Solutions to OpenIntro Statistics Exercises
Fourth Edition

Prepared by

Mine Çetinkaya-Rundel
mine@openintro.org

Redistributing this document is not permitted due to the sensitivity of solutions to even exercises.

Version date: August 10th, 2019.

This document is only available to OpenIntro Verified teachers.

OpenIntro was built on openness, so why are solutions locked down?
OpenIntro's mission is to increase access to high-quality, low-cost education materials. Without teachers choosing to use our resources, we cannot fulfill our mission, and many teachers have told us it is critical to have some exercises for graded homework or exam questions where solutions are not readily available to students. Without a large collection of such exercises, many teachers would drop OpenIntro and instead require a more expensive and lower-quality textbook. For this reason, we do not permit sharing of this solution manual. We know this creates challenges for some students and self-study learners who could benefit from a full solution guide, and this decision was not made lightly.

What can I do to keep the solutions safe?
If you downloaded this file from OpenIntro's site as a verified teacher, do not share this file with anyone. If you got access to this document elsewhere, please report it to admin@openintro.org.

Can I share this solutions guide with another teacher?
Instead, please invite other teachers to OpenIntro so they can download it directly and also get access to all other teacher-only options, e.g. so they can also request a free desk copy:

openintro.org/invite

Teachers invited through our website by a Verified Teacher are able to skip the manual verification process.

Please report solution typos at openintro.org/os (click on Typos and Feedback).

Contents

Chapter 1

Introduction to data

1.1

(a) Percent pain free in the treatment group: $10/43 = 0.23 \rightarrow 23\%$.

(b) Percent pain free in the control group: $2/46 = 0.04 \rightarrow 4\%$.

(c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture.

(d) It is possible that the observed difference between the two group percentages is due to chance.

1.2

(a) Percent of patients in the treatment group with self-reported significant improvement in symptoms: $66/85 = 0.78 \rightarrow 78\%$.
Percent of patients in the control group with self-reported significant improvement in symptoms: $65/81 = 0.80 \rightarrow 80\%$.

(b) 78% of patients in the treatment and 80% of patients in the control group reported significant improvement in symptoms; therefore at a first glance the non-antibiotic treatment appears to be more effective for the treatment of sinusitis.

(c) Even if the two treatments had the same effect on the improvement rates of sinusitis symptoms, typically we would not get the exact same rates in symptom improvement in each group. The difference of 2%, which separates the groups by just one or two successful treatments, seems like it could be due to just chance. (The best answers will include a component regarding uncertainty.)

1.3

(a) The research question is "Is there an association between air pollution exposure and preterm births?".

(b) The cases are 143,196 births in Southern California between 1989 and 1993.

(c) The variables are measurements of carbon monoxide (CO), nitrogen dioxide, ozone, and particulate matter less than 10μm (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. All of these variables are continuous numerical variables.

1.4

(a) The research question is "Do asthmatic patients who practice the Buteyko method experience improvement in their condition?".

(b) The cases are 600 adult patients aged 18-69 years diagnosed and currently treated for asthma.

(c) The variables are whether or not the patient practiced the Buteyko method (categorical) and measures of quality of life, activity, asthma symptoms and medication reduction of the patients (categorical, ordinal). It may also be reasonable to treat the ratings on a scale of 0 to 10 as discrete numerical variables.

1.5

(a) The research question is "Does explicitly telling children not to cheat affect their likelihood to cheat?".

(b) The subjects are 160 children between the ages of 5 and 15.

(c) Four variables were recorded for each subject in the study: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), and (4) whether they cheated or not (categorical).

1.6

(a) The research question is "Is there a difference between the unethical behaviors of people who identify themselves as having low and high social-class rank?"

(b) The cases are 129 University of California at Berkeley undergraduates.

(c) Two variables: (1) social-class rank (categorical) and (2) number of candies taken (numerical, discrete).

1.7 Explanatory: acupuncture or not.
Response: if the patient was pain free or not.

1.8 Explanatory: antibiotics or placebo.
Response: symptoms of acute sinusitis.

1.9

(a) There are 50 * 3 = 150 cases included in this data set.

(b) There are 4 numerical variables in the data: sepal length, sepal width, petal length and petal width, and they are all continuous numerical variables.

(c) There is only one categorical variable in the data, which is species. It has 3 levels: setosa, versicolor, and virginica.

1.10

(a) Each row of the data matrix represents a participant in the survey.

(b) There are 1,691 participants in the survey.

(c) The table below lists the types of the variables included in the study.

variable	**type**
sex	categorical
age	numerical, continuous (recorded as rounded to whole years)
maritalStatus	categorical
grossIncome	ordinal*
smoke	categorical
amtWeekends	numerical, discrete
amtWeekdays	numerical, discrete

*Note that income is a continuous numerical variable, but in this survey it is recorded as an ordinal variable.

1.11

(a) Airport ownership status (public/private), airport usage status (public/private), latitude, and longitude.

(b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

1.12

(a) We see the order of the categories and the relative frequencies in the bar plot.

(b) There are no features that are apparent in the pie chart but not in the bar plot.

(c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

1.13

(a) The population of interest is all births. The sample consists of the 143,196 births between 1989 and 1993 in Southern California.

(b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

1.14

(a) The population of interest is all children between the ages of 5 and 15. The sample consists of 160 children between these ages.

(b) If the children in this sample, who are likely not randomly sampled, can be considered to be representative of all children between the ages of 5 and 15, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.15

(a) The population of interest is all asthma patients aged 18-69 who rely on medication for asthma treatment. The sample consists of 600 such patients.

(b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.16

(a) The population of interest is likely all people, however since the sample only consists of 129 UC Berkeley undergraduates, the target population is limited to UC Berkeley undergraduates.

(b) If the students in this sample, who are likely not randomly sampled, can be considered to be representative of UC Berkeley undergraduates, then the results are generalizable to the UC Berkeley undergraduate population. If students from this school can be considered to be representative of all college students, or all people, we could potentially generalize to these populations as well. However the latter (being representative of all people) is quite unlikely. Additionally, since the study is observational (participants were not randomly assigned to lower and upper class groups), the findings cannot be used to establish causal relationships.

1.17

(a) Observation.
(b) Variable.
(c) Sample statistic (mean).
(d) Population parameter (mean).

1.18

(a) Population parameter (proportion).
(b) Sample statistic (proportion).
(c) Observation.
(d) Variable.

1.19

(a) This is an observational study.

(b) The professor suspects students in a given section may have similar feelings about a course. To ensure each section is reasonably represented, she may choose to randomly select a fixed number of students, say 10, from each section for a total sample size of 40 students. Since a random sample of fixed size was taken within each section in this scenario, this represents a stratified sampling.

1.20

(a) This is an observational study.

(b) To ensure students from each year are reasonably represented, we might choose to randomly sample a fixed number of students, say 60, from each part of the campus (east and west). Since a random sample of fixed size was taken within each part in this scenario, this represents a stratified sampling.

1.21

(a) The two variables are positively associated. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies. However rise in life expectancy trails off before around 80 years old, hence resulting in a non-linear relationship between the two variables. The relationship is also somewhat strong.

(b) This is an observational study.

(c) Wealth is one confounding variable. Countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

1.22

(a) Observational.

(b) No, since the study is observational we cannot infer causation.

(c) Caffeine and lack of sleep are potential confounding variables that might explain the observed relationship between stress and muscle cramps, i.e. students might be experiencing muscle cramps due to increased caffeine consumption or lack of sleep.

1.23

(a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method!

(b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable.

(c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

1.24 Sampling from the phone book would introduce bias since it would miss unlisted phone numbers. If the unlisted phone numbers are missing at random this may not be a problem, but if people who choose to not list their numbers share a certain characteristic, our sample would not be able to capture such people and therefore would not be representative of the population.

1.25

(a) The cases are 200 randomly sampled men and women.

(b) The response variable is attitude towards a fictional microwave oven.

(c) The explanatory variable is dispositional attitude.

(d) Yes, the cases are sampled randomly.

(e) This is an observational study since there is no random assignment to treatments.

(f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational.

(g) Yes, the results of the study can be generalized to the population at large since the sample is random.

1.26 Yes, the estimate will be biased, and it will tend to overestimate the true family size. Notice that families without children cannot be sampled using the sampling strategy. Additionally, if a family has two children, it is twice as likely to be sampled than a family with one child since there are twice as many children included in the sample (on average).

1.27

(a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population.

(b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey.

(c) Convenience sample. This will have a similar issues to handing out surveys to friends.

(d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

1.28

(a) No, this is an observational study.

(b) This statement is not justified; it implies a causal association between sleep disorders and bullying. However, this was an observational study. A better conclusion would be "School children identified as bullies are more likely to suffer from sleep disorders than non-bullies."

1.29

(a) Exam performance.

(b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(c) Sex: man, woman.

1.30

(a) Experiment, since the researchers randomly assigned different treatments to the participants.

(b) Response variable: Duration of the cold.
Explanatory variable: Treatment, with 4 levels; placebo, 1g, 3g, 3g with additives.

(c) The patients were blinded as they did not know which treatment they received.

(d) The study was double-blind with respect to the researchers evaluating the patients, but the nurses who briefly interacted with patients during the distribution of the medication were not blinded. (It was partially double-blind.)

(e) Since the patients were randomly assigned to the treatment groups and they are blinded we would expect about an equal number of patients in each group to not adhere to the treatment. While this means that final results of the study will be based on fewer number of participants, non-adherence does not introduce a confounding variable to the study.

1.31

(a) Exam performance.

(b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise).

(c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

1.32 Recruit 30 friends and randomly assign them to three groups: no music, instrumental music, and music with lyrics. Have each participant read a passage to learn about a new concept, and then give them a short quiz assessing what they have learned. Compare the number of questions participants got correct on average across the three groups.

1.33 For a good experiment we need randomization and blinding. Here is one possible outline:

- Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!)
- Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double-blind experiment. (Answers may vary.)

1.34

(a) This is an experiment.

(b) The treatment is exercise twice a week and control is no exercise.

(c) Yes, the blocking variable is age.

(d) No, the study is not blinded since the patients will know whether or not they are exercising.

(e) Since this is an experiment, we can make a causal statement. Since the sample is random, the causal statement can be generalized to the population at large. However, we should be cautious about making a causal statement because of a possible placebo effect.

(f) It would be very difficult, if not impossible, to successfully conduct this study since randomly sampled people cannot be required to participate in a clinical trial.

1.35

(a) These data collected as part of an observational study.

(b) The most common Dog name is Lucy, and the most common cat name is Luna.

(c) Oliver and Lily are more common for cats than dog.

(d) The relationship is positive, which means that as the popularity of a name for dogs increases, so does the popularity of that name for cats.

1.36

(a) This is an experiment.

(b) Yes, since the study is an experiment, we can infer causation.

1.37

(a) This is an experiment.

(b) The treatment is 25 grams of chia seeds twice a day and the control is a placebo.

(c) Yes, the blocking variable is gender.

(d) Yes, the study is single blind since the patients were blinded to the treatment they received.

(e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

1.38

(a) Simple random sampling. This is usually an effective method as it assigns equal probability to each household to be picked.

(b) Stratified sampling. This is an effective method in this setting since neighborhoods are unique and this method allows us to sample from each neighborhood.

(c) Cluster sampling. This is not an effective method in this setting since the resulting sample will not contain households from certain neighborhoods and we are told that some neighborhoods are very different from others.

(d) Multi-stage sampling. This method will suffer from the same issue discussed in part (c).

(e) Convenience sampling. This is not an effective method since it will result in a biased sample for households that are similar to each other (in the same neighborhood) and the sample will not contain any houses from neighborhoods far from the city council offices.

1.39

(a) Non-responders may have a different response to this question. The parents who returned the surveys are probably those who do not have difficulty spending time with their kids after school. Parents who work might not have returned the surveys since they probably have a busier schedule.

(b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might have a lower socio-economic status than the respondents.

(c) First, there is no control group in this study. It may be that if we looked at 30 patients with joint problems, 20 of them regularly go running as well. Second, this is an observational study. Third, there may be confounding variables. For example, these people may go running because they are generally healthier and/or do other exercises.

1.40

(a) The explanatory variable is percent of population with a bachelor's degree and the response variable is per capita income (in thousands). This is implied by the plot since the x-axis is typically set as the explanatory variable.

(b) There is a strong positive linear relationship between the two variables. As the percentage of population with a bachelor's degree increases, the per capita income increases as well. There are very few counties where more than 60% of the population have a bachelor's degree and very few countries that have a more than $50,000 in per capita income.

(c) This is an observational study so we cannot make a causal statement based on the results. We can only infer that having a higher percentage of population with bachelor's degree is associated with a higher per capita income.

1.41

(a) The study is a randomized controlled experiment.

(b) The explanatory variable is treatment group (categorical, with 3 levels), and the response variable is psychological well-being.

(c) The results of the study cannot be generalized to the population as the participants were volunteers.

(d) The results of the study can be used to establish causal relationships since it was an experiment.

(e) The statement should say, "The results of this study provide **evidence** that giving young adults fresh fruit and vegetables to eat can have psychological benefits, even over a brief period of time."

1.42

(a) This is an observational study.

(b) The explanatory variables are screen time, child's sex and age and the mother's education, ethnicity, psychological distress, and employment.

(c) The response variable is psychological well-being.

(d) The results of the study can be generalized to the population since the data are a representative sample.

(e) The results of the study cannot be used to establish causal relationships since the study was observational.

1.43

(a) For each stop the following variables were collected: county, state, driver's race, whether the car was searched or not, whether the driver was arrested or not.

(b) None of the variables are numerical. They are all categorical, non-ordinal.

(c) The response variable would be whether the car was searched or not and the explanatory variable would be the race of the driver.

1.44

(a) For each launch the following variables were collected: year, launching agency, whether the launch was a success or failure.

(b) Year may be numerical (discrete) or categorical (not ordinal), with levels between 1957 and 1999 and between 2000 and 2018. Launching agency and whether the launch was a success or failure are both categorical (nor ordinal) variables.

(c) The response variable would be whether the launch was a success or failure, and the explanatory variables would be year and launching agency.

Chapter 2

Summarizing data

2.1

(a) The association is positive association, mammals with longer gestation periods tend to live longer as well.

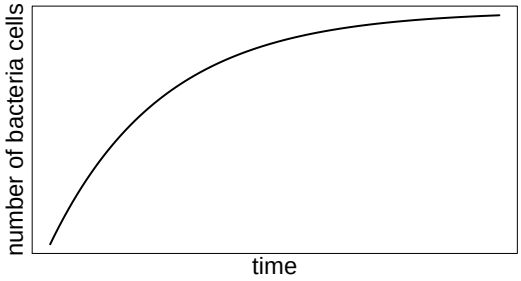
(b) Association would still be positive.

(c) No, they are not independent. See part (a).

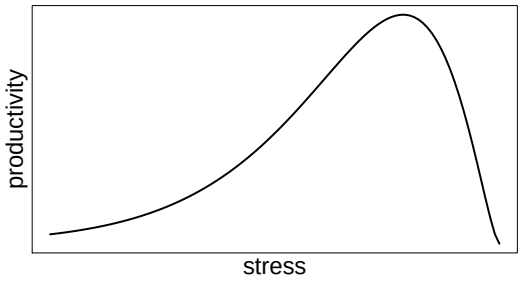
2.2

(a) 1: linear. 3: nonlinear.
(b) 4: linear.
(c) 2.

2.3 The graph below shows a ramp up period. There may also be a period of exponential growth at the start before the size of the petri dish becomes a factor in slowing growth.



2.4



2.5

(a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$.

(b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

2.6 Population mean = 5.5. Sample mean = 6.25.

2.7 In order to increase the average number of days off, the manager should fire any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant. However, firing the 10 employees with the minimum number of days off will have the biggest impact on the average.

2.8

(a) Both distributions have the same median and IQR.
(b) Second distribution has a higher median and higher IQR.
(c) Second distribution has higher median. IQRs are equal.
(d) Second distribution has higher median and larger IQR.

2.9

(a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13.

(b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20.

(c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means.

(d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

2.10

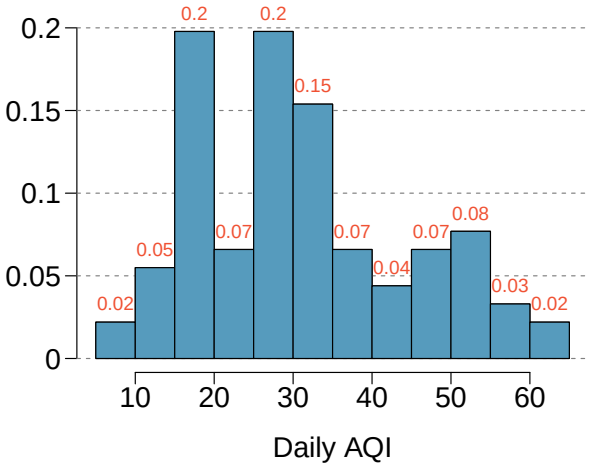
(a) The distribution is unimodal and symmetric, and about 95% of the data falls within about 7 units of the center, so the standard deviation will be about 3 or 4. This matches box plot (2).

(b) The distribution is uniform and values range from 0 to 100. This matches box plot (3) which shows a symmetric distribution in this range. Also, each 25% chunk of the box plot has about the same width and there are no suspected outliers.

(c) The distribution is unimodal and right skewed with a median between 1 and 2. 25th and 75th percentile are near 1 and 2, so the IQR is roughly 1. This matches box plot (1).

2.11

(a) The median is the 50% of the distribution, which is about 30.



(b) Since the distribution is right skewed the mean is higher than the median.

(c) Q1: between 15 and 20.
Q3: between 35 and 40.
IQR: about 20.

(d) Values that are considered to be unusually low or high lie more than 1.5× IQR away from the quartiles.
Upper fence: Q3 + 1.5 × IQR = 37.5 + 1.5 × 20 = 67.5
Lower fence: Q1 - 1.5 × IQR = 17.5 + 1.5 × 20 = −12.5
The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

2.12 Median is around 80. We would expect the mean to be slightly lower since the distribution is left skewed.

2.13 The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

2.14 The statement "50% of Facebook users have over 100 friends" means that the median number of friends is 100, which is lower than the mean number of friends (190), which suggests a right skewed distribution for the number of friends of Facebook users.

2.15

(a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR.

(b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR.

(c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

2.16

(a) The distribution is right skewed with potential outliers on the positive end, therefore the median and the IQR are preferable measures of center and spread.

(b) The distribution is somewhat symmetric and has few, if any, extreme observations, therefore the mean and the standard deviation are preferable measures of center and spread.

(c) The distribution would be right skewed. There would be some students who did not consume any alcohol, but this is the minimum since students cannot consume fewer than 0 drinks. There would be a few students who consume many more drinks than their peers, giving the distribution a long right tail. Due to the skew, the median and IQR would be preferable measures of center and spread.

(d) The distribution would be right skewed. Most employees would make something on the order of the median salary, but we would anticipate upper management makes much more. The distribution would have a long right tail, and the median and the IQR would be preferable measures of center or spread.

2.17

(a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers.

(b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

2.18 No, the outliers are likely the maximum and the minimum of the distribution so a statistic based on these values cannot be robust to outliers.

2.19

(a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. Since the distribution is already unimodal and symmetric, a log transformation is not necessary.

(b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

2.20

(a) The distribution of percentage of population that is Hispanic is extremely right skewed with majority of counties with less than 10% Hispanic residents. However there are a few counties that have more than 90% Hispanic population. It might be preferable to, in certain analyses, to use the log-transformed values since this distribution is much less skewed.

(b) The map reveals that counties with higher proportions of Hispanic residents are clustered along the Southwest border, all of New Mexico, a large swath of Southwest Texas, the bottom two-thirds of California, and in Southern Florida. In the map all counties with more than 40% of Hispanic residents are indicated by the darker shading, so it is impossible to discern the how high Hispanic percentages go. The histogram reveals that there are counties with over 90% Hispanic residents. The histogram is also useful for estimating measures of center and spread.

(c) Both visualizations are useful, but if we could only examine one, we should examine the map since it explicitly ties geographic data to each county's percentage.

2.21

(a) We see the order of the categories and the relative frequencies in the bar plot.

(b) There are no features that are apparent in the pie chart but not in the bar plot.

(c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

2.22

(a) $\frac{372}{910} = 0.41 \rightarrow 41\%$

(b) $\frac{278}{910} = 0.31 \rightarrow 31\%$

(c) $\frac{57}{910} = 0.06 \rightarrow 6\%$

(d) Conservatives: $\frac{57}{372} = 0.15 \rightarrow 15\%$
Moderates: $\frac{120}{363} = 0.33 \rightarrow 33\%$
Liberals: $\frac{101}{175} = 0.58 \rightarrow 58\%$

(e) The percentages of Tampa, FL conservatives, moderates, and liberals who are in favor of illegal immigrants working in the US staying and applying for citizenship are quite different from one another. Therefore, the two variables appear to be dependent.

2.23 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

2.24 The vertical locations at which the party groups break into the categories of the question on raising taxes differ, which indicates that likelihood of supporting the idea of raising taxes on the poor vs. the rich varies by political party. This suggests that the two variables may be dependent.

2.25

(a) i. False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems.

ii. True.

iii. False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. (We cannot say changing the drug a person is on affects her risk, which is why part (b) is true. The difference in these statements is subtle.)

iv. True.

(b) Proportion of all patients who had a heart attack: $\frac{7,979}{227,571} \approx 0.035$

(c) Expected number of heart attacks in the rosiglitazone group if having cardiovascular problems and treatment were independent can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$.

(d) i. H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance.
H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems.

ii. A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems.

iii. In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

2.26

(a) Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. These data suggest that survival is not independent of whether or not the patient got a transplant.

(b) The shape of the distribution of survival times in both groups is right skewed with one very clear outlier for the control group and other possible outliers in both groups on the high end. The median survival time for the control group is much lower than the median survival time for the treatment group; patients who got a transplant typically lived longer. Tying this together with the much lower variability in the control group, evident by a much smaller IQR than the treatment group (about 50 days versus 500 days), and we can see that patients who did not get a heart transplant tended to consistently die quite early relative to those who did have a transplant. Overall, very few patients without transplants made it beyond a year while nearly half of the transplant patients survived at least one year. It should also be noted that while the first and third quartiles of the treatment group is higher than those for the control group, the IQR for the treatment group is much bigger, indicating that there is more variability in survival times in the treatment group.

(c) Proportion of patients who in the treatment group that died: $\frac{45}{69} = 0.652$
Proportion of patients who in the control group that died: $\frac{30}{34} = 0.882$

(d) i. H_0: The variables group and outcome are independent. They have no relationship, and the difference in survival rates between the control and treatment groups was due to chance. In other words, heart transplant is not effective.
H_A: The variables group and outcome are not independent. The difference in survival rates between the control and treatment groups was not due to chance and the heart transplant is effective.

ii. 28, 75, 69, 34, 0, -0.23 or lower.

iii. Under the independence model, only 2 out of 100 times (2%) did we get a difference of -0.23 or lower between the proportions of patients that died in the treatment and control groups. Since this is a low probability, we can reject the claim of independence in favor of the alternate model. There is convincing evidence to suggest that the transplant program is effective.

2.27

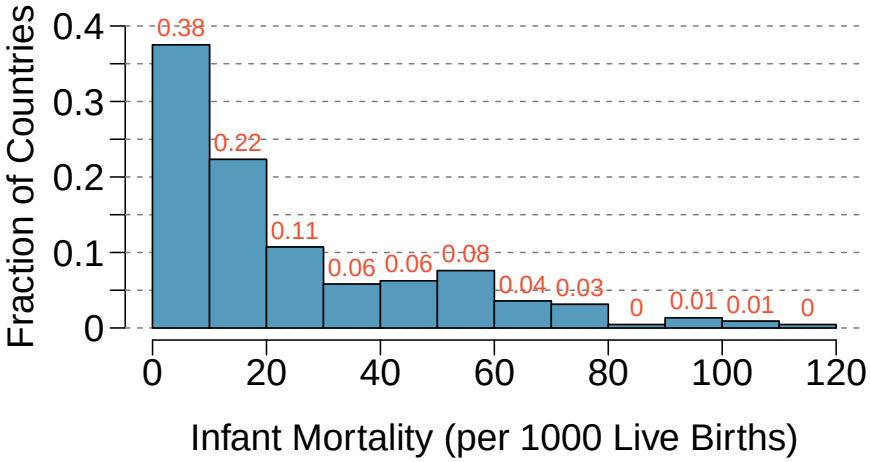
(a) Since the new score is smaller than the mean of the 24 previous scores, the new mean should be smaller than the old mean.

(b) We are given that $n = 24, \quad \bar{x} = 74, \quad s_x = 8.9$.

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_{24}}{24} = 74 \\ x_1 + x_2 + \cdots + x_{24} &= 74 * 24 = 1776 \\ x_1 + x_2 + \cdots + x_{24} + x_{25} &= 1776 + 64 = 1840 \\ \bar{x}_{new} &= \frac{x_1 + x_2 + \cdots + x_{24} + x_{25}}{25} = \frac{1840}{25} = 73.6 \end{aligned}$$

(c) The new score, x_{25}, is more than 1 standard deviation away from the previous mean, and this will tend to increase the standard deviation of the data. While possible, it is mathematically tedious to calculate the new standard deviation.

2.28



(a) First we eyeball the heights of the bars in the relative frequency histogram (as shown on the right). Remember that these relative frequencies must add up to 1. Then, find out which bin the 25^{th}, 50^{th}, and 75^{th} percentiles fall in and estimate Q1, median (Q2), and Q3 as the midpoint of those bins, respectively.

- Q1 $\approx$ 5
- Q2 $\approx$ 15
- Q3 $\approx$ 35

(b) Since the distribution is right skewed, we would expect the mean to be higher than the median.

2.29 No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

2.30

(a) If $midrange = 1$, then $\bar{x} = median$. This is most likely to be the case for symmetric distributions.

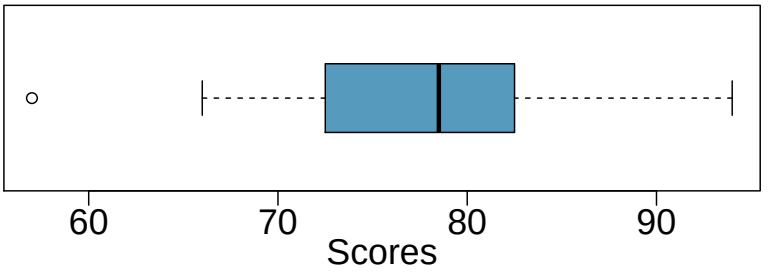
(b) If $midrange < 1$, then $\bar{x} < median$. This is most likely to be the case for left skewed distributions, since the mean is affected (and pulled down) by the lower values more so than the median.

(c) If $midrange > 1$, then $\bar{x} > median$. This is most likely to be the case for right skewed distributions, since the mean is affected (and pulled up) by the higher values more so than the median.

2.31 The distribution of ages of best actress winners are right skewed with a median around 30 years. The distribution of ages of best actress winners is also right skewed, though less so, with a median around 40 years. The difference between the peaks of these distributions suggest that best actress winners are typically younger than best actor winners. The ages of best actress winners are more variable than the ages of best actor winners. There are potential outliers on the higher end of both of the distributions.

2.32 No, we would expect this distribution to be left skewed. There are two reasons for this: (1) there is a natural boundary at 100 (it is not possible to score above 100 on the exam), (2) the standard deviation of the distribution is very large compared to the mean.

2.33



2.34

(a) From the histogram we can see the the distribution is bimodal. In the box plot the more extreme observations, many of which could be considered outliers, are easier to identify.

(b) Gender may be the reason, it is likely that men and women have different average marathon times.

(c) The median marathon time for men is about 2.2 hrs while it is about 2.5 hrs for women; therefore, men are faster on average. The minimum marathon time for men is about 2.1 hrs whereas it is about 2.4 hrs for women. The maximum marathon time for men is about 2.5 hrs for men whereas it is about 3.1 hrs for women. Both distributions have apparent outliers on the high end.

(d) It appears that marathon times decreased greatly between 1970-1975 and remained somewhat steady thereafter. Males consistently had shorter marathon times than females throughout the years. From the box plots of males and females, we could tell that males ran faster "on average", however, we could not tell that the winning male time for each year was better than the winning female time. We also could not tell from the histogram or the box plot that marathon times have been decreasing for males and females throughout the years.

Chapter 3

Probability

3.1

(a) False. Since these are independent trials the probability does not change from trial to trial and does not depend on what the outcome was on the previous trial(s). Therefore, in the next toss the probability of getting a head will still be 0.5.

(b) False. There are red face cards (jack, queen, or king of hearts and diamonds). Since a card can be both a face card and a red card, the two events are not mutually exclusive.

(c) True. A card cannot be both a face card and an ace.

3.2

(a) P(red on 4^{th} spin) = 18 / 38 = 9 / 19

(b) P(red on 301^{st} spin) = 18 / 38 = 9 / 19

(c) Theoretically this probability is also 18/38 in both cases however it is very unlikely that a fair roulette wheel will land on a red 300 consecutive times. It is possible that the the wheel in part (b) has been rigged, and hence the probability of red might be different that 18/38. We should be less confident of the answer to part (b).

3.3

(a) 10 tosses. Since the desired outcome is larger than the expected proportion of heads (50%), we would want fewer trials (flips). With a low number of flips the variability in the number of heads observed is much larger. For example, we wouldn't be very surprised if 7 out of 10 flips were heads. On the other hand, with 100 flips, we would often be surprised if 70 out of 100 flips were heads. With 100 flips it would often be more likely to get something very close to 50% heads and 50% tails, which is not what we want.

(b) 100 tosses. The expected proportion of 50% is greater than 40%, and with more flips the observed proportion of heads would often be closer to the expected value than if we used a sample size of 10.

(c) 100 tosses. The expected proportion of 50% is between 40% and 60%. With more flips the observed proportion of heads would often be closer to the expected value than if we used a sample size of 10.

(d) 10 tosses. The desired proportion of heads (less than 30%) is below the expected proportion, and with fewer flips it would often be more likely to observe a proportion that is not close to the expected value.

