Appendix A

Exercise solutions

1 Introduction to data

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. (b) Control: $2/46 = 0.04 \rightarrow 4\%$. (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.

1.3 (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.

1.5 (a) "Does explicitly telling children not to cheat affect their likelihood to cheat?". (b) 160 children between the ages of 5 and 15. (c) Four variables: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), (4) whether they cheated or not (categorical).

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

1.9 (a) $50 \times 3 = 150$. (b) Four continuous numerical variables: sepal length, sepal width, petal length, and petal width. (c) One categorical variable, species, with three levels: *setosa*, *versicolor*, and *virginica*.

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.

1.13 (a) Population: all births, sample: 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.

1.15 (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 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.

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

1.19 (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

1.21 (a) Positive, non-linear, somewhat strong. 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. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

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.)

1.25 (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.

1.27 (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.

1.29 (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

1.31 (a) Experiment. (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.

1.33 Need randomization and blinding. One possible outline: (1) 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!) (2) 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.)

1.35 (a) Observational study. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive, as the popularity of a name for dogs increases, so does the popularity of that name for cats.

1.37 (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, 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.

1.39 (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children. (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 be in a lower socio- economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

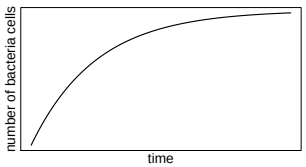
1.41 (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

1.43 (a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched or not. Explanatory: race of the driver.

2 Summarizing data

2.1 (a) 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).

2.3 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.



2.5 (a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

2.7 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.

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.

2.11 (a) 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 = $37.5 + 1.5 \times 20 = 67.5$; Lower fence: Q1 - 1.5 × IQR = $17.5 - 1.5 \times 20 = -12.5$; 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.

2.13 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.

2.15 (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.

2.17 (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.

2.19 (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.

2.21 (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.

2.23 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.

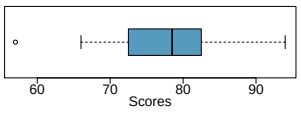
2.25 (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$
(c) The 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.

2.27 (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24+1) = 73.6$. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

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.

2.31 The distribution of ages of best actress winners are right skewed with a median around 30 years. The distribution of ages of best actor 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.

2.33



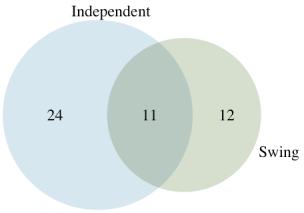
3 Probability

3.1 (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

3.3 (a) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

3.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

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



(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) 0.47. (e) 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.

3.9 (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.

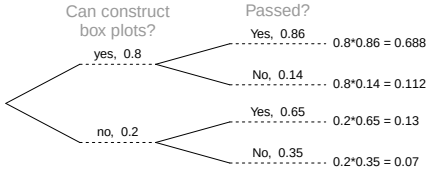
3.11 (a) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

3.13 (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

3.15 (a) No, 0.18 of respondents fall into this combination. (b) $0.60+0.20-0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.

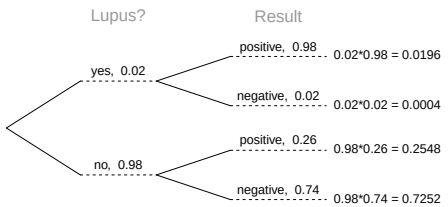
3.17 (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

3.19 (a)



(b) 0.84

3.21 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.



3.23 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

3.25 (a) $2/9 \approx 0.22$. (b) $3/9 \approx 0.33$. (c) $\frac{3}{10} \times \frac{2}{9} \approx 0.067$. (d) No, e.g. in this exercise, removing one chip meaningfully changes the probability of what might be drawn next.

3.27 $P(^1\text{leggings}, {}^2\text{jeans}, {}^3\text{jeans}) = \frac{5}{24} \times \frac{7}{23} \times \frac{6}{22} = 0.0173$. However, the person with leggings could have come 2nd or 3rd, and these each have this same probability, so $3 \times 0.0173 = 0.0519$.

3.29 (a) 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 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

3.31 (a) E(X) = 3.59. SD(X) = 9.64. (b) E(X) = -1.41. SD(X) = 9.64. (c) No, the expected net profit is negative, so on average you expect to lose money.

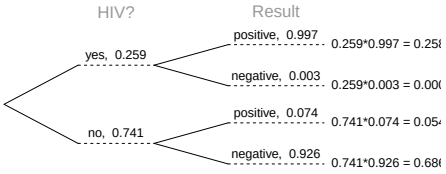
3.33 5% increase in value.

