amount of tar for a cigarette that has 4.75 mg of nicotine. The formula was not designed to plug in 4.75 for X. 4.75 is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

1f. Answers may vary. Tar is a dangerous substance to put into the body. Cigarette smoking has been linked to lung cancer and other diseases.

2a.



Correlation Coefficient r = 0.8633
There is a strong positive correlation between the amount of nicotine and carbon monoxide. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting carbon monoxide.

2b.

Regression Line Equation: $Y = 0.7950 + 12.3057 X$

2c.

0.1 mg nicotine ≤ scope of X values ≤ 1.4 mg of nicotine

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept may not make sense. Though the formula is not designed to plug in zero for x, the number may have some meaning here. If a cigarette with zero mg of nicotine is lighted, it will still release 0.7950 ppm of carbon monoxide.

2d.

Prediction for X = 1.2 mg nicotine:
$Y = 0.7950 + 12.3057 X$

$Y = 0.7950 + 12.3057 (1.2)$

$Y = 0.7950 + 14.76684$

$Y = 15.56184 \approx 15.6$ PPM of carbon monoxide

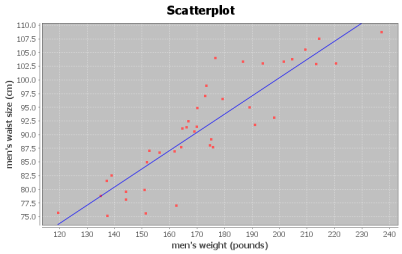
A cigarette that has 1.2 mg of nicotine would be predicted to release 15.6 ppm of carbon monoxide when burned. This prediction could have an average error of 2.2961 parts per million (ppm) (Se). So our prediction of 15.6 ppm of carbon monoxide could be about 2.3 ppm too high or too low.

2e.

Prediction for X = 4.75 mg nicotine:
We should not use this formula to predict the amount of tar for a cigarette that has 4.75 mg of nicotine. The formula was not designed to plug in 4.75 for X. 4.75 is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

2f. Answers may vary. Carbon Monoxide is a very dangerous gas to take into the lungs. Cigarette smoke has been linked to lung cancer and other diseases.

3a.



Correlation Coefficient r = 0.8889
There is a strong positive correlation between the weight and waist size of these men. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting waist size.

3b.

Regression Line Equation: $Y = 33.8291 + 0.3330 X$

3c.

120 pounds ≤ scope of X values ≤ 237 pounds

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a man to have a weight of zero pounds.

3d.

Prediction for X = 200 pounds:
$Y = 33.8291 + 0.3330 X$

$Y = 33.8291 + 0.3330 (200)$

$Y = 33.8291 + 66.6$

$Y = 100.4291 \approx 100.4$ cm

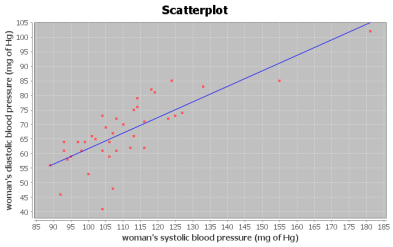
A man that weighs 200 pounds would be predicted to have a waist size of about 100.4 cm. This prediction could have an average error of 4.5763 cm (Se). So our prediction of 100.4 cm could be about 4.6 cm too high or too low.

3e.

Prediction for X = 400 pounds:
We should not use this formula to predict the waist size of a man that is 400 pounds. The formula was not designed to plug in 400 for X. 400 pounds is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

3f. Answers may vary. This could have applications in the medical field and clothing industry.

4a.



Correlation Coefficient r = 0.7854
There is a strong positive correlation between the women's systolic and diastolic blood pressure in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting diastolic BP from systolic BP.

4b.

Regression Line Equation: $Y = 8.3079 + 0.5335 X$

4c.

89 pounds ≤ scope of X values ≤ 181 pounds

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a living woman to have a systolic blood pressure of zero.

4d.

Prediction for X = 135 mm of Hg:
$Y = 8.3079 + 0.5335 X$

$Y = 8.3079 + 0.5335 (135)$

$Y = 8.3079 + 72.0225$

$Y = 80.3304 \approx 80.3$ mm of Hg

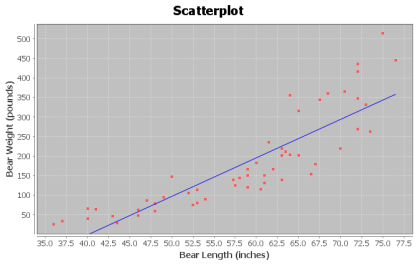
A woman with a systolic blood pressure of 135 mm of Hg would be predicted to have a diastolic blood pressure of about 80.3 mm of Hg. This prediction could have an average error of 7.2912 mm of Hg (Se). So our prediction of 80.3 mm of Hg could be about 7.3 mm of Hg too high or too low.

4e.

Prediction for X = 240 mm of Hg:
We should not use this formula to predict the diastolic blood pressure for a woman with a systolic blood pressure of 240. The formula was not designed to plug in 240 for X. 240 mm of Hg is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

4f. Answers may vary. High blood pressure is a dangerous condition and needs to be studied to better understand how to help people.

5a.



Correlation Coefficient r = 0.8644
There is a strong positive correlation between the length and weight of bears in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting the weight of bears from the length.

5b.

Regression Line Equation: $Y = -393.8391 + 9.8390 X$

5c.

36 inches ≤ scope of X values ≤ 76.5 inches

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a bear to be zero inches long and it is impossible for a bear to weigh -393.8 pounds.

5d.