3.4

$$\begin{aligned}
P(your\ rolls) &= P(six\ and\ six\ on\ the\ first\ roll\ and\ six\ and\ six\ on\ the\ second\ roll) \\
&= P(six\ and\ six\ on\ the\ first\ roll) \times P(six\ and\ six\ on\ the\ second\ roll) \\
&= (1/6 \times 1/6) \times (1/6 \times 1/6) = 0.00077 \\
P(friend's\ rolls) &= P(one\ and\ one\ on\ the\ first\ roll\ and\ three\ and\ three\ on\ the\ second\ roll) \\
&= P(one\ and\ one\ on\ the\ first\ roll) \times P(three\ and\ three\ on\ the\ second\ roll) \\
&= (1/6 \times 1/6) \times (1/6 \times 1/6) = 0.00077
\end{aligned}$$

Both outcomes have equal probability.

3.5

(a) P(all tails) = $0.5^{10} = 0.00098$.

(b) P(all heads) = $0.5^{10} = 0.00098$.

(c) P(at least one tails) = 1 - P(no tails) = $1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

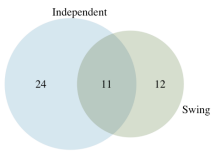
3.6

(a) P(sum of 1) = 0. Since there are two dice being rolled, the minimum possible sum is 2.
(b) P(sum of 5) = P(1,4) + P(2,3) + P(3,2) + P(4,1) = $\left(\frac{1}{6} \times \frac{1}{6}\right) \times 4 = \frac{4}{36} \approx 0.11$.
(c) P(sum of 12) = P(6,6) = $\left(\frac{1}{6} \times \frac{1}{6}\right) = \frac{1}{36} \approx 0.0278$.

3.7

(a) No, there are voters who are both independent and swing voters.

(b) The Venn diagram is shown below.



(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters.

(d) Use the General Addition Rule:

$$\begin{aligned} P(\text{Independent or swing}) &= P(\text{Independent}) + P(\text{swing}) - P(\text{Independent and swing}) \\ &= 0.35 + 0.23 - 0.11 = 0.47 \end{aligned}$$

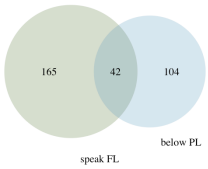
(e) P(neither Independent nor swing) = 1 - P(Independent or swing) = 1 - 0.47 = 0.53.

(f) P(Independent) $\times$ P(swing) = $0.35 \times 0.23 = 0.08$, which does not equal P(Independent and swing) = 0.11, so the events are dependent.

3.8 Poverty line: PL, foreign language: FL

(a) No, there are people who are both living below the poverty line and speak a language other than English at home.

(b) The Venn diagram is shown below.



(c) Each person living below the poverty line either speaks only English at home or doesn't. Since 14.6% of Americans live below the poverty line and 4.2% speak a language other than English at home, the other 10.4% only speak English at home.

(d) Using the General Addition Rule:

$$\begin{aligned} P(\textit{below PL or speak FL}) &= P(\textit{below PL}) + P(\textit{speak FL}) - P(\textit{both}) \\ &= 0.146 + 0.207 - 0.042 = 0.311 \end{aligned}$$

(e) P(neither below PL nor speak FL) = 1 - P(below PL or speak FL) = 1 - 0.311 = 0.689.

(f) Two approaches:
(1) Using the multiplication rule: $P(\textit{below PL}) \times P(\textit{speak FL}) = 0.146 \times 0.207 = 0.030$, which does not equal $P(\textit{below PL and speak FL}) = 0.042$, therefore the events are dependent.
(2) Using Bayes' theorem: If the two events are independent, then $P(\textit{below PL} \mid \textit{speak FL}) = P(\textit{below PL})$. Using Bayes' theorem,

$$\begin{aligned} P(\textit{below PL} \mid \textit{speak FL}) &= \frac{P(\textit{below PL and speak FL})}{P(\textit{speak FL})} \\ &= \frac{0.042}{0.207} \approx 0.203 \end{aligned}$$

Since this probability is different than $P(\textit{below PL}) = 0.146$, we determine that the two events are dependent.

3.9

(a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c).

(b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related.

(c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

3.10

(a) Since she is randomly guessing, the probability of getting each question right is $p = 0.25$. Then,
P(Wrong, Wrong, Wrong, Wrong, Right) $= 0.75^4 \times 0.25 = 0.07910156 \approx 0.0791$.

(b) P(Right, Right, Right, Right, Right) $= 0.25^5 = 0.0009765625 \approx 0.0010$.

(c) P(at least 1 Right) = 1 - P(none Right) $= 1 - 0.75^5 = 1 - 0.2373 = 0.7627$.

3.11

(a) P(at least a Bachelor's degree | male) = 0.16 + 0.09 = 0.25

(b) P(at least a Bachelor's degree | female) = 0.17 + 0.09 = 0.26

(c) Assuming that the education level of the husband and wife are independent:
P(man and woman both have at least a Bachelor's degree) = $0.25 \times 0.26 = 0.065$

(d) The independence assumption may not be reasonable since people often marry others with a comparable level of education.

3.12

(a) P(no misses) = 1 - (0.25 + 0.15 + 0.28) = 0.32

(b) P(at most 1 miss) = P(no misses) + P(1 miss) = 0.32 + 0.25 = 0.57

(c) P(at least 1 miss) = P(1 miss) + P(2 misses) + P(3+ misses) = 1 - P(no misses)
= 1 - 0.32 = 0.68

(d) For parts (d) and (e) assume that whether or not one kid misses school is independent of the other.
P(neither miss any) = P(no miss) × P(no miss) = $0.32^2 = 0.1024$

(e) P(both miss some) = P(at least 1 miss) × P(at least 1 miss) = $0.68^2 = 0.4624$

(f) These kids are siblings, and if one gets sick it probably raises the chance that the other one will get sick as well. So whether or not one misses school due to sickness is probably not independent of the other.

3.13

(a) No, we cannot compute $P(A \textit{ and } B)$ since we do not know if A and B are independent. We could if A and B were independent.

(b) i. P(A and B) = P(A) $\times$ P(B) = 0.21.

ii. P(A or B) = P(A) + P(B) - P(A and B) = 0.3 + 0.7 - 0.21 = 0.79.

iii. P(A|B) = P(A) = 0.3.

(c) No, because $0.1 \neq 0.21$.

(d) P(A|B) = P(A and B) / P(B) = 0.1 / 0.7 = 0.143.

3.14 $P(J \mid PB) = \frac{P(PB \text{ and } J)}{P(PB)} = \frac{0.78}{0.80} = 0.975$

3.15

(a) No, 0.18 of respondents fall into this combination.

(b) P(earth is warming or liberal Democrat) =
= P(earth is warming) + P(liberal Democrat) - P(earth is warming and liberal Democrat)
= 0.60 + 0.20 - 0.18 = 0.62

(c) P(earth is warming | liberal Democrat) = $\frac{0.18}{0.20} = 0.9$

(d) P(earth is warming | conservative Republican) = $\frac{0.11}{0.33} = 0.33$

(e) No, the two appear to be dependent. The percentages of conservative Republicans and liberal Democrats who believe that there is solid evidence that the average temperature on earth has been getting warmer over the past few decades are very different.

(f) P(moderate/liberal Republican | not warming) = $\frac{0.06}{0.34} = 0.18$

3.16

(a) No, there are individuals who are both excellent in health and have health coverage.

(b) P(excellent health) = 0.2329

(c) P(excellent health | health coverage) = 0.2099 / 0.8738 = 0.24

(d) P(excellent health | no health coverage) = 0.0230 / 0.1262 = 0.18

(e) No, because the probability that a person has excellent health varies between the two health coverage categories (24% vs 18%). That is, knowing something about someone's health coverage provides useful information in predicting whether the person has excellent health, which means the variables are not independent.

3.17

(a) No. There are 6 females who like Five Guys Burgers.

(b) P(In-N-Out | male) = 162 / 248 $\approx$ 0.65.

(c) P(In-N-Out | female) = 181 / 252 $\approx$ 0.72.

(d) Under the assumption of independence of gender and hamburger preference: P(man and woman dating both like In-N-Out burgers the best) = 0.65 $\times$ 0.72 = 0.468. While it is possible there is some mysterious connection between burger choice and finding a partner, independence is probably a reasonable assumption.

(e) P(Umami or female) = P(Umami) + P(female) - P(Umami and female) = $\frac{6+252-1}{500} = 0.514$.

3.18

(a) P(man or partner has blue eyes) = (108 + 114 - 78) / 204 = 0.7059

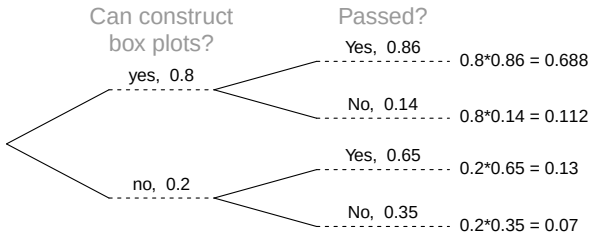
(b) P(partner with blue eyes | man with blue eyes) = 78 / 114 = 0.6842

(c) P(partner with blue eyes | man with brown eyes) = 19 / 54 = 0.3519
P(partner with blue eyes | man with green eyes) = 11 / 36 = 0.3056

(d) It is much more likely for a man with blue eyes to have a partner with blue eyes than a man with another eye color to have a partner with blue eyes. Therefore it appears that eye colors of males and their partners are not independent.

3.19

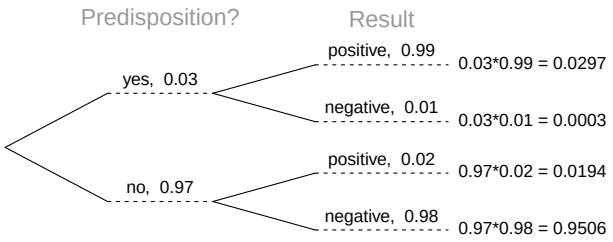
(a) A tree diagram of this scenario is below:



(b)

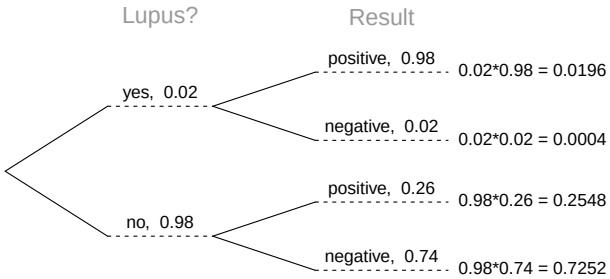
$$\begin{aligned} P(can\ construct \mid pass) &= \frac{P(pass\ and\ can\ construct)}{P(pass)} \\ &= \frac{0.80 \times 0.86}{0.80 \times 0.86 + 0.20 \times 0.65} \\ &= \frac{0.688}{0.818} \\ &\approx 0.84 \end{aligned}$$

3.20



$$
\begin{aligned}
P(pre \mid positive) &= \frac{P(pre \text{ and } positive)}{P(positive)} \\
&= \frac{0.0297}{0.0297 + 0.0194} \\
&= 0.6049
\end{aligned}
$$

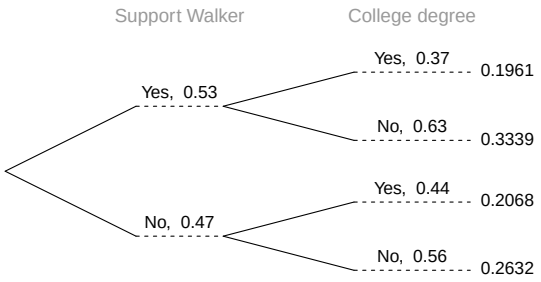
3.21



$$\begin{aligned} P(lupus|positive) &= \frac{P(lupus\ and\ positive)}{P(positive)} \\ &= \frac{0.0196}{0.0196 + 0.2548} \\ &= 0.0714 \end{aligned}$$

Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.

3.22



$$
\begin{aligned}
P(support \mid college) &= \frac{P(\text{support and college})}{P(college)} \\
&= \frac{0.1961}{0.1961 + 0.2068} \\
&= 0.49
\end{aligned}
$$

3.23

(a) $P(1^{st} \text{ marble } B) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$

(b) $P(2^{nd} \text{ marble } B | 1^{st} \text{ marble } B) = \frac{3}{10} = 0.3$

(c) $P(2^{nd} \text{ marble } B | 1^{st} \text{ marble } O) = \frac{3}{10} = 0.3$

(d) $P(1^{st} \text{ marble } B) \cdot P(2^{nd} \text{ marble } B | 1^{st} \text{ marble } B) = 0.3 \times 0.3 = 0.09$

(e) Yes, each draw is from the same set of marbles..

3.24

(a) Blue Blue $\rightarrow \frac{4}{12} \times \frac{3}{11} = 0.0909$

(b) Not Gray Not Gray $\rightarrow \frac{7}{12} \times \frac{6}{11} = 0.3182$

(c) P(at least 1 black) = 1 - P(no black)
P(no black) $\rightarrow$ Not Black Not Black $\rightarrow \frac{9}{12} \times \frac{8}{11} = 0.5455$
P(at least 1 black) = 1 - 0.5455 = 0.4545.

(d) 0, there are no green socks in the drawer.

(e) Blue Blue $\rightarrow 0.0909$
Black Black $\rightarrow \frac{3}{12} \times \frac{2}{11} = 0.0455$
Gray Gray $\rightarrow \frac{5}{12} \times \frac{4}{11} = 0.1515$
P(matching socks) = 0.0909 + 0.0455 + 0.1515 = 0.2879.

3.25

(a) $P(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } B) = \frac{2}{9} = 0.22$

(b) $P(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } O) = \frac{3}{9} = 0.33$

(c) $P(1^{st} \text{ chip } B) \cdot P(2^{nd} \text{ chip } B \mid 1^{st} \text{ chip } B) = 0.3 \times 0.22 = 0.067$

(d) They are dependent. Notice that the probability of a blue chip on the second draw depended on the chip color in the first draw in parts (a) and (b).

3.26

(a) P(first hardcover, second paperback fiction) $= \frac{28}{95} \times \frac{59}{94} = 0.1850$

(b) Break this into two disjoint statements (let $F = fiction$, $N = nonfiction$, $H = hardcover$, $P = paperback$):

$$\begin{aligned} P(\text{1st F, 2nd H}) &= P((\text{1st FH and 2nd H}) \text{ OR } (\text{1st FP and 2nd H})) \\ &= P(\text{1st FH and 2nd H}) + P(\text{1st FP and 2nd H}) \\ &= \frac{13}{95}\frac{27}{94} + \frac{59}{95}\frac{28}{94} = 0.2243 \end{aligned}$$

(c) Same probability statements as part (c), except now the calculations are $\frac{13}{95}\frac{28}{95} + \frac{59}{95}\frac{28}{95} = 0.2234$.

(d) There are so many books on the bookcase and we are only drawing two books, so the probability associated with the second book will be almost entirely unaffected by whatever book is drawn first. This makes the second draw under the "without replacement" setting almost independent of the first draw, meaning it is about equivalent to drawing with replacement.

3.27 The number of students wearing leggings is $24-(7+4+8)=5$. When selecting 3 students, there are three scenarios under which we would get one student with leggings and two with jeans:

$$\begin{aligned}\text{Scenario 1} &= \text{P(L,J,J)} = \tfrac{5}{24}\times\tfrac{7}{23}\times\tfrac{6}{22} = 0.0173\\ \text{Scenario 2} &= \text{P(J,L,J)} = \tfrac{7}{24}\times\tfrac{5}{23}\times\tfrac{6}{22} = 0.0173\\ \text{Scenario 3} &= \text{P(J,J,L)} = \tfrac{7}{24}\times\tfrac{6}{23}\times\tfrac{5}{22} = 0.0173\end{aligned}$$

These scenarios exhaust the ways where we can get one student with leggings and two students with jeans, and these scenarios are also disjoint. Therefore, to determine the probability that one of the selected students is wearing leggings and the other two are wearing jeans, we add up the calculated probabilities:

$$0.0173 + 0.0173 + 0.0173 = 0.0519$$

Note that each scenario has the same probability.

3.28

(a) P(first two people share a birthday) = 1/365 = 0.0027.

(b) P(at least one pair of people share a birthday)
= 1- P(none of the three people share a birthday) = 1 - (365/365)(364/365)(363/365) = 0.0082.

3.29

(a) $E(X) = 100 \times 0.13 = 13$

(b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

3.30

(a) The probability model and the calculation of the expected value and standard deviation are shown below:

Event	X	$P(X)$	$X \cdot P(X)$	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
Red	0	$\frac{26}{52}$	$0 \times \frac{26}{52} = 0$	$(0 - 4.14)^2 = 17.1396$	$17.1396 \times \frac{26}{52} = 8.5698$
Spade	5	$\frac{13}{52}$	$5 \times \frac{13}{52} = 1.25$	$(5 - 4.14)^2 = 0.7396$	$0.7396 \times \frac{13}{52} = 0.1849$
Club	10	$\frac{12}{52}$	$10 \times \frac{12}{52} = 2.31$	$(10 - 4.14)^2 = 34.3396$	$34.3396 \times \frac{12}{52} = 7.9245$
Ace of clubs	30	$\frac{1}{52}$	$30 \times \frac{1}{52} = 0.58$	$(30 - 4.14)^2 = 668.7396$	$668.7396 \times \frac{1}{52} = 12.8604$
			$E(X) = 4.14$		$V(X) = 29.5396$
					$SD(X) = \sqrt{V(X)} = 5.4350$

(b) If you are playing with the goal of making money, you should not play unless the cost of the game is less than \$4.14.

3.31

(a) The probability model and the calculation of the expected value are shown below:

Event	X	P(X)	X · P(X)	$(X - E(X))^2 \cdot P(X)$
3 hearts	50	$\frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} = 0.0129$	0.65	$(50 - 3.59)^2 \times 0.0129 = 27.87$
3 blacks	25	$\frac{26}{52} \times \frac{25}{51} \times \frac{24}{50} = 0.1176$	2.94	$(25 - 3.59)^2 \times 0.1176 = 53.93$
Else	0	$1 - (0.0129 + 0.1176) = 0.8695$	0	$(0 - 3.59)^2 \times 0.8695 = 11.21$
			E(X) = \$3.59	$V(X) = 93.01$
				$SD(X) = \sqrt{V(X)} = 9.64$

(b) Let X denote winnings, then net profit can be denoted as $X - 5$.

$$E(X - 5) = E(X) - 5 = 3.59 - 5 = -\$1.41$$
$$SD(X - 5) = SD(X) = 9.64$$

(c) No, the expected net profit is negative, so on average you expect to lose money.

3.32

(a) The probability model for the amount Andy can profit at this game and the calculation of the expected profits is as follows:

Event	X	$P(X)$	$X \cdot P(X)$
Number	-2	$\frac{36}{52} = 0.6923$	$-2 \times \frac{36}{52} = -1.38$
J, Q, K	1	$\frac{12}{52} = 0.2308$	$1 \times \frac{12}{52} = 0.23$
Ace	3	$\frac{3}{52} = 0.0577$	$3 \times \frac{3}{52} = 0.17$
Ace of clubs	23	$\frac{1}{52} = 0.0192$	$23 \times \frac{1}{52} = 0.44$
			$E(X) = -0.54$

(b) No, he is expected to lose money on average.

3.33

Event	X	$P(X)$	$X \cdot P(X)$
Boom	0.18	$\frac{1}{3}$	$0.18 \times \frac{1}{3} = 0.06$
Normal	0.09	$\frac{1}{3}$	$0.09 \times \frac{1}{3} = 0.03$
Recession	-0.12	$\frac{1}{3}$	$-0.12 \times \frac{1}{3} = -0.04$
			$E(X) = 0.05$

The expected return is a 5% increase in value.

3.34

(a) The probability model and the calculation of average revenue per passenger (expected value) are as follows:

Event	X	P(X)	X · P(X)	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
No baggage	0	0.54	0	$(0 - 15.70)^2 = 246.49$	$246.49 \times 0.54 = 133.10$
1 checked bag	25	0.34	8.5	$(25 - 15.70)^2 = 86.49$	$86.49 \times 0.34 = 29.41$
2 checked bags	60	0.12	7.2	$(60 - 15.70)^2 = 1962.49$	$1962.49 \times 0.12 = 235.50$
			E(X) = \$15.70		V(X) = \$398.01
					$SD(X) = \sqrt{V(X)} = \$19.95$

(b) We assume independence between individual fliers. This probably is not exactly correct, but it would provide a helpful first approximation for the true revenue.

$$E(X_1 + \cdots + X_{120}) = E(X_1) + \cdots + E(X_{120}) = 120 \times 15.70 = \$1,884$$
$$V(X_1 + \cdots + X_{120}) = V(X_1) + \cdots + V(X_{120}) = 120 \times 398.01 = \$47,761.20$$
$$SD(X_1 + \cdots + X_{120}) = \sqrt{47,761.20} = \$218.54$$

3.35

X	$P(X)$	$X \cdot P(X)$	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
1	$\frac{18}{38}$	$1 \times \frac{18}{38} = \frac{18}{38}$	$(1 + 0.0526)^2 = 1.1080$	$1.1080 \times \frac{18}{38} = 0.5248$
-1	$\frac{20}{38}$	$-1 \times \frac{20}{38} = -\frac{20}{38}$	$(-1 + 0.0526)^2 = 0.8976$	$0.8976 \times \frac{20}{38} = 0.4724$
		$E(X) = -\frac{2}{38} = -0.0526$		$V(X) = 0.9972$
				$SD(X) = \sqrt{V(X)} = 0.9986$

3.36

(a) Let X indicate winnings.

X	$P(X)$	$X \cdot P(X)$	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
3	$\frac{18}{37}$	$3 \times \frac{18}{37} = 1.46$	$(3 + 0.081)^2 = 9.4926$	$9.4926 \times \frac{18}{37} = 4.62$
-3	$\frac{19}{37}$	$-3 \times \frac{19}{37} = -1.54$	$(-3 + 0.081)^2 = 8.5206$	$8.5206 \times \frac{19}{37} = 4.38$
		$E(X) = -0.081$		$V(X) = 9$
				$SD(X) = \sqrt{V(X)} = 3$

(b) First, we need to solve for the expected value and variance of the winnings in one round. Let $Y = 1/3\ X$.

$$E(Y) = E(1/3\ X) = 1/3\ E(X) = 1/3 \times -0.081 = -0.027$$
$$V(Y) = V(1/3\ X) = 1/9\ V(X) = 1/9 \times 9 = 1$$
$$SD(Y) = \sqrt{V(Y)} = 1$$

Then,

$$E(Y) + E(Y) + E(Y) = 3\ E(Y) = -0.081$$
$$V(Y) + V(Y) + V(Y) = 1 + 1 + 1 = 3$$
$$SD(Y) + SD(Y) + SD(Y) = \sqrt{3} = 1.73$$

(c) The expected values of the two games are the same, but the standard deviation is much higher for the first game, meaning that playing once with \$3 is riskier than playing three times with \$1.

3.37 Approximate answers are OK.

(a) (29 + 32) / 144 = 0.4236

(b) 21 / 144 = 0.1458

(c) (26 + 12 + 15) / 144 = 0.3681

3.38

(a) The distribution is right skewed, with a median between \$35,000 and \$49,999. The IQR of the distribution is very roughly about \$30,000. The distribution is skewed to the high end, and there are probably outliers on the high end due to the nature of the data.

(b) P(less than \$50,000) = 2.2 + 4.7 + 15.8 + 18.3 + 21.2 = 62.2%

(c) Assuming that gender and income are independent:
P(less than \$50,000 and female) = P(less than \$50,000) × P(female) = 0.622 × 0.41 = 0.255 = 25.5%

(d) If these variables were independent, then the percentage of females who earn less than \$50,000 (71.8%) would equal the percentage of all people who make less than \$50,000 (62.2%). Since this is not the case, gender and income are dependent.

3.39

(a) Invalid. The sum of the probabilities is greater than 1 ($0.3 + 0.3 + 0.3 + 0.2 + 0.1 = 1.2$).

(b) Valid. The probabilities are all between 0 and 1, and they sum to 1. In this class, every person gets a C!

(c) Invalid. The sum of the probabilities is less than 1 ($0.3 + 0.3 + 0.3 + 0 + 0 = 0.9$).

(d) Invalid. There is a negative probability listed (-0.1 for F).

(e) Valid. The probabilities are all between 0 and 1, and they sum to 1. In this distribution, 80% of students get an A, B, or C.

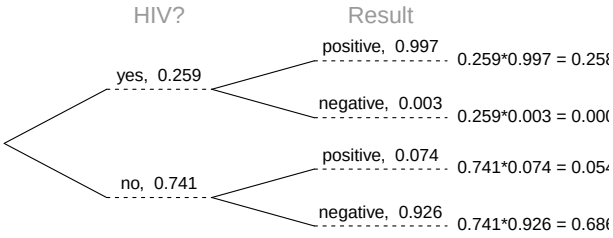
(f) Invalid. There is a negative probability listed (-0.1 for B).

3.40

(a) P(excellent health and no health coverage) = $459/20,000 \approx 0.023$

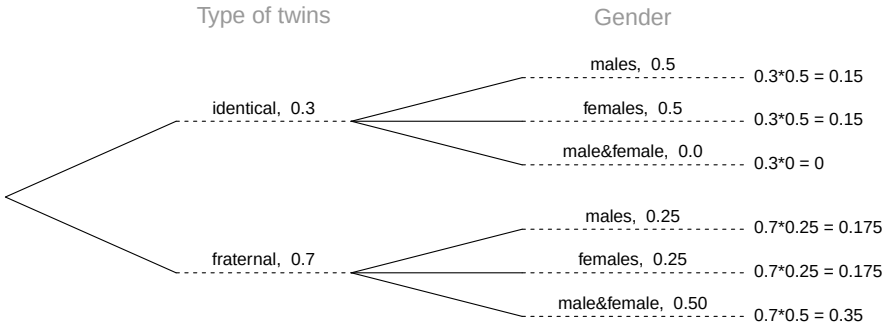
(b) P(excellent health or no health coverage) = P(excellent health) + P(no health coverage) - P(excellent health and no health coverage)
$= \frac{4,657+2,524-459}{20,000} \approx 0.336$

3.41



$$
\begin{aligned}
P(HIV \mid +) &= \frac{P(HIV \text{ and } +)}{P(+)} \\
&= \frac{0.259 \times 0.997}{0.259 \times 0.997 + (1 - 0.259) \times (1 - 0.926)} \\
&= \frac{0.2582}{0.3131} \\
&= 0.8247
\end{aligned}
$$

3.42



$$\begin{aligned} P(identical \mid females) &= \frac{P(identical\ and\ females)}{P(females)} \\ &= \frac{0.15}{0.15 + 0.175} \\ &= 0.46 \end{aligned}$$

3.43

(a) Let X represent the amount Sally spends on coffee (in ¢), and Y represent the amount she spends on muffins (in ¢).

$$\begin{array}{ll} E(X) = 140 & E(Y) = 250 \\ SD(X) = 30 & SD(Y) = 15 \\ V(X) = 30^2 = 900 & V(Y) = 15^2 = 225 \end{array}$$

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) = 140 + 250 = 390¢ = \$3.90 \\ V(X+Y) &= V(X) + V(Y) = 900 + 225 = 1125¢^2 \\ SD(X+Y) &= \sqrt{V(X) + V(Y)} = \sqrt{1125} = 34¢ = \$0.34 \end{aligned}$$

(b) Let W represent the amount Sally spends on coffee and breakfast each week. Then,

$$W = (X_1 + Y_1) + \cdots + (X_7 + Y_7)$$

$$\begin{aligned} E(W) &= E((X_1 + Y_1) + \cdots + (X_7 + Y_7)) = \underbrace{3.90 + \cdots + 3.90}_{7} = 7 \times 3.90 = \$27.30 \\ V(W) &= V((X_1 + Y_1) + \cdots + (X_7 + Y_7)) = \underbrace{1125 + \cdots + 1125}_{7} = 7 \times 1125 = 7875¢^2 \\ SD(W) &= \sqrt{7875} = 89¢ = \$0.89 \end{aligned}$$

3.44

(a) $E(X + Y_1 + Y_2 + Y_3) = E(X) + 3 * E(Y) = 48 + 3 \times 2 = 54$
$V(X + Y_1 + Y_2 + Y_3) = V(X) + 3 \times V(Y) = 1 + 3 \times 0.0625 = 1.1875$
$SD(X + Y_1 + Y_2 + Y_3) = \sqrt{V(X + Y_1 + Y_2 + Y_3)} = \sqrt{1.1875} \approx 1.09$

(b) $E(X - Y) = E(X) - E(Y) = 48 - 2 = 46$
$V(X - Y) = V(X) + V(Y) = 1 + 0.0625 = 1.0625$
$SD(X - Y) = \sqrt{V(X - Y)} = \sqrt{1.0625} \approx 1.03$

(c) Initially we do not know exactly how much ice cream is in the box. Then we scoop out an unknown amount. We should now be even more unsure about the amount of ice cream that is left in the box.