3.35 E = -0.0526. SD = 0.9986.

3.37 Approximate answers are OK.
(a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$. (c) $(26 + 12 + 15)/144 = 0.37$.

3.39 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

3.41 0.8247.



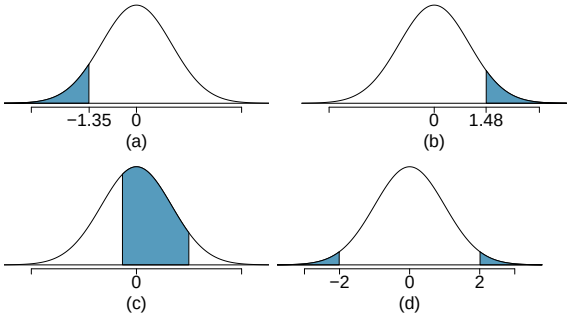
3.43 (a) E = \$3.90. SD = \$0.34.
(b) E = \$27.30. SD = \$0.89.

3.45
$$\begin{aligned} & Var\left(\frac{X_1+X_2}{2}\right) \\ &= Var\left(\frac{X_1}{2} + \frac{X_2}{2}\right) \\ &= \frac{Var(X_1)}{2^2} + \frac{Var(X_2)}{2^2} \\ &= \frac{\sigma^2}{4} + \frac{\sigma^2}{4} \\ &= \sigma^2/2 \end{aligned}$$

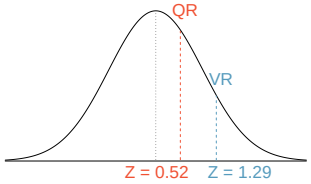
3.47
$$\begin{aligned} & Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= Var\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right) \\ &= \frac{Var(X_1)}{n^2} + \frac{Var(X_2)}{n^2} + \cdots + \frac{Var(X_n)}{n^2} \\ &= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2} \text{ (there are } n \text{ of these terms)} \\ &= n\frac{\sigma^2}{n^2} \\ &= \sigma^2/n \end{aligned}$$

4 Distributions of random variables

4.1 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



4.3 (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.



(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) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (f) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (g) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (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 (d)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

4.5 (a) $Z = 0.84$, which corresponds to approximately 159 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

4.7 (a) $Z = 1.2$, $P(Z > 1.2) = 0.1151$.
(b) $Z = -1.28 \rightarrow 70.6^\circ$F or colder.

4.9 (a) $N(25, 2.78)$. (b) $Z = 1.08$, $P(Z > 1.08) = 0.1401$. (c) The answers are very close because only the units were changed. (The only reason why they differ at all because 28°C is 82.4°F, not precisely 83°F.) (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.

4.11 (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 simplify to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be necessary.

4.13 (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

4.15 If p is the probability of a success, then the mean of a Bernoulli random variable X is given by
$\mu = E[X] = P(X = 0) \times 0 + P(X = 1) \times 1$
$= (1 - p) \times 0 + p \times 1 = 0 + p = p$

4.17 (a) 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) 0.203. (c) 0.203. (d) 0.167. (e) 0.997.

4.19 (a) $\mu = 35$, $\sigma = 3.24$ (b) $Z = \frac{45-35}{3.24} = 3.09$. 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) Using the normal approximation, 0.0010. With 0.5 correction, 0.0017.

4.21 (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

4.23 (a) Geometric distribution: 0.109. (b) Binomial: 0.219. (c) Binomial: 0.137. (d) $1 - 0.875^6 = 0.551$. (e) Geometric: 0.084. (f) Using a binomial distribution with $n = 6$ and $p = 0.75$, we see that $\mu = 4.5$, $\sigma = 1.06$, and $Z = 2.36$. Since this is not within 2 SD, it may be considered unusual.

4.25 (a) $\overset{Anna}{1/5} \times \overset{Ben}{1/4} \times \overset{Carl}{1/3} \times \overset{Damian}{1/2} \times \overset{Eddy}{1/1} = 1/5! = 1/120$. (b) Since the probabilities must add to 1, there must be $5! = 120$ possible orderings. (c) $8! = 40{,}320$.

4.27 (a) Geometric, 0.0804. (b) Binomial, 0.0322. (c) Negative binomial, 0.0193.

4.29 (a) Negative binomial with $n = 4$ and $p = 0.55$, where a success is defined here as a female student. The negative binomial setting is appropriate since the last trial is fixed but the order of the first 3 trials is unknown. (b) 0.1838. (c) $\binom{3}{1} = 3$. (d) In the binomial model there are no restrictions on the outcome of the last trial. 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.