Prediction for X = 72 inches:
$Y = -393.8391 + 9.8390 X$

$Y = -393.8391 + 9.8390 (72)$

$Y = -393.8391 + 708.408$

$Y = 314.5689 \approx 314.6$ pounds

A bear that is 72 inches long would have a predicted weight of 314.6 pounds.

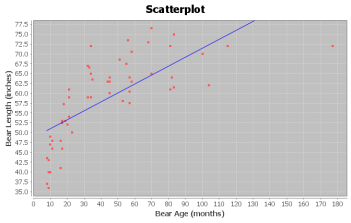
This prediction could have an average error of 61.8272 pounds (Se). So our predicted bear weight of 314.6 pounds could be about 61.8 pounds too high or too low.

5e.

Prediction for X = 18 inches:
We should not use this formula to predict the weight of a young bear that is 18 inches long. The formula was not designed to plug in 18 for X. 18 inches is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

5f. Answers may vary. This analysis may be important to anyone studying bears.

6a.



Correlation Coefficient r = 0.7188
There is a strong positive correlation between the age and length of bears in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting the length of bears from the age.

6b.

Regression Line Equation: $Y = 48.6903 + 0.2281 X$

6c.

$8 \text{ months} \leq \text{scope of X values} \leq 177 \text{ months}$

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. A bear zero months old (newborn) is not 48.7 inches long.

6d.

Prediction for X = 120 months:
$Y = 48.6903 + 0.2281 X$

$Y = 48.6903 + 0.2281 (120)$

$Y = 48.6903 + 27.372$

$Y = 76.0623$

A bear that is 120 months (10 years) old would have a predicted length of 76.1 inches.

This prediction could have an average error of 7.5109 inches (Se). So our predicted bear length of 76.1 inches could be about 7.5 inches too high or too low.

6e.

Prediction for X = 0 months:
We should not use this formula to predict the weight of a newborn bear at 0 months. The formula was not designed to plug in 0 for X. Zero months is way out of the scope of the x values and would be an extrapolation. It would result in dramatic error.

6f. Answers may vary. This analysis may be important to anyone studying bears.

Answers to Problems from Chapter 6 Review Sheet

1.

Explanatory Variable: the "X" variable

Response Variable: This is the "Y" variable and should respond to the explanatory variable. This is the focus of the correlation study and the variable you want to make predictions about.

Correlation Coefficient "r": A statistic between -1 and +1 that measures the strength and direction of the correlation.

r-squared: The square of the correlation coefficient tells us the percent of variability in the Y variable that can be explained by the linear relationship with the X variable.

Slope: A rate of change that measures the amount of increase or decrease in the y variable per unit of x.

Y-intercept: The predicted Y value when X is zero.

Residual: The vertical distance that each point in the scatterplot is above or below the regression line.

Standard Deviation of the Residual Errors (Se): A statistic that measure the average distance that the points in the scatterplot are from the regression line. It also measures the average amount of error if the regression line is used to make a prediction.

2. The response variable Y should respond to the explanatory variable x, but the key to choosing the variables is to know which variable you plan to make predictions about. The focus of your correlation study and the variable you want to make predictions about should be the response variable (Y). The variable you are not making predictions about is the explanatory variable (x).

3a. Both variables may respond to each other. Sanvi should make the hours of sleep the explanatory variable (x) since she wants to predict the number of migraines.

3b. Both variables may respond to each other. Sanvi should make the number of migraines the response variable (y) since the focus of her study is migraine headaches and she wants to predict the number of migraines.

3c. She will <u>not</u> be able to prove the lack of sleep causes migraines because correlation is <u>not</u> causation. There are many confounding variables that may influence migraines besides sleep.

4.

Scatterplot C: No Correlation (r = 0.023)
Scatterplot D: Strong Negative Correlation (r = -0.993)
Scatterplot B: Moderate Positive Correlation (r = 0.592)

5a. Slope = 2.489075718 (about 2.49)
5b. Y intercept = 34.80952773 (about 34.81)

6. The scatterplot and the correlation coefficient r indicate that there is a moderate positive correlation between the wrist size and the BMI of these women.

7. 4.2 in < scope of x values (wrist circumference) < 5.8 in

8. Slope = 10.9407

For every 1 inch increase in wrist circumference, the Body mass index is increasing 10.9407 kg/m^2.

9. r^2 = 0.3446 x 100% = 34.46%

10.

34.46% of the variability in the women's body mass index can be explained by the linear relationship with wrist size.

11. (Answers may vary) Weight, Height, Muscle Mass, Diet, Exercise, Genetics

12. No. Correlation does <u>not</u> prove causation.

13. Se = 5.0568 kg/m^2 (BMI units)

14.

The average distance the points are from the line is 5.0568 kg/m^2.

If we predict a woman's body mass index from her wrist size, we could have an average error of 5.0568 kg/m^2.

15.

The residual plot is V shaped (fan shaped).

16.

There does not appear to be any curve pattern in the residual plot.

17.

No. The histogram does not look very bell shaped.

18.

No. The graph does not appear to be centered at zero.

19.

Y = 48.802 - 8.367 (4.5) = 48.802 – 37.6515 = 11.1505 kg/m^2

20.

Prediction could have an average error of 5.0568 kg/m^2. (Standard Deviation of the Residuals)

21.

No. The scope of the X values is 4.2 to 5.8 inches. A child's wrist of 3.1 inches is out of the scope and would be an extrapolation if we used it to predict BMI. It would have a large error in the prediction.