3.45

$$\begin{aligned} Var\left(\frac{X_1 + X_2}{2}\right) &= Var\left(\frac{X_1}{2} + \frac{X_2}{2}\right) \\ &= \frac{Var(X_1)}{2^2} + \frac{Var(X_2)}{2^2} \\ &= \frac{\sigma^2}{4} + \frac{\sigma^2}{4} \\ &= \frac{\sigma^2}{2} \end{aligned}$$

3.46

$$\begin{aligned} Var\left(\frac{X_1+X_2+X_3}{3}\right) &= Var\left(\frac{X_1}{3}+\frac{X_2}{3}+\frac{X_3}{3}\right) \\ &= \frac{Var(X_1)}{3^2}+\frac{Var(X_2)}{3^2}+\frac{Var(X_3)}{3^2} \\ &= \frac{\sigma^2}{9}+\frac{\sigma^2}{9}+\frac{\sigma^2}{9} \\ &= \frac{\sigma^2}{3} \end{aligned}$$

3.47

$$\begin{aligned}
Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) &= Var\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right) \\
&= \frac{Var(X_1)}{n^2} + \frac{Var(X_2)}{n^2} + \cdots + \frac{Var(X_n)}{n^2} \\
&= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2} \quad \text{(there are } n \text{ of these terms)} \\
&= n\frac{\sigma^2}{n^2} \\
&= \sigma^2/n
\end{aligned}$$

Chapter 4

Distributions of random variables

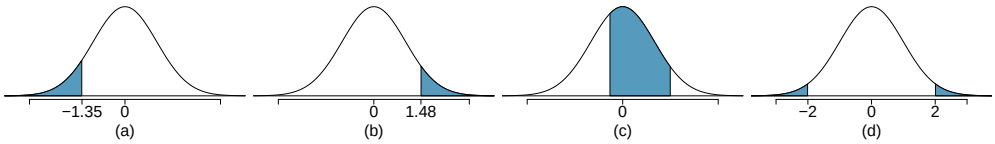
4.1

(a) $P(Z < -1.35) = 0.0885 \ \rightarrow \ 9\%$

(b) $P(Z > 1.48) = 1 - 0.9306 = 0.0694 \ \rightarrow \ 7\%$

(c) $P(-0.4 < Z < 1.5) = P(Z < 1.5) - P(Z < -0.4) = 0.9332 - 0.3446 = 0.5886 \ \rightarrow \ 59\%$

(d) $P(|Z| > 2) = P(Z < -2) + P(Z > 2) = 0.0228 + (1 - 0.9772) = 0.0456 \ \rightarrow \ 5\%$



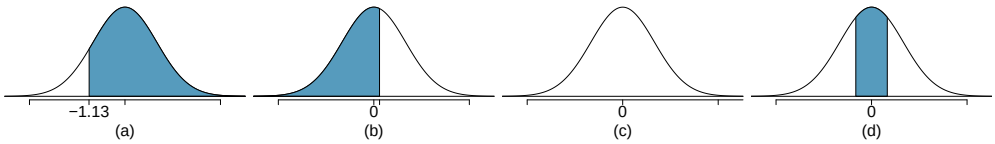
4.2

(a) $P(Z > -1.13) = 1 - 0.1292 = 0.8708 \ \rightarrow \ 87\%$

(b) $P(Z < 0.18) = 0.5714 \ \rightarrow \ 57\%$

(c) $P(Z > 8) \approx 0 \ \rightarrow \ 0\%$

(d) $P(|Z| < 0.5) = P(-0.5 < Z < 0.5) = P(Z < 0.5) - P(Z < -0.5)$
$= 0.6915 - 0.3085 = 0.3830 \ \rightarrow \ 38\%$



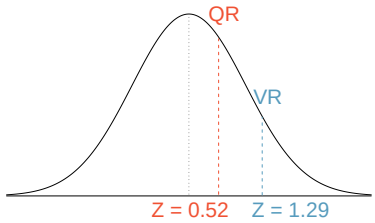
4.3

(a) Let X denote scores on the Verbal Reasoning section and Y denote scores on the Quantitative Reasoning section. Then,

$$Verbal: \ N(\mu = 151, \sigma = 7)$$
$$Quant: \ N(\mu = 153, \sigma = 7.67)$$

(b) The Z scores can be calculated as follows:

$$Z_{VR} = \frac{x - \mu}{\sigma} = \frac{160 - 151}{7} = 1.29$$
$$Z_{QR} = \frac{y - \mu}{\sigma} = \frac{157 - 153}{7.67} = 0.52$$

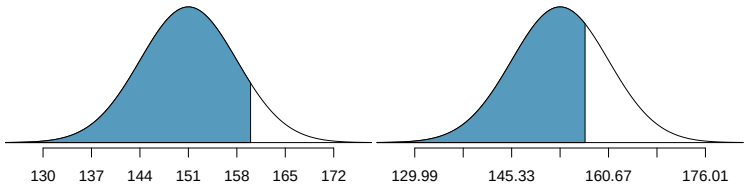


(c) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section.

(d) She did better on the Verbal Reasoning section since her Z score on that section was higher.

(e) The percentile scores can be calculated as follows:

$$Perc_{VR} = P(Z < 1.29) = 0.9014 \approx 10\%$$
$$Perc_{QR} = P(Z < 0.52) = 0.6984 \approx 30\%$$



(f) $100 - 90 = 10\%$ did better than her on the Verbal Reasoning section and $100 - 70 = 30\%$ did better than her on the Quantitative Reasoning section.

(g) We cannot compare the raw scores since they are on different scales. Her scores will be measured relative to the merits of other students on each exam, so it is helpful to consider the Z score. Comparing her percentile scores is a more appropriate way of determining how well she did compared to others taking the exams.

(h) Answer to part (b) would not change as Z scores can be calculated for distributions that are not normal. However, we could not answer parts (c)-(f) since we cannot use the Z table to calculate probabilities and percentiles without a normal model.

4.4

(a) Let X denote the finishing times of *Men, Ages 30 - 34* and Y denote the finishing times of emphWomen, Ages 25 - 29. Then,

$$X \sim N(\mu = 4313, \sigma = 583)$$
$$Y \sim N(\mu = 5261, \sigma = 807)$$

(b) The Z scores can be calculated as follows:

$$Z_{Leo} = \frac{x - \mu}{\sigma} = \frac{4948 - 4313}{583} = 1.09$$
$$Z_{Mary} = \frac{y - \mu}{\sigma} = \frac{5513 - 5261}{807} = 0.31$$

Leo finished 1.09 standard deviations above the mean of his group's finishing time and Mary finished 0.31 standard deviations above the mean of her group's finishing time.

(c) Mary ranked better since she she has a lower Z score indicating that her finishing time is relatively shorter.

(d) Leo:

$$\begin{aligned} P(Z > 1.09) &= 1 - P(Z < 1.09) \\ &= 1 - 0.8621 \\ &= 0.1379 \ \rightarrow \ 13.79\% \end{aligned}$$

(e) Mary:

$$\begin{aligned} P(Z > 0.31) &= 1 - P(Z < 0.31) \\ &= 1 - 0.6217 \\ &= 0.3783 \ \rightarrow \ 37.83\% \end{aligned}$$

(f) Answer to part (b) would not change as Z scores can be calculated for distributions that are not normal. However, we could not answer parts (c)-(e) since we cannot use the Z table to calculate probabilities and percentiles without a normal model.

4.5

(a) The Z score corresponding to the 80^{th} percentile of the normal distribution is approximately 0.84. Then,

$$Z = 0.84 = \frac{y - 153}{7.67} \;\rightarrow\; y = 0.84 \times 7.67 + 153 = 159.44 \approx 159$$

(b) If a student scored worse than 70% of the test takers, it means she scored in the 30^{th} percentile. The Z score corresponding to the 30^{th} percentile of the normal distribution is approximately -0.52. Then,

$$Z = -0.52 = \frac{x - 151}{7} \;\rightarrow\; x = -0.52 \times 7 + 151 = 147.36 \approx 147$$

4.6

(a) The fastest 5% are in the 5^{th} percentile of the distribution. The Z score corresponding to the 5^{th} percentile of the normal distribution is approximately -1.64. Then,

$$Z = -1.65 = \frac{x - 4313}{583} \;\rightarrow\; x = -1.65 \times 583 + 4313 = 3351 \; sec$$

The fastest 5% of males in this age group finished in less than 56 minutes.

(b) The slowest 10% are in the 90^{th} percentile of the distribution. The Z score corresponding to the 90^{th} percentile of the normal distribution is approximately 1.28. Then,

$$Z = 1.28 = \frac{y - 5261}{807} \;\rightarrow\; y = 1.28 \times 807 + 5261 = 6294 \; sec$$

The slowest 10% of females in this age group took 1 hour, 45 minutes or longer to finish.

4.7

(a) Let X denote temperatures, then $X \sim N(\mu = 77, \sigma = 5)$.

$$\begin{aligned}(X > 83) &= P\left(Z > \frac{83 - 77}{5}\right) \\ &= P(Z > 1.2) \\ &= 1 - 0.8849 = 0.1151\end{aligned}$$

(b) The Z score corresponding to the bottom 10% (or 10^{th} percentile) is -1.28.

$$Z = -1.28 = \frac{x - 77}{5} \;\rightarrow\; x = -1.28 \times 5 + 77 = 70.6^{\circ}F \text{ or colder}$$

4.8

(a) Let X denote returns on this portfolio, then $X \sim N(\mu = 14.7, \sigma = 33)$.

$$P(X < 0) = P\left(Z < \frac{0 - 14.7}{33}\right) = P(Z < -0.45) = 0.3264 \ \rightarrow \ 32.64\%$$

(b) The Z score corresponding to the top 15% (or 85^{th} percentile) is 1.04.

$$Z = 1.04 = \frac{x - 14.7}{33} \ \rightarrow \ x = 1.04 \times 33 + 14.7 = 49.02$$

4.9

(a) The parameters of the distribution in °C can be calculated as follows:

$$\begin{aligned} \mu_C &= (\mu_F - 32) \times \tfrac{5}{9} \\ &= (77 - 32) \times \tfrac{5}{9} \\ &= 25^\circ C \end{aligned} \qquad \begin{aligned} \sigma_C &= \sigma_F \times \tfrac{5}{9} \\ &= 5 \times \tfrac{5}{9} \\ &= 2.78^\circ C \end{aligned}$$

Temperatures in °C $\sim N(25, 2.78)$

(b) Probability of observing a 28 °C temperature or higher in June in LA can be calculated as follows:

$$Z = \frac{28 - 25}{2.78} = 1.08$$
$$P(X > 28) = P(Z > 1.08) = 1 - P(Z < 1.08) = 1 - 0.8599 = 0.1401$$

(c) We got the same answer and this was expected. It doesn't matter which scale we use, °F or °C, once we standardize we get the same Z scores and therefore the probabilities of observing a 82.4 °F or 28 °C or higher temperature are equal.

(d) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26. 86 - 23.13 = 3.73.

4.10 The Z score corresponding to the top 18.5% (or the 81.5^{th} percentile) is approximately 0.90.

$$Z = 0.90 = \frac{220 - 185}{\sigma} \;\rightarrow\; \sigma = \frac{220 - 185}{0.90} = 38.9 \; mg/dl$$

4.11

(a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified.

(b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6

4.12

(a) With replacement: $P(F) \times P(F) = \frac{5}{10} \times \frac{5}{10} = 0.25$
Without replacement: $P(F) \times P(F) = \frac{5}{10} \times \frac{4}{9} = 0.2222$

(b) With replacement: $P(F) \times P(F) = \frac{5{,}000}{10{,}000} \times \frac{5{,}000}{10{,}000} = 0.25$
Without replacement: $P(F) \times P(F) = \frac{5{,}000}{10{,}000} \times \frac{4{,}999}{9{,}999} = 0.249975$

(c) When the population size is small, the results of sampling with and without replacement are quite different. However when the population size is large, the two probabilities are almost identical because the fraction of females remaining in the population when drawing with replacement stays very nearly the same regardless of the first person sampled.

4.13

(a) Since we are asked for the probability of a certain number of trials until the first success we use a geometric distribution with $p = 0.125$. Let X be the trial at which the first success (first blue-eyed child) occurs. Then,

$$P(3^{rd} \text{ child is the first blue-eyed}) = P(X = 3) = 0.875^2 \times 0.125 = 0.096$$

(b) The mean and the standard deviation for a geometric distribution with $p = 0.125$ can be calculated as follows:

$$\mu = \frac{1}{p} = \frac{1}{0.125} = 8, \qquad \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.125}{0.125^2}} = 7.48$$

4.14

(a) Since we are asked for the probability of a certain number of trials until the first success we use a geometric distribution with $p = 0.02$. Let X be the trial at which the first success (defective transistor) occurs. Then,

$$P(\text{defective transistor on the } 10^{th} \text{ try}) = 0.98^9 \times 0.02 = 0.0167$$

(b) P(no defective transistors in 100) $= 0.98^{100} = 0.1326$.

(c) The mean and the standard deviation for a geometric distribution with $p = 0.02$ can be calculated as follows:

$$\mu = \frac{1}{p} = \frac{1}{0.02} = 50, \qquad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{0.98}{0.02^2}} \approx 49.5$$

(d) The mean and the standard deviation for a geometric distribution with $p = 0.05$ can be calculated as follows:

$$\mu = \frac{1}{p} = \frac{1}{0.05} = 20, \qquad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{0.95}{0.05^2}} \approx 19.49$$

(e) As p gets larger, i.e. the event occurs more often, the expected number of trials before a success and the standard deviation decrease.

4.15 If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = E[X] &= P(X = 0) \times 0 + P(X = 1) \times 1 \\ &= (1 - p) \times 0 + p \times 1 = 0 + p = p\end{aligned}$$

4.16 Letting X be Bernoulli with a probability of success p, the variance can be computed:

$$\begin{aligned}\sigma^2 &= P(X=0)(0-p)^2 + P(X=1)(1-p)^2 \\ &= (1-p)p^2 + p(1-p)^2 = p(1-p)\end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

4.17

(a) In order to determine if we can we use the binomial distribution to calculate the probability of finding exactly six people out of a random sample of ten 18-20 year olds who consumed alcoholic beverages, we need to check if the binomial conditions are met:

1. Independent trials: In a random sample, whether or not one 18-20 year old has consumed alcohol does not depend on whether or not another one has.
2. Fixed number of trials: $n = 10$.
3. Only two outcomes at each trial: Consumed or did not consume alcohol.
4. Probability of a success is the same for each trial: $p = 0.697$.

(b) Let X = number of people who have consumed alcohol in the group, then, using a binomial distribution with $n = 10$ and $p = 0.697$:

$$P(X = 6) = \binom{10}{6} \times 0.697^6 \times 0.303^4 = 210 \times 0.697^6 \times 0.303^4 = 0.203$$

(c) P(6 out of 10 have consumed alcohol) = P(4 out of 10 have not consumed alcohol) = 0.203

(d) P(at most 2) = P(less than or equal to 2)
Let X be the number of people who have consumed alcohol in the group. Then, using a binomial distribution with $n = 5$ and $p = 0.697$:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{5}{0} \times 0.697^0 \times 0.303^5 + \binom{5}{1} \times 0.697^1 \times 0.303^4 + \binom{5}{2} \times 0.697^2 \times 0.303^3 \\ &= 0.0026 + 0.0293 + 0.1351 \\ &= 0.167 \end{aligned}$$

(e) P(at least 1) = P(greater than or equal to 1):

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + \cdots + P(X = 5) \\ &= 1 - P(X = 0) \\ &= 1 - 0.003 \\ &= 0.997 \end{aligned}$$

4.18

(a) In order to determine if we can we use the binomial distribution to calculate the probability of finding exactly 97 people out of a random sample of 100 American adults had chickenpox in childhood, we need to check if the binomial conditions are met:

1. Independent trials: In a random sample, whether or not one adult has had chickenpox does not depend on whether or not another one has.
2. Fixed number of trials: $n = 100$.
3. Only two outcomes at each trial: Have or have not had chickenpox.
4. Probability of a success is the same for each trial: $p = 0.90$.

(b) Let X be number of people who have had chickenpox in childhood, using a binomial distribution with $n = 100$ and $p = 0.90$:

$$P(X = 97) = \binom{100}{97} \times 0.90^{97} \times 0.10^{3} = 0.0059$$

(c) P(97 out of 100 did have chickenpox) = P(3 out of 100 did not have chickenpox in childhood) = 0.0059

(d) P(at least 1) = P(greater than or equal to 1):

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + \cdots + P(X = 10) \\ &= 1 - P(X = 0) \\ &= 1 - 0.10^{10} \\ &\approx 1 \end{aligned}$$

(e) P(at most 3 did not have chickenpox) = P(less than or equal to 3 where p = 0.10)

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \binom{10}{0} \times 0.10^{0} \times 0.90^{10} + \binom{10}{1} \times 0.10^{1} \times 0.90^{9} + \binom{10}{2} \times 0.10^{2} \times 0.90^{8} + \binom{10}{3} \times 0.10^{3} \times 0.90^{7} \\ &= 0.3487 + 0.3874 + 0.1937 + 0.0574 \\ &= 0.9872 \end{aligned}$$

4.19

(a) Since we are asked for the expected number of successes in a given number of trials, we use the binomial distribution with $n = 50$ and $p = 0.697$:

$$\mu = np = 50 \times 0.697 = 34.85$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{50 \times 0.697 \times (1 - 0.697)} = 3.25$$

(b) $Z = \frac{45 - 34.85}{3.25} = 3.12$. 45 is more than 3 standard deviations away from the mean, we can assume that it is an unusual observation. Therefore yes, we would be surprised.

(c) $P(X \geq 45) = P(X = 45) + P(X = 46) + \cdots + P(X = 50)$
Since this is a bit tedious to solve using the binomial distribution, we can instead use the normal approximation to binomial to estimate this probability. But first we must verify that np and $n(1-p)$ are at least 10.

$$np = 50 \times 0.697 = 34.85 > 10 \qquad \textit{and} \qquad n(1-p) = 50 \times (1 - 0.697) = 15.15 > 10$$

Since the conditions are met, we can use the normal model, $N(\mu = 34.85, \sigma = 3.25)$.

$$\begin{aligned} P(X > 45) &= P\left(Z > \frac{45 - 34.85}{3.25}\right) \\ &= P(Z > 3.12) \\ &= 1 - 0.9991 \\ &= 0.0009 \end{aligned}$$

In part (b) we had determined that it would be unusual to observe 45 or more 18-20 year olds who have consumed alcoholic beverages among a random sample of 50, which is consistent with the calculated probability in this part.

If we were to apply a 0.5 correction, the calculations would change very slightly, still yielding a low probability.

$$\begin{aligned} P(X > 45 - 0.5) = P(X > 44.5) &= P\left(Z > \frac{44.5 - 34.85}{3.25}\right) \\ &= P(Z > 2.97) \\ &= 1 - 0.9985 \\ &= 0.0015 \end{aligned}$$

4.20

(a) Since we are asked for the expected number of successes in a given number of trials, we use the binomial distribution with $n = 120$ and $p = 0.90$:

$$\mu = np = 120 \times 0.90 = 108$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{120 \times 0.90 \times (1 - 0.90)} = 3.29$$

(b) $Z = \frac{105-108}{3.29} = -0.91$; Since 105 is less than 2 standard deviations away from the mean we wouldn't consider this to be an unusual observation or be surprised.

(c) $P(X \leq 102) = P(X = 0) + P(X = 1) + \cdots + P(X = 102)$
Since this is a bit tedious to solve using binomial distribution, we can instead use the normal approximation to binomial to estimate this probability. But first we must verify that np and $n(1-p)$ are at least 10.

$$np = 120 \times 0.90 = 108 > 10 \qquad \textit{and} \qquad n(1-p) = 120 \times (1 - 0.90) = 12 > 10$$

Since the conditions are met, we can use the normal model, $N(\mu = 108, \sigma = 3.29)$.

$$\begin{aligned} P(X < 105) &= P\left(Z < \frac{105 - 108}{3.29}\right) \\ &= P(Z < -0.91) \\ &= 0.1814 \end{aligned}$$

In part (b) we had determined that it would not necessarily be considered unusual to observe 105 American adults who have had chickenpox among a random sample of 120. Here we calculated a somewhat high probability for this event, so the results from parts (b) and (c) agree.

If we were to apply a 0.5 correction, the calculations would change very slightly, still yielding a high probability.

$$\begin{aligned} P(X < 105 + 0.5) = P(X < 105.5) &= P\left(Z < \frac{105.5 - 108}{3.29}\right) \\ &= P(Z < -0.76) \\ &= 0.2236 \end{aligned}$$

4.21

(a) P(at least one *nun*) = 1 - P(no *nun*s) $= 1 - (0.75^3) = 1 - 0.4219 = 0.5781$

(b) P(exactly 2 *nun*s) $= \binom{3}{2} \times 0.25^2 \times 0.75^1 = 3 \times 0.25^2 \times 0.75^1 = 0.1406$

(c) P(exactly 1 *hei*) $= \binom{3}{1} \times 0.25^1 \times 0.75^2 = 3 \times 0.25^1 \times 0.75^2 = 0.4219$

(d) P(at most 2 *gimel*s) = 1 - P(3 *gimel*s) $= 1 - (0.25^3) = 1 - 0.0156 = 0.9844$

4.22

(a) P(at least one afraid) = 1 - P(none afraid) = $1 - (1 - 0.07)^{10}) = 1 - 0.484 = 0.516$

(b) P(exactly 2 afraid) = $\binom{10}{2} \times 0.07^2 \times 0.93^8 = 45 \times 0.07^2 \times 0.93^8 = 0.1234$

(c) P(at most 1 afraid) = P(none afraid) + P(1 afraid)
$= 0.4840 + \binom{10}{1} \times 0.07^1 \times 0.93^9$
$= 0.4840 + 0.3643$
$= 0.8483$

(d) If the group of 10 teenagers in a tent is a random sample, then there is about 84.83% chance that at most one of them will be afraid of spiders, and hence about 15.17% chance that 2 or more teenagers will be afraid of spiders in any particular tent. Because this probability is somewhat large (usually we use 0.05 as a somewhat arbitrary cutoff rule), he probably should not randomly assign the teenagers to their tents.

4.23

(a) $P(1^{st}$ has green eyes, 2^{nd} does not$) = 0.125 \times 0.875 = 0.109$

(b) Let X be the number of children with green eyes. Then, using a binomial distribution with $n = 2$ and $p = 0.125$;

$$P(X = 1) = \binom{2}{1} \times 0.125^1 \times 0.875^1 = 0.219$$

(c) Using a binomial distribution with $n = 6$ and $p = 0.125$;
$P(X = 2) = \binom{6}{2} \times 0.125^2 \times 0.875^4 = 15 \times 0.125^2 \times 0.875^4 = 0.137$

(d) Using a binomial distribution with $n = 6$ and $p = 0.125$;
$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.449 = 0.551$

(e) Let X = trial at which the first success (green eyes) occurs. Then, using a geometric distribution with $p = 0.125$;
$P(X = 4) = 0.875^3 \times 0.125 = 0.084$

(f) Using a binomial distribution with $n = 6$ and $p = 0.75$;
$\mu = np = 6 \times 0.75 = 4 \qquad \sigma = \sqrt{npq} = \sqrt{6 \times 0.75 \times 0.25} = 1.06$
Since 2 is $\frac{2-4}{1.06} = -1.89$ standard deviations below the expected number of brown eyed children, strictly speaking this would not be considered unusual, since $|-1.89| < 2$, however, it should be noted that the Z score for this value is pretty close to 2, making this observation somewhat unusual.

4.24

(a) Using the binomial distribution with $n = 3$ and $p = 0.25$;

$$P(X = 2) = \binom{3}{2} \times 0.25^2 \times 0.75^1 = 3 \times 0.25^2 \times 0.75^1 = 0.1406$$

(b) Using the binomial distribution with $n = 3$ and $p = 0.25$;

$$P(X = 0) = \binom{3}{0} \times 0.25^0 \times 0.75^3 = 0.4219$$

(c) Using the binomial distribution with $n = 3$ and $p = 0.25$;

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.4219 = 0.5781$$

(d) Let X be the trial at which the first success (disease) occurs. Then, using a geometric distribution with $p = 0.25$;

$$P(X = 3) = 0.75^2 \times 0.25 = 0.1406$$

4.25

(a) For the cars to be placed in alphabetical order we need the following ordering: Anna, Ben, Carl, Damian, Eddy. The probability of choosing Anna first is 1 / 5. The probability of choosing Ben first is 1 / 4, and so on.

$$\underbrace{1/5}_{Anna} \times \underbrace{1/4}_{Ben} \times \underbrace{1/3}_{Carl} \times \underbrace{1/2}_{Damian} \times \underbrace{1}_{Eddy} = \frac{1}{5 \times 4 \times 3 \times 2 \times 1} = \frac{1}{5!} = \frac{1}{120}$$

(b) Since the sum of all probabilities need to add up to 1, $\frac{1}{1/5!} = 5! = 120$.

(c) $8! = 40,320$.

4.26

(a) Binomial distribution with $n = 3$ and $p = 0.51$;
$P(X = 2) = \binom{3}{2} \times 0.51^2 \times 0.49 = 3 \times 0.51^2 \times 0.49 = 0.3823$

(b) $P(B, B, G) = 0.51 \times 0.51 \times 0.49 = 0.12744$
$P(B, G, B) = 0.51 \times 0.49 \times 0.51 = 0.12744$
$P(G, B, B) = 0.49 \times 0.51 \times 0.51 = 0.12744$
$P(2 \text{ out of 3 children are boys}) = 0.12744 + 0.12744 + 0.12744 = 3 \times 0.12744 = 0.3823$

(c) There are now $\binom{8}{3} = 56$ scenarios. It would be tedious to write them all out, and if we didn't know how many scenarios there are we might actually miss some of them. Using the binomial model to calculate this probability is a much more efficient approach.

4.27

(a) Geometric distribution with $p = \frac{1}{6}$: First success on a certain trial. $P(X = 5) = \left(\frac{5}{6}\right)^4 \times \left(\frac{1}{6}\right) = 0.0804$

(b) Binomial distribution with $n = 5$ and $p = \frac{1}{6}$: Given number of successes in a given number of trials.
$P(X = 3) = \binom{5}{3} \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = 10 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = 0.0322$

(c) Negative binomial distribution with $n = 5$ and $p = \frac{1}{6}$: A certain success on a certain trial.
$P(X = 3) = \binom{4}{2} \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = 6 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = 0.0193$

4.28

(a) Negative binomial distribution with $n = 15$ and $p = 0.65$: A certain success on a certain trial.
$P(X = 10) = \binom{14}{9} \times 0.65^{10} \times 0.35^5 = 2002 \times 0.65^{10} \times 0.35^5 = 0.1416$

(b) Binomial distribution with $n = 15$ and $p = 0.65$: Given number of successes in a given number of trials.
$P(X = 10) = \binom{15}{10} \times 0.65^{10} \times 0.35^5 = 3003 \times 0.65^{10} \times 0.35^5 = 0.2123$

(c) Geometric distribution with $p = 0.65$: First success on a certain trial.
$P(X = 3) = 0.35^2 \times 0.65 = 0.0796$

4.29

(a) Negative binomial with $n = 4$ and $p = 0.55$: Of the four trials considered here, the last trial must be a success and there were exactly 2 successes.

(b) $P(X = 2) = \binom{4-1}{2-1} \times 0.55^2 \times 0.45^2 = 3 \times 0.55^2 \times 0.45^2 = 0.1838$

(c) $\binom{3}{1} = \frac{3!}{1! \times 2!} = 3$.

(d) In the binomial model we have no restrictions on the outcome of the last trial while in the negative binomial model the last trial is fixed. Therefore we are interested in the number of ways of orderings of the other $k - 1$ successes in the first $n - 1$ trials.

4.30

(a) Negative binomial with $n = 10$ and $p = 0.15$;

$$P(X = 2) = \binom{10-1}{3-1} \times 0.15^3 \times 0.85^7 = 36 \times 0.15^3 \times 0.85^7 = 0.0390$$

(b) 0.15.

(c) In part (a) we considered the probability of an entire sequence. In part (b) we only considered the probability of one component of that sequence.

4.31

(a) Poisson with $\lambda = 75$. We may have some concerns about observations being independent if people commonly come to the coffee shop in pairs. Regardless, the Poisson distribution will provide a distribution that will be a reasonable first approximation.

(b) Mean $= \lambda = 75$, standard deviation $= \sqrt{\lambda} = \sqrt{75} = 8.66$.

(c) $\frac{60-75}{8.66} = -1.73$. Since 60 customers is within 2 standard deviations of the mean, this would not be considered unusual.

(d) Using Poisson with $\lambda = 70$:

$$P(X = 70) = \frac{75^{70}e^{-75}}{70!} = 0.0402$$

4.32

(a) Poisson with $\lambda = 1$.

(b) Mean $= \lambda = 1$, standard deviation $= \sqrt{\lambda} = \sqrt{1} = 1$.

(c) $\frac{4-1}{1} = 3$. Since 4 typos is more than 2 standard deviations above the mean, this would be considered an unusually high number of typos for this stenographer.

(d) Using Poisson with $\lambda = 1$:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} + \frac{1^2 e^{-1}}{2!} \\ &= 0.3679 + 0.3679 + 0.1839 = 0.9197 \end{aligned}$$

4.33

(a) $\frac{\lambda^k \times e^{-\lambda}}{k!} = \frac{6.5^5 \times e^{-6.5}}{5!} = 0.1454$

(b) $P(0) + P(1) + P(2) = \frac{6.5^0 \times e^{-6.5}}{0!} + \frac{6.5^1 \times e^{-6.5}}{1!} + \frac{6.5^2 \times e^{-6.5}}{2!} = 0.0015 + 0.0098 + 0.0318 = 0.0431$ (Or 0.0430 without the rounding error.)

(c) The number of people per car is $11.7/6.5 = 1.8$, meaning people are coming in small clusters. That is, if one person arrives, there's a chance that they brought someone else in her vehicle. This means individuals (the people) are not independent, even if the car arrivals are independent, and this breaks a core assumption for the Poisson distribution, so no, the number of people visiting between 2pm and 3pm would not follow a Poisson distribution.

4.34

(a) $\frac{\lambda^k \times e^{-\lambda}}{k!} = \frac{2.2^0 \times e^{-2.2}}{0!} = 0.1108$

(b)

$$\begin{aligned} P(0) + P(1) + P(2) &= \frac{2.2^0 \times e^{-2.2}}{0!} + \frac{2.2^1 \times e^{-2.2}}{1!} + \frac{2.2^2 \times e^{-2.2}}{2!} \\ &= 0.1108 + 0.2438 + 0.2681 \\ &= 0.6227 \end{aligned}$$

(c) It would certainly not be reasonable to model it using a Poisson distribution of 2.2, since the rate of bags lost may have changed. It may be reasonable to model the number of bags lost as a Poisson distribution after re-estimating the parameter. A good first step would be looking at the distribution of bags lost per day over a few recent months and see if it looks like a Poisson distribution with λ equal to the average number of bags lost per day.

4.35 First we need to calculate the possible amounts one can win or lose when the game is played 3 times. Let X = number of wins. Then,

$$\text{P(3 wins)} = P(X=3) = \binom{3}{3}\left(\frac{18}{38}\right)^3\left(\frac{20}{38}\right)^0 = \left(\frac{18}{38}\right)^3 = 0.1063$$

$$\text{P(2 wins, 1 loss)} = P(X=2) = \binom{3}{2}\left(\frac{18}{38}\right)^2\left(\frac{20}{38}\right)^1 = 3\times\left(\frac{18}{38}\right)^2\times\left(\frac{20}{38}\right) = 0.3543$$

$$\text{P(1 win, 2 losses)} = P(X=1) = \binom{3}{1}\left(\frac{18}{38}\right)^1\left(\frac{20}{38}\right)^2 = 3\times\left(\frac{18}{38}\right)\times\left(\frac{20}{38}\right)^2 = 0.3936$$

$$\text{P(3 losses)} = P(X=0) = \binom{3}{0}\left(\frac{18}{38}\right)^0\left(\frac{20}{38}\right)^3 = \left(\frac{20}{38}\right)^3 = 0.1458$$

- If you win 3 times, the total winnings is $Y = 1+1+1 = 3$.
- If you win twice but lose once the total winnings is $Y = 1+1-1 = 1$.
- If you win once and lose twice the total winnings is $Y = 1-1-1 = -1$.
- Lastly, if you lose three times the total winnings is $Y = -1+-1+-1 = -3$.