4.31 (a) Poisson with $\lambda = 75$. (b) $\mu = \lambda = 75$, $\sigma = \sqrt{\lambda} = 8.66$. (c) $Z = -1.73$. Since 60 is within 2 standard deviations of the mean, it would not generally be considered unusual. Note that we often use this rule of thumb even when the normal model does not apply. (d) Using Poisson with $\lambda = 75$: 0.0402.

4.33 (a) $\frac{\lambda^k \times e^{-\lambda}}{k!} = \frac{6.5^5 \times e^{-6.5}}{5!} = 0.1454$
(b) The probability will come to $0.0015 + 0.0098 + 0.0318 = 0.0431$ (0.0430 if no 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 one or more other people in their 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. That is, the number of people visiting between 2pm and 3pm would not follow a Poisson distribution.

4.35 0 wins (-\$3): 0.1458. 1 win (-\$1): 0.3936. 2 wins (+\$1): 0.3543. 3 wins (+\$3): 0.1063.

4.37 Want to find the probability that there will be 1,787 or more enrollees. Using the normal approximation, with $\mu = np = 2,500 \times 0.7 = 1750$ and $\sigma = \sqrt{np(1-p)} = \sqrt{2,500 \times 0.7 \times 0.3} \approx 23$, $Z = 1.61$, and $P(Z > 1.61) = 0.0537$. With a 0.5 correction: 0.0559.

4.39 (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

4.41 (a) $(1-0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

4.43 $Z = 1.56$, $P(Z > 1.56) = 0.0594$, i.e. 6%.

4.45 (a) $Z = 0.73$, $P(Z > 0.73) = 0.2327$. (b) If you are bidding on only one auction and set a low maximum bid price, someone will probably outbid you. If you set a high maximum bid price, you may win the auction but pay more than is necessary. If bidding on more than one auction, and you set your maximum bid price very low, you probably won't win any of the auctions. However, if the maximum bid price is even modestly high, you are likely to win multiple auctions. (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 use the 10^{th} percentile: $Z = -1.28 \rightarrow \$69.80$.

4.47 (a) $Z = 3.5$, upper tail is 0.0002. (More precise value: 0.000233, but we'll use 0.0002 for the calculations here.)
(b) $0.0002 \times 2000 = 0.4$. We would expect about 0.4 10 year olds who are 76 inches or taller to show up.
(c) $\binom{2000}{0}(0.0002)^0(1-0.0002)^{2000} = 0.67029$.
(d) $\frac{0.4^0 \times e^{-0.4}}{0!} = \frac{1 \times e^{-0.4}}{1} = 0.67032$.

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, which 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, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

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) Estimate the parameter using the data: $\hat{p} = \frac{27}{212} = 0.127$. (d) *Standard error* (or SE). (e) Compute the SE using $\hat{p} = 0.127$ in place of p: $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$. (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) Recomputed standard error using $p = 0.1$: $SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$. 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!).

5.5 (a) Sampling distribution. (b) If the population proportion is in the 5-30% range, the success-failure condition would be satisfied and the sampling distribution would be symmetric. (c) We use the formula for the standard error: $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$. (d) Standard error. (e) The distribution will tend to be more variable when we have fewer observations per sample.

5.7 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: $45\% \pm 1.96 \times 1.2\% \rightarrow (42.6\%, 47.4\%)$ 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%.

5.9 (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 5.9, 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.

5.11 (a) False. The point estimate is always in the confidence interval, and this is a non-sensical use of a confidence interval with a point estimate (because the point estimate is, by design, listed within the confidence interval). (b) True. (c) False. The confidence interval is not about a sample mean. (d) False. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (e) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (f) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

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 $z^\star = 1.6449$. The confidence interval is $0.085 \pm 1.6449 \times 0.010 \rightarrow (0.0683, 0.1017)$. We are 90% confident that 6.83% to 10.17% of first-time site visitors will register using the new design.

5.15 (a) $H_0 : p = 0.5$ (Neither a majority nor minority of students' grades improved) $H_A : p \neq 0.5$ (Either a majority or a minority of students' grades improved)
(b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.) $H_A : \mu \neq 15$ (The average amount of company time each employee spends not working is different than 15 minutes for March Madness.)

5.17 (1) The hypotheses should be about the population proportion (p), not the sample proportion. (2) The null hypothesis should have an equal sign. (3) The alternative hypothesis should have a not-equals sign, and (4) it should reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is: $H_0 : p = 0.6$ and $H_A : p \neq 0.6$.

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.