Introduction to Data Analysis (2nd Edition)
Chapter 7 Answer Keys

Section 7A Answers

1.

a)

The scatterplot shows an exponential growth pattern. The scope of X values is from 0 years to 14 years since 1995. This would represent years 1995 – 2009.

b)

The exponential growth curve fits the data very well. The points are very close to the curve. The data appears to have a strong exponential relationship.

c)

There were 8 ordered pairs in the data.

d)

Exponential Curve Equation: $\hat{y} = 4945.11427\ (1.28424)^x$

Y-intercept (Predicted Y value when x = 0): 4945.11427 MW

Base: 1.28424

Base is greater than 1, confirming that this is an exponential growth curve. If the base were less than 1, it would be a decay curve.

e)

$r^2 = 0.9965$

r-squared sentence: 99.65% of the variability in wind power can be explained by the exponential relationship with the years since 1995.

The r-squared value also confirms that there is an extremely strong exponential relationship between the variables.

f)

The points in the scatterplot are about 4039.2 megawatts (MW) from the exponential curve on average.

If we use the exponential curve equation and a year in the scope of the x values to make a prediction, our prediction could have an average error of 4039.2 MegaWatts (MW).

g)

y = 4945.11427 (1.28424)$^{\mathbf{x}}$

y = 4945.11427 (1.28424)^(7)

y = 4945.11427 (5.761337922)

Y = 28490.47437 ≈ 28,490.5 MW

The exponential curve predicts that by 2002 (year 7) there should be about 28,490.5 MW of wind power generated worldwide.

This prediction could have an average error of about 4039.2 MW (Se) too high or too low.

h)

y = 4945.11427 (1.28424)^{x}

y = 4945.11427 (1.28424)^(13)

y = 4945.11427 (25.84642641)

Y = 127813.5321 ≈ 127,813.5 MW

The exponential curve predicts that by 2008 (year 13) there should be about 127,813.5 MW of wind power generated worldwide.

This prediction could have an average error of about 4039.2 MW (Se) too high or too low.

i)

No. We should not plug in 70 into the equation. Year 70 would be an extremely bad extrapolation and could result in a huge error. The standard deviation of the residuals would not apply as the prediction error since 70 is way out of the scope of the x values.

2.

a)

The scatterplot shows an exponential decay pattern. The scope of the X-value are from 1 – 25 months since January 2010. These represent February 2010 to March 2012.

b)

The exponential decay curve fits the data moderately well. The points are somewhat close to the curve. It appears to have a moderate exponential relationship.

c)

There were 25 ordered pairs in the data.

d)

Exponential Curve: **y =** 63.85340 (0.97985)^{x}

Y-intercept (predicted y value when X = 0) is 63.85340 thousand dollars ($65,853.40)

The base is 0.97985. Notice the base is less than 1. This indicates that this is an exponential decay curve.

e)

r^2 = 0.8337

r-squared sentence: 83.37% of the variability in the retirement account balance can be explained by the exponential relationship with the months since January 2010.

We thought by the graph that there was only a moderate relationship, but the r-squared value indicates that there is a strong exponential relationship between the variables.

f)

The points in the scatterplot are about 4.182 thousand dollars from the exponential curve on average.

If we use the exponential curve equation and a month in the scope of the x values to make a prediction, our prediction could have an average error of 4.182 thousand dollars.

g)

y = 63.85340 $(0.97985)^{x}$

y = 63.85340 (0.97985)^(11.5)

y = 63.85340 (0.791289431)

Y = 50.52652055 ≈ 50.5 thousand dollars

The exponential curve predicts that by month 11.5 there should be about 50.5 thousand dollars left in the retirement account.

This prediction could have an average error of about 4.182 thousand dollars too high or too low.

h)

y = 63.85340 $(0.97985)^{x}$

y = 63.85340 (0.97985)^(24.5)

y = 63.85340 (0.607309571)

Y = 38.77878096 ≈ 38.8 thousand dollars

The exponential curve predicts that by month 24.5 there should be about 38.8 thousand dollars left in the retirement account.

This prediction could have an average error of about 4.182 thousand dollars too high or too low.

i)

No. We should not plug in 480 into the equation. Month 480 would be an extremely bad extrapolation and could result in a huge error. Without adding to the account, the retirement account balance would have run out long before then. The standard deviation of the residuals would not apply as the prediction error.

3.

a) The scatterplot shows an exponential growth pattern. The scope of x values is from 0 years to 22 years since 1990. These represent the years 1990 to 2012.

b)

The exponential growth curve fits the data very well. The points are very close to the curve. It appears to have a strong exponential relationship.

c)

There were 8 ordered pairs in the data.

d)

Exponential Curve: **y =** 497.44019 (1.07211)^{x}

Y-intercept (predicted Y value when X = 0): $497.44019

Base = 1.07211

The base is greater than 1. This indicates that this equation describes an exponential growth curve.

e)

$r^2 = 0.9806$

r-squared sentence: 98.06% of the variability in the savings account balance can be explained by the exponential relationship with the years since 1990.

The r-squared value indicates that there is a very strong exponential relationship between the variables.

f)

The points in the scatterplot are about $110.94 from the exponential curve on average.

If we use the exponential curve equation and a year in the scope of the x values to make a prediction, our prediction could have an average error of $110.94.

g)

y = 497.44019 (1.07211)^{x}

y = 497.44019 (1.07211)^(16)

y = 497.44019 (3.046698999)

Y = 1515.550529 ≈ $1515.55

The exponential curve predicts that by 2006 (year 16) there should be about $1515.55 in the savings account.

This prediction could have an average error of about $110.94 too high or too low.

h)

y = 497.44019 (1.07211)^{x}

y = 497.44019 (1.07211)^(21)

y = 497.44019 (4.315451939)

Y = 2146.679233 ≈ $2146.68

The exponential curve predicts that by 2011 (year 21) there should be about $2146.68 in the savings account.