These are the possible values Y can take, and the probabilities with which Y takes these values are calculated above. Then, the probability model for Y is:

Y	-3	-1	1	3
$P(Y)$	0.1458	0.3936	0.3543	0.1063

4.36

(a) Let X represent the speeds of the cars traveling on this stretch of the I-5. Then, $X \sim N(\mu = 72.6, \sigma = 4.78)$.

$$P(X < 80) = P\left(Z < \frac{80 - 72.6}{4.78}\right) = P(Z < 1.55) = 0.9394 \ \rightarrow \ 93.94\%$$

(b) $P(60 < X < 80) = P(Z < 1.55) - P\left(Z < \frac{60-72.6}{4.78}\right) = P(Z < 1.55) - P(Z < -2.64) = 0.9394 - 0.0041 = 0.9353 \ \rightarrow \ 93.53\%$

(c) The fastest 5% of cars travel in the 95^{th} percentile. The Z score corresponding to the 95^{th} percentile is approximately 1.65.

$$Z = 1.65 = \frac{X - 72.6}{4.78} \ \rightarrow \ X = 1.65 \times 4.78 + 72.6 = 80.49 \ miles/hour$$

(d) The probability can be calculated as follows:

$$\begin{aligned} P(X > 70) &= 1 - P\left(Z > \frac{70 - 72.6}{4.78}\right) \\ &= 1 - P(Z < -0.54) \\ &= 1 - 0.2946 \\ &= 0.7054 \ \rightarrow \ 70.54\% \end{aligned}$$

4.37 Let X be the number of students among the 2,500 who decide to attend this university. X has a binomial distribution with number of trials $n = 2,500$ and $p = 0.70$. The university will not have enough spots if $X \geq 1787$, and we use the normal approximation to the binomial to calculate this probability. The mean and standard deviation of this distribution are

$$\mu = np = 2,500 \times 0.7 = 1750 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{2,500 \times 0.7 \times 0.3} \approx 23$$

Then, the probability can be calculated as follows:

$$\begin{aligned} P(X \geq 1787) &= P\left(Z > \frac{1787 - 1750}{23}\right) \\ &= P(Z > 1.61) \\ &= 1 - 0.9463 \\ &= 0.0537 \end{aligned}$$

With such large numbers, applying the continuity correction of 0.5 barely makes a difference.

$$\begin{aligned} P(X > 1786.5) &= P\left(Z > \frac{1786.5 - 1750}{23}\right) \\ &= P(Z > 1.59) \\ &= 1 - 0.9441 \\ &= 0.0559 \end{aligned}$$

4.38

(a) The probability a car is speeding is 0.7054 (see solution to Exercise **??**(d)). Since we are asked for the probability of a certain number of trials until the first success we use a geometric distribution with $p = 0.7054$. Let X be the trial at which the first success (first car that is speeding) occurs. Then,

$$P(\text{no cars are speeding}) = (1 - 0.7054)^5 = 0.0022$$

(b) The mean and the standard deviation for a geometric distribution with $p = 0.7088$ can be calculated as follows:

$$\mu = \frac{1}{p} = \frac{1}{0.7054} = 1.42, \qquad \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1 - 0.7054}{0.7054^2}} = 0.77$$

4.39

(a) The Z score corresponding to the upper 25% of the distribution is 0.67.

(b) x = 1800, $\mu = 1650$.

(c) The standard deviation of the distribution (σ) can be calculated as follows:

$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} \\ 0.67 &= \frac{1800 - 1650}{\sigma} \\ \sigma &= \frac{1800 - 1650}{0.67} \\ \sigma &= \$223.88 \end{aligned}$$

4.40

$$
\begin{aligned}
P(X > 1500 | X > 1350) &= \frac{P(X > 1500 \text{ and } X > 1350)}{X > 1350} \\
&= \frac{P(X > 1500)}{P(X > 1350)} \\
P(X > 1500) &= P\left(\frac{1500 - 1100}{200}\right) \\
&= P(Z > 2) = 1 - 0.9772 = 0.0228 \\
P(X > 1350) &= P(Z > 1.25) = 1 - 0.8944 = 0.1056 \\
P(X > 1500 | X > 1350) &= \frac{0.0228}{0.1056} = 0.216 \;\rightarrow\; 21.6\% \text{ of students}
\end{aligned}
$$

4.41

(a) Since we are asked for the probability of a certain number of trials until the first success we use a geometric distribution with $p = 0.471$. Let X be the trial at which the first married woman is selected. Then,

$$P(\text{first married woman is the } 3^{rd} \text{ selected}) = (1 - 0.471) \times (1 - 0.471) \times 0.471 = 0.1318.$$

(b) P(all three women are married) $= 0.471^3 = 0.1045$.

(c) Use the mean and standard deviation of a geometric distribution with $p = 0.471$:

$$\mu = \frac{1}{0.471} = 2.12, \qquad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{(1-0.471)}{0.471^2}} = \sqrt{2.38} \approx 1.54$$

(d) Use the mean and standard deviation of a geometric distribution with $p = 0.471$:

$$\mu = \frac{1}{0.30} = 3.33, \qquad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{0.70}{0.30^2}} \approx 2.79$$

(e) When p is smaller, i.e. the event is rarer, the expected number of trials before a success and the standard deviation are higher.

4.42 Let X be the number of households who decide to respond. X has a binomial distribution with number of trials $n = 15,000$ and $p = 0.09$. We are interested in the probability $P(X \geq 1,500)$, and we use the normal approximation to the binomial to calculate this probability. The mean and standard deviation of this distribution are

$$\mu = np = 15,000 \times 0.09 = 1,350 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{15,000 \times 0.09 \times 0.91} = 35$$

Then, the probability can be calculated as follows:

$$\begin{aligned} P(X \geq 1,500) &= P\left(Z > \frac{1,500 - 1,350}{35}\right) \\ &= P(Z > 4.29) \\ &\approx 0 \end{aligned}$$

With such large numbers applying the continuity correction of 0.5 would barely make a difference.

4.43 Let X represent the weight of the checked bag, then $X \sim N(\mu = 45, \sigma = 3.2)$.

$$\begin{aligned} P(X > 50) &= P\left(Z < \frac{50 - 45}{3.2}\right) \\ &= P(Z > 1.56) \\ &= 1 - 0.9406 \\ &= 0.0594 \end{aligned}$$

Roughly 6% of passengers incur overweight baggage fees.

4.44

(a) Let X represent the heights of 10 year olds. Then,

$$X \sim N(\mu = 55, \sigma = 6)$$

$$\begin{aligned} P(X < 48) &= P\left(Z < \frac{48 - 55}{6}\right) \\ &= P(Z < -1.17) \\ &= 0.1210 \end{aligned}$$

(b) In order to calculate the probability that a randomly chosen 10 year old is between 60 and 65 inches, we need to calculate $P(X < 65)$ and $P(X < 60)$ and take the difference of the two probabilities:

$$\begin{aligned} P(60 < X < 65) &= P\left(\frac{60 - 55}{6} < Z < \frac{65 - 55}{6}\right) \\ &= P(0.83 < Z < 1.67) \\ &= P(Z < 1.67) - P(Z < 0.83) \\ &= 0.9525 - 0.7967 \\ &= 0.1558 \end{aligned}$$

(c) The Z score for the cutoff for the upper 10%, or the 90% percentile, is 1.28 as $P(Z < 1.28) = 0.90$. We use this value to solve for the unknown height cutoff.

$$Z = 1.28 = \frac{x - 55}{6} \;\rightarrow\; x = 62.68 \; \textit{inches}$$

4.45

(a) Let X represent the closing bid of a randomly selected auction. Then,

$$\begin{aligned} P(X > 100) &= P\left(Z > \frac{100 - 89}{15}\right) \\ &= P(Z > 0.73) \\ &= 1 - 0.7673 \\ &= 0.2327 \end{aligned}$$

(b) If you are bidding on only one auction and set a maximum bid price that is low, chances are someone will outbid you and you won't win the auction. If your maximum bid price is high, you may win the auction but you may be paying more than you need to for the item. If you are bidding on more than one auction and your maximum bid price is very low, chances are you won't win any of the auctions. However, if your maximum bid price is high, you may win more than one auction and end up with multiple copies of the book.

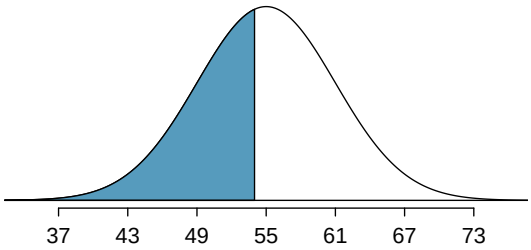
(c) An answer roughly equal to the 10th percentile would be reasonable. Regrettably, no percentile cutoff point guarantees beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction.

(d) Answers will vary a little but should correspond to the answer in part (c). We can think of this as the cutoff point for the cheapest 10% of auctions and can be calculated as follows:

$$\begin{aligned} P(X < x) = 0.10 \;\rightarrow\; P(Z < -1.28) &= 0.10 \\ -1.28 = \frac{x - 89}{15} \\ x &= 69.8 \end{aligned}$$

4.46

(a) This is a normal probability problem. We want the lower tail, shown below:



Next, find the Z-score:

$$Z = \frac{X - \mu}{\sigma} = \frac{54 - 55}{6} = -0.17$$

The lower tail area corresponding to this Z-score is 0.4325.

(b) We want 2, 3, or 4 of them to be able to ride ($p = 1 - 0.4325 = 0.5675$), and we need to find these probabilities separately. Each is individually a binomial probability problem, *if* we can assume they are independent (which we'll do). Here's the first one:

$$\begin{pmatrix} P(x=2) = n \\ xp^x(1-p)^{n-x} = \frac{4!}{2!2!}0.5675^2(1-0.5675)^{4-2} = 0.3615 \end{pmatrix}$$

Similarly, $P(x = 3) = 0.3162$ and $P(x = 4) = 0.1037$. Adding these up, we get the answer: 0.7814. We might also report that there is some uncertainty as to whether the 10 year olds were independent.

(c) We'll suppose independence is again reasonable. The first two 10 year olds seen would not be able to ride and the third would:

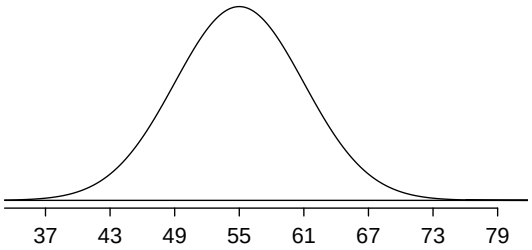
$$P(^1No, ^2No, ^3Yes) = P(^1No) \times P(^2No) \times P(^3Yes) = 0.4325 \times 0.4325 \times 0.5675 = 0.1062$$

(d) This is a negative binomial problem, and we'll again suppose independence. Another way to look at this is that the very last (12th) 10 year old is a success ($p = 0.5675$) and $x = 4$ of the $n = 11$ prior 10 year olds are successes. This means we can multiple the binomial probability for the first 11 kids with the probability that the 12th wsa a success (which also is how we get the negative binomial formula):

$$P(\text{4 of 11 successes \& 12th is success}) = \left[\frac{11!}{4!7!}p^4(1-p)^7\right] \times p = 0.0550$$

4.47

(a) This is a normal probability problem. First, draw a picture:



The tail is so small that the shaded region is not visible. Next, find the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{76 - 55}{6} = 3.5$$

Next, we find the tail area, e.g. using software: 0.00023263. The fraction of 10 year olds who are taller than 76 inches is 0.00023263. We'll use 0.0002 for the calculations below while reporting values for other rounded answers.

(b) About 0.0002 of them will be at least 76 inches tall: $0.0002 \times 2000 = 0.4$. That is, we expect 0.4 10 year olds to at least 76 inches tall.

If we used different rounding:

- $0.00023 \to 0.46$.
- $0.000233 \to 0.466$.
- $0.0002326 \to 0.4652$.
- $0.00023263 \to 0.46526$.

(c) For the binomial probability, $n = 2000$, $p = 0.0002$, and $x = 0$:

$$P(x = 0) = \frac{2000!}{0!2000!}0.0002^0 \times (1 - 0.0002)^2000 = \frac{1}{1}1 \times (1 - 0.0002)^2000 = 0.67029$$

If we used different rounding:

- $0.00023 \to 0.63125$.
- $0.000233 \to 0.62747$.
- $0.0002326 \to 0.62798$.
- $0.00023263 \to 0.62794$.

(d) For the Poisson probability, $\lambda = 0.4$ and $x = 0$:

$$P(x = 0) = \frac{0.4^0 e^{-0.4}}{0!} = \frac{1 \times e^{-0.4}}{1} = 0.67032$$

If we used different rounding:

- $0.46 \to 0.63128$.
- $0.466 \to 0.62751$.
- $0.4652 \to 0.62801$.
- $0.46526 \to 0.62797$.

4.48

(a) Since she is randomly guessing, probability of getting each question right is $p = 0.25$. Let X = trial at which the first success occurs. Then, using a geometric distribution with $p = 0.25$:

$$P(X = 3) = 0.75^{3-1} \times 0.25 \approx 0.1406$$

(b) Let X = number of questions she answers correctly. Then, using a binomial distribution with $n = 5$ and $p = 0.25$:

$$P(X = 3) = \binom{5}{3} \times 0.25^3 \times 0.75^2 = 10 \times 0.25^3 \times 0.75^2 = 0.08789062 = 0.0879$$

$$P(X = 4) = \binom{5}{4} \times 0.25^4 \times 0.75^1 = 5 \times 0.25^4 \times 0.75 = 0.01464844 = 0.0146$$

$$P(X = 3 \text{ or } 4) = 0.0879 + 0.0146 = 0.1025$$

(c) Majority means that she gets more than half of the questions right. Getting 3 or 4 or 5 of the questions right would all satisfy this condition. Then,

$$\begin{aligned} P(\text{majority right}) &= P(X \geq 3) \\ &= P(X = 3) + P(X = 4) = P(X = 5) \\ &= 0.0879 + 0.0146 + \binom{5}{5} \times 0.25^5 \times 0.75^0 \\ &= 0.0879 + 0.0146 + 0.0010 \\ &= 0.1035 \end{aligned}$$

Chapter 5

Foundations for inference

5.1

(a) Mean. Each student reports a numerical value: a number of hours.

(b) Mean. Each student reports a number (it is a percentage, and we can average over these percentages).

(c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion.

(d) Mean. Each student reports a number (a percentage).

(e) Proportion. Each student reports whether or not he got a job, so this is a categorical variable and we use a proportion.

5.2

(a) Proportion. Each respondent reports whether or not they worry a great deal about federal spending and the budget deficit, so this is a categorical variable and we use a proportion.

(b) Mean. Each TV news program and newspaper report a number: revenue.

(c) Proportion. Each student reports whether or not they use geolocation services on their smart phones, so this is a categorical variable and we use a proportion.

(d) Proportion. Each user reports whether or not they use a web-based taxi service, so this is a categorical variable and we use a proportion.

(e) Mean. Each user reports a number: how many times they used a web-based taxi service over the last year.

5.3

(a) The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time.

(b) The fraction of computer chips manufactured at the factory during the week of production that had defects.

(c) We estimate the parameter by computing the observed value in the data:

$$\hat{p} = \frac{27}{212} = 0.127$$

(d) We quantify this uncertainty using the *standard error*, which may be abbreviated as SE.

(e) Compute the standard error using the SE formula and plugging in the point estimate $\hat{p} = 0.127$ for p:

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$$

(f) The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (Usually beyond about 2 standard errors is a good rule of thumb.) The engineer should not be surprised.

(g) The recomputed standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$$

This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!).

5.4

(a) The sample is from all adults in the United States, so US adults is the population under consideration.

(b) The fraction of US adults who could not cover a $400 expense without borrowing money or selling something.

(c) We estimate the parameter by computing the observed value in the data:

$$\hat{p} = \frac{322}{765} = 0.421$$

(d) We quantify this uncertainty using the *standard error*, which may be abbreviated as SE.

(e) We can compute the standard error using the SE formula and plugging in the point estimate $\hat{p} = 0.421$ for p:

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.421(1-0.421)}{765}} = 0.0179$$

(f) The standard error can be thought of as the standard deviation of $\hat{p}$. A value of 0.50 would be over 4 standard errors from the observed value, which would represent a very uncommon observation. The news pundit should be surprised by the data.

(g) The recomputed standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{765}} = 0.0177$$

This value is hardly different at all, since 0.4 isn't too different from 0.421.

5.5

(a) Sampling distribution.

(b) To know whether the distribution is skewed, we need to know the proportion. We've been told the proportion is likely above 5% and below 30%, and the success-failure condition would be satisfied for any of these values. If the population proportion is in this range, the sampling distribution will be symmetric.

(c) We use the standard error to describe the variability:

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$$

(d) Standard error.

(e) The distribution will tend to be more variable when we have fewer observations per sample.

5.6

(a) Sampling distribution.

(b) Since the proportion is $p = 0.16$ and $n = 40$, the success-failure condition is *not* satisfied, with the expected number of successes being just $40 \times 0.16 = 6.4$. When we have too few expected successes, the sampling distribution of $\hat{p}$ is right-skewed.

(c) The variability can still be calculated, even if we cannot model $\hat{p}$ using a normal distribution:

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.16(1-0.16)}{40}} = 0.058$$

(d) Standard error.

(e) When there are more observations in the sample, the point estimate tends to be less variable. This means the distribution will tend to be less variable when we have more observations per sample. Beyond the required answer: the success-failure condition would be satisfied with this larger sample, so the distribution would also be symmetric in this scenario.

5.7 Recall that the general formula is *point estimate* $\pm z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula:

$$45\% \pm 1.96 \times 1.2\% \quad \rightarrow \quad (42.6\%, 47.4\%)$$

We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

5.8 Recall that the general formula is *point estimate* $\pm z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 2.58$ for a 99% confidence level, and $SE = 2.4\%$. Then, plug the values into the formula:

$$52\% \pm 2.58 \times 2.4\% \quad \rightarrow \quad (45.8\%, 58.2\%)$$

We are 99% confident that 45.8% to 58.2% of U.S. adult Twitter users get some news on Twitter.

5.9

(a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time.

(b) True. Notice that the description focuses on the true population value.

(c) True. If we examine the 95% confidence interval computed in Exercise **??**, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5.

(d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

5.10

(a) False. 50% is included in the 99% confidence interval, hence a null hypothesis of $p = 0.50$ would not be rejected at this level.

(b) False. The standard error measures the variability of the sample proportion, and is unrelated to the proportion of the population included in the study.

(c) False. We need to increase the sample size to decrease the standard error.

(d) False. As the confidence level decreases so does the margin of error, and hence the width of the confidence interval.

5.11

(a) False, even if the population distribution is not normal, with a large enough sample size we can assume that the sampling distribution is nearly Normal and calculate a confidence interval. It would however be ideal to see a histogram or Normal probability plot of the population distribution to assess the level of skewness. If the distribution is very strongly skewed, the sample size of 64 may be insufficient for the sample mean to be approximately normally distributed.

(b) True, this is the correct interpretation of the confidence interval.

(c) False, the confidence interval is not about a sample mean. The true interpretation of the confidence level would be that 95% of random samples produce confidence intervals that include the true population mean.

(d) False. To be more confident that we capture the parameter, we need a wider interval.

(e) True, since the normal model was used to model the sample mean. The margin of error can be calculated as half the width of the interval, and the sample mean is the midpoint of the interval.

$$ME = \frac{146 - 128}{2} = \frac{19}{2} = 9.5 \qquad \bar{x} = 128 + 9.5 = 137.5$$

(f) False, since in calculation of the standard error we divide the standard deviation by square root of the sample size. In order to cut the standard error in half (and hence the margin of error) we would need to sample $2^2 = 4$ times the number of people in the initial sample.

$$ME_{\text{half as big}} = Z^\star \frac{s}{\sqrt{n}} \;\rightarrow\; \frac{1}{2} ME_{original} = Z^\star \frac{s}{\sqrt{4n}} = Z^\star \frac{s}{2\sqrt{n}} = \frac{1}{2} Z^\star \frac{s}{\sqrt{n}} = \frac{1}{2} ME_{original}$$

5.12

(a) We are 90% confident that US residents experience poor mental health 3.40 to 4.24 days per month.

(b) 90% of random samples of size 1,151 will yield a confidence interval that contains the true average number of bad mental health days that US residents experience per month.

(c) To be more sure they capture the actual mean, they require a wider interval, unless they collect more data.

(d) Less data means less precision. The estimate will probably be less accurate with less data, so the interval will be larger.

5.13

(a) The visitors are from a simple random sample, so independence is satisfied. The success-failure condition is also satisfied, with both 64 and $752 - 64 = 688$ above 10. Therefore, we can use a normal distribution to model $\hat{p}$ and construct a confidence interval.

(b) The sample proportion is $\hat{p} = \frac{64}{752} = 0.085$. The standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.085(1-0.085)}{752}} = 0.010$$

(c) For a 90% confidence interval, use $z^\star = 1.65$. The confidence interval is therefore

$$\begin{aligned} &\text{point estimate} \pm z^\star \times SE \\ &0.085 \pm 1.65 \times 0.010 \\ &(0.0685, 0.1015) \end{aligned}$$

We are 90% confident that 6.85% to 10.15% of first-time site visitors will register using the new design.

5.14 First, we check conditions:

Independence. Since the data come from a random sample, independence is satisified.

Success-failure. We also have at least 10 successes (142) and 10 failures ($603 - 142 = 461$), so the success-failure condition is satisfied.

With the conditions satisfied, $\hat{p} = 142/603 = 0.235$ can be modeled using a normal distribution. Next, we compute the standard error:

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.235(1-0.235)}{603}} = 0.0173$$

For a 95% confidence level, $z^\star = 1.96$, and the confidence interval is

$$0.235 \pm 1.96 \times 0.0173 \rightarrow (0.201, 0.269)$$

We are 95% confident that 20.1% to 26.9% of the store's shoppers during the year made their visit because they had received a coupon in the mail.

5.15

(a) $H_0 : p = 0.5$ (The number of students with grades that improve after the program is equal to the number of students who do not have their grades improve)
$H_A : p \neq 0.5$ (Either a majority or a minority of students' grades improved)

(b) $H_0 : \mu = 15$ (The average amount of company time spent not working is 15 minutes)
$H_A : \mu \neq 15$ (The average amount of company time spent not working is different than 15 minutes)

5.16

(a) $H_0 : \mu = 1100$ (The current average calorie intake is 1100 calories)
$H_A : \mu \neq 1100$ (The current average calorie intake is different than 1100 calories.)

(b) $H_0 : p = 0.7$ (The fraction of Wisconsin adults who consume alcohol is 0.7)
$H_A : p \neq 0.7$ (The fraction of Wisconsin adults who consume alcohol is different from 0.7)

5.17 First, the hypotheses should be about the population proportion (p), not the sample proportion. Second, the null hypothesis should have an equal sign. Third, the alternative hypothesis should have a not-equals sign and reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is:

$$H_0 : p = 0.6$$
$$H_A : p \neq 0.6$$

5.18 First, the hypotheses should be about the population proportion (p), not the sample proportion. Second, the null value should be what we are testing (0.25), not the observed value (0.24). The correct way to set up these hypotheses is:

$$H_0 : p = 0.25$$
$$H_A : p \neq 0.25$$

5.19

(a) This claim is reasonable, since the entire interval lies above 50%.

(b) The value of 70% lies outside of the interval, so we have convincing evidence that the researcher's conjecture is wrong.

(c) A 90% confidence interval will be narrower than a 95% confidence interval. Even without calculating the interval, we can tell that 70% would not fall in the interval, and we would reject the researcher's conjecture based on a 90% confidence level as well.

5.20

(a) This claim does is not supported since 3 hours (180 minutes) is not in the interval.

(b) 2.2 hours (132 minutes) is in the 95% confidence interval, so we do not have evidence to say she is wrong.

(c) A 99% confidence interval will be wider than a 95% confidence interval. Even without calculating the interval, we can tell that 132 minutes would be in it, and we would not reject her claim based on a 99% confidence level as well.

5.21 Here's the answer in the Prepare, Check, Calculate, and Conclude framework.

Prepare. Set up hypotheses.

H_0: $p = 0.5$

H_A: $p \neq 0.5$

We will use a significance level of $\alpha = 0.05$.

Check. Simple random sample gets us independence, and the success-failure conditions is satisfied since $0.42 \times 1000 = 420$ and $(1 - 0.42) \times 1000 = 580$ are both at least 10.

Calculate. $SE = \sqrt{0.5(1 - 0.5)/1000} = 0.016$. $Z = \frac{0.42 - 0.5}{0.016} = -5$, which has a one-tail area of about 0.0000003, so the p-value is twice this one-tail area at 0.0000006.

Conclude. Because the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that the fraction of US adults who believe raising the minimum wage will help the economy is not 50%. Because the observed value is less than 50% and we have rejected the null hypothesis, we can conclude that this belief is held by fewer than 50% of US adults.

For reference, the survey also explores support for changing the minimum wage, which is itself a different question.

5.22 Here's the answer in the Prepare, Check, Calculate, and Conclude framework.

Prepare. Set up hypotheses. H_0: $p = 0.5$, H_A: $p \neq 0.5$. We are told to use a significance level of $\alpha = 0.01$.

Check. Simple random sample gets us independence, and the success-failure conditions is satisfied since 289 and $400 - 289 = 111$ are both at least 10.

Calculate. $\hat{p} = 289/400 = 0.7225$. $SE = \sqrt{0.5(1-0.5)/400} = 0.025$. $Z = \frac{0.7225-0.5}{0.025} = 8.9$, which has a one-tail area of about 3e-19 (0.0000000000000000003). so the p-value is twice this one-tail area at 6e-19 (0.000000000000000006).

Conclude. Because the p-value is less than $\alpha = 0.01$, we reject the null hypothesis and conclude that the the fraction of students who don't get enough sleep is different than 50%. Because the observed value is greater than 50% and we have rejected the null hypothesis, we can conclude that a majority of students at the surveyed university don't get enough sleep.

5.23 If the p-value is 0.05, this means the test statistic would be either $Z = -1.96$ or $Z = 1.96$. We'll show the calculations for $Z = 1.96$. The standard error would be $SE = \sqrt{0.3(1-0.3)/90} = 0.048$. Finally, we set up the test statistic formula and solve for $\hat{p}$:

$$\begin{aligned} Z &= \frac{\hat{p} - 0.3}{SE} \\ 1.96 &= \frac{\hat{p} - 0.3}{0.048} \\ \hat{p} &= 0.394 \end{aligned}$$

Alternatively, if $Z = -1.96$ was used: $\hat{p} = 0.206$.

5.24 If the p-value is 0.01, this means the test statistic would be either $Z = -2.58$ or $Z = 2.58$. If either of these Zs is chosen, it is okay. We'll use $Z = 2.58$ for our calculations. The standard error would be

$$SE = \sqrt{0.9(1 - 0.9)/90} = 0.0079$$

Finally, we set up the test statistic formula and solve for $\hat{p}$:

$$\begin{aligned} Z &= \frac{\hat{p} - 0.9}{SE} \\ 2.58 &= \frac{\hat{p} - 0.9}{0.0079} \\ \hat{p} &= 0.92 \end{aligned}$$

Alternatively, if $Z = -2.58$ was used: $\hat{p} = 0.88$.

5.25

(a) H_0: Anti-depressants do not affect the symptoms of Fibromyalgia. H_A: Anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming).

(b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither.

(c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.

5.26

(a) Scenario I is higher. Recall that a sample mean based on less data tends to be less accurate and have larger standard errors.

(b) Scenario I is higher. The higher the confidence level, the higher the corresponding margin of error.

(c) They are equal. The sample size does not affect the calculation of the p-value for a given Z-score.

(d) Scenario I is higher. If the null hypothesis is harder to reject (lower α), then we are more likely to make a Type 2 Error when the alternative hypothesis is true.

5.27

(a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy.

(b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases.

(c) The new margin of error will be smaller, since as the sample size increases, the standard error decreases, which will decrease the margin of error.

5.28

- We could note the data and context provided: $\hat{p} = 0.42$, $n = 1000$, and we're using a 99% confidence level.
- Checking conditions. While it doesn't explicitly say here whether the sample is random, we will assume the survey by this company was in fact random, which would satisfy the independence condition. The success-failure condition is also satisfied since 0.42×1000 and 0.58×1000 are both at least 10.
- Next, we compute the standard error for the confidence interval using $\hat{p} = 0.42$:

$$SE = \sqrt{\frac{0.42(1 - 0.42)}{1000}} = 0.016$$

  Next, we can compute the confidence interval itself, where we use $z^\star = 2.58$ for a 99% confidence level:

$$0.42 \pm 2.58 \times 0.016 \rightarrow (0.379, 0.461)$$

- We are 99% confident that 37.9% to 46.1% of US adults believe that increasing the minimum wage would help the economy.

5.29

(a) H_0: The restaurant meets food safety and sanitation regulations.
H_A: The restaurant does not meet food safety and sanitation regulations.

(b) The food safety inspector concludes that the restaurant does not meet food safety and sanitation regulations and shuts down the restaurant when the restaurant is actually safe.

(c) The food safety inspector concludes that the restaurant meets food safety and sanitation regulations and the restaurant stays open when the restaurant is actually not safe.

(d) A Type 1 Error may be more problematic for the restaurant owner since his restaurant gets shut down even though it meets the food safety and sanitation regulations.

(e) A Type 2 Error may be more problematic for diners since the restaurant deemed safe by the inspector is actually not.

(f) A diner would prefer strong evidence as any indication of evidence might mean there may be an issue with the restaurant meeting food safety and sanitation regulations, and diners would rather a restaurant that meet the regulations get shut down than a restaurant that doesn't meet the regulations not get shut down.

5.30

(a) True.

(b) False. The significance level is the probability of the Type 1 Error.

(c) False. Failure to reject H_0 only means there wasn't sufficient evidence to reject it, not that it has been confirmed.

(d) True.

5.31

(a) $H_0 : p_{unemp} = p_{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are equal.
$H_A : p_{unemp} \neq p\mathit{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are different.