5.21 (i) Set up hypotheses. H_0: $p = 0.5$, H_A: $p \neq 0.5$. We will use a significance level of $\alpha = 0.05$. (ii) Check conditions: simple random sample gets us independence, and the success-failure conditions is satisfied since $0.5 \times 1000 = 500$ for each group is at least 10. (iii) Next, we 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. (iv) Make a conclusion: 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 a different question than if it will help the economy.)

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 $Z = 1.96$. Standard error: $SE = \sqrt{0.3(1-0.3)/90} = 0.048$. Finally, set up the test statistic formula and solve for $\hat{p}$: $1.96 = \frac{\hat{p}-0.3}{0.048} \rightarrow \hat{p} = 0.394$ Alternatively, if $Z = -1.96$ was used: $\hat{p} = 0.206$.

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.

5.27 (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.

5.29 (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) Strong evidence. 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.

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

5.33 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.

5.35 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).

5.37 (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:

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

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.

6 Inference for categorical data

6.1 (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

6.3 (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

6.5 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% ± 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

6.7 With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

6.9 (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true proportion. (d) Yes. The interval lies entirely above 50%.

6.11 (a) We want to check for a majority (or minority), so we use the following hypotheses:

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

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
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).
Therefore, 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 $2 \times 0.0062 = 0.0124$.
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.

6.13 (a) $H_0 : p = 0.5$. $H_A : p \neq 0.5$. Independence (random sample) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ the one tail area is 0.0018, so the p-value is 0.0036. Since the p-value < 0.05, 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.

6.15 Because 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 $1.6449 \times SE = 1.6449 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.6449 \times \sqrt{\frac{0.55(1-0.55)}{n}} \leq 0.01$$
$$1.6449^2 \frac{0.55(1-0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6697.

6.17 This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers. That is, independence may not hold. Additionally, there are only 5 interventions under the provocative scenario, so the success-failure condition does not hold. Even if we consider a hypothesis test where we pool the proportions, the success-failure condition will not be satisfied. Since one condition is questionable and the other is not satisfied, the difference in sample proportions will not follow a nearly normal distribution.

6.19 (a) False. The entire confidence interval is above 0. (b) True. (c) True. (d) True. (e) False. It is simply the negated and reordered values: (-0.06,-0.02).

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.

6.23 (a) College grads: 23.7%. Non-college grads: 33.7%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college graduates who responded "do not know". $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 235/827 = 0.284$), is also satisfied. $Z = -3.18 \rightarrow$ p-value $= 0.0014$. 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. The data also indicate that fewer college grads say they "do not know" than non-college grads (i.e. the data indicate the direction after we reject H_0).

6.25 (a) College grads: 35.2%. Non-college grads: 33.9%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college grads who support offshore drilling. $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 286/827 = 0.346$), is also satisfied. $Z = 0.39 \rightarrow$ p-value $= 0.6966$. 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.

6.27 Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ p-value $= 0.0989$. 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 are different for non-transportation workers and truck drivers.

6.29 (a) Summary of the study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) $H_0 : p_N = p_L$. 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) 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. The success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value $= 0.0039$. 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.

6.31 (a) False. The chi-square distribution has one parameter called 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.

6.33 (a) 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) 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. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, 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.

6.35 (a) Two-way table:

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total) \times (col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total) \times (col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

6.37 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.

$$E_{row\ 1,col\ 1} = 151.5 \qquad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \qquad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \qquad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow$ p-value $= 0.003$. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for offshore drilling and having a college degree.

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.

6.41 (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

6.43 (a) Independence is satisfied (random sample), as is the success-failure condition (40 smokers, 160 non-smokers). The 95% CI: (0.145, 0.255). We are 95% confident that 14.5% to 25.5% of all students at this university smoke. (b) We want $z^\star SE$ to be no larger than 0.02 for a 95% confidence level. We use $z^\star = 1.96$ and plug in the point estimate $\hat{p} = 0.2$ within the SE formula: $1.96\sqrt{0.2(1-0.2)/n} \leq 0.02$. The sample size n should be at least 1,537.

6.45 (a) Proportion of graduates from this university who found a job within one year of graduating. $\hat{p} = 348/400 = 0.87$. (b) This is a random sample, so the observations are independent. Success-failure condition is satisfied: 348 successes, 52 failures, both well above 10. (c) (0.8371, 0.9029). 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 such 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) (0.8267, 0.9133). Similar interpretation as before. (f) 99% CI is wider, as we are more confident that the true proportion is within the interval and so need to cover a wider range.

6.47 Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: (1/3) * 99 = 33 for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow$ 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.

6.49 (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample) and the success-failure condition are satisfied. $Z = -20.5 \rightarrow$ p-value ≈ 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 used their cell phones as a primary access point to 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) (0.1545, 0.1855). We are 95% confident that approximately 15.5% to 18.6% of all Americans primarily use their cell phones to browse the internet.