This prediction could have an average error of about $110.94 too high or too low.

i)

No. We should not plug in 50 into the equation. Year 50 would be an extremely bad extrapolation and could result in a huge error. If money in this account is invested, the investment may go bad at some point or there may be recession. Also, the standard deviation of the residuals would not apply as the prediction error.

4.

4a. The scatterplot shows an exponential decay pattern. The scope of x values (metal distance) are from 0.5 mm to 6 mm.

b)

The exponential decay curve fits the data very well. The points are very close to the curve. It appears to have a strong exponential relationship.

c)

There were 214 ordered pairs in the data.

d)

Exponential Curve: $\mathbf{y =}$ 74.91083 $(0.61685)^{\mathbf{x}}$

Y-intercept (predicted Y value when X is zero) = 74.91083 Watts per square cm

Base = 0.61685

Notice the base is less than 1. This indicates that the exponential equation corresponds to an exponential decay curve.

e)

$r^2 = 0.9126$

r-squared sentence: 91.26% of the variability in ultrasonic response (Watts per square cm) can be explained by the exponential relationship with the metal distance in mm.

The r-squared value indicates that there is a very strong exponential relationship between the variables.

f)

The points in the scatterplot are about 8.239 Watts per cm^2 from the exponential curve on average.

If we use the exponential curve equation and a metal distance in the scope of the x values to make a prediction, our ultrasound response prediction could have an average error of 8.239 Watts per cm^2.

g)

$\mathbf{y =}$ 74.91083 $(0.61685)^{\mathbf{x}}$

y = 74.91083 (0.61685)^(2.83)

y = 74.91083 (0.254805137)

Y = 19.08766435 ≈ 19.1 Watts per cm^2

The exponential curve predicts that if the metal is 2.83 mm away, the ultrasound response will be about 19.1 Watts per cm^2.

This prediction could have an average error of about 8.239 Watts per cm^2 too high or too low.

h)

y = 74.91083 (0.61685)^{x}

y = 74.91083 (0.61685)^(4.51)

y = 74.91083 (0.113164408)

Y = 8.477239743 ≈ 8.5 Watts per cm^2

The exponential curve predicts that if the metal is 4.51 mm away, the ultrasound response will be about 8.5 Watts per cm^2.

This prediction could have an average error of about 8.239 Watts per cm^2 too high or too low.

i)

No. We should not plug in 12.75 mm into the equation for x. 12.75 mm would be an extremely bad extrapolation and could result in a huge error. The standard deviation of the residuals would not apply as the prediction error.

5.

Exponential curves are only defined for bases that are positive and not equal to 1. Raising a positive number to an exponent will always give you a positive result. Hence you can plug in zero or negative numbers for x, but it is impossible for the exponential curve to give negative Y values or a Y value of zero.

Section 7B Answers

1.

a)

The scatterplot shows a decay pattern. A logarithmic decay curve may work well. The scope of the X values are between 2 years and 25 years since 1980. This represents years 1982 – 2005.

b)

The logarithmic decay curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There were 24 ordered pairs in the data.

d)

r^2 = 0.8688

86.88% of the variability in the number of drunk driving fatal accidents can be explained by the logarithmic relationship with the year since 1980.

e)

Standard Deviation of the Residuals (Se) = 1036.1736

The points in the scatterplot are 1036.2 drunk driving fatal accidents from the logarithmic curve on average.

If we use the logarithmic curve and a year since 1980 in the scope of the x values in order to predict the number of drunk driving fatal accidents, our prediction could have an average error of 1036.2 accidents too few or too many.

f)

Logarithmic Curve Equation: **y =** 23389.29331 **+ (-**3753.61219) **Ln (x)**

g)

The number in front of Ln(x) is negative. This indicates that the equation describes a logarithmic decay curve. This does agree with part (a).

h)

y = 23389.29331 **+ (-**3753.61219) **Ln (x)**

y = 23389.29331 **+ (-**3753.61219) **Ln (12.5)**

y = 23389.29331 **+ (-**3753.61219) 2.525728644

y = 23389.29331 **+** (-9480.605828)

Y = 13908.687 ≈ 13909

The logarithmic curve predicts that in year 12.5 we would have about 13,909 fatal car crashes due to drunk driving.

This prediction could have an average error of 1036.1736 (Se) too low or too high.

i)

y = 23389.29331 **+ (-**3753.61219) **Ln (x)**

y = 23389.29331 **+ (-**3753.61219) **Ln (23.75)**

y = 23389.29331 **+ (-**3753.61219) 3.16758253

y = 23389.29331 **+** (-11889.8764)

Y = 11499.41691 ≈ 11499

The logarithmic curve predicts that in year 23.75 we would have about 11,499 fatal car crashes due to drunk driving.

This prediction could have an average error of 1036.1736 (Se) too low or too high.

j)

I think extrapolation in this circumstance would be very bad. The logarithmic decay will quickly go below zero and start predicting negative number of drunk driving fatal crashes, which is impossible. We should not plug in 70 into the formula. We will not be able to use the Se for the prediction error in 70 is not in the scope of the x-values. It will likely have much greater error.

2.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-values (bear ages) is between 8 months and 177 months.

b)

The logarithmic growth curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There are 54 ordered pairs in the data.

d)

$r^2 = 0.7539$

75.39% of the variability in bear lengths can be explained by the logarithmic relationship with the bears age.

e)

Standard Deviation of the Residuals (Se) = 5.3594 inches (bear length)

The points in the scatterplot are about 5.4 inches from the logarithmic curve on average.