(b) If in fact the two population proportions are equal, the probability of observing at least a 2% difference between the sample proportions is approximately 0.35. Since this is a high probability we fail to reject the null hypothesis. The data do not provide convincing evidence that the proportion of of unemployed and underemployed people who are having relationship problems are different.

5.32 Here's the answer in the Prepare, Check, Calculate, and Conclude framework.

Prepare. Set up hypotheses. H_0: $p = 0.08$, H_A: $p \neq 0.08$. We will use a significance level of $\alpha = 0.05$.

Check. Simple random sample gets us independence, and the success-failure condition is satisfied since 21 and $194 - 21 = 173$ are both at least 10.

Calculate. Several calculations:

$$\hat{p} = 21/194 = 0.108$$
$$SE = \sqrt{0.08(1 - 0.08)/194} = 0.0195$$
$$Z = \frac{0.108 - 0.08}{0.0195} = 1.44$$

which has a one-tail area of about 0.075, so the p-value is twice this one-tail area at 0.15.

Conclude. Because the p-value is bigger than $\alpha = 0.05$, we do not reject the null hypothesis. The sample does not provide convincing evidence that the fraction of children who are nearsighted is different from 0.08.

5.33 Because 130 is inside the confidence interval, we do not have convincing evidence that the true average is any different than what the nutrition label suggests.

5.34 The sampling distribution is the distribution of sample proportions from samples of the same size randomly sampled from the same population. As the same size increases, the shape of the sampling distribution (when $p = 0.1$) will go from being right-skewed to being more symmetric and resembling the normal distribution. With larger sample sizes, the spread of the sampling distribution gets smaller. Regardless of the sample size, the center of the sampling distribution is equal to the true mean of that population, provided the sampling isn't biased.

5.35 True. If the sample size gets ever larger, then the standard error will become ever smaller. Eventually, when the sample size is large enough and the standard error is tiny, we can find statistically significant yet very small differences between the null value and point estimate (assuming they are not exactly equal).

5.36 As the sample size increases the standard error will decrease, the sample statistic will increase, and the p-value will decrease.

5.37

(a) In effect, we're checking whether men are paid more than women (or vice-versa), and we'd expect these outcomes with either chance under the null hypothesis:

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

We'll use p to represent the fraction of cases where men are paid more than women.

(b) Below is the completion of the hypothesis test.

- There isn't a good way to check independence here since the jobs are not a simple random sample. However, independence doesn't seem unreasonable, since the individuals in each job are different from each other. The success-failure condition is met since we check it using the null proportion: $p_0 n = (1 - p_0)n = 10.5$ is greater than 10.
- We can compute the sample proportion, SE, and test statistic:

$$\hat{p} = 19/21 = 0.905$$
$$SE = \sqrt{\frac{0.5 \times (1 - 0.5)}{21}} = 0.109$$
$$Z = \frac{0.905 - 0.5}{0.109} = 3.72$$

The test statistic Z corresponds to an upper tail area of about 0.0001, so the p-value is 2 times this value: 0.0002.

- Because the p-value is smaller than 0.05, we reject the notion that all these gender pay disparities are due to chance. Because we observe that men are paid more in a higher proportion of cases and we have rejected H_0, we can conclude that men are being paid higher amounts in ways not explainable by chance alone.

If you're curious for more info around this topic, including a discussion about adjusting for additional factors that affect pay, please see the following video by Healthcare Triage: youtu.be/aVhgKSULNQA.

Chapter 6

Inference for categorical data

6.1

(a) False. For the distribution of $\hat{p}$ to be approximately normal, we need to have at least 10 successes and 10 failures in our sample (on the average).

(b) True. The success-failure condition is not satisfied

$$np = 50 \times 0.08 = 4 \; and \; n(1-p) = 50 \times 0.92 = 46,$$

therefore we know that the distribution of $\hat{p}$ is not approximately normal. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be as high as 1 (though we would expect this to effectively never happen), it can only go as low as 0. Therefore the distribution would probably take on a right-skewed shape. Plotting the sampling distribution would confirm this suspicion.

(c) False. Standard error of $\hat{p}$ in samples with $n = 125$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08 \times 0.92}{125}} = 0.0243$$

A $\hat{p}$ of 0.12 is only $\frac{0.12-0.08}{0.0243} = 1.65$ standard errors away from the mean, which would not be considered unusual.

(d) True. Standard error of $\hat{p}$ in samples with $n = 125$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08 \times 0.92}{250}} = 0.0172$$

A $\hat{p}$ of 0.12 is $\frac{0.12-0.08}{0.0172} = 2.32$ standard errors away from the mean, which might be considered unusual.

(e) False. Since n appears under the square root sign in the formula for the standard error, increasing the sample size from 125 to 250 would decrease the standard error of the sample proportion only by a factor of $\sqrt{2}$.

6.2

(a) True. The success-failure condition is not satisfied

$$np = 20 \times 0.77 = 15.4 \; and \; n(1-p) = 20 \times 0.23 = 4.6,$$

therefore we know that the distribution of $\hat{p}$ is not approximately normal. In most samples we would expect $\hat{p}$ to be close to 0.77, the true population proportion. While $\hat{p}$ can be as low as 0 (though we would expect this to happen very rarely), it can only go as high as 1. Therefore, since 0.77 is closer to 1, the distribution would probably take on a left skewed shape. Plotting the sampling distribution would confirm this suspicion.

(b) False. Unlike with means, for the sampling distribution of proportions to be approximately normal, we need to have at least 10 successes and 10 failures in our sample. We do not use $n \geq 30$ as a condition to check for the normality of the distribution of $\hat{p}$.

(c) False. Standard error of $\hat{p}$ in samples with $n = 60$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.77 \times 0.23}{120}} = 0.0384$$

A $\hat{p}$ of 0.85 is only $Z = \frac{0.85-0.77}{0.0384} = 2.08$ standard errors away from the mean, which would be considered unusual.

(d) True. Standard error of $\hat{p}$ in samples with $n = 120$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.77 \times 0.23}{120}} = 0.046$$

A $\hat{p}$ of 0.85 is $Z = \frac{0.85-0.77}{0.046} = 1.73$ standard errors away from the mean, which would not be considered unusual.

6.3

(a) True. The success-failure condition is not satisfied

$$np = 30 \times 0.90 = 27 \; and \; n(1 - \hat{p}) = 30 \times 0.10 = 3,$$

therefore we know that the distribution of $\hat{p}$ is not nearly normal. In most samples we would expect $\hat{p}$ to be close to 0.90, the true population proportion. While $\hat{p}$ can be as low as 0 (thought we would expect this to happen very rarely), it can only go as high as 1. Therefore the distribution would probably take on a left-skewed shape. Plotting the sampling distribution would confirm this suspicion.

(b) True. Since n appears in a square root for SE, using a sample size that is 4 times as large will reduce the SE by half.

(c) True. The success-failure condition is satisfied

$$np = 140 \times 0.90 = 126 \; and \; n(1 - p) = 140 \times 0.10 = 14,$$

therefore the distribution of $\hat{p}$ is nearly normal.

(d) True. The success-failure condition is satisfied

$$np = 280 \times 0.90 = 252 \; and \; n(1 - p) = 70 \times 0.10 = 28,$$

therefore the distribution of $\hat{p}$ is nearly normal.

6.4

(a) True. The success-failure condition is not satisfied

$$np = 12 \times 0.25 = 4 \; and \; n(1-p) = 12 \times 0.75 = 8,$$

therefore we know that the distribution of $\hat{p}$ is not approximately normal. In most samples we would expect $\hat{p}$ to be close to 0.25, the true population proportion. While $\hat{p}$ can be as high as 1 (though we would expect this to happen very rarely), it can only go as low as 0. Therefore, since 0.25 is closer to 0, the distribution would probably take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion.

(b) True. The minimum sample size that yields at least 10 successes and 10 failures can be calculated as

$$n = max\left(\frac{10}{0.25}, \frac{10}{0.75}\right) = min(40, 13.3) = 40$$

We need a sample of at least $n = 40$ to meet the success failure condition.

(c) False. Standard error of $\hat{p}$ in samples with $n = 50$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.25 \times 0.75}{50}} = 0.0612$$

A $\hat{p}$ of 0.20 is only $Z = \frac{0.20-0.25}{0.0612} = -0.82$ standard errors away from the mean, which would not be considered unusual.

(d) False. Standard error of $\hat{p}$ in samples with $n = 150$ can be calculated as:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.25 \times 0.75}{150}} = 0.0354$$

A $\hat{p}$ of 0.20 is $Z = \frac{0.20-0.25}{0.0354} = -1.41$ standard errors away from the mean, which would not be considered unusual.

(e) False. Since n appears under the square root sign in the formula for the standard error, doubling the sample size would decrease the standard error of the sample proportion only by a factor of $\sqrt{3}$.

6.5

(a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

(b) True. This is the correct interpretation of the confidence interval, which can be calculated as $0.82 \pm 0.02 = (0.80, 0.84)$.

(c) True. This is the correct interpretation of the confidence level.

(d) True. Since the sample size appears under the square root sign in calculation of the standard error, in order to halve the margin of error we would need to quadruple the sample size.

$$ME_{old} = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad ME_{new} = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{4n}} = z^\star \frac{1}{2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{1}{2} ME_{old}$$

(e) True. The confidence interval lies entirely above 50%.

6.6

(a) With a random sample from $< 10\%$ of the population, independence is satisfied. The success-failure condition is also satisfied. Hence, the margin of error can be calculated as follows:

$$ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.66 \times 0.34}{1018}} = 0.029 \approx 3\%$$

(b) A 95% confidence interval for the proportion of adults who think that licensed drivers should be required to re-take their road test once they reach 65 years of age can be calculated as

$$0.66 \pm 0.03 = (0.63, 0.69).$$

Since two thirds (roughly 67%) is contained in the interval we wouldn't reject a null hypothesis where $p = 0.67$. Therefore, the data do not provide evidence that more than two thirds of the population think that drivers over the age of 65 should re-take their road test.

6.7 With a random sample, independence is satisfied. The success-failure condition is also satisfied. Hence, the margin of error can be calculated as follows:

$$ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56 \times 0.44}{600}} = 0.0397 \approx 4\%$$

6.8

(a) The population parameter of interest is the proportion of all Greeks who would rate their lives poorly enough to be considered "suffering", p. The point estimate for this parameter is the proportion of Greeks in this sample who would rate their lives as such, $\hat{p} = 0.25$.

(b) 1. Independence: The sample is random, and $1,000 < 10\%$ of all Greeks, therefore the life rating of one Greek in this sample is independent of another.

2. Success-failure: $1,000 \times 0.25 = 250 > 10$ and $1,000 \times 0.75 = 750 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal.

(c) A 95% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.25 \pm 1.96 \times \sqrt{\frac{0.25 \times 0.75}{1000}} \\ &= 0.25 \pm 1.96 \times 0.0137 \\ &= 0.25 \pm 0.0269 \\ &= (0.2231, 0.2769) \end{aligned}$$

We are 95% confident that approximately 22% to 28% of Greeks would rate their lives poorly enough to be considered "suffering".

(d) Increasing the confidence level would increase the margin of error hence widen the interval.

(e) Increasing the sample size would decrease the margin of error hence make the interval narrower.

6.9

(a) No. The sample only represents students who took the SAT, and this was also an online survey.

(b) A 90% confidence interval can be calculated as follows:

$$\begin{aligned}\hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.55 \pm 1.65 \sqrt{\frac{0.55 \times 0.45}{1509}} \\ &= 0.55 \pm 1.65 \times 0.0128 \\ &= 0.55 \pm 0.0211 \\ &(0.5289, 0.5711)\end{aligned}$$

We are 95% confident that 53% to 57% of high school seniors are fairly certain that they will participate in a study abroad program in college.

(c) 90% of random samples of 1,509 high school seniors would produce a 90% confidence interval that includes the true proportion of high school seniors who took the SAT are fairly and who certain that they will participate in a study abroad program in college.

(d) Yes. The interval lies entirely above 50%. Therefore, at 90% confidence level, it would be appropriate to claim the majority of high school seniors who took the SAT who are fairly certain they will participate in a study abroad program in college.

6.10

(a) 61% is a sample statistic, it's the observed sample proportion.

(b) A 95% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.61 \pm 1.96 \sqrt{\frac{0.61 \times (1-0.61)}{1578}} \\ &= 0.61 \pm 1.96 \times 0.012 \\ &= 0.61 \pm 0.024 \\ &(0.586, 0.634) \end{aligned}$$

We are 95% confident that approximately 58.6% to 63.4% of Americans think marijuana should be legalized.

(c) 1. Independence: The sample is random, and comprises less than 10% of the American population, therefore we can assume that the individuals in this sample are independent of each other

2. Success-failure: The number of successes (people who said marijuana should be legalized: $1578 \times 0.61 = 962.58$) and failures (people who said it shouldn't be: $1578 \times 0.39 = 615.42$) are both greater than 10, therefore the success-failure condition is met as well.

Therefore the distribution of the sample proportion is expected to be approximately normal.

(d) Yes, the interval is above 50%, suggesting, with 95% confidence, that the true population proportion of Americans who think marijuana should be legalized is greater than 50%.

6.11

(a) (i) Let's prepare for running a hypothesis test. We want to check for a majority (or minority), so we use the following hypotheses:

$$H_0 : p = 0.5 \qquad\qquad H_A : p \neq 0.5$$

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
(ii) Next, we check whether the conditions are met to proceed. Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1 - 0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test).
(iii) Next, we can start performing calculations. We can model $\hat{p}$ using a normal distribution with a standard error of

$$SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$$

(We use the null proportion, $p_0 = 0.5$, to compute the standard error for a one-proportion hypothesis test.) Next, we compute the test statistic:

$$Z = \frac{0.55 - 0.5}{0.02} = 2.5$$

This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$.
(iv) Lastly, we make a conclusion. Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit.

(b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g. a 99% confidence level and a $\alpha = 0.05$ significance level), then this is no longer generally true.

6.12

(a) The hypotheses are as follows:

$H_0 : p = 0.5$ (50% of Americas who decide not to go to college because they cannot afford it do so because they cannot afford it)

$H_A : p < 0.5$ (Less than 50% of Americas who decide not to go to college because they cannot afford it do so because they cannot afford it)

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The sample is representative and we can safely assume that $331 < 10\%$ of all American adults who decide not to go to college, therefore whether or not one person in the sample decided not to go to college because they can't afford it is independent of another.
2. Success-failure: $331 \times 0.5 = 165.5 > 10$ and $331 \times 0.5 = 165.5 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal. The test statistic can be calculated as follows:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
$$= \frac{0.48 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{331}}} = \frac{-0.02}{0.0275} = -0.73$$
$$p - value = P(\hat{p} < 0.48 | p = 0.5) = P(Z < -0.73) = 0.2327$$

Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence that less than half of American adults who decide not to go to college make this decision because they cannot afford college.

(b) Yes, since we failed to reject $H_0 : p = 0.5$.

6.13

(a) The hypotheses are as follows:

$H_0 : p = 0.5$ (Results are equivalent to randomly guessing)
$H_A : p \neq 0.5$ (Results are different than just randomly guessing)

Before conducting the hypothesis test, we must first check that the conditions for inference are satisfied.

1. Independence: The sample is random, therefore whether or not one person in the sample can identify a soda correctly in independent of another.
2. Success-failure: $80 \times 0.5 = 40 > 10$ and $80 \times 0.5 = 40 > 10$. Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal.

The test statistic and the p-value can be calculated as follows:

$$\hat{p} = \frac{53}{80} = 0.6625$$
$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.6625 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{80}}} = \frac{0.1625}{0.0559} = 2.91$$
$$p - value = 2 \times P(\hat{p} > 0.6625 \mid p = 0.5) = 2 \times P(Z > 2.91) = 2 \times (1 - 0.9982) = 0.0036$$

Since the p-value $< \alpha$ (use $\alpha = 0.05$ since not given), we reject the null hypothesis. Since we rejected H_0 and the point estimate suggests people are better than random guessing, we can conclude the rate of correctly identifying a soda for these people is significantly better than just by random guessing.

(b) If in fact people cannot tell the difference between diet and regular soda and they were randomly guessing, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly (or 53 or more identify a soda incorrectly) would be 0.0036.

6.14

(a) We have previously confirmed that the independence condition is satisfied. We need to recheck the success-failure condition using the sample proportion: $331 \times 0.48 = 158.88 > 10$ and $331 \times 0.52 = 172.12 > 10$. An 80% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.48 \pm 1.65 \times \sqrt{\frac{0.48 \times 0.52}{331}} \\ &= 0.48 \pm 1.65 \times 0.0275 \\ &= 0.48 \pm 0.045 \\ &= (0.435, 0.525) \end{aligned}$$

We are 90% confident that the 43.5% to 52.5% of all Americans who decide not to go to college do so because they cannot afford it. This agrees with the conclusion of the earlier hypothesis test since the interval includes 50%.

(b) We are asked to solve for the sample size required to achieve a 1.5% margin of error for a 90% confidence interval and the point estimate is $\hat{p} = 0.48$.

$$\begin{aligned} ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ \rightarrow \ 0.01 &\geq 1.65 \sqrt{\frac{0.48 \times 0.52}{n}} \\ 0.015^2 &\geq 1.65^2 \frac{0.48 \times 0.52}{n} \\ n &\geq \frac{1.65^2 \times 0.48 \times 0.52}{0.015^2} \\ n &\geq 3020.16 \approx 3121 \end{aligned}$$

The sample size n should be at least 3,121.

6.15 Since a sample proportion ($\hat{p} = 0.55$) is available, we use this for the sample size calculations. The margin of error for a 90% confidence interval is $1.65 \times SE = 1.65 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.65 \times \sqrt{\frac{0.55(1-0.55)}{n}} \leq 0.01$$
$$1.65^2 \frac{0.55(1-0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6739.

6.16 We are asked to solve for the sample size required to achieve a 2% margin of error for a 95% confidence interval and the point estimate is $\hat{p} = 0.61$.

$$
\begin{aligned}
ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ \rightarrow \ 0.02 &\geq 1.96 \sqrt{\frac{0.61 \times 0.39}{n}} \\
0.02^2 &\geq 1.96^2 \frac{0.61 \times 0.39}{n} \\
n &\geq \frac{1.96^2 \times 0.61 \times 0.39}{0.02^2} \\
n &\geq 2284.792 \\
n &\geq 2285
\end{aligned}
$$

The sample size n should be at least 2,285.

6.17 The conditions that need to be met to ensure the approximately normal distribution of this difference are:

1. Independence: This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers.
2. Success-failure: There are only 5 interventions under the provocative scenario.

Since the success-failure condition is not met, the distribution of $(\hat{p}_P - \hat{p}_C)$ is not approximately normal. Note that even if we were conducting a hypothesis test and using a pooled proportion for the success- failure check, it would not have been satisfied.

6.18 Before we can calculate a confidence interval, we must first check that the conditions are met.

1. Independence: If patients are randomly assigned into the two groups, whether or not one patient in the treatment group survives is independent of another, and whether or not one patient in the control group survives is independent of another as well.
2. Success-failure: There are only 4 deaths in the control group.

Since the success-failure condition is not met, $(\hat{p}_C - \hat{p}_T)$ is not expected to be approximately normal and therefore cannot calculate a confidence interval for the difference between the proportion of patients who survived in the treatment and control groups using large sample techniques and a critical Z score.

6.19

(a) False. Since $(p_{male} - p_{female})$ is positive, the proportion of males whose favorite color is black is higher than the proportion of females.

(b) True.

(c) True.

(d) True.

(e) False. To get the 95% confidence interval for $(p_{female} - p_{male})$, all we have to do is to swap the bounds of the confidence interval for $(p_{male} - p_{female})$ and take their negatives: (-0.06,-0.02).

6.20

(a) False. The confidence interval includes 0.

(b) False. We are 95% confident that 16% fewer to 2% Americans who make less than $40,000 per year are not at all personally affected by the government shutdown compared to those who make $40,000 or more per year.

(c) False. As the confidence level decreases the width of the confidence level decreases as well.

(d) True.

6.21

(a) Standard error:

$$SE = \sqrt{\frac{0.79(1-0.79)}{347} + \frac{0.55(1-0.55)}{617}} = 0.03$$

Using $z^\star = 1.96$, we get:

$$0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$$

We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan.

(b) True.

6.22 Before calculating the confidence interval we should check that the conditions are satisfied.

1. Independence: Both samples are random, and 11,545 $<$ 10% of all Californians and 4,691 $<$ 10% of all Oregonians, therefore how much one Californian sleeps is independent of how much another Californian sleeps and how much one Oregonian sleeps is independent of how much another Oregonian sleeps. In addition, the two samples are independent of each other.
2. Success-failure:

$$11,545 \times 0.08 = 923.6 > 10 \qquad 11,545 \times 0.92 = 10621.4 > 10$$

$$4,691 \times 0.088 = 412.8 > 10 \qquad 4,691 \times 0.912 = 4278.2 > 10$$

Since the observations are independent and the success-failure condition is met, $\hat{p}_{CA} - \hat{p}_{OR}$ is expected to be approximately normal. A 95% confidence interval for the difference between the population proportions can be calculated as follows:

$$\begin{aligned}(\hat{p}_{CA} - \hat{p}_{OR}) \pm z^\star \sqrt{\frac{\hat{p}_{CA}(1-\hat{p}_{CA})}{n_{CA}} + \frac{\hat{p}_{OR}(1-\hat{p}_{OR})}{n_{OR}}} &= (0.08 - 0.088) \pm 1.96 \sqrt{\frac{0.08 \times 0.92}{11,545} + \frac{0.088 \times 0.912}{4,691}} \\ &= -0.008 \pm 0.009 \\ &= (-0.017, 0.001)\end{aligned}$$

We are 95% confident that the difference between the proportions of Californians and Oregonians who are sleep deprived is between -1.7% and 0.1%. In other words, we are 95% confident that 1.7% less to 0.1% more Californians than Oregonians are sleep deprived.

6.23

(a) The percentages can be calculated as follows:

Fraction of college grads who do not know $= \frac{104}{438} = 0.237 \rightarrow 23.7\%$.

Fraction of non-college grads who do not know $= \frac{131}{389} = 0.337 \rightarrow 33.7\%$.

(b) Let p_{CG} represent the proportion of college graduates who responded "do not know", and p_{NCG} represent the proportion of non-college graduates who responded "do not know". Then,

$H_0 : p_{CG} = p_{NCG}$
$H_A : p_{CG} \neq p_{NCG}$

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: Both samples are random, so observations within each group are independent. Additionally, each sample is independent of the other.
2. Success-failure: First we need to find $\hat{p}_{pool}$ and then use that to calculate the numbers of expected successes and failures in each group.

$$\hat{p}_{pool} = \frac{success_{CG} + success_{NCG}}{n_{CG} + n_{NCG}} = \frac{104 + 131}{438 + 389} = \frac{235}{827} = 0.284$$
$$1 - \hat{p}_{pool} = 1 - 0.284 = 0.716$$

$438 \times 0.284 = 124.392 > 10 \qquad 438 \times 0.716 = 313.608 > 10$

$389 \times 0.284 = 110.476 > 10 \qquad 389 \times 0.716 = 278.524 > 10$

Since the observations are independent and the success-failure condition is met, $\hat{p}_{CG} - \hat{p}_{NCG}$ is expected to be approximately normal. The test statistic and the p-value can be calculated as follows:

$$Z = \frac{\hat{p}_{CG} - \hat{p}_{NCG}}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{CG}} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{NCG}}}}$$
$$= \frac{0.237 - 0.337}{\sqrt{\frac{0.284 \times 0.716}{438} + \frac{0.284 \times 0.716}{389}}} = \frac{-0.1}{0.0314} = -3.18$$
$$p - value = P(|Z| > 3.18) = 0.0007 \times 2 = 0.0014$$

Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

6.24

(a) The hypotheses are:

$H_0 : p_{CA} = p_{OR}$
$H_A : p_{CA} \neq p_{OR}$

We have confirmed in Exercise **??** that the independence condition is satisfied but we need to recheck the success-failure condition using $\hat{p}_{pool}$ and expected counts.

$success_{CA} = n_{CA} \times p_{CA} = 11,545 \times 0.08 = 923.6 \approx 924$
$success_{OR} = n_{OR} \times p_{OR} = 4,691 \times 0.088 = 412.8 \approx 413$

$$\hat{p}_{pool} = \frac{success_{CA} + success_{OR}}{n_{CA} + n_{OR}} = \frac{924 + 413}{11,545 + 4,691} = \frac{1,337}{16,236} \approx 0.0821 - \hat{p}_{pool} \quad = 1 - 0.082 = 0.918$$

$$11,545 \times 0.082 = 946.69 > 10 \qquad 11,545 \times 0.918 = 10598.31 > 10$$
$$4,691 \times 0.082 = 384.662 > 10 \qquad 4,691 \times 0.918 = 4306.338 > 10$$

Since the observations are independent and the success-failure condition is met, $\hat{p}_{CA} - \hat{p}_{OR}$ is expected to be approximately normal. Next we calculate the test statistic and the p-value:

$$\begin{aligned} Z &= \frac{(\hat{p}_{CA} - \hat{p}_{OR}) - (p_{CA} - p_{OR})}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{CA}} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{OR}}}} \\ &= \frac{(0.08 - 0.088) - 0}{\sqrt{\frac{0.082 \times 0.918}{11,545} + \frac{0.082 \times 0.918}{4,691}}} \\ &= \frac{-0.008}{0.00475} = -1.68 \end{aligned}$$

$p - value = P(|\hat{p}_{CA} - \hat{p}_{OR}| > 0.008 \mid (p_{CA} - p_{OR}) = 0) = 2 \times P(|Z| > 1.68) = 2 \times 0.0465 = 0.093$

Since the p-value $> \alpha$ (use $\alpha = 0.05$ since not given), we fail to reject H_0 and conclude that the data do not provide strong evidence that the rate of sleep deprivation is different for the two states.

(b) Type II, since we may have incorrectly failed to reject H_0.

6.25

(a) College grads: $\frac{154}{438} = 0.352$
Non-college grads: $\frac{132}{389} = 0.339$

(b) Let p_{CG} represent the proportion of college graduates who support offshore drilling, and p_{NCG} represent the proportion of non-college graduates who do so. Then,

$$H_0 : p_{CG} = p_{NCG}$$
$$H_A : p_{CG} \neq p_{NCG}$$

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: Both samples are random and unrelated, so independence is satisfied.
2. Success-failure: First we need to find $\hat{p}_{pool}$ and then use that to calculate the numbers of expected successes and failures in each group.

$$\hat{p}_{pool} = \frac{success_{CG} + success_{NCG}}{n_{CG} + n_{NCG}} = \frac{154 + 132}{438 + 389} = \frac{286}{827} = 0.346$$
$$1 - \hat{p}_{pool} = 1 - 0.346 = 0.654$$

$$438 \times 0.346 = 151.548 > 10 \qquad 438 \times 0.654 = 286.452 > 10$$
$$389 \times 0.346 = 134.594 > 10 \qquad 389 \times 0.654 = 254.406 > 10$$

Since the observations are independent and the success-failure condition is met, $\hat{p}_{CG} - \hat{p}_{NCG}$ is expected to be approximately normal. Next we calculate the test statistic and the p-value:

$$Z = \frac{(\hat{p}_{CG} - \hat{p}_{NCG}) - 0}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{CG}} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{NCG}}}}$$
$$= \frac{(0.352 - 0.339)}{\sqrt{\frac{0.346 \times 0.654}{438} + \frac{0.346 \times 0.654}{389}}} = \frac{0.013}{0.033} = 0.39$$

$$p - value = P(|\hat{p}_{CG} - \hat{p}_{NCG}| > 0.013 \mid (p_{CG} - p_{NCG}) = 0) = P(|Z| > 0.39) = 2 \times 0.3483 = 0.6966$$

Since the p-value $> \alpha$ (0.05), we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling in California.

6.26

(a) The hypotheses are

$H_0 : p_R = p_D$ (Proportions of Republicans and Democrats who support the use of full-body scans are equal.)

$H_A : p_R \neq p_D$ (Proportions of Republicans and Democrats who support the use of full-body scans are different.)

The pooled proportion can be calculated as follows:

$$\hat{p}_{pool} = \frac{success_R + success_D}{n_R + n_D} = \frac{264 + 299}{318 + 369} = \frac{563}{687} \approx 0.82$$

Next we calculate the test statistic and the p-value:

$$\hat{p}_R = \frac{264}{318} = 0.83 \qquad \hat{p}_D = \frac{299}{369} = 0.81$$

$$Z = \frac{(\hat{p}_R - \hat{p}_D)}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_R} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_D}}}$$

$$= \frac{(0.83 - 0.81)}{\sqrt{\frac{0.82 \times 0.18}{318} + \frac{0.82 \times 0.18}{369}}} = \frac{0.02}{0.0294} = 0.68$$

$$p - value = P(|\hat{p}_R - \hat{p}_D)| > 0.02 \mid (p_R - p_D) = 0) = P(|Z| > 0.68) = 2 \times 0.2483 = 0.4966$$

Since the p-value is high, we fail to reject H_0. The data do not provide strong evidence that the proportions of Republicans and Democrats who support the use of full-body scans are different.

(b) We may have made a Type II error, since we may have incorrectly failed to reject H_0.

6.27 The hypotheses are

$H_0 : p_C = p_T$ (Rate of sleep deprivation is the same for the non-transportation workers and truck drivers.)

$H_A : p_C \neq p_T$ (Rate of sleep deprivation is different for the non-transportation workers and truck drivers.)

Before conducting the hypothesis test, we must first check that the conditions for inference are satisfied.

1. Independence: Both samples are random, so observations are independent within each sample. The samples are also unrelated and so are independent.
2. Success-failure: First we need to find $\hat{p}_{pool}$ and then use that to calculate the numbers of expected successes and failures in each group.

$$\hat{p}_{pool} = \frac{success_C + success_T}{n_C + n_T} = \frac{35 + 35}{292 + 203} = \frac{70}{495} = 0.1414$$
$$1 - \hat{p}_{pool} = 1 - 0.1414 = 0.8586$$

$$292 \times 0.1414 = 41.29 > 10 \qquad 292 \times 0.8586 = 250.71 > 10$$
$$203 \times 0.1414 = 28.70 > 10 \qquad 203 \times 0.8586 = 174.30 > 10$$

Since the observations are independent and the success-failure condition is met, $(\hat{p}_C - \hat{p}_T)$ is expected to be approximately normal.