7 Inference for numerical data

7.1 (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

7.3 (a) 0.085, do not reject H_0. (b) 0.003, reject H_0. (c) 0.438, do not reject H_0. (d) 0.042, reject H_0.

7.5 The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

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) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) 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 strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) No, since the p-value is smaller than $1 - 0.90 = 0.10$.

7.9 T is either -2.09 or 2.09. Then $\bar{x}$ is one of the following:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

7.11 (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.

7.13 If the sample is large, then the margin of error will be about $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

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.

7.17 (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.

7.19 (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.

7.21 (a) $SE = 1.23$ and $t^\star = 1.65$. $2.9 \pm 1.65 \times 1.23 \rightarrow (0.87, 4.93)$.
(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.

7.23 (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) $T = 4.93$ for $df = 10 - 1 = 9 \rightarrow$ 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.

7.25 (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value = 0.042. Since p-value < 0.05, reject H_0. The data provide strong 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) (-6.49, -0.17).
(c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13^{th} has a higher chance of harm than on any other night.

7.27 (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed.
(b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$.
We leave the conditions to you to consider.
$T = 3.02$, $df = min(11, 9) = 9 \rightarrow$ p-value = 0.014. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chickens that were fed linseed and horsebean.
(c) Type 1 Error, since we rejected H_0.
(d) Yes, since p-value > 0.01, we would not have rejected H_0.

7.29 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ p-value = 0.007. 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 (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

7.31 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} \neq 0$: Treatment has an effect on P.D.T. scores, either positive or negative. Conditions: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied. 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. For all three groups: $df = 13$. $T_1 = 1.89 \rightarrow$ p-value = 0.081, $T_2 = 1.35 \rightarrow$ p-value = 0.200), $T_3 = -1.40 \rightarrow$ (p-value = 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.

7.33 Difference we care about: 40. Single tail of 90%: $1.28 \times SE$. Rejection region bounds: $\pm 1.96 \times SE$ (if 5% significance level). Setting $3.24 \times SE = 40$, subbing in $SE = \sqrt{\frac{94^2}{n} + \frac{94^2}{n}}$, and solving for the sample size n gives 116 plots of land for each fertilizer.

7.35 Alternative.

7.37 H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$. H_A: The average weight varies across some (or all) groups. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable. Approx. normal: the distributions of weights within each feed type appear to be fairly symmetric. Constant variance: Based on the side-by-side box plots, the constant variance assumption appears to be reasonable. There are differences in the actual computed standard deviations, but these might be due to chance as these are quite small samples. $F_{5,65} = 15.36$ 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.

7.39 (a) H_0: The population mean of MET for each group is equal to the others. H_A: At least one pair of means is different. (b) 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. In practice, we would inquire for more details. Normality: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable. Constant variance: This condition is sufficiently met, as the standard deviations are reasonably consistent across groups. (c) See below, with the last column omitted:

	Df	Sum Sq	Mean Sq	F value
coffee	4	10508	2627	5.2
Residuals	50734	25564819	504	
Total	50738	25575327		

(d) Since p-value is very small, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

7.41 (a) H_0: Average GPA is the same for all majors. 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 $197 + 1 = 198$.

7.43 (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.

7.45 (a) 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 use the pooled SD, so we can compute $SE = 3.7$ with the pooled $df = 39$. The p-value for Trmt 1 vs. Trmt 3 is the only one under 0.05: p-value = 0.035 (or 0.024 if 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!

7.47 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow$ p-value $= 0.036$. Since p-value < 0.05, reject H_0. The data provide strong 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.

7.49 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.

7.51 (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.

7.53 (a) We should set 1.0% equal to 2.8 standard errors: $2.8 \times SE_{desired} = 1.0\%$ (see Example 7.37 on page 282 for details). This means the standard error should be about $SE = 0.36\%$ 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 $1.97^2 = 3.88$.
(d) The team should run an experiment 3.88 times larger, so they should have a random sample of 3.88% of their users in each of the experiment arms in the new experiment.

7.55 Independence: it is a random sample, so 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). The sample size is at least 30, and there are no particularly extreme outliers, so the normality condition is reasonable. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

7.57 The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu \neq 10 \; hours$$

A two-sided test allows us to consider the possibility that the data show us something that we would find surprising.

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.

8.3 (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.

8.5 (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.

8.7 (a) $r = -0.7 \rightarrow (4)$. (b) $r = 0.45 \rightarrow (3)$. (c) $r = 0.06 \rightarrow (1)$. (d) $r = 0.92 \rightarrow (2)$.