If we use the logarithmic curve and a bears age in the scope of the x-values to predict the length of the bear, our bear length prediction could have an average error of 5.4 inches too low or too high.

f)

Logarithmic Curve Equation: **y =** 19.12622 **+** 11.37504 **Ln (x)**

g)

The number in front of the Ln(x) is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

y = 19.12622 **+** 11.37504 **Ln (x)**

y = 19.12622 **+** 11.37504 **Ln (48)**

y = 19.12622 **+** 11.37504 (3.871201011)

y = 19.12622 + 44.03506635

y = 63.16128635 ≈ 63.2 inches

The logarithmic curve predicts that a 4 year old bear (48 months) would have a length of about 63.2 inches.

This prediction could have an average error of 5.4 inches (Se) too low or too high.

i)

y = 19.12622 **+** 11.37504 **Ln (x)**

y = 19.12622 **+** 11.37504 **Ln (120)**

y = 19.12622 **+** 11.37504 (4.787491743)

y = 19.12622 + 54.45791007

y = 73.58413007 ≈ 73.6 inches

The logarithmic curve predicts that a 10 year old bear (120 months) would have a length of about 73.6 inches.

This prediction could have an average error of 5.4 inches (Se) too low or too high.

j)

This logarithmic growth pattern seems to fit the bears well even for older bears. It is still not good to extrapolate. 50 years (600 months) is way out of the scope of the x values. We will not be able to use the Se for the prediction error. It may have more error.

3.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-value temperatures is between about 14 degrees Kelvin and 852 degrees Kelvin.

b)

The logarithmic growth curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There are 236 ordered pairs in the data.

d)

$r^2 = 0.9628$

96.28% of the variability in copper expansion (cubic cm) can be explained by the logarithmic relationship with temperature (Kelvin).

r-squared also indicates a very strong relationship between the variables.

e)

Standard Deviation of the Residuals (Se) = 1.1149 cubic cm

The points in the scatterplot are about 1.1149 cubic cm from the logarithmic curve on average.

If we use the logarithmic curve and the temperature to predict the copper expansion, our prediction could have an average error of 1.1149 cubic cm too low or too high.

f)

Logarithmic Curve Equation: **y =** -16.99737 **+** 5.77086 **Ln (x)**

g)

The number in front of the Ln(x) is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

y = -16.99737 **+** 5.77086 **Ln (x)**

y = -16.99737 **+** 5.77086 **Ln (400)**

y = -16.99737 **+** 5.77086 (5.991464547)

y = -16.99737 **+** 34.5759031

y = 17.5785331 ≈ 17.58 cm^3

The logarithmic curve predicts that at a temperature of 400 degrees Kelvin, copper would expand about 17.58 cubic centimeters.

This prediction could have an average error of about 1.11 cm^3 (Se) too low or too high.

3h

y = -16.99737 **+** 5.77086 **Ln (x)**

y = -16.99737 **+** 5.77086 **Ln (600)**

y = -16.99737 **+** 5.77086 (6.396929655)

y = -16.99737 **+** 36.91578547

y = 19.91841547 ≈ 19.92 cm^3

The logarithmic curve predicts that at a temperature of 600 degrees Kelvin, copper would expand about 19.92 cubic centimeters.

This prediction could have an average error of about 1.11 cm^3 (Se) too low or too high.

i)

The logarithmic curve fits the data very well for higher temperatures, yet without knowing about copper expansion, it is hard to determine if it will follow this pattern beyond the scope of the x values. I would not plug in 1000 degrees into this formula. It may have more error. The standard deviation of the residuals would not apply.

4.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-values is between about 0.25 u and 0.63 u.

b)

The logarithmic growth curve fits the data moderately well. The points look somewhat close to the curve. There seems to be a moderate logarithmic relationship between the variables.

c)

There are 25 ordered pairs in the data set.

d)

$r^2 = 0.9410$

94.10% of the variability in the atomic energy released can be explained by the logarithmic relationship with the atomic defect.

The graph showed only a moderate relationship, but the r-squared indicates there is a very strong logarithmic relationship between the variables.

e)

Standard Deviation of the Residuals (Se) = 341.8581 MeV

The points in the scatterplot are 341.8581 MeV from the logarithmic curve on average.

If we use the logarithmic curve and the atomic defect in the scope to predict the atomic energy released, our prediction could have an average error of 341.8581 MeV too low or too high.

f)

Logarithmic Curve Equation: **y =** 2241.60751 **+** 4720.83078 **Ln (x)**

g)

The number in front of the Ln(x) is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

y = 2241.60751 **+** 4720.83078 **Ln (x)**

y = 2241.60751 **+** 4720.83078 **Ln (0.37)**

y = 2241.60751 **+** 4720.83078 (-0.994252273)

y = 2241.60751 **+** (-4693.696735)

y = -2452.089225 ≈ -2452.1 Mega Electron Volts (MeV)

The logarithmic curve predicts that if the atomic defect was 0.37 atomic defect units (u), we would have about -2452.1 Mega Electron Volts (MeV) of energy released.

This prediction could have an average error of 341.8581 MeV too low or too high.

i)

y = 2241.60751 **+** 4720.83078 **Ln (x)**

y = 2241.60751 **+** 4720.83078 **Ln (0.56)**

y = 2241.60751 **+** 4720.83078 (-0.579818495)

y = 2241.60751 **+** (-2737.22499)

y = -495.6174892 ≈ -495.6 Mega Electron Volts (MeV)

The logarithmic curve predicts that if the atomic defect was 0.56 atomic defect units (u), we would have about -495.6 Mega Electron Volts (MeV) of energy released.

This prediction could have an average error of 341.8581 MeV too low or too high.

j)

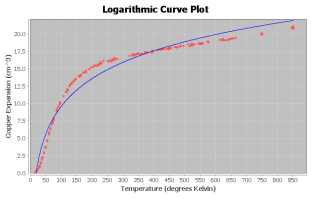
0.9 is not in the scope of the x-values. We should not extrapolate. It may have more error in the prediction. The standard deviation of the residuals would not apply.