Next, we calculate the test statistic and the p-value:

$$\hat{p}_C = \frac{35}{292} = 0.1199 \qquad \hat{p}_T = \frac{35}{203} = 0.1724$$
$$Z = \frac{(\hat{p}_C - \hat{p}_T)}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_C} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_T}}} = \frac{(0.1199 - 0.1724)}{\sqrt{\frac{0.1414 \times 0.8586}{292} + \frac{0.1414 \times 0.8586}{203}}} = \frac{-0.0525}{0.0318} = -1.65$$
$$p - value = P(|\hat{p}_C - \hat{p}_T| > 0.0525 \mid (p_C - p_T) = 0) = P(|Z| > 1.65) = 2 \times 0.0495 = 0.0989$$

Since the p-value is high (default to alpha = 0.05), we fail to reject H_0. The data do not provide strong evidence that the rates of sleep deprivation (defined as getting less than 6 hours of sleep per night) are different for non-transportation workers and truck drivers.

6.28

(a) The hypotheses are as follows:

$H_0 : p_V = p_{NV}$ (There is no difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.)
$H_A : p_V \neq p_{NV}$ (There is some difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.)

(b) Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The sample is random, and we can safely assume that $254 < 10\%$ of all mothers of autistic children and $229 < 10\%$ of all mothers of children with a typical development, therefore whether or not one mother took prenatal vitamins during the three months before pregnancy is independent of another.
2. Success-failure: First we need to find $\hat{p}_{pool}$ and then use that to calculate the numbers of expected successes and failures in each group.

$$\hat{p}_{pool} = \frac{success_V + success_{NV}}{n_V + n_{NV}} = \frac{111 + 143}{181 + 302} = \frac{254}{483} = 0.53$$
$$1 - \hat{p}_{pool} = 1 - 0.47 = 0.86$$

$$181 \times 0.53 = 95.93 > 10 \qquad 181 \times 0.47 = 85.07 > 10$$
$$302 \times 0.53 = 160.06 > 10 \qquad 302 \times 0.47 = 141.94 > 10$$

Since the observations are independent and the success-failure condition is met, $(\hat{p}_V - \hat{p}_{NV})$ is expected to be approximately normal.

Next we calculate the test statistic and the p-value:

$$\hat{p}_V = \frac{143}{302} = 0.47 \qquad \hat{p}_{NV} = \frac{111}{181} = 0.61$$
$$Z = \frac{(\hat{p}_V - \hat{p}_{NV})}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_V} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{NV}}}}$$
$$= \frac{(0.47 - 0.61)}{\sqrt{\frac{0.53 \times 0.47}{181} + \frac{0.53 \times 0.47}{302}}} = \frac{-0.14}{0.0469} = -2.99$$
$$p - value = P(|\hat{p}_V - \hat{p}_{NV}| > 0.14 \mid (p_V - p_{NV}) = 0) = P(|Z| > 2.99) = 2 \times 0.0014 = 0.0028$$

Since the p-value $< \alpha$, we reject H_0. There is strong evidence of a difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.

(c) The title of this newspaper article makes it sound like using prenatal vitamins can prevent autism, which is a causal statement. Since this is an observational study, we cannot make causal statements based on the findings of the study. A more accurate title would be "Mothers who use prenatal vitamins before pregnancy are found to have children with a lower rate of autism".

6.29

(a) The two-way table below presents the results of this study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) The hypotheses are as follows:

$H_0 : p_N = p_L$: There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups.

$H_A : p_N \neq p_L$: There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups.

(c) First, check conditions.

1. Independence: Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population.
2. Success-failure: First we need to find $\hat{p}_{pool}$ and then use that to calculate the numbers of expected successes and failures in each group.

$$\hat{p}_{pool} = \frac{success_N + success_L}{n_N + n_L} = \frac{26 + 10}{120 + 120} = \frac{36}{240} \approx 0.15$$
$$1 - \hat{p}_{pool} = 1 - 0.15 = 0.85$$
$$120 \times 0.15 = 18 > 10 \qquad 120 \times 0.85 = 102 > 10$$

Since the observations are independent and the success-failure condition is met, $(\hat{p}_C - \hat{p}_T)$ is approximately normal. We should keep in mind that it is unclear if the findings generalize to the population. Next we calculate the test statistic and the p-value:

$$\hat{p}_N = \frac{26}{120} = 0.2167 \qquad \hat{p}_L = \frac{10}{120} = 0.0833$$
$$Z = \frac{(\hat{p}_C - \hat{p}_T)}{\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_C} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_T}}}$$
$$= \frac{(0.2167 - 0.0833)}{\sqrt{\frac{0.15 \times 0.85}{120} + \frac{0.15 \times 0.85}{120}}} = \frac{0.1334}{0.0461} = 2.89$$
$$p - value = P(|\hat{p}_C - \hat{p}_T| > 0.1334 \mid (\hat{p}_C - \hat{p}_T) = 0) = P(|Z| > 2.89) = 2 \times 0.0019 = 0.0039$$

Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups. Treatment and virologic failure do not appear to be independent.

6.30 No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

6.31

(a) False. The chi-square distribution has only one parameter, the degrees of freedom.

(b) True.

(c) True.

(d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

6.32

(a) True.

(b) True.

(c) False. The chi-square distribution is right skewed, and the p-value is defined as $P(\chi^2 > X^2)$, therefore we only shade the right tail.

(d) False. The variability increases as the degrees of freedom increases.

6.33

(a) The hypotheses are as follows:

H_0: The distribution of the format of the book used by the students follows the professor's predictions.

H_A: The distribution of the format of the book used by the students does not follow the professor's predictions.

(b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$
$E_{print} = 126 \times 0.25 = 31.5$
$E_{online} = 126 \times 0.15 = 18.9$

(c) 1. Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable.

2. Sample size: All expected counts are at least 5.

(d) The chi-squared statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(71-75.6)^2}{75.6} + \frac{(30-31.5)^2}{31.5} + \frac{(25-18.9)^2}{18.9} = 2.32$$
$$df = 2$$
$$p - value = 0.313$$

(e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

6.34

(a) The hypotheses are as follows:

H_0: Distribution of foraging preference follows distribution of available land type.

H_A: Distribution of foraging preference follows distribution of available land type.

(b) Use a chi-squared goodness of fit test.

(c) 1. Independence: We are not told if these plots are sampled randomly.

2. Sample size: Expected counts can be calculated as follows: $E_{Wood} = 426 \times 0.048 = 20$ $E_{Cultivated\ grassplot} = 426 \times 0.147 = 63$ $E_{Deciduous\ forests} = 426 \times 0.396 = 169$ $E_{Other} = 426 \times (1 - (0.048 + 0.147 + 0.396)) = 174$ These are all above 5.

(d) The chi-squared statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(4-20)^2}{20} + \frac{(16-63)^2}{63} + \frac{(67-169)^2}{169} + \frac{(345-174)^2}{174} = 277.48$$

$$df = 4 - 1 = 3$$

$$p - value = P(\chi_3^2 > 277.48) < 0.001$$

Since the p-value is less than 5%, we reject H_0. The data provide strong evidence that barking deer prefer to forage in certain habitats over others.

6.35

(a) The two-way table below presents the results of this study:

		Quit		
		Yes	No	Total
Treatment	Patch + support group	40	110	150
	Only patch	30	120	150
	Total	70	230	300

(b) i. $E_{row_1,col_1} = \frac{(row\ 1\ total)*(col\ 1\ total)}{table\ total} = \frac{150 \times 70}{300} = 35$

ii. $E_{row_2,col_2} = \frac{(row\ 2\ total)*(col\ 2\ total)}{table\ total} = \frac{150 \times 230}{300} = 115$

6.36

(a) $E_{row_2,col_1} = \frac{(row\ 2\ total)*(col\ 1\ total)}{table\ total} = \frac{(38+55+77)\times 318}{(318+369+450)} = \frac{170\times 318}{1137} = 47.5$

(b) $E_{row_1,col_2} = \frac{(row\ 1\ total)*(col\ 2\ total)}{table\ total} = \frac{(264+299+351)\times 369}{1137} = \frac{914\times 369}{1137} = 296.6$

(c) $E_{row_3,col_3} = \frac{(row\ 3\ total)*(col\ 3\ total)}{table\ total} = \frac{(16+15+22)\times 450}{1137} = \frac{53\times 450}{1137} = 21.0$

6.37 Use a chi-squared test for independence. The hypotheses are as follows:

H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California.

H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence is reasonable.
2. Sample size: Under H_0 the expected counts can be calculated as follows:

$$E_{row\ 1,col\ 1} = \frac{438 \times (154+132)}{827} = 151.5 \qquad E_{row\ 1,col\ 2} = \frac{389 \times (154+132)}{827} = 134.5$$
$$E_{row\ 2,col\ 1} = \frac{438 \times (180+126)}{827} = 162.1 \qquad E_{row\ 2,col\ 2} = \frac{389 \times (180+126)}{827} = 143.9$$
$$E_{row\ 3,col\ 1} = \frac{438 \times (104+131)}{827} = 124.5 \qquad E_{row\ 3,col\ 2} = \frac{389 \times (104+131)}{827} = 110.5$$

All expected counts are at least 5.

The chi-squared statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(154-151.5)^2}{151.5} + \frac{(132-134.5)^2}{134.5} + \frac{(180-162.1)^2}{162.1}$$
$$+ \frac{(126-143.9)^2}{143.9} + \frac{(104-124.5)^2}{124.5} + \frac{(131-110.5)^2}{110.5} = 11.47$$
$$df = (R-1) \times (C-1) = (3-1) \times (2-1) = 2$$
$$p-value = P(\chi_2^2 > 11.47) \rightarrow p-value = 0.003$$

Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is some difference in the rate of support for drilling for oil and natural gas off the Coast of California based on whether or not the respondent graduated from college. Support for off- shore drilling and having graduated from college do not appear to be independent.

6.38

(a) The hypotheses are:

H_0: The outcomes are independent of which treatment a patient received.

H_A: There is some dependence of the patient outcome with the treatment.

(b) The p-value is smaller than 0.05, so we reject the null hypothesis. That is, there is strong evidence that some treatments were better / worse than others at treating lymphatic filariasis.

6.39 No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

6.40

(a) Multiply each percent by 701:

	Download	No Download
Position 1	97	128
Position 2	102	130
Position 3	85	159

(b) **Prepare.** It is convenient to construct the actual table we care about, which will represent the totals in the three experiment groups:

Group	Total
Position 1	225
Position 2	232
Position 3	244

This now looks like a goodness of fit chi-square problem. As part of our Prepare step, we should also construct the hypotheses.

H_0: The chance a user is in any experiment group is equal among the three groups.

H_A: The chance of being in one group or another varies by the group.

Check. If H_0 were true, then we'd expect 1/3 of the 701 visitors to be in each group, i.e. 233.67 in each group. The 233.67 is greater than 5 for all three groups, satisfying one of the required conditions. The independence condition is also satisfied since each visitor was randomly assigned to a group and only counted once in the table. With the conditions satisfied, we can move forward with conducting a goodness of fit chi-square test.
Calculate. We can compute the test statistic and degrees of freedom:

$$\begin{aligned} X^2 &= \frac{(225-233.67)^2}{233.67} + \frac{(232-233.67)^2}{233.67} + \frac{(244-233.67)^2}{233.67} \\ &= 0.79 \\ df &= 3-1 = 2 \end{aligned}$$

Using software, we can identify the tail area as 0.67.
Conclude. Because the p-value is larger than 0.05, we do not reject H_0. That is, we do not see any evidence that visitor randomization to the groups is imbalanced / malfunctioning.

(c) **Prepare.**

H_0: No difference in download rate across the experiment groups.

H_A: There is some difference in download rate across the groups.

Check. Each visitor was randomly assigned to a group and only counted once in the table, so the observations are independent. The expected counts can also be computed by constructing row and column totals, then multiplying these (and dividing by the table total) to get the expected counts under H_0. Those expected counts are (reading down the first column then down the second): 91.2, 94.0, 98.9, 133.8, 138.0, 145.2. All of these expected counts are at least 5. Therefore we can use the chi-square test.
Calculate. We can calculate the test statistic as

$$X^2 = \frac{(97-91.2)^2}{91.2} + \cdots + \frac{(159-145.2)^2}{145.2} = 5.04$$

with $df = (3-1) \times (2-1) = 2$. This results in a p-value of 0.08, which we can find using software.
Conclude. Because the p-value is less than 0.05, we do not reject H_0. That is, we have not found evidence that there is any difference in the download rate depending on the download link position.

6.41

(a) The hypotheses are as follows:

H_0: There is no difference in rates of preferred shipping method and age among Los Angeles residents.

H_A: There is some difference in rates of preferred shipping method and age among Los Angeles residents.

(b) The conditions are not satisfied since some expected counts are below 5.

6.42

(a) The hypotheses are as follows:

$H_0 : p = 0.5$ (50% of Americans think the Civil War is still relevant)

$H_A : p \neq 0.5$ (A majority of Americans hold an opinion (for or against) that the Civil War is still relevant.)

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The sample is random, and $1,507 < 10\%$ of all Americans, therefore whether or not one American in the sample thinks the Civil War is still relevant is independent of another.
2. Success-failure: $1,507 \times 0.50 = 753.5 > 10$ and $1,507 \times 0.50 = 753.5 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal. The test statistic and the p-value can be calculated as follows:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
$$= \frac{0.56 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1,507}}} = \frac{0.06}{0.0129} = 4.65$$
$$p - value = P(\hat{p} > 0.56 \mid p = 0.5) = P(|Z| > 4.65) \approx 0$$

Since the p-value is very small, we reject H_0. Since the data show $\hat{p} > 0.5$ and we've rejected H_0, the data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) If in fact only 50% of Americans thought the Civil War is still relevant, the probability of obtaining a ransom sample of 1,507 Americans where 56% or more think it is still relevant would be approximately 0.

(c) First we need to recheck the success-failure condition using the sample proportion: $1,507 \times 0.56 = 843.92 > 10$ and $1,507 \times 0.44 = 663.08 > 10$. A 90% confidence interval can be calculated as follows:

$$\hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.56 \pm 1.65 \times \sqrt{\frac{0.56 \times 0.44}{1,507}}$$
$$= 0.56 \pm 1.65 \times 0.0128$$
$$= 0.56 \pm 0.02$$
$$= (0.54, 0.58)$$

We are 90% confident that 54% to 58% of all Americans think that the Civil War is still relevant. This agrees with the conclusion of the earlier hypothesis test since the interval lies above 50%.

6.43

(a) Before constructing the confidence interval we should check that the conditions are satisfied.

1. Independence: The sample is random, therefore whether or not not one student smokes is independent of the smoking status of another student in this sample.
2. Success-failure: $200 \times 0.2 = 40 > 10$ and $200 \times 0.8 = 160 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal. A 95% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} &= \frac{40}{200} = 0.2 \\ \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.2 \pm 1.96 \times \sqrt{\frac{0.2 * 0.8}{200}} \\ &= 0.2 \pm 0.055 \\ &= (0.145, 0.255) \end{aligned}$$

We are 95% confident that 14.5% to 25.5% of all students at this university smoke.

(b) We are asked to solve for the sample size required to achieve a 2% margin of error for a 95% confidence interval and the point estimate is $\hat{p} = 0.2$.

$$\begin{aligned} ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad &\rightarrow \quad 0.02 \geq 1.96 \sqrt{\frac{0.2 \times 0.8}{n}} \\ 0.02^2 &\geq 1.96^2 \frac{0.2 \times 0.8}{n} \\ n &\geq \frac{1.96^2 \times 0.2 \times 0.8}{0.02^2} \\ n &\geq 1536.64 \\ n &\geq 1537 \end{aligned}$$

The sample size n should be at least 1,537.

6.44

(a) We are asked to solve for the sample size required to achieve a 2% margin of error for a 98% confidence interval. Since we are not given information from a previous study to use as an estimate for $\hat{p}$, we use 0.5 as it will yield the most conservative estimate.

$$
\begin{aligned}
ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad &\rightarrow \quad 0.02 \geq 2.33 \sqrt{\frac{0.5 \times 0.5}{n}} \\
0.02^2 &\geq 2.33^2 \frac{0.5 \times 0.5}{n} \\
n &\geq \frac{2.33^2 \times 0.5 \times 0.5}{0.02^2} \\
n &\geq 3393.062 \\
n &\geq 3394
\end{aligned}
$$

The researcher needs a minimum of 3,394 subjects, and therefore needs to set aside a minimum of $\$3,394 \times 20 = \$67,880$.

(b) Decreasing the sample size would increase the margin of error hence make the interval wider, i.e. the interval would lose precision.

6.45

(a) The population parameter of interest is the proportion of all graduates from this university who found a job within one year of graduating, p. The point estimate of this parameter is the proportion of graduates in the sample who found a job within one year of graduating, $\hat{p} = \frac{348}{400} = 0.87$.

(b) 1. Independence: This is a simple random sample, so the observations are independent.

2. Success-failure:$400 \times 0.87 = 348 > 10$ and $400 \times 0.13 = 52 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal.

(c) A 95% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.87 \pm 1.96 \times \sqrt{\frac{0.87 \times 0.13}{400}} \\ &= 0.87 \pm 1.96 \times 0.0168 \\ &= 0.87 \pm 0.0329 \\ &= (0.8371, 0.9029) \end{aligned}$$

We are 95% confident that approximately 84% to 90% of graduates from this university found a job within one year of completing their undergraduate degree.

(d) 95% of random samples would produce a 95% confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college.

(e) A 99% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.87 \pm 1.96 \times \sqrt{\frac{0.87 \times 0.13}{400}} \\ &= 0.87 \pm 2.58 \times 0.0168 \\ &= 0.87 \pm 0.0433 \\ &= (0.8267, 0.9133) \end{aligned}$$

We are 99% confident that approximately 83% to 91% of graduates from this university found a job within one year of completing their undergraduate degree.

(f) The width of the 95% confidence interval is $90\% - 84\% = 6\%$ and the width of the 99% confidence interval is $91\% - 83\% = 8\%$. Since these intervals are based on the same data, i.e. sample sizes are the same, increasing the confidence level decreases the precision.

6.46

(a) The two-way table below presents the results of this study:

		Diabetes		
		Yes	No	Total
Employment	Employed	629	41,290	41,919
	Unemployed	146	5,709	5,855
	Total	775	46,999	47,774

(b) The hypotheses are as follows:

$H_0 : p_E = p_U$: There is no difference in the rate of diabetes between employed and unemployed 18-29 year olds.

$H_A : p_E \neq p_U$: There is some difference in the rate of diabetes between employed and unemployed 18-29 year olds.

(c) With a small p-value we reject H_0 and conclude that the data provide strong evidence of a difference in the diabetes rates of employed and unemployed people. However the observed difference between the sample proportions is very small (1%) and the sample sizes are very large. Therefore while the results are statistically significant, they may not be practically significant.

6.47 Use a chi-squared goodness of fit test. The hypotheses are as follows:

H_0: Each option is equally likely.

H_A: Some options are preferred over others.

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: We are checking to see if the decisions for rock, paper, scissors are independent or not, so we can't validate this condition.
2. Sample size: Total sample size is 99, and expected counts are $(1/3) \times 99 = 33$ for each option. These are all above 5, so conditions are satisfied.

The chi-squared statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52$$
$$df = 3 - 1 = 2$$
$$p - value = P(\chi_2^2 > 7.52) \rightarrow p - value = 0.023$$

Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

6.48

(a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

(b) True. This is the correct interpretation of the confidence interval, which can be calculated as $0.46 \pm 0.03 = (0.43, 0.49)$.

(c) False. The confidence interval does not tell us what we might expect to see in another random sample.

(d) False. As the confidence level decreases, the margin of error decreases as well.

6.49

(a) The hypotheses are as follows:

$H_0 : p = 0.38$ (38% of Americans use only their cell phones for browsing the internet)
$H_A : p \neq 0.38$ (Some percentage different than 38% of Americans use only their cell phones for browsing the internet)

Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The sample is random, therefore whether or not one American in the sample uses only their phone to browse the internet is independent of another.
2. Success-failure: $2,254 \times 0.38 = 856.52 > 10$ and $2,254 \times 0.62 = 1397.48 > 10$.

Since the observations are independent and the success-failure condition is met, $\hat{p}$ is expected to be approximately normal. The test statistic and the p-value can be calculated as follows:

$$\begin{aligned} Z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.17 - 0.38}{\sqrt{\frac{0.38 \times 0.62}{2,254}}} = \frac{-0.21}{0.0105} = -20 \end{aligned}$$

$$p - value = P(\hat{p} > 0.17 | p = 0.38) = P(|Z| > 20) \approx 0$$

Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%, and the data indicate that the proportion is lower in the US.

(b) If in fact 38% of Americans only used only their cell phones to access the internet, the probability of obtaining a random sample of 2,254 Americans where 17% or less or 59% or more use their only their cell phones to access the internet would be approximately 0.

(c) A 95% confidence interval can be calculated as follows:

$$\begin{aligned} \hat{p} \pm z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.17 \pm 1.96 \times \sqrt{\frac{0.17 \times 0.83}{2,254}} \\ &= 0.17 \pm 1.96 \times 0.0079 \\ &= 0.17 \pm 0.0155 \\ &= (0.1545, 0.1855) \end{aligned}$$

We are 95% confident that approximately 15% to 19% of all Americans use only their cell phones to browse the internet.

6.50

(a) Chi-squared test of independence.

(b) The hypotheses are:

H_0: Caffeinated coffee consumption and depression in women are independent.

H_A: Caffeinated coffee consumption and clinical are in women dependent/associated.

(c) Depression: $2607/50739 = 0.0514$
No depression: $1 - 0.0514 = 0.9486$

(d) $E = \frac{2607*6617}{50739} = 339.9854 \approx 340$
$\frac{(O-E)^2}{E} = \frac{(373-340)^2}{340} = 3.20$

(e) $df = (R-1) \times (C-1) = 1 \times 4 = 4$, and $p - value < 0.001$.

(f) p-value is small and we reject H_0. The data provide convincing evidence to suggest that caffeinated coffee consumption and depression in women are associated.

(g) Yes, this is an observational study. Based on this study we can't deduce that drinking more coffee leads to less depression. There may be other factors, lurking variables, that cause decreased depression in women who drink more coffee.

Chapter 7

Inference for numerical data

7.1

(a) $n = 6$, CL $= 90\%$, $df = 6 - 1 = 5$, $t^\star_5 = 2.02$
(b) $n = 21$, CL $= 98\%$, $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$
(c) $n = 29$, CL $= 95\%$, $df = 29 - 1 = 28$, $t^\star_{28} = 2.05$
(d) $n = 12$, CL $= 99\%$, $df = 12 - 1 = 11$, $t^\star_{11} = 3.11$

7.2 The dotted line is the t-distribution with 1 degree of freedom, the dashed line is the t-distribution with 5 degrees of freedom, and the solid line is the standard normal distribution. As the degrees of freedom increases the t-distribution approaches the normal distribution. Another valid justification is that lower the degrees of freedom, thicker the tails.

7.3

(a)	$n = 11$	$T = 1.91$	$df = 11 - 1 = 10$	$p - value = 0.085$	Do not reject H_0
(b)	$n = 17$	$T = -3.45$	$df = 17 - 1 = 16$	$p - value = 0.003$	Reject H_0
(c)	$n = 7$	$T = 0.83$	$df = 7 - 1 = 6$	$p - value = 0.438$	Do not reject H_0
(d)	$n = 28$	$T = 2.13$	$df = 28 - 1 = 27$	$p - value = 0.042$	Reject H_0

7.4

(a) $n = 26$ $T = 2.485$ $df = 26 - 1 = 25$ $p - value = 0.020$ Do not reject H_0

(a) $n = 18$ $T = 0.5$ $df = 18 - 1 = 17$ $p - value = 0.623$ Do not reject H_0

7.5 The sample mean is the mid-point of the confidence interval, i.e. the average of the upper and lower bounds:

$$\bar{x} = \frac{18.985 + 21.015}{2} = 20$$

The margin of error is $21.015 - 20 = 1.015$. Since $n = 36$, $df = 35$, and the critical t-score is $t^\star_{35} = 2.03$. Then,

$$1.015 = 2.03\frac{s}{\sqrt{36}} \;\rightarrow\; s = 3$$

7.6 The sample mean is the mid-point of the confidence interval, i.e. the average of the upper and lower bounds:

$$\bar{x} = \frac{65 + 77}{2} = 71$$

The margin of error is $77 - 71 = 6$. Since $n = 25$, $df = 24$, and the critical t-score is $t^\star_{35} = 1.71$. Then,

$$6 = 1.71\frac{s}{\sqrt{25}} \ \rightarrow s \approx 17.54$$

7.7

(a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.)
H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.)

(b) Before calculating the test statistic we should check that the conditions are satisfied.

1. Independence: The sample is random.
2. Normality: All observations are within three standard deviations of the mean. While this is encouraging, it would be useful to see the raw data. However, for now we will proceed while acknowledging that we are assuming there aren't any clear outliers.

The test statistic and degrees of freedom can be calculated as follows:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = \frac{-0.27}{0.154} = -1.75$$
$$df = 25 - 1 = 24$$

(c) p-value $= 2 \times P(T_{24} < -1.75) \rightarrow p - value = 0.093$. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093.

(d) Since p-value > 0.05, do not reject H_0. The data do not provide convincing evidence that New Yorkers sleep more or less than 8 hours per night on average.

(e) Yes, the hypothesis test didn't reject the null hypothesis, therefore we'd expect 8 hours to be in the interval.

7.8

(a) Use the sample mean to estimate the population mean: 171.1. Likewise, use the sample median to estimate the population median: 170.3.

(b) Use the sample standard deviation (9.4) and sample IQR ($177.8 - 163.8 = 14$).

(c) In order to determine if 180 cm or 155 cm are considered unusual observations we need to calculate how many standard deviations away from the mean this observation is, i.e. calculate the Z-score.

$$Z = \frac{180 - 171.1}{9.4} = 0.95 \qquad Z = \frac{155 - 171.1}{9.4} = -1.71$$

Neither of these observations is more than two standard deviations away from the mean, so neither would be considered unusual.

(d) No, sample point estimates only estimate the population parameter, and they vary from one sample to another. Therefore we cannot expect to get the same mean and standard deviation with each random sample.

(e) We use the standard error of the mean to measure the variability in means of random samples of same size taken from a population. The variability in the means of random samples is quantified by the standard error. Based on this sample, $SE_{\bar{x}} = \frac{9.4}{\sqrt{507}} = 0.417$.

7.9 For the single tails to each be 0.025 at $n-1 = 20-1 = 19$ degrees of freedom, T score must equal to be either -2.09 or +2.09. Then, either:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

7.10 With a larger critical value, the confidence interval ends up being wider. This makes intuitive sense as when we have a small sample size and the population standard deviation is unknown, we should have a wider interval than if we knew the population standard deviation, or if we had a large enough sample size.

7.11

(a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu \neq 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.4919$. The test statistic is $T = (4.6 - 5)/SE = -0.81$. $df = 20 - 1 = 19$. The one-tail area is about 0.21, so the p-value is about 0.42, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject the notion that the average is 5 years.

(b) Using $SE = 0.4919$ and $t^\star_{df=19} = 2.093$, the confidence interval is (3.57, 5.63). We are 95% confident that the average number of years a child takes piano lessons in this city is 3.57 to 5.63 years.

(c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

7.12

(a) $H_0 : \mu = 35$, $H_A : mu \neq 35$.

(b) 1. Independence: if we can assume that these 52 officers represent a random sample (big assumption), then independence would be satisfied, but we cannot check this.

2. Normality: We don't have a plot of the distribution that we can use to check this condition. We at least have more than 30 observations, so the distribution would have to be extremely skewed to be an issue. We again cannot check this, but this seems like a less concerning issue than the independence consideration.

(c) The test statistic and the p-value can be calculated as follows:

$$T = \frac{124.32 - 35}{\frac{37.74}{\sqrt{52}}} \approx 17.07$$

$$df = 52 - 1 = 51$$

$$p - value = 2 \times P(T_{51} > 17.07) < 0.001$$

The hypothesis test yields a very small p-value, so we reject H_0. Given the direction of the data, there is very convincing evidence that the police officers have been exposed to a higher concentration of lead than individuals living in a suburban area.

7.13

$$10 \geq 1.96 \times \frac{100}{\sqrt{n}} \;\rightarrow\; n \geq \left(\frac{1.96 \times 100}{10}\right)^2 \approx 384.16$$

He should survey at least 385 customers. Note that we need to round up the calculated sample size.

7.14

(a) $25 \geq 1.65 \times \frac{250}{\sqrt{n}} \rightarrow n \geq \left(\frac{1.65 \times 250}{25}\right)^2 \approx 272.25$. Raina should collect a sample of at least 273 students.

(b) If Luke had the same sample size as Raina but used a higher confidence level, he would end up with wider interval. To keep the width of his confidence interval the same as Raina's Luke will need a higher sample size.

(c) $25 \geq 2.58 \times \frac{250}{\sqrt{n}} \rightarrow n \geq \left(\frac{2.58 \times 250}{25}\right)^2 \approx 665.64$. Luke should collect a sample of at least 666 students.

7.15 Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point.

7.16

(a) True.

(b) True.

(c) True.

(d) False. We find the difference of each pair of observations, and then we do inference on these differences.

7.17

(a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent.

(b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group.

(c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent.

(d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

7.18

(a) Paired, on the same day the stock prices may be dependent on external factors that affect the price of both stocks.

(b) Paired, the prices are for the same items.

(c) Not paired, these are two independent random samples, individual students are not matched.

7.19

(a) For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired.

(b) $H_0 : \mu_{\text{diff}} = 0$ (There is no difference in average number of days exceeding 90°F in 1948 and 2018 for NOAA stations.) $H_A : \mu_{\text{diff}} \neq 0$ (There is a difference.)

(c) Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). Therefore, the conditions are satisfied.

(d) $SE = 17.2/\sqrt{197} = 1.23$. $T = \frac{2.9-0}{1.23} = 2.36$ with degrees of freedom $df = 197 - 1 = 196$. This leads to a one-tail area of 0.0096 and a p-value of about 0.019.

(e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed more 90°F days in 2018 than in 1948.

(f) Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe a decrease, but the sample we took just so happened to make it appear that this was the case.

(g) No, since we rejected H_0, which had a null value of 0.

7.20

(a) The median writing score is slightly higher but it's difficult to tell if the average scores on the two tests are different or not.