8.9 (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.

8.11 (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: $r = 0.636$.

8.13 (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.

8.15 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.

8.17 Correlation: no units. Intercept: kg. Slope: kg/cm.

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.

8.21 (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.

8.23 (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (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 in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

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) $\sqrt{0.7052} = 0.8398$.

8.27 (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.

8.29 (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.

8.31 (a) The relationship is positive, moderate-to-strong, 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 (about 2.2 pounds).
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) 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. 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.

8.33 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. The p-value, as reported in the table, is incredibly small and is smaller than 0.05, so we reject H_0. The data provide convincing evidence that wives' and husbands' heights are positively correlated.
(b) $\widehat{height}_W = 43.5755 + 0.2863 \times height_H$.
(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) 63.33. Since R^2 is low, the prediction based on this regression model is not very reliable.
(f) No, we should avoid extrapolating.

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.

8.37 (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

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) $\widehat{age}_W = 1.5740 + 0.9112 \times age_H$. (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) $R = \sqrt{0.88} = 0.94$. The regression of wives' ages on husbands' ages has a positive slope, so the correlation coefficient will be positive. (e) $\widehat{age}_W = 1.5740 + 0.9112 \times 55 = 51.69$. 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 an 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.

8.41 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.

8.43 (a) $r = -0.72 \rightarrow$ (2) (b) $r = 0.07 \rightarrow$ (4) (c) $r = 0.86 \rightarrow$ (1) (d) $r = 0.99 \rightarrow$ (3)

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 babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected H_0, we can conclude that smoking is associated with lower birth weights.

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 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 complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

9.5 (a) (-0.32, 0.16). 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).

9.7 Remove age.

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.)

9.11 She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

9.13 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.
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.
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.
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.

9.15 (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.

9.17 (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i - 1.8057 \times tail\_length_i$. Only `total_length` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)

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 $n < 30$.

9.21 (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.

9.23 (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:

$$\begin{aligned}&\log\left(\frac{p}{1-p}\right) \\ &\quad = -0.8124 - 2.6351 \times 0 + 1.6272 \times 1 \\ &\qquad - 1.5881 \times 0 - 3.0467 \times 0 \\ &\quad = 0.8148 \\ &\frac{p}{1-p} = e^{0.8148} \quad \rightarrow \quad p = 0.693\end{aligned}$$

(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.

Appendix B

Data sets within the text

Each data set within the text is described in this appendix, and there is a corresponding page for each of these data sets at **openintro.org/data**. This page also includes additional data sets that can be used for honing your skills. Each data set has its own page with the following information:

- List of the data set's variables.
- CSV download.
- R object file download.

B.1 Introduction to data

1.1 `stent30`, `stent365` → The stent data is split across two data sets, one for days 0-30 results and one for days 0-365 results.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY Times article: www.nytimes.com/2011/09/08/health/research/08stent.html.

1.2 `loan50`, `loans_full_schema` → This data comes from Lending Club (lendingclub.com), which provides a large set of data on the people who received loans through their platform. The data used in the textbook comes from a sample of the loans made in Q1 (Jan, Feb, March) 2018.

1.2 `county`, `county_complete` → These data come from several government sources. For those variables included in the county data set, only the most recent data is reported, as of what was available in late 2018. Data prior to 2011 is all from census.gov, where the specific Quick Facts page providing the data is no longer available. The more recent data comes from USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), and American Community Survey (census.gov/programs-surveys/acs).

1.3 Nurses' Health Study → For more information on this data set, see www.channing.harvard.edu/nhs

1.4 The study we had in mind when discussing the simple randomization (no blocking) study was Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

B.2 Summarizing data

2.1 `loan50`, `county` → These data sets are described in Data Appendix B.1.

2.2 `loan50`, `county` → These data sets are described in Data Appendix B.1.

2.3 `malaria` → Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. www.pnas.org/content/114/10/2711

B.3 Probability

3.1 `loan50`, `county` → These data sets are described in Data Appendix B.1.

3.1 `playing_cards` → Data set describing the 52 cards in a standard deck.

3.2 `family_college` → Simulated data based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

3.2 `smallpox` → Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

3.2 Mammogram screening, probabilities → The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

3.2 Jose campus visits, probabilities → This example was made up.

3.3 No data sets were described in this section.

3.4 Course material purchases and probabilities → This example was made up.

3.4 Auctions for TV and toaster → This example was made up.

3.4 `stocks_18` → Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

3.5 `fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

B.4 Distributions of random variables

4.1 SAT and ACT score distributions → The SAT score data comes from the 2018 distribution, which is provided at
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
The ACT score data is available at
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

4.1 Male heights → The distribution is based on the USDA Food Commodity Intake Database.

4.1 `possum` → The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala)*. Australian Journal of Zoology 43: 449-458.