5.

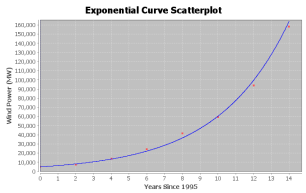
To use a logarithmic curve, the explanatory variable (x) cannot be zero or negative. You can only plug in positive values into a logarithm. However the once you plug in a positive number for x, the Ln(x) value can be negative. So in logarithmic equations, the x must be positive, but the y can be zero or negative.

6.

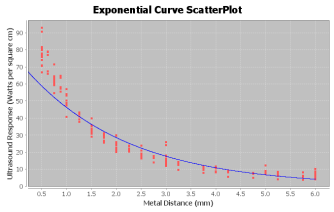
Log Growth



Exponential Growth



Logarithmic and Exponential Decay (Same basic shape)



Logarithmic curves and exponential curves are inverses of each other. If you switch the x and y variables in an exponential curve and then solve for y, you would get a logarithmic curve. If you switch the x and y variables in a logarithmic curve and then solve for y, you would get an exponential curve.

Section 7C Answers

1.

a)

The data shows an opening down parabolic shape indicating there is a maximum height of the rock. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data very well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 $\mathbf{x +}$ (−23.38095) $\mathbf{x^2}$

The leading coefficient (number in front of $\mathbf{x^2}$) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum y value at the vertex.

d)

$r^2 = 0.9870$

98.70% of the variability in the rock height can be explained by the quadratic relationship with the time in seconds.

The r-squared also tells us there is a very strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 8.7402 ft.

The average distance that the points are from the quadratic curve is 8.7402 feet.

If we use the quadratic curve and the time in seconds to predict the rock height, we could have an average error of 8.7402 feet too high or too low.

f)

$\mathbf{y =}$ c $\mathbf{+}$ b $\mathbf{x +}$ a $\mathbf{x^2}$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 $\mathbf{x +}$ (−23.38095) $\mathbf{x^2}$

Number of Seconds for max =

x coordinate of vertex =

−1b/2a = −1(34.88095) / 2(−23.38095)

= −34.88095 / −46.7619

= 0.745926705 seconds.

The maximum height will happen in about 0.746 seconds.

Max Height = Y value when x = 0.745926705

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 $\mathbf{x +}$ (-23.38095) $\mathbf{x^2}$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (0.745926705) $\mathbf{+}$ (-23.38095) $(0.745926705)^2$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (0.745926705) $\mathbf{+}$ (-23.38095) (0.556406649)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 26.0186321 $\mathbf{+}$ (-13.00931604)

$\mathbf{y =}$ 261.5093161 ≈ 261.5 feet

The maximum predicted height of the rock is 261.5 feet.

g)

0 seconds ≤ Scope of X values ≤ 3.5 seconds

h)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 $\mathbf{x +}$ (-23.38095) $\mathbf{x^2}$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (1.8) $\mathbf{+}$ (-23.38095) $(1.8)^2$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (1.8) $\mathbf{+}$ (-23.38095) (3.24)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 62.78571 $\mathbf{+}$ (-75.754278)

$\mathbf{y =}$ 235.531432 ≈ 235.5 feet

At 1.8 seconds the rock will be predicted to have a height of 235.5 feet

This prediction could have an average error of 8.7 feet too high or too low.

i)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 $\mathbf{x +}$ (-23.38095) $\mathbf{x^2}$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (3.2) $\mathbf{+}$ (-23.38095) $(3.2)^2$

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 34.88095 (3.2) $\mathbf{+}$ (-23.38095) (10.24)

$\mathbf{y =}$ 248.50000 $\mathbf{+}$ 111.61904 $\mathbf{+}$ (-239.420928)

y = 120.698112 ≈ 120.7 feet

At 3.2 seconds the rock will be predicted to have a height of 120.7 feet

This prediction could have an average error of 8.7 feet too high or too low.

j)

The rock will not follow this pattern very long. It will hit the ground. We should not plug in 20 into the equation. The equation will start predicting a negative height, which is impossible.

2.

a)

The data shows an opening down parabolic shape indicating there is a predicted maximum solar energy. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data very well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

y = 84.17045 **+** 425.60047 **x +** −33.22890 $\mathbf{x^2}$

The leading coefficient (number in front of $\mathbf{x^2}$) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum y value at the vertex.

d)

$r^2 = 0.8202$

82.02% of the variability in solar energy can be explained by the quadratic relationship with the month.

The r-squared also tells us there is a very strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 189.8436 kWh.

The average distance that the points are from the quadratic curve is about 189.8 kWh.

If we use the quadratic curve and the month to predict the amount of solar energy, we could have an average error of 189.8 kWh too high or too low.

f)

y = c **+** b **x +** a $\mathbf{x^2}$

y = 84.17045 **+** 425.60047 **x +** −33.22890 $\mathbf{x^2}$

Number of Months for max =

x coordinate of vertex =

−1b/2a = −1(425.60047) / 2(−33.22890)

= −425.60047 / −66.4578

= 6.404071004 month ≈ 6.4 months.

The predicted maximum solar energy will happen at about month 6.4 (mid june).