(b) No, the score of one student on the reading test is not independent of their score on the writing test.

(c) Let $diff = read - write$. Then the hypotheses are:

$$H_0 : \mu_{diff} = 0$$
$$H_A : \mu_{diff} \neq 0$$

(d) The conditions for the sampling distribution of $\bar{x}_{diff}$ to be nearly normal and the estimate of the standard error to be sufficiently accurate are as follows:

1. Independence: Students are randomly sampled and $200 < 10\%$ of all students who take this survey, therefore we can assume that the reading and writing scores of one student are independent of another.
2. Normality: The distribution of the differences appears fairly symmetric, so we can assume that the sampling distribution of average differences will be approximately normal.

(e) The test statistic and the p-value can be calculated as follows:

$$\begin{aligned} T &= \frac{\bar{x}_{diff} - \mu_{diff}}{\frac{s_{diff}}{\sqrt{n}}} \\ &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} \approx -0.87 \\ df &= 200 - 1 = 199 \\ p - value &= P(|T_{199}| > 0.87) > 0.20 \end{aligned}$$

Since the p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average reading and writing scores.

(f) We may have made a Type 2 error, i.e. we may have incorrectly failed to reject H_0. In this context a Type 2 error means deciding that the data do not provide convincing evidence of a difference between the average reading and writing scores of students when in reality there is a difference.

(g) Since we failed to reject H_0, which claimed the average difference is equal to 0, we would expect a confidence interval to include this value.

7.21

(a) We checked conditions in an earlier exercise, so we'll jump right to the calculations. First we compute the standard error and find $z^\star$:

$$SE = \frac{17.2}{\sqrt{197}} = 1.23$$
$$z^\star = 1.65$$

Next, we can compute the confidence interval:

$$\bar{x} \pm z^\star \times SE \quad \rightarrow \quad 2.9 \pm 1.65 \times 1.23 \quad \rightarrow \quad (0.87, 4.93)$$

(b) We are 90% confident that there was an increase of 0.87 to 4.93 in the average number of days that hit 90°F in 2018 relative to 1948 for NOAA stations.

(c) Yes, since the interval lies entirely above 0.

7.22

(a) A 95% confidence interval can be calculated as follows:

$$\begin{aligned}\bar{x}_{diff} \pm t^{\star}_{df}\frac{s_{diff}}{\sqrt{n}} &= -0.545 \pm 1.98 \times \frac{8.887}{\sqrt{200}} \\ &= -0.545 \pm 1.98 \times 0.6284 \\ &= -0.545 \pm 1.244 \\ &= (-1.79, 0.70)\end{aligned}$$

(b) We are 95% confident that on the reading test students score, on average, 1.79 points lower to 0.70 points higher than they do on the writing test.

(c) No, since 0 is included in the interval.

7.23

(a) These data are paired. For example, the Friday the 13th in say, September 1991, would probably be more similar to the Friday the 6th in September 1991 than to Friday the 6th in another month or year.

(b) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$.

$H_0 : \mu_{diff} = 0$.

$H_A : \mu_{diff} \neq 0$.

(c) **Independence:** The months selected are not random. However, if we think these dates are roughly equivalent to a simple random sample of all such Friday 6th/13th date pairs, then independence is reasonable. To proceed, we must make this strong assumption, though we should note this assumption in any reported results.
Normality: With fewer than 10 observations, we would need to see clear outliers to be concerned. There is a borderline outlier on the right of the histogram of the differences, so we would want to report this in formal analysis results.

(d) First we compute the standard error, then the test statistic:

$$SE = \frac{1176}{\sqrt{10}} = 371.9T \qquad = \frac{1835 - 0}{371.9} = 4.93$$

The degrees of freedom are $df = 10 - 1 = 9$, and using software, we can then find the test statistic $T = 4.93$ with $df = 9$ corresponds to p-value = 0.001.

(e) Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of cars at the intersection is higher on Friday the 6^{th} than on Friday the 13^{th}. (We should exercise caution about generalizing the interpretation to all intersections or roads.)

(f) If the average number of cars passing the intersection actually was the same on Friday the 6^{th} and 13^{th}, then the probability that we would observe a test statistic so far from zero is less than 0.01.

(g) We might have made a Type 1 Error, i.e. incorrectly rejected the null hypothesis.

7.24 The hypotheses are: $H_0 : \mu_{0.99} = \mu_1$ and $H_A : \mu_{0.99} \neq \mu_1$. The conditions that need to be satisfied for the sampling distribution of $(\bar{x}_{0.99} - \bar{x}_1)$ to be nearly normal and the estimate of the standard error to be sufficiently accurate are:

1. Independence: Both samples are random and represent less than 10% of their respective populations. Also, we have no reason to think that the 0.99 carats are not independent of the 1 carat diamonds since they are both sampled randomly.
2. Normality: The distributions are not extremely skewed, hence we can assume that the distribution of the average differences will be nearly normal as well.

The test statistic and the p-value are calculated as follows:

$$\begin{aligned} T &= \frac{(\bar{x}_{0.99} - \bar{x}_1) - (\mu_{0.99} - \mu_1)}{\sqrt{\frac{s_{0.99}^2}{n_{0.99}} + \frac{s_1^2}{n_1}}} \\ &= \frac{(44.51 - 56.81) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{16.13^2}{23}}} = \frac{-12.3}{4.36} = -2.82 \\ df &= 23 - 1 = 22 \\ p - value &= P(|T_{22}| > 2.82) = 0.01 \end{aligned}$$

Since p-value < 0.05, reject H_0. The data provide convincing evidence that the average standardized price of 0.99 carats and 1 carat diamonds are different.

7.25

(a) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$. Then,

$H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$

$$T = \frac{\bar{x}_{diff} - \mu_{diff}}{\frac{s_{diff}}{\sqrt{n}}}$$
$$= \frac{-3.33 - 0}{\frac{3.01}{\sqrt{6}}} = \frac{-3.33}{1.23} = -2.71$$
$$df = n - 1 = 6 - 1 = 5$$
$$p - value = P(|T_5| > 2.71) \rightarrow 0.02 < p - value < 0.05$$

Since p-value < 0.05, reject H_0. The data provide convincing evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}.

(b) The 95% confidence interval can be constructed as follows:

$$t^\star_{df} = t^\star_{6-1=5} = 2.57$$
$$\bar{x}_{diff} \pm t^\star_{df} \frac{s_{diff}}{\sqrt{n}} = -3.33 \pm 2.57 \frac{3.01}{\sqrt{6}}$$
$$= -3.33 \pm 2.57 \times 1.23$$
$$= -3.33 \pm 3.16$$
$$= (-6.49, -0.17)$$

We are 95% confident that the average number of traffic accident related emergency room admissions on Friday the 6^{th} is 0.17 to 6.49 less than that of Friday the 13^{th}.

(c) This is an observational study, not an experiment, so we cannot infer causation based on the results of this study. While we found a significant difference between the average numbers of traffic accident related emergency room admissions, we have no way of telling if the difference is due to the day being Friday the 13^{th} or some other reason.

7.26 The 95% confidence interval can be calculated as follows:

$$
\begin{aligned}
t^\star_{df} &= t^\star_{23-1=22} = 2.07 \\
(\bar{x}_{0.99} - \bar{x}_1) \pm t^\star_{df}\sqrt{\frac{s^2_{0.99}}{n_{0.99}} + \frac{s^2_1}{n_1}} &= (56.81 - 44.51) \pm 2.07\sqrt{\frac{s^2_{0.99}}{n_{0.99}} + \frac{s^2_1}{n_1}} \\
&= (44.51 - 56.81) \pm 2.07\sqrt{\frac{13.32^2}{23} + \frac{16.13^2}{23}} \\
&= -12.3 \pm 2.07 \times 4.36 \\
&= -12.3 \pm 9.03 \\
&= (-21.33, -3.27)
\end{aligned}
$$

We are 95% confident that the average standardized price of a 0.99 carat diamond is \$3.27 to \$21.33 lower than the average standardized price of a 1 carat diamond.

7.27

(a) Chicken that were fed linseed on average weigh 218.75 grams while those that were given horsebean weigh on average 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken that were given linseed.

(b) The hypotheses are $H_0 : \mu_{ls} = \mu_{hb}$ and $H_0 : \mu_{ls} \neq \mu_{hb}$.

Before calculating the test statistic we should check that the conditions for the sampling distribution of $(\bar{x}_{ls} - \bar{x}_{hb})$ to be nearly normal and the estimate of the standard error to be sufficiently accurate are as follows:

1. Independence: Chickens are randomly assigned to feed groups, and 12 and $10 < 10\%$ of all chickens fed linseed and horsebean, respectively. Therefore we can assume that the weights of chicken fed linseed are independent of each other, as well as the weights of chicken fed horsebean.
2. Normality: The distributions look fairly symmetric, therefore we can assume that the distribution of average differences will be nearly normal.

Since population standard deviations are unknown and samples are small, we calculate a T score.

$$T = \frac{(\bar{x}_{ls} - \bar{x}_{hb}) - (\mu_{ls} - \mu_{hb})}{\sqrt{\frac{s_{ls}^2}{n_{ls}} + \frac{s_{hb}^2}{n_{hb}}}}$$

$$= \frac{(218.75 - 160.20) - 0}{\sqrt{\frac{52.24^2}{12} + \frac{38.63^2}{10}}} = \frac{58.55}{19.41} = 3.02$$

$$df = min(n_1 - 1, n_2 - 1) = min(11, 9) = 9$$

$$p - value = P(|T_9| > 3.02) \rightarrow 0.01 < p - value < 0.02$$

Since p-value < 0.05, we reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chicken that were fed linseed and horsebean.

(c) Type 1, since we rejected H_0.

(d) Yes, since p-value > 0.01, we would fail to reject H_0 and conclude that the data do not provide convincing evidence of a difference between the average weights of chickens that were fed linseed and horsebean.

7.28 The hypotheses are as follows:

$H_0 : \mu_{A,c} - \mu_{M,c}$

$H_0 : \mu_{A,c} \neq \mu_{M,c}$

We are told to assume that conditions for inference are satisfied.

Then, the test statistic and the p-value can be calculated as follows:

$$T = \frac{(\bar{x}_{A,c} - \bar{x}_{M,c}) - (\mu_{A,c} - \mu_{M,c})}{\sqrt{\frac{s^2_{A,c}}{n_{A,c}} + \frac{s^2_{M,c}}{n_{M,c}}}} = \frac{(16.12 - 19.85) - 0}{\sqrt{\frac{3.58^2}{26} + \frac{4.51^2}{26}}} = \frac{-3.73}{1.13} = -3.3$$

$$df = min(n_{M,c} - 1, n_{A,c} - 1) = min(26 - 1, 26 - 1) = 25$$

$$p - value = P(|T_{25}| > 3.3) < 0.01$$

Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a difference in the average city mileage between cars with automatic and manual transmissions.

7.29 The hypotheses are $H_0 : \mu_C = \mu_S$ and $H_0 : \mu_C \neq \mu_S$.

We are told to assume that conditions for inference are satisfied.

$$\begin{aligned} T &= \frac{(\bar{x}_{cs} - \bar{x}_{sb}) - (\mu_{cs} - \mu_{sb})}{\sqrt{\frac{s_{cs}^2}{n_{sb}} + \frac{s_{cs}^2}{n_{sb}}}} \\ &= \frac{(323.58 - 246.43) - 0}{\sqrt{\frac{64.43^2}{12} + \frac{54.13^2}{14}}} = \frac{82.15}{23.56} = 3.48 \\ df &= min(n_1 - 1, n_2 - 1) = min(11, 13) = 11 \\ p - value &= P(|T_{11}| > 3.27) < 0.01 \end{aligned}$$

Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. Since this is a randomized experiment the type of diet is the only difference between the various groups of chicken, and hence the observed differences are can be attributed to differences in diet.

7.30

$$df = min(n_1 - 1, n_2 - 1) = min(26 - 1, 26 - 1) = 25 \ \rightarrow \ t^\star_{25} = 2.49$$

$$\begin{aligned}(\bar{x}_{A,hwy} - \bar{x}_{M,hwy}) \pm t^\star_{df}\sqrt{\frac{s^2_{A,hwy}}{n_{A,hwy}} + \frac{s^2_{M,hwy}}{n_{M,hwy}}} &= (22.92 - 27.88) \pm 2.49 * \sqrt{\frac{5.29^2}{26} + \frac{5.01^2}{26}} \\ &= -4.96 \pm 2.49 \times 1.43 \\ &= -4.96 \pm 3.56 \\ &= (-8.52, -1.4)\end{aligned}$$

We are 98% confident that on the highway cars with manual transmissions get on average 1.4 to 8.52 MPG more than cars with automatic transmissions.

7.31 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. Then, for each treatment the hypotheses are:

$H_0 : \mu_{diff} = 0$: Treatment has no effect.

$H_A : \mu_{diff} > 0$: Treatment has an effect on P.D.T. scores, either positive or negative.

The conditions that need to be satisfied are:

1. Independence: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied.
2. Normality: All three sample sizes are smaller than 30, so we look for clear outliers. There is a borderline outlier in the first treatment group. Since it is borderline, we will proceed, but we should report this caveat with any results.

The test statistics is calculated as $T_{df} = \frac{\bar{x}_{diff} - \mu_{diff}}{\frac{s_{diff}}{\sqrt{n}}}$ and the associated degrees of freedom is $df = n - 1$. The calculation of the test statistics and p-values are shown below.

Treatment 1:

$$T = \frac{6.21 - 0}{\frac{12.3}{\sqrt{14}}} = \frac{6.21}{3.29} = 1.89$$
$$df = 14 - 1 = 13$$
$$p - value = 2 \times P(T_{13} > 1.89) = 0.081$$

Treatment 2:

$$T_{df} = \frac{2.86 - 0}{\frac{7.94}{\sqrt{14}}} = \frac{2.86}{2.12} = 1.35$$
$$df = 14 - 1 = 13$$
$$p - value = 2 \times P(T_{13} > 1.35) = 0.200$$

Treatment 3:

$$T_{df} = \frac{-3.21 - 0}{\frac{8.57}{\sqrt{14}}} = \frac{2.86}{2.29} = -1.40$$
$$df = 14 - 1 = 13$$
$$p - value = 2 \times P(T_{13} > -1.40) = 0.185$$

We do not reject the null hypothesis for any of these groups. As earlier noted, there is some uncertainty about if the method applied is reasonable for the first group.

7.32

(a) False, in order to be able to use a Z test both sample sizes need to be above 30.

(b) True.

(c) False, we use the pooled standard deviation when the variability in groups is constant.

7.33 Difference we care about: 40. Single tail of 90%: $1.28 \times SE$. At the 5% significance level the rejection region bounds span $\pm 1.96 \times SE$.

$$40 = 1.28 \times SE + 1.96 \times SE = 3.24 \times SE$$
$$SE = 12.35 = \sqrt{\frac{94^2}{n} + \frac{94^2}{n}}$$
$$n = 115.86$$

We will need 116 plots of land for each fertilizer.

7.34 Difference we care about: 0.5. Single tail of 90%: $0.84 \times SE$. At the 5% significance level the rejection region bounds span $\pm 1.96 \times SE$.

$$0.5 = 0.84 \times SE + 1.96 \times SE = 2.8 \times SE$$
$$SE = 0.1786 = \sqrt{\frac{2.2^2}{n} + \frac{2.2^2}{n}}$$
$$n = 303.47$$

We will need 304 plots of land for each fertilizer.

7.35 Alternative.

7.36 ANOVA, since we're comparing means of more than two groups.

7.37 The conditions that need to be satisfied for ANOVA are:

1. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable.
2. Approximately normal: The distributions of weights within each feed type appear to be fairly symmetric, with the possible exception of the sunflower group.
3. Constant variance: Based on the side-by-side box plots shown in Exercise **??**, the constant variance assumption appears to be reasonable.

The hypotheses are:
H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$
H_A: The average weight varies across some (or all) groups.

$F_{5,65} = 15.36$ and the p-value is approximately 0. With such a small p- value, we reject H_0. The data provide convincing evidence that the average weight of chicks varies across some (or all) feed supplement groups.

7.38

(a) The hypotheses are:
H_0: $\mu_{lec+disc} = \mu_{textbook} = \mu_{text+lec} = \mu_{comp} = \mu_{comp+lec}$
H_A: The average score varies across some (or all) groups.

(b) $df_G = 5 - 1 = 4$, $df_E = 44 - 6 = 40$

(c) Since the p-value is small (assuming $\alpha = 0.05$), the data provide convincing evidence that the average score varies across some (or all) groups.

7.39

(a) H_0: The mean MET for each group is equal to each other.

$$\mu_{\leq 1 \text{ cup/week}} = \mu_{\text{2-6 cups/week}} = \mu_{\text{1 cup/day}} = \mu_{\text{2-3 cups/day}} = \mu_{\geq 4 \text{ cups/day}}$$

H_A: At least one pair of means is different.

(b) The conditions that need to be satisfied for ANOVA are:

1. Independence: We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent.
2. Approximately normal: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable.
3. Constant variance: This condition appears to be met, standard deviations are pretty consistent across groups.

In order to proceed with the test we will need to assume independence.

(c)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee	5 - 1 = 4	25575327 - 25564819 = 10508	10508 / 4 = 2627	2627 / 505 = 5.2	0.0003
Residuals	50739 - 5 = 50734	25564819	25564819 / 50624 = 505		
Total	50739 - 1 = 50738	25575327			

(d) Since p-value is low, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

7.40 The conditions that need to be satisfied for ANOVA are:

1. Independence: We are not told if the students are randomly assigned to discussion sections, so we cannot be sure of independence of observations. That is, these data will not allow us to attribute any differences in scores to the teaching assistants since these data are observational. Students with different academic strengths may have tended to enroll in different sections.
2. Approximately normal: We are not given plots of the distributions of grades by discussion section, therefore we cannot check this condition. Also, the sample sizes are not all large, so we can't relax this condition.
3. Constant variance: Based on the standard deviations given in the summary table, the constant variance assumption appears to be reasonable for most groups, but not all. Also, the sample sizes aren't consistent so we can't relax this condition.

In order to proceed with the test we will need to assume independence and approximately normal distributions within groups.

The hypotheses are as follows:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_8$

H_A: The average score varies across some (or all) groups.

$F_{7,198} = 1.87$ and the p-value is 0.0767. With a p-value > 0.05, we fail to reject H_0. The data do not provide convincing evidence that the average score varies across some (or all) groups.

7.41

(a) H_0: Average GPA is the same for all majors: $\mu_{\text{AH}} = \mu_{\text{NS}} = \mu_{\text{SS}}$ H_A: At least one pair of means are different.

(b) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average GPAs across three groups of majors.

(c) The total degrees of freedom is $195 + 2 = 197$, so the sample size is $197 + 1 = 198$.

7.42

(a) H_0: The mean number of hours worked per week is equal for all groups.

$$\mu_{\leq \text{Less than HS}} = \mu_{\text{HS}} = \mu_{\text{Jr Coll}} = \mu_{\text{Bachelor's}} = \mu_{\geq \text{Graduate}}$$

H_A: At least one pair of means are different.

(b) The conditions that need to be satisfied for ANOVA are:

1. Independence: The data are random and the sample makes up less than 10% of all US residents, so independence is satisfied.
2. Approximately normal: The distributions are fairly symmetric. In addition, sample sizes are sufficiently large such that minor deviations from normality will not affect the reliability of the test.
3. Constant variance: Variability seems fairly constant across all groups except for Bachelor's.

In order to proceed with the test we will need to assume independence and constant variance across groups.

(c)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	5-1=4	501.54×4=2,006.16	501.54	501.54/229.12=2.189	0.0682
Residuals	1,171-4=1,167	267,382	267,382/1,167=229.12		
Total	1,172-1=1,171	2,006.16+267,382=269,388.16			

(d) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence that the average number of hours worked per week is different across educational attainment levels.

7.43

(a) False. As the number of groups increases, so does the number of comparisons and hence the modified significance level decreases.

(b) True.

(c) True.

(d) False. We need observations to be independent regardless of sample size.

7.44

(a) H_0: Average child care hours is the same for all majors: $\mu_{\text{C}} = \mu_{\text{LMS}} = \mu_{\text{PS}} = \mu_{\text{TV}} = \mu_{\text{UMS}}$
H_A: At least one pair of means are different.

(b) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average number of hours spent on child care across educational attainment levels.

7.45

(a) The hypotheses are:

H_0: Average score difference is the same for all treatments.

H_A: At least one pair of means are different.

(b) We should check conditions. If we look back to the earlier exercise, we will see that the patients were randomized, so independence is satisfied. There are some minor concerns about skew, especially with the third group, though this may be acceptable. The standard deviations across the groups are reasonably similar.

Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average reduction in score among treatments.

(c) We determined that at least two means are different in part (b), so we now conduct $K = 3 \times 2/2 = 3$ pairwise t-tests that each use $\alpha = 0.05/3 = 0.0167$ for a significance level. Use the following hypotheses for each pairwise test.

H_0: The two means are equal.

H_A: The two means are different.

The sample sizes are equal and we will use the pooled SD, so we can compute the standard error for comparing two groups, each with a sample size of 14, as follows:

$$SE = \sqrt{\frac{9.793^2}{14} + \frac{9.793^2}{14}} = 3.70$$

It would also be acceptable to not use the pooled standard deviation and instead compute the SE for each pair of groups based on the group standard deviations.

Next we would compute the Z-scores for each, and using the pooled $df = 39$, find the p-values. For example, comparing Trmt 1 to Trmt 2:

$$T = \frac{(6.21 - 2.86) - 0}{3.7} = 0.91$$

This has a one-tail area (using a t-distribution with $df = 39$) of 0.184, so the p-value is 0.368.

The p-value for Trmt 1 vs. Trmt 3 is the only one under 0.05: p-value = 0.024 (or 0.035 if not using s_{pooled} in place of s_1 and s_3, though this won't affect the final conclusion). The p-value is larger than $0.05/3 = 1.67$, so we do not have strong evidence to conclude that it is this particular pair of groups that are different. That is, we cannot identify if which particular pair of groups are actually different, even though we've rejected the notion that they are all the same!

7.46

(a) False, we conclude that at least one pair of means are different.

(b) True.

(c) False, it is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons.

(d) False, the Bonferroni correction requires dividing the original α by the number of pairwise tests to be conducted, not the number of groups. With 4 groups to start with, $k = 4$, the number of pairwise tests will be $K = \frac{k(k-1)}{2} = \frac{4 \times 3}{2} = 6$, and hence the new significance level will be $\alpha^\star = \frac{\alpha}{K} = \frac{0.05}{6} = 0.0083$.

7.47 The hypotheses are $H_0 : \mu_T = \mu_C$ and $H_A : \mu_T \neq \mu_C$.

We are told to assume that conditions for inference are satisfied.

$$\begin{aligned} T &= \frac{(\bar{x}_T - \bar{x}_C) - (\mu_T - \mu_C)}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}} \\ &= \frac{(57.1 - 27.1) - 0}{\sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}} = \frac{30}{11.14} = 2.69 \\ df &= min(n_1 - 1, n_2 - 1) = min(22 - 1, 22 - 1) = 21 \\ p - value &= P(|T_{21}| > 2.69) \rightarrow 0.01 < p - value < 0.02 \end{aligned}$$

Since p-value < 0.05, we reject H_0. The data provide convincing evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

7.48 The hypotheses are $H_0 : \mu_T = \mu_C$ and $H_A : \mu_T \neq \mu_C$.

We are told to assume that conditions for inference are satisfied.

$$\begin{aligned} T &= \frac{(\bar{x}_T - \bar{x}_C) - (\mu_T - \mu_C)}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}} \\ &= \frac{(4.9 - 6.1) - 0}{\sqrt{\frac{1.8^2}{22} + \frac{1.8^2}{22}}} = \frac{-1.2}{0.543} = -2.21 \\ df &= min(n_1 - 1, n_2 - 1) = min(22 - 1, 22 - 1) = 21 \\ p - value &= P(|T_{21}| > 2.21) \rightarrow 0.02 < p - value < 0.05 \end{aligned}$$

Since p-value < 0.05, we reject H_0. The data provide convincing evidence that the average number of food items recalled by the patients in the treatment and control groups are different.

7.49 False. While it is true that paired analysis requires equal sample sizes, only having the equal sample sizes isn't, on its own, sufficient for doing a paired test. Paired tests require that there be a special correspondence between each pair of observations in the two groups.

7.50

(a) Based on this sample, the point estimate for the average number of credits taken per semester by students at this college is 13.65. The point estimate for the median is 14.

(b) Based on this sample, the point estimate for the standard deviation of the number of credits taken per semester by students at this college is 1.91. The point estimate for the IQR is 15 - 13 = 2.

(c) In order to determine if a load of 16 or 18 credits is unusually high, we need to calculate how many standard deviations away from the mean this observation is, i.e. calculate the Z-score.

$$Z = \frac{16 - 13.65}{1.91} = 1.23 \qquad Z = \frac{18 - 13.65}{1.91} = 2.28$$

16 credits is not unusually high but 18 credits is since it is more than two standard deviations away from the mean.

(d) No, sample point estimates only approximate the population parameter, and they vary from one sample to another. Therefore we cannot expect to get the same mean with each random sample.

(e) We use the standard error of the mean to measure the variability in means of random samples of same size taken from a population. The variability in the means of random samples is quantified by the standard error. Based on this sample, $SE_{\bar{x}} = \frac{1.91}{\sqrt{100}} = 0.191$.

7.51

(a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution.

(b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric.

(c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$.

(d) The sample means will be more variable with the smaller sample size.

7.52 The data are paired, since this is a before-after measurement of the same trees, so we will construct a confidence interval using the differences summary statistics. But before we proceed with a confidence interval, we must first check conditions:
Independent: this is satisfied since the trees were randomly sampled.
Normality: since $n = 50 \geq 30$, we only need consider whether there are any particularly extreme outliers. None are mentioned, and it doesn't seem like we'd expect to observe any such cases from data of this type, so we'll consider this condition to be satisfied.
With the conditions satisfied, we can proceed with calculations. First, compute the standard error and degrees of freedom: $SE = \frac{7.2}{\sqrt{50}} = 1.02$, $df = 50 - 1 = 49$.
Next, we find $t^\star = 2.68$ for a 99% confidence interval using a t-distribution with 49 degrees of freedom, and then we construct the confidence interval:

$$12.5 \pm 2.68 \times 1.02 \quad \rightarrow \quad (9.77, 15.23)$$

We are 99% confident that the average growth of young trees in this area during the 10-year period was 9.77 to 15.23 feet.

7.53

(a) We should set 1.0% equal to 2.84 standard errors: $2.84 \times SE_{desired} = 1.0\%$ (see the last example in the section on power for details). This means the standard error should be about $SE = 0.35\%$ to achieve the desired statistical power.

(b) The margin of error was $0.5 \times (2.6\% - (-0.2\%)) = 1.4\%$, so the standard error in the experiment must have been $1.96 \times SE_{original} = 1.4\% \rightarrow SE_{original} = 0.71\%$.

(c) The standard error decreases with the square root of the sample size, so we should increase the sample size by a factor of $2.03^2 = 4.12$.

(d) The team should run an experiment 4.12 times larger, so they should have a random sample of 4.12% of their users in each of the experiment arms in the new experiment.

7.54

(a) We need to check independence, approximately normal, and constant variance. If the bolts were randomly assigned (not explicitly stated), then independence is satisfied; we'll suppose this is the case. We don't see any major outliers, so we'll consider it reasonable to assume the sample means are approximately normally distributed. We also don't see any clear changes in variation across the groups, so the constant variance condition also seems reasonable.

(b) The hypotheses are

H_0: The torque required does not, on average, differ among the different treatments.

H_A: Some methods make it easier or more difficult to loosen a bolt (requiring less torque), on average.

The p-value of 0.0056 is smaller than 0.05, so we would reject H_0. That is, there is evidence that at least some of the methods perform better or worse than others at reducing the torque required to loosen rusty bolts.

(c) We'll use a Bonferroni correction with $\alpha = 0.05$, so we're looking for any p-values less than $0.05/28 = 0.00179$. Only one difference is statistically significant when using a Bonferroni correction: none vs Liquid Wrench. If we look at the plot of the data, we can also see the sample mean for Liquid Wrench was less than that of "none". That is, since we've inferred there's a difference and the sample means show Liquid Wrench required less torque on average, we can conclude that Liquid Wrench did in fact help reduce the torque required vs doing nothing. That said, we cannot confidently say Liquid Wrench was any better than any of the other methods.

7.55 The assumptions that need to be satisfied in order to proceed with the confidence interval are as follows:

1. Independence: The sample is from less than 10% of all undergraduates, and it is a random sample. We can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported)
2. Normality: The data are strongly skewed, but the sample is relatively large.

If the assumption about the sample being reasonably equivalent to a simple random sample is acceptable, then the confidence interval can be calculated as follows:

$$\begin{aligned} \bar{x} \pm Z^\star SE_{\bar{x}} &= 3.2 \pm 1.65 \times \frac{1.97}{\sqrt{203}} \\ &= 3.2 \pm 0.23 \\ &= (2.97, 3.43) \end{aligned}$$

We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

7.56 The conditions that need to be satisfied in order to proceed with the confidence interval are as follows:

1. Independence: The sample is random and 5,534 women make up less than 10% of women in the US, so the age at first marriage of all of the women are independent.
2. Normality: The data are only moderately skewed.

The confidence interval can be calculated as follows:

$$\begin{aligned}\bar{x} \pm Z^\star SE_{\bar{x}} &= 23.44 \pm 1.96 \times \frac{4.72}{\sqrt{5534}} \\ &= 23.44 \pm 0.12 \\ &= (23.32, 23.56)\end{aligned}$$

We are 95% confident that the average age at first marriage for women is between 23.32 years old and 23.56 years old.

7.57 First, the hypotheses should be about the population mean (μ), not the sample mean. Second, the null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. The correct way to set up these hypotheses is shown below:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu \neq 10 \; hours$$

A two-sided test allows us to consider the possibility that the data show us something that we would find surprising.

7.58 First, the hypotheses should be about the population mean (μ), not the sample mean. Second, the = and $\neq$ signs are flipped; she should be skeptical about there being a change. The correct way to set up these hypotheses is shown below:

$$H_0 : \mu = 23.44 \; years \; old$$
$$H_A : \mu \neq 23.44 \; years \; old$$

Chapter 8

Introduction to linear regression

8.1

(a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant.