4.2 Exceeding insurance deductible → These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 Exceeding insurance deductible → These statistics were made up but are possible values one might observe for low-deductible plans.

4.3 Smoking friends → Unfortunately, we don't currently have additional information on the source for the 30% statistic, so don't consider this one as fact since we cannot verify it was from a reputable source.

4.3 US smoking rate → The 15% smoking rate in the US figure is close to the value from the Centers for Disease Control and Prevention website, which reports a value of 14% as of the 2017 estimate:
cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

4.4 Football kicker → This example was made up.

4.4 Heart attack admissions → This example was made up, though the heart attack admissions are realistic for some hospitals.

4.5 `ami_occurrences` → This is a simulated data set but resembles actual AMI data for New York City based on typical AMI incidence rates.

B.5 Foundations for inference

5.1 pew_energy_2018 → The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The `pew_energy_2018` data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

5.2 pew_energy_2018 → See the details for this data set above in the Section 5.1 data section.

5.2 ebola_survey → In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014. Poll ID NY141026 on maristpoll.marist.edu.

5.3 pew_energy_2018 → See the details for this data set above in the Section 5.1 data section.

5.3 Rosling questions → We noted much smaller samples than the Roslings' describe in their book, Factfulness, The samples we describe are similar but not the same as the actual rates. The approximate rates for the correct answers for the two questions for (sometimes different) populations discussed in the book, as reported in Factfulness, are

- 80% of the world's 1 year olds have been vaccinated against some disease: 13% get this correct (17% in the US). gapm.io/q9
- Number of children in the world in 2100: 9% correct. gapm.io/q5

Here are a few more questions and a rough percent of people who get them correct:

- In all low-income countries across the world today, how many girls finish primary school: 20%, 40%, or 60%? Answer: 60%. About 7% of people get this question correct. gapm.io/q1
- What is the life expectancy of the world today: 50 years, 60 years, or 70 years? Answer: 70 years. In the US, about 43% of people get this question correct. gapm.io/q4
- In 1996, tigers, giant pandas, and black rhinos were all listed as endangered. How many of these three species are more critically endangered today: two of them, one of them, none of them? Answer: none of them. About 7% of people get this question correct. gapm.io/q11
- How many people in the world have some access to electricity? 20%, 50%, 80%. Answer: 80%. About 22% of people get this correct. gapm.io/q12

For more information, check out the book, Factfulness.

5.3 pew_energy_2018 → See the details for this data set above in the Section 5.1 data section.

5.3 nuclear_survey → A simple random sample of 1,028 US adults in March 2013 found that 56% of US adults support nuclear arms reduction.
www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

5.3 Car manufacturing → This example was made up.

5.3 stent30, stent365 → These data sets are described in Data Appendix B.1.

B.6 Inference for categorical data

6.1 Payday loans → The statistics come from the following source: pewtrusts.org/-/media/assets/2017/04/payday-loan-customers-want-more-protections-methodology.pdf

6.1 Tire factory → This example was made up.

6.2 `cpr` → Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

6.2 `fish_oil_18` → Manson JE, et al. 2018. *Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer.* NEJMoa1811403.

6.2 `mammogram` → Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

6.2 `drone_blades` → The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the `drone_blades` data set.

6.3 `jury` → The jury data set for examining discrimination is a made-up data set an example. We provide the simulated data in the `jury` data set.

6.3 `sp500_1950_2018` → Data is sourced from finance.yahoo.com.

6.4 `ask` → Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.* opim.wharton.upenn.edu/DPlab/papers/workingPapers/Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

6.4 `diabetes2` → Zeitler P, et al. 2012. *A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.* N Engl J Med.

B.7 Inference for numerical data

7.1 Risso's dolphins → Endo T and Haraguchi K. 2009. *High mercury levels in hair samples from residents of Taiji, a Japanese whaling town.* Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we assumes these 19 dolphins reasonably represent a simple random sample from those dolphins.

7.1 Croaker white fish → fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1 `run17` → www.cherryblossom.org

7.2 `textbooks`, `ucla_textbooks_f18` → Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved: sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

7.3 `stem_cells` → Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012.

7.3 `ncbirths` → Birth records released by North Carolina in 2004. Unfortunately, we don't currently have additional information on the source for this data set.

7.3 Exam versions → This example was made up.

7.4 Blood pressure statistics → The blood pressure standard deviation for patients with blood pressure ranging from 140 to 180 mmHg is guessed and may be a little (but likely not dramatically) imprecise from what we'd observe in actual data.