Max Energy = Y value when x = 6.404071004

y = 84.17045 **+** 425.60047 **x +** -33.22890 **x^2**

y = 84.17045 **+** 425.60047 (6.404071004) **+** −33.22890 (6.404071004)**²**

y = 84.17045 **+** 425.60047 (6.404071004) **+** −33.22890 (41.01212543)

y = 84.17045 **+**(2725.575629) **+** (−1362.787815)

y = 1446.958264 ≈ 1447.0 kWh

The maximum predicted solar energy is 1447.0 kWh.

g)

1 month ≤ Scope of X values ≤ 12 months

h)

y = 84.17045 **+** 425.60047 **x +** −33.22890 **x^2**

y = 84.17045 **+** 425.60047 (3.5) **+** −33.22890 (3.5)**²**

y = 84.17045 **+** 425.60047 (3.5) **+** −33.22890 (12.25)

y = 84.17045 **+**(1489.601645) **+** (−407.054025)

y = 1166.71807 ≈ 1166.7 kWh

The predicted solar energy in mid-march (month 3.5) is 1166.7 kWh.

This prediction could have an average error of 189.8 kWh too high or too low.

i)

y = 84.17045 **+** 425.60047 **x +** −33.22890 **x^2**

y = 84.17045 **+** 425.60047 (10.5) **+** −33.22890 (10.5)**²**

y = 84.17045 **+** 425.60047 (10.5) **+** −33.22890 (110.25)

y = 84.17045 **+**(4468.804935) **+** (−3663.486225)

y = 889.48916 ≈ 889.5 kWh

The predicted solar energy in mid-october (month 10.5) is 889.5 kWh.

This prediction could have an average error of 189.8 kWh too high or too low.

j)

The solar energy will not follow this pattern long. We should not plug in month 240. Not only is it an extrapolation, but the equation will also predict a negative solar energy. This would be impossible.

3.

a)

The data shows an opening up parabolic shape indicating there may be a minimum cost possible. A quadratic curve may fit the data.

b)

The quadratic curve fits the data moderately. The points are somewhat close to the curve. There appears to be a moderate quadratic relationship.

c)

$\mathbf{y =}$ 124202.34526 **+** (−4926.25365) $\mathbf{x +}$ 61.72503 $\mathbf{x^2}$

The leading coefficient (number in front of $\mathbf{x^2}$) is positive. This tells us that the equation corresponds to a quadratic curve that opens up and has a minimum y value at the vertex.

d)

$r^2 = 0.6404$

64.04% of the variability in transmission company monthly costs can be explained by the quadratic relationship with the number of hours worked.

The r-squared also tells us there is a strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = $1634.0807

The average distance that the points are from the quadratic curve is about $1634.08 .

If we use the quadratic curve and the average number of hours worked to predict the transmission company costs, we could have an average error of $1634.08 too high or too low.

f)

$\mathbf{y =}$ c **+** b $\mathbf{x +}$ a $\mathbf{x^2}$

$\mathbf{y =}$ 124202.34526 **+** (−4926.25365) $\mathbf{x +}$ 61.72503 $\mathbf{x^2}$

Number of Hours work for min cost =

x coordinate of vertex =

−1b/2a = −1(−4926.25365) / 2(61.72503)

= 4926.25365 / 123.45006

= 39.90482994 ≈ 39.9 hours.

The transmission company should have its employees work 39.9 hours per week to minimize costs.

Min Cost = Y value when x = 39.90482994

$\mathbf{y =}$ 124202.34526 **+** (−4926.25365) $\mathbf{x +}$ 61.72503 $\mathbf{x^2}$

$\mathbf{y =}$ 124202.34526 **+** (−4926.25365) (39.90482994) **+** 61.72503 (39.90482994) **²**

$\mathbf{y =}$ 124202.34526 **+** (−4926.25365) (39.90482994) **+** 61.72503 (1592.395452)

y = 124202.34526 **+** (−196581.3141) **+** 98290.65706

y = $25911.68818 ≈ $25911.69

The predicted minimum costs of the company is $25,911.69 if the employees work 39.9 hours per week.

g)

32 hours ≤ Scope of X values ≤ 50 hours

h)

y = 124202.34526 **+** (−4926.25365) **x +** 61.72503 $\mathbf{x^2}$

y = 124202.34526 **+** (−4926.25365) (44) **+** 61.72503 (44) $^{\mathbf{2}}$

y = 124202.34526 **+** (−4926.25365) (44) **+** 61.72503 (1936)

y = 124202.34526 **+** (−216755.1606) **+** 119499.6581

y = $26946.84276 ≈ $26,947

If the employees work 44 hours a week, the predicted monthly costs would be about $26,947.

This prediction could have an average error of $1634.08 too high or too low.

i)

y = 124202.34526 **+** (−4926.25365) **x +** 61.72503 $\mathbf{x^2}$

y = 124202.34526 **+** (−4926.25365) (35) **+** 61.72503 (35) $^{\mathbf{2}}$

y = 124202.34526 **+** (−4926.25365) (35) **+** 61.72503 (1225)

y = 124202.34526 **+** (−172418.8778) **+** 75613.16175

y = $27396.62926 ≈ $27,397

If the employees work 35 hours a week, the predicted monthly costs would be about $27,397.

This prediction could have an average error of $1634.08 too high or too low.

j)

It seems that if employees work too little or too much, the costs begin to rise. This trend will probably continue even out of the scope of the x values. I still would not plug in 120 into the equation. It is an excessive extrapolation and will probably result in a large error. The standard deviation of the residuals will not apply.

4.

a)

The data shows a opening down parabolic shape indicating there is a maximum length of the bears. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

$\mathbf{y =}$ 41.93861 **+** 0.54530$\mathbf{x}$ **+** (−0.00234)$\mathbf{x^2}$

The leading coefficient (number in front of $\mathbf{x^2}$) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum y value at the vertex.

d)

$r^2 = 0.6884$

68.84% of the variability in bear length can be explained by the quadratic relationship with the bears age.

The r-squared also tells us there is a strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 6.0897 inches.

The average distance that the points are from the quadratic curve is about 6.1 inches.