(b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

8.2

(a) There is a fan shape in the residual plot. Variability around the regression line increases as x increases. Since there is a trend in the residual plot, a linear model would using the methods we have described would not be appropriate for these data.

(b) There is an apparent curvature in the residual plot. A linear model would not be appropriate for these data.

8.3

(a) Strong relationship, but a straight line would not fit the data.
(b) Strong relationship, and a linear fit would be reasonable.
(c) Weak relationship, and trying a linear fit would be reasonable.
(d) Moderate relationship, but a straight line would not fit the data.
(e) Strong relationship, and a linear fit would be reasonable.
(f) Weak relationship, and trying a linear fit would be reasonable.

8.4

(a) Strong relationship, but a straight line would not fit the data.
(b) Strong relationship, but a straight line would not fit the data.
(c) Strong relationship, and a linear fit would be reasonable.
(d) Weak relationship, and trying a linear fit would be reasonable.
(e) Weak relationship, and trying a linear fit would be reasonable.
(f) Moderate relationship, and a linear fit would be reasonable.

8.5

(a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear.

(b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary.

8.6

(a) The association between husbands' and wives' ages is positive, weak, and linear. There are a few potential outliers: a man in his early- to-mid 30s marrying a woman in her mid 40s, a man in his early 40s marrying a woman in her early-to-mid 50s, and a man in his early 50s marrying a woman in her early 30s.

(b) The association between husbands' and wives' heights is weak but positive.

(c) The age plot, where there is a more evident linear trend in the data.

(d) No. Correlation is unaffected by such rescaling.

8.7

(a) $r = -0.7 \rightarrow (4)$.
(b) $r = 0.45 \rightarrow (3)$.
(c) $r = 0.06 \rightarrow (1)$.
(d) $r = 0.92 \rightarrow (2)$.

8.8

(a) $r = 0.49 \rightarrow (2)$
(b) $r = -0.48 \rightarrow (4)$
(c) $r = -0.03 \rightarrow (3)$
(d) $r = -0.85 \rightarrow (1)$

8.9

(a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot.

(b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion.

(c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

8.10 Their correlation coefficients will be the same, since the correlation coefficient is unitless.

8.11

(a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of.

(b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours.

(c) Changing units doesn't affect correlation: $r = 0.636$.

8.12

(a) The relationship is moderate, negative, and linear. There is also an outlying month when the average temperature is about 53°F and average crawling age of about 28.5 weeks.

(b) Changing the units will not change the form, direction or strength of the relationship between the two variables. After all, if higher temperatures measured in °F are associated with lower average crawling age measured in weeks, higher temperatures measured in °C will be associated with lower average crawling age measured in months.

(c) Since changing units doesn't affect correlation, $R = -0.70$.

8.13

(a) There is a moderate, positive, and linear relationship between shoulder girth and height.

(b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

8.14

(a) The relationship is strong and positive. However there appears to be some departure from linearity as the scatterplot shows somewhat of a fan shape. There is less variability in weights for people with lower hip girth measurements than for people with larger hip girth measurements. One explanation for the fan shape might be that the data are composed of two groups of people (males and females) who are likely to have a different relationship between their weights and hip girths.

(b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

8.15 In each part, we can write the husband ages as a linear function of the wife ages.

(a) $age_H = age_W + 3$.

(b) $age_H = age_W - 2$.

(c) $age_H = 2 \times age_W$.

Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the husband ages for each wife in each part and create a scatterplot.

8.16 In each part, we may write the pay for men as a linear function of the women's pay:

(a) $pay_M = pay_W + 5000$.

(b) $pay_M = 1.25 \times pay_W$.

(c) $pay_M = 0.85 \times pay_W$.

Therefore, the correlation will be exactly 1 in all three parts. Alternatively, we can create a mock data set, such as a data set of 3 women with salaries \$15,000, \$25,000, and \$35,000. Then, based on the regression equation from part (a), men who have similar positions will be predicted to make \$20,000, \$30,000, and \$40,000. Plotting these against each other we will see that the data line up on a straight line, yielding a correlation of 1. The same approach can be applied to the other pars as well.

8.17 Correlation: no units. Intercept: kg. Slope: kg/cm.

8.18 I is higher, the more the scatter the lower the correlation coefficient, and hence the higher the uncertainty around the regression line.

8.19 Over-estimate. Since the residual is calculated as *observed* − *predicted*, a negative residual means that the predicted value is higher than the observed value.

8.20 Under-estimate. Since the residual is calculated as *observed* − *predicted*, a positive residual means that the predicted value is lower than the observed value.

8.21

(a) There is a positive, very strong, linear association between the number of tourists and spending.

(b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars).

(c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism.

(d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

8.22

(a) There is a positive, moderate, linear association between number of calories and amount of carbohydrates. In addition, the amount of carbohydrates is more variable for menu items with higher calories, indicating non-constant variance. There also appear to be two clusters of data: a patch of about a dozen observations in the lower left and a larger patch on the right side.

(b) Explanatory: number of calories. Response: amount of carbohydrates (in grams).

(c) With a regression line, we can predict the amount of carbohydrates for a given number of calories. This may be useful if only calorie counts for the food items are posted but the amount of carbohydrates in each food item is not readily available.

(d) Even though the relationship appears linear in the scatterplot, the constant variability assumption is violated. We should not fit a least squares line to these data.

8.23

(a) First calculate the slope.

$$b_1 = \frac{s_y}{s_x} R = \frac{113}{99} \times 0.636 = 0.726$$

Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$.

$$\begin{aligned} \bar{y} &= b_0 + b_1 \times \bar{x} \\ 129 &= b_0 + 0.726 \times 107 \\ b_0 &= 129 - 0.726 \times 107 = 51 \end{aligned}$$

The regression line can be written as

$$\widehat{travel\ time} = 51 + 0.726 \times distance$$

(b) b_1 = For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time.
b_0 = When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.

(c) $R^2 = 0.636^2 = 0.40$.
Approximately 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled.

(d) In order to calculate the predicted time it takes to travel from Santa Barbara to Los Angeles, we plug in 103 for distance in the model:

$$\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126\ minutes$$

(e) The residual can be calculated as $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time.

(f) No, we should not use this linear model to predict the travel time for a distance of 500 miles as this would be extrapolation. The data we used to create the model is for approximately 10 to 350 miles of distance. The linear model may no longer hold outside the range of the data.

8.24

(a) First calculate the slope.

$$b_1 = \frac{s_y}{s_x} R = \frac{9.41}{10.37} \times 0.666 = 0.604$$

Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$.

$$\begin{aligned} \bar{y} &= b_0 + b_1 \bar{x}) \\ 171.14 &= b_0 + 0.604 \times 107.20 \\ b_0 &= 171.14 - 0.604 \times 107.20 = 106.39 \end{aligned}$$

The regression line can be written as

$$\widehat{height} = 106.39 + 0.604 \times \textit{shoulder girth}$$

(b) b_1 = For each centimeter increase in shoulder girth, we would expect height to increase on average by 0.604 centimeters.
b_0 = People who have a shoulder girth of 0 cm are expected to be on average 106.39 cm tall. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.

(c) $R^2 = 0.666^2 = 0.44$.
Approximately 44% of the variation in heights is accounted for by the model, i.e. explained by shoulder girth.

(d) In order to calculate height from shoulder girth, plug in 100 for shoulder girth in the model:

$$\widehat{height} = 106.39 + 0.604 \times \textit{shoulder\_girth} = 106.39 + 0.604 \times 100 \approx 167 \textit{ cm}$$

(e) The residual can be calculated as $e_i = y_i - \hat{y}_i = 160 - 167 = -7$ cm. A negative residual means that the model overestimated this student's height.

(f) No. Predicting the height of a child would require extrapolation. The data used to create the model is for people with approximately 90 to 130 cm shoulder girth. The linear model may no longer hold outside the range of the data.

8.25

(a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$.

(b) Expected murder rate in metropolitan areas with no poverty is -29.901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line.

(c) For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559.

(d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas.

(e) $\sqrt{0.7052} = 0.8398$.

8.26

(a) $\widehat{heart\ wt} = -0.357 + 4.034 \times body\ wt$

(b) Expected heart weight of cats with 0 body weight is -0.357. This is obviously not a meaningful value, it just serves to adjust the height of the regression line.

(c) For each additional kilogram in cats' weights, we expect their hearts to be heavier by 4.034 grams, on average.

(d) Body weight explains 64.66% of the variability in weights of cats' hearts.

(e) $\sqrt{0.6466} = 0.8041$

8.27

(a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.

(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.

(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

8.28

(a) The outlier is in the upper-left corner. Since it is horizontally far from the center of the data, it is an influential point. Additionally, since the fit of the regression line is greatly influenced by this point, it is a point with high leverage.

(b) The outlier is located in the lower-left corner. It is horizontally far from the rest of the data, so it is a high-leverage point. The regression line also would fall relatively far from this point if the fit excluded this point, meaning it the outlier is influential.

(c) The outlier is in the upper-middle of the plot. Since it is near the horizontal center of the data, it is not a high-leverage point. This means it also will have little or no influence on the slope of the regression line.

8.29

(a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot.

(b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

8.30 The outlier is located in the bottom center of the plot. The point does not have high leverage since it is near the center of the data on the horizontal axis. Since it is not a point of high leverage, it is also not an influential point.

8.31

(a) The relationship appears to be moderate-to-strong, positive, and linear. There are a few outliers but no points that appear to be influential.

(b) $\widehat{weight} = -105.0113 + 1.0176 \times height$
Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms.
Intercept: People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself.

(c) The hypotheses are as follows:

H_0: The true slope coefficient of height is zero ($\beta_1 = 0$).
H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$).

The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. With such a small p-value, we reject H_0 and conclude that the data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0.

(d) $R^2 = 0.72^2 = 0.52$
Approximately 52% of the variability in weight can be explained by the height of individuals.

8.32

(a) The relationship appears to be strong, positive and linear. There is one potential outlier, the student who had 9 cans of beer.

(b) $\widehat{BAC} = -0.0127 + 0.0180 \times beers$
Slope: For each additional can of beer consumed, the model predicts an additional 0.0180 grams per deciliter BAC.
Intercept: Students who don't have any beer are expected to have a blood alcohol content of -0.0127.

(c) The hypotheses are as follows:

H_0: The true slope coefficient of number of beers is zero ($\beta_1 = 0$).
H_A: The true slope coefficient of number of beers is different than zero ($\beta_1 \neq 0$).

The p-value is approximately 0. (Note that this output doesn't mean the p-value is exactly zero, only that when rounded to four decimal places it is zero.) With such a small p-value and since the data showed a positive relationship, we reject H_0 and conclude that the data provide convincing evidence that number of cans of beer consumed and blood alcohol content are positively correlated and the true slope parameter is greater than 0.

(d) $R^2 = 0.89^2 = 0.79$
Approximately 79% of the variability in blood alcohol content can be explained by number of cans of beer consumed.

(e) It would probably be weaker. This study had people of very similar ages, and they also had identical drinks. In bars and elsewhere, drinks vary widely in the amount of alcohol they contain.

8.33

(a) The hypotheses are as follows:

H_0: The true slope coefficient of the husband's height is zero ($\beta_1 = 0$).

H_A: The true slope coefficient of the husband's height is greater than zero ($\beta_1 > 0$).

The test-statistic is 4.17 (with df = 170 - 2 = 168), and p-value for a two-sided alternative hypothesis is approximately 0. The p-value for a one- tailed alternative will also be very low. With such a low p-value, we reject H_0 and conclude the data provide convincing evidence that wives' and husbands' heights are positively correlated and the true slope parameter is indeed greater than 0.

(b) The equation of the regression line is

$$\widehat{height}_W = 43.5755 + 0.2863 \times height_H.$$

(c) Slope: For each additional inch in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average.
Intercept: Men who are 0 inches tall are expected to have wives who are on average 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line.

(d) The slope is positive, so R must also be positive. $R = \sqrt{0.09} = 0.30$.

(e) Using the equation of the regression line:

$$\widehat{height}_W = 43.5755 + 0.2863 \times 69 = 63.2612.$$

Since R^2 is low, the prediction based on this regression model is not very reliable.

(f) No, we shouldn't use the same model to predict the height of the wife of a man who is 79 inches tall. The scatterplot shows that husbands in this data set are approximately 60 to 75 inches tall. The regression model may not be reasonable outside of this range.

8.34

(a) The regression of percent homeownership and percent of the population living in an urban area has a negative slope, so the correlation coefficient will be negative as well. Since $R^2 = 0.28$, $\sqrt{0.28} = 0.53$, $R = -0.53$.

(b) There is a fan shaped pattern apparent in this plot, which indicates non-constant variability in the residuals (little variability when x is small, more variability when x is large). Since the residuals have changing variability, we should seek more appropriate statistical methods if we want to obtain a reliable estimate of the best fitting straight line.

8.35

(a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

(b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate.

(c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million.
(d) Yes, we rejected H_0 and the confidence interval does not include 0.

8.36

(a) The predicted head circumference for a baby whose gestational age is 28 weeks can be calculated as follows:

$$\widehat{head\ circumference} = 3.91 + 0.78 \times 28 = 25.75$$

(b) The hypotheses are as follows:

H_0: The true slope coefficient of gestational age is zero ($\beta_1 = 0$).

H_A: The true slope coefficient of gestational age is different than zero ($\beta_1 \neq 0$).

The test statistic can be calculated as:

$$T = \frac{estimate - null\ value}{SE} = \frac{0.78 - 0}{0.35} = 2.23$$

with $df = n - 2 = 25 - 2 = 23$. Using the t-table, the two-tailed p-value is between 0.02 and 0.05, which is lower than the significance level of 0.05. With such a low p-value, we reject H_0 and conclude that there is an association between gestational age and head circumference and that the true slope parameter is not 0.

8.37

(a) True.

(b) False, correlation is a measure of the linear association between any two numerical variables.

8.38

(a) There is a weak-to-moderate, positive, linear association between height and volume. There also appears to be some non-constant variance since the volume of trees is more variable for taller trees.

(b) There is a very strong, positive association between diameter and volume. The relationship may include slight curvature.

(c) Since the relationship is stronger between volume and diameter, using diameter would be preferred. However, as mentioned in part (b), the relationship between volume and diameter may not be, and so we may benefit from a model that properly accounts for nonlinearity.

8.39

(a) The point estimate and standard error are $b_1 = 0.9112$ and $SE = 0.0259$. We can compute a T-score: $T = (0.9112 - 1)/0.0259 = -3.43$. Using $df = 168$, the p-value is about 0.001, which is less than $\alpha = 0.05$. That is, the data provide strong evidence that the average difference between husbands' and wives' ages has actually changed over time.

(b) The equation of the regression line is

$$\widehat{age}_W = 1.5740 + 0.9112 \times age_H.$$

(c) Slope: For each additional year in husband's age, the model predicts an additional 0.9112 years in wife's age. This means that wives' ages tend to be lower for later ages, suggesting the average gap of husband and wife age is larger for older people.
Intercept: Men who are 0 years old are expected to have wives who are on average 1.5740 years old. The intercept here is meaningless and serves only to adjust the height of the line.

(d) The regression of wives' ages on husbands' ages has a positive slope, so the correlation coefficient will be positive. Since the $R^2 = 0.88$, $R = \sqrt{0.88} = 0.94$.

(e) Using the equation of the regression line:

$$\widehat{age}_W = 1.5740 + 0.9112 \times 55 = 51.69.$$

Since R^2 is pretty high, the prediction based on this regression model is reliable.

(f) No, we shouldn't use the same model to predict a 85 year old man's wife's age. This would require extrapolation. The scatterplot from an earlier exercise shows that husbands in this data set are approximately 20 to 65 years old. The regression model may not be reasonable outside of this range.

8.40

(a) The hypotheses are as follows:

H_0: The true slope coefficient of body weight is zero ($\beta_1 = 0$).

H_A: The true slope coefficient of body weight is different than zero ($\beta_1 \neq 0$).

(b) The p-value is extremely small (zero to 4 decimal places), which is lower than the significance level of 0.05. The data provide strong evidence that the true slope coefficient of body weight is greater than zero and that body weight is positively associated with heart weight in cats.

(c) The confidence interval can be calculated as follows:

$$\begin{aligned} b_1 \pm t^\star SE &= 4.034 \pm 1.98 \times 0.250 \\ &= (3.539, 4.529) \end{aligned}$$

We are 95% confident that for each additional kilogram in cats' weights, we expect their hearts to be heavier by 3.539 to 4.529 grams, on average.

(d) Yes, we rejected the null hypothesis and the confidence interval lies above 0.

8.41 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

8.42

(a) $r = \sqrt{R^2} = -0.849$

(b) $b_1 = \frac{s_y}{s_x} r = \frac{16.9}{26.7}(-0.849) = -0.537$
$b_0 = \bar{y} - \bar{x} b_1 = 38.8 - 30.8(-0.537) = 55.34$

(c) For a neighborhood with 0% reduced-fee lunch, we would expect 55.34% of the bike riders to wear helmets.

(d) For every additional percentage point of lunch there is a decrease of 0.537 percentage points in helmet.

(e) $e = 40 - \hat{y} = 40 - (40 \times (-0.537) + 55.34) = 6.14$
There are 6.14% more bike riders wearing helmets than predicted by the regression model in this neighborhood.

8.43

(a) $r = -0.72 \rightarrow (2)$
(b) $r = 0.07 \rightarrow (4)$
(c) $r = 0.86 \rightarrow (1)$
(d) $r = 0.99 \rightarrow (3)$

8.44

(a) The slope can be calculated as follows:

$$\begin{aligned} \bar{y} &= b_0 + b_1 \bar{x} \\ 3.9983 &= 4.010 + b_1 \times -0.0883 \\ -0.0117 &= b_1 \times -0.0883 \\ b_1 &= 0.133 \end{aligned}$$

(b) $H_0 : \beta_{beauty} = 0$
$H_A : \beta_{beauty} > 0$
With a T-score of 4.13 and a p-value of approximately 0, we reject H_0. The data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive.

(c) 1. Nearly normal residuals: The distribution of residuals look unimodal but somewhat left skewed, this assumption may not be satisfied.
2. Constant variance of residuals: The residuals vs. beauty score plot shows residuals that are randomly distributed around 0.
3. Independent observations/residuals: We are told the sample is random, so the observations/residuals are independent.
4. Linear relationship: Beauty score seems to be linearly associated with teaching evaluation score, or at least, the relationship doesn't appear non-linear.

Chapter 9

Multiple and logistic regression

9.1

(a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$.

(b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than those who are born to non-smoking mothers.
Smoker: $\widehat{baby\_weight} = 123.05 - 8.94 \times 1 = 114.11$ ounces.
Non-smoker: $\widehat{baby\_weight} = 123.05 - 8.94 \times 0 = 123.05$ ounces.

(c) H_0: The true slope coefficient for the variable `smoke` is zero ($\beta_1 = 0$).
H_A: The true slope coefficient for the variable `smoke` is not zero ($\beta_1 \neq 0$).
The question is asking whether any relationship exists between smoking and birth weight, which requires a two-sided alternative. $T = -8.65$, and the p-value is approximately 0. Since p-value is very small, we reject H_0. The data provide convincing evidence that the true slope parameter is different than 0 and that the linear relationship between birth weight and smoking is real.

9.2

(a) $\widehat{bwt} = 120.07 - 1.93 \times parity$.

(b) The estimated body weight of first borns is 1.93 ounces lower than others.
Firstborn: $\widehat{bwt} = 120.07 - 1.93 \times 1 = 118.14$ *ounces*.
Not firstborn: $\widehat{bwt} = 120.07 - 1.93 \times 0 = 120.07$ *ounces*.

(c) H_0: The true coefficient for `parity` is zero ($\beta_1 = 0$).
H_A: The true coefficient for `parity` is not zero ($\beta_1 \neq 0$).
The question is asking whether any relationship exists between parity and birth weight, therefore we use a two-sided alternative. $T = -1.62$, and the p-value is approximately 0.1052. If using a 5% significance level, since p-value > 0.05, we fail to reject H_0. The data do not provide convincing evidence that the true slope parameter is different than 0, and hence there does not appear to be a statistically significant relationship between birth weight and parity.

9.3

(a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$.

(b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day in length of pregnancy, all else held constant.
β_{age}: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant.

(c) Parity might be correlated with one of the other variables in the model, which introduces collinearity and complicates model estimation.

(d) $\widehat{baby\_weight} = -80.41 + 0.44 \times 284 - 3.33 \times 0 - 0.01 \times 27 + 1.15 \times 62 + 0.05 \times 100 - 8.40 \times 0 = 120.58$
$e = baby\_weight - \widehat{baby\_weight} = 120 - 120.58 = -0.58$

(e) The R^2 and the adjusted R^2 can be calculated as follows:

$$R^2 = 1 - \frac{Var(e_i)}{Var(y_i)} = 1 - \frac{249.28}{332.57} = 0.2504$$

$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-p-1)}{Var(y_i)/(n-1)} = 1 - \frac{249.28/(1236-6-1)}{332.57/(1236-1)} = 0.2468$$

9.4

(a) $\widehat{days} = 18.93 - 9.11 \times eth + 3.10 \times sex + 2.15 \times lrn$.

(b) b_1: The estimated absence days of students who are not aboriginal is 9.11 lower than the aboriginal students.
b_2: The estimated absence days of male students is 3.10 days higher than females.
b_3: The estimated absence days of students who are slow learners is 2.15 days higher than students who are average learners.

(c) $\widehat{days} = 18.93 - 9.11 \times 0 + 3.10 \times 1 + 2.15 \times 1 = 24.18$
$e = days - \widehat{days} = 2 - 24.18 = -22.18$

(d) The R^2 and the adjusted R^2 can be calculated as follows:

$$R^2 = 1 - \frac{Var(e_i)}{Var(y_i)} = 1 - \frac{240.57}{264.17} = 0.0893$$

$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-p-1)}{Var(y_i)/(n-1)} = 1 - \frac{240.57/(146-3-1)}{264.17/(146-1)} = 0.0701$$

9.5

(a) A 95% confidence interval for the slope of gender can be calculated as follows:

$$\begin{aligned} df &= n - p - 1 = 55 - 4 - 1 = 50 \\ t^\star_{50} &= 2.01 \\ b_{gender} \pm t^\star_{df} SE_{gender} &= -0.08 \pm 2.01 \times 0.12 \\ &= -0.08 \pm 0.24 \\ &= (-0.32, 0.16) \end{aligned}$$

We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model.

(b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

9.6

(a) A 95% confidence interval for the slope of height can be calculated as follows:

$$\begin{aligned} df &= n - p - 1 = 31 - 2 - 1 = 28 \\ t^\star_{28} &= 2.05 \\ b_{height} \pm t^\star_{df} SE_{height} &= 0.34 \pm 2.05 \times 0.13 \\ &= 0.34 \pm 0.27 \\ &= (0.07, 0.61) \end{aligned}$$

We are 95% confident that for each foot increase in the height of a tree the volume is expected to increase on average by 0.083 to 0.617 cubic feet when controlling for the other variables in the model.

(b) $\hat{y} = -57.99 + 0.34 \times 79 + 4.71 \times 11.3 = 22.093 < 24.2$.
The model underestimates the volume of the tree by $24.2 - 22.093 = 2.107$ cubic feet.

9.7 Remove age.

9.8 Remove learner status.

9.9 Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted R^2 for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward- selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted R^2.)

9.10 Based on both the p-value and R^2_{adj} ethnicity should be added to the model first.

9.11 She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

9.12 She should use adjusted R^2 selection since she is interested in optimizing predictions.

9.13

1. Nearly normal residuals: With so many observations in the data set, we look for particularly extreme outliers in the histogram and do not see any.
2. Constant variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
3. Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
4. Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.

All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

9.14 The residuals are right skewed (skewed to the high end). Horror movies seem to show a much different pattern than the other genres. While the residuals plots show a random scatter over years and in order of data collection, there is a clear pattern in residuals for various genres, which signals that this regression model is not appropriate for these data.

9.15

(a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large.

(b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

9.16

(a) More damaged O-rings are observed for lower temperatures, with fewer damaged O-rings seen for higher temperatures. The lowest-temperature shuttle mission had five damaged O-rings, much more than any other shuttle launch, so this observation will be especially influential on the results of an analysis.

(b) We are generally most interested interested in the coefficients of variables, so in this case, the *Temperature* row. The coefficient of this term is negative, indicating that increasing temperatures are associated with a lower probability of O-ring damage. This coefficient was statistically significant with a p-value near 0 (H_0: $\beta_{Temp.} = 0$, H_A: $\beta_{Temp.} \neq 0$), indicating that the data provide strong evidence that the coefficient is less than 0.

(c) $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$, where $\hat{p}$ represents the probability of damage to an O-ring.

(d) Yes. While the data are observational, the relationship between temperature and damage to O-rings is very strong. Since lives are at stake, such a strong association indicates something important is going on that must be carefully investigated. (Ultimately, O-rings were cited as the cause for the shuttle disaster, though this investigation required much more investigation than what was completed in this exercise to come to a causal conclusion.)

9.17

(a) The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 33.5095 - 1.4207 \times sex\_male - 0.2787 \times skull\_width + 0.5687 \times total\_length - 1.8057 \times tail\_length$$

Only `total_length` has a positive association with a possum being from Victoria.

(b) The reduced model may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 33.5095 - 1.4207 \times 1 - 0.2787 \times 63 + 0.5687 \times 83 - 1.8057 \times 37 \quad = -5.0781$$

Next, we solve for $\hat{p}$

$$\begin{aligned} e^{\log\left(\frac{\hat{p}}{1-\hat{p}}\right)} &= e^{-5.0781} \\ \left(\frac{\hat{p}}{1-\hat{p}}\right) &= e^{-5.0781} \\ \hat{p} &= e^{-5.0781} \times (1-\hat{p}) \\ \hat{p} &= e^{-5.0781} - e^{-5.0781}\hat{p} \\ (1+e^{-5.0781})\hat{p} &= e^{-5.0781} \\ \hat{p} &= \frac{e^{-5.0781}}{(1+e^{-5.0781})} \\ &= 0.0062 \end{aligned}$$

While the probability is very near zero, we have not run diagnostics on the model. We should also have a little skepticism that the model will hold for a possum found in a US zoo. However, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)

9.18

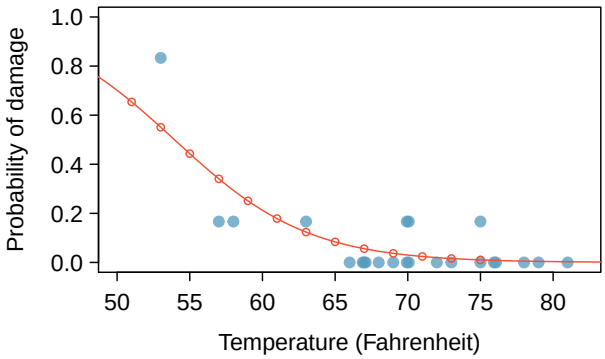
(a) The calculation of the probabilities are as follows:

$$log\left(\frac{\hat{p}_{51}}{1-\hat{p}_{51}}\right) = 11.6630 - 0.2162 \times 51 \longrightarrow \hat{p}_{51} = \frac{e^{11.6630-0.2162\times 51}}{1+e^{11.6630-0.2162\times 51}} = 0.654$$

$$log\left(\frac{\hat{p}_{53}}{1-\hat{p}_{53}}\right) = 11.6630 - 0.2162 \times 53 \longrightarrow \hat{p}_{53} = \frac{e^{11.6630-0.2162\times 53}}{1+e^{11.6630-0.2162\times 53}} = 0.551$$

$$log\left(\frac{\hat{p}_{55}}{1-\hat{p}_{55}}\right) = 11.6630 - 0.2162 \times 55 \longrightarrow \hat{p}_{55} = \frac{e^{11.6630-0.2162\times 55}}{1+e^{11.6630-0.2162\times 55}} = 0.443$$

(b) The plot is shown below:



(c) For logistic regression to be appropriate in this case, each O-ring must be independent of the others. We must assume independence is true (or very nearly so) if we are to believe the model. For example, if the O-ring manufacturing process has changed dramatically over the time of the different shuttle launches, then we should be skeptical of independence and therefore also of the model.

9.19

(a) False. When predictors are collinear, it means they are correlated, and the inclusion of one variable can have a substantial influence on the point estimate (and standard error) of another.

(b) True.

(c) False. This would only be the case if the data was from an experiment and x_1 was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.)

(d) False. We should check normality like we would for inference for a single mean: we look for particularly extreme outliers if $n \geq 30$ or for clear outliers if $n < 30$.

9.20

(a) False. The log-odds will change the same amount, but these changes are not generally proportional to a change in probability.

(b) True.

(c) False. Independent observations is a condition for applying standard logistic regression methods. More advanced methods are needed when observations are not independent.

(d) False. We typically use AIC.

9.21

(a) `exclaim_subj` should be removed, since it's removal reduces AIC the most (and the resulting model has lower AIC than the None Dropped model).

(b) Removing any variable will increase AIC, so we should not remove any variables from this set.

9.22 While the model is not doing a good fit for any genre, it is under-predicting return-on-investment for horror movies a lot more than other genres. This is in line with the FiveThirtyEight article, since it suggests the margins are unusually high for horror movies.

9.23

(a) The equation is:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) = &-0.8124 - 2.6351 \times \texttt{to\_multiple} \\ &+ 1.6272 \times \texttt{winner} - 1.5881 \times \texttt{format} \\ &- 3.0467 \times \texttt{re\_subj}\end{aligned}$$

(b) First find $\log\left(\frac{p}{1-p}\right)$, then solve for p:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= -0.8124 - 2.6351 \times 0 + 1.6272 \times 1 - 1.5881 \times 0 - 3.0467 \times 0 \\ &= 0.8148 \\ \frac{p}{1-p} &= e^{0.8148} \rightarrow p = 0.693\end{aligned}$$

(c) It should probably be pretty high, since it could be very disruptive to the person using the email service if they are missing emails that aren't spam. Even only a 90% chance that a message is spam is probably enough to warrant keeping it in the inbox. Maybe a probability of 99% would be a reasonable cutoff.

As for other ideas to make it even better, it may be worth building a second model that tries to classify the importance of an email message. If we have both the spam model and the importance model, we now have a better way to think about cost-benefit tradeoffs. For instance, perhaps we would be willing to have a lower probability-of-spam threshold for messages we were confident were not important, and perhaps we want an even higher probability threshold (e.g. 99.99%) for emails we are pretty sure are important.