If we use the quadratic curve and the bears age to predict the bears length, we could have an average error of about 6.1 inches too high or too low.

f)

$\mathbf{y =}$ c **+** b $\mathbf{x}$ **+** a $\mathbf{x^2}$

$\mathbf{y =}$ 41.93861 **+** 0.54530$\mathbf{x}$ **+** (−0.00234)$\mathbf{x^2}$

Age of bear (months) for max length =

x coordinate of vertex =

−1b/2a = −1(0.54530) / 2(−0.00234)

= −0.54530 / −0.00468

= 116.517094 ≈ 116.5 months old

The maximum length will happen when a bear is about 116.5 months old.

Max Bear Length = Y value when x = 116.517094

$\mathbf{y =}$ 41.93861 **+** 0.54530$\mathbf{x}$ **+** (−0.00234)$\mathbf{x^2}$

$\mathbf{y =}$ 41.93861 **+** 0.54530 (116.517094) **+** (−0.00234) $(116.517094)^2$

$\mathbf{y =}$ 41.93861 **+** 0.54530 (116.517094) **+** (−0.00234) (13576.23319)

$\mathbf{y =}$ 41.93861 **+** 63.53677136 **+** (−31.76838567)

Y = 73.70699569 ≈ 73.7 inches

The maximum predicted length of the bears is 73.7 inches.

g)

8 months ≤ Scope of X values ≤ 177 months

h)

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\mathbf{x} \mathbf{+} (-0.00234)\mathbf{x^2}$

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\ (48) \mathbf{+} (-0.00234)\ (48)^{\mathbf{2}}$

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\ (48) \mathbf{+} (-0.00234)\ (2304)$

$\mathbf{y =} 41.93861 \mathbf{+} 26.1744 \mathbf{+} (-5.39136)$

Y = 62.72165 ≈ 62.7 inches

A bear 48 months old will be predicted length of 62.7 inches.

This prediction could have an average error of 6.0897 inches too high or too low.

i)

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\mathbf{x} \mathbf{+} (-0.00234)\mathbf{x^2}$

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\ (150) \mathbf{+} (-0.00234)\ (150)^{\mathbf{2}}$

$\mathbf{y =} 41.93861 \mathbf{+} 0.54530\ (150) \mathbf{+} (-0.00234)\ (22500)$

$\mathbf{y =} 41.93861 \mathbf{+} 81.795 \mathbf{+} (-52.65)$

Y = 71.08361 ≈ 71.1 inches

We predict that the length of a bear 150 months old will be 71.1 inches.

This prediction could have an average error of 6.0897 inches too high or too low.

j)

We should not plug in 360 months. The equation will start predicting that the length of the bear will shrink, which is impossible. It will result in a huge error. The standard deviation of the residuals will not apply since 360 is out of the scope of the x values.

5.

The quadratic curves have a formula of the form $\mathbf{y =} c \mathbf{+} b\ \mathbf{x} \mathbf{+} a\ \mathbf{x^2}$. The key is that the highest power of x is an x-squared in the formula. This tells us it is quadratic. If the number in front of the x-squared is positive, the curve will open up. If the number in front of the x-squared is negative, the curve will open down.

6.

If the number in front of the x-squared is positive, the curve will open up and will therefore have a minimum Y value at the vertex. If the number in front of the x-squared is negative, the curve will open down and will therefore have a maximum Y value at the vertex.

(Answers may vary)
There are many applications when maximum and minimum are useful. Maximizing profits and minimizing costs are vital in business. Minimizing cases of disease and maximizing the number of people vaccinated are vital in the medical field. Maximizing productivity and minimizing errors are vital in sports and business.

Answers for Chapter 6 Review Sheet Problems

1. Logarithmic Growth (b)
2. Open Down Quadratic (e)
3. Decay (Exponential or Logarithmic) (c)
4. Exponential Growth (a)

5a.

-1b / 2a = -1(5.868) / 2(-0.163) = -5.868 / -0.326 = 18 breaks

5b.

Max occurs when x = 18

Y = 41.8 + 5.868(18) + (-0.163)(18)^2

Y = 41.8 + 5.868(18) + (-0.163)(324)

Y = 41.8 + 105.624 + (-52.812)

Y = 94.612

The maximum efficiency for the company is 94.6 if the employees are given 18 breaks each week.

6.

18.08% (r-squared converted to %)

7.

356.8787 grams (Se)

8.

356.8787 grams (Se)

9.

Predicted Baby Weight if mother is 33 years old: 1763.154711 (about 1763 grams)

10.

The log curve does not fit the data that well since the r-squared is low at 18.08% and the standard deviation is high at 356.8787 grams. I would classify it as a weak to moderate relationship.

11.

No. Just because there is a relationship does not imply causation.

12. (Answers my vary)

Confounding Variables: Health of the mother, Genetics of mother, genetics of the father, diet of mother during pregnancies, alcohol or drugs during pregnancy

13.

37.46% (r-squared for exponential curve)

14.

53.97% (r-squared for quadratic curve)

15.

The quadratic curve is the better fit and the stronger relationship since the r-squared percentage was higher.

16.

48.297 CHD deaths (standard deviation of residuals for exponential)

17.

47.5847 CHD deaths (standard deviation of residuals for quadratic)

18.

The points were closer to quadratic curve since the standard deviation of the residuals was smaller for the quadratic.

19.

48.297 CHD deaths (standard deviation of residuals for exponential)

20.

47.5847 CHD deaths (standard deviation of residuals for quadratic)

21.

The quadratic had less prediction error since the standard deviation of the residuals was smaller for the quadratic.

22.

The quadratic curve was the better fit since it had the higher r-square percentage and the lower standard deviation of the residuals.