7.5 `toy_anova` → Data used for Figure 7.19, where this data was made up.

7.5 `mlb_players_18` → Data were retrieved from mlb.mlb.com/stats. Only players with at least 100 at bats were considered during the analysis.

7.5 `classdata` → This example was made up.

B.8 Introduction to linear regression

8.1 `simulated_scatter` → Fake data used for the first three plots. The perfect linear plot uses group 4 data, where `group` variable in the data set (Figure 8.1). The group of 3 imperfect linear plots use groups 1-3 (Figure 8.2). The sinusoidal curve uses group 5 data (Figure 8.3). The group of 3 scatterplots with residual plots use groups 6-8 (Figure 8.8). The correlation plots uses groups 9-19 data (Figures 8.9 and 8.10).

8.1 `possum` → This data set is described in Data Appendix B.4.

8.2 `elmhurst` → These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

8.2 `simulated_scatter` → The plots for things that can go wrong uses groups 20-23 (Figure 8.12).

8.2 `mariokart` → Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

8.3 `simulated_scatter` → The plots for types of outliers uses groups 24-29 (Figure 8.18).

8.4 `midterms_house` → Data was retrieved from Wikipedia.

B.9 Multiple and logistic regression

9.1 `loans_full_schema` → This data set is described in Data Appendix B.1.

9.2 `loans_full_schema` → This data set is described in Data Appendix B.1.

9.3 `loans_full_schema` → This data set is described in Data Appendix B.1.

9.4 `mariokart` → This data set is described in Data Appendix B.8.

9.5 `resume` → Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

We did omit discussion of some structure in the data for the analysis presented: the experiment design included blocking, where typically four resumes were sent to each job: one for each inferred race/sex combination (as inferred based on the first name). We did not worry about this blocking aspect, since accounting for the blocking would *reduce* the standard error without notably changing the point estimates for the `race` and `sex` variables versus the analysis performed in the section. That is, the most interesting conclusions in the study are unaffected even when completing a more sophisticated analysis.

Appendix C

Distribution tables

C.1 Normal Probability Table

A **normal probability table** may be used to find percentiles of a normal distribution using a Z-score, or vice-versa. Such a table lists Z-scores and the corresponding percentiles. An abbreviated probability table is provided in Figure C.1 that we'll use for the examples in this appendix. A full table may be found on page 410.

	Second decimal place of Z									
Z	*0.00*	0.01	0.02	0.03	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure C.1: A section of the normal probability table. The percentile for a normal random variable with $Z = 1.00$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

When using a normal probability table to find a percentile for Z (rounded to two decimals), identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation. For instance, the percentile of $Z = 0.45$ is shown in row 0.4 and column 0.05 in Figure C.1: 0.6736, or the 67.36^{th} percentile.

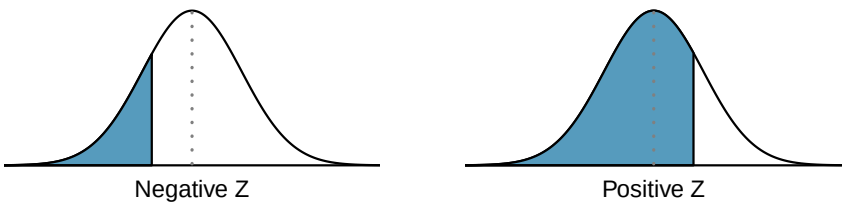
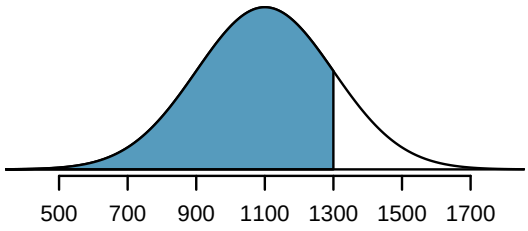


Figure C.2: The area to the left of Z represents the percentile of the observation.

EXAMPLE C.1

SAT scores follow a normal distribution, $N(1100, 200)$. Ann earned a score of 1300 on her SAT with a corresponding Z-score of $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than her. We shade the area representing those individuals in the following graph:



The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in the graph. We find this area by looking in row 1.0 and column 0.00 in the normal probability table: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

EXAMPLE C.2

How do we find an upper tail area?

The normal probability table *always* gives the area to the left. This means that if we want the area to the right, we first find the lower tail and then subtract it from 1. For instance, 84.13% of SAT takers scored below Ann, which means 15.87% of test takers scored higher than Ann.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z-score for the 80^{th} percentile by combining the row and column Z values: 0.84.

EXAMPLE C.3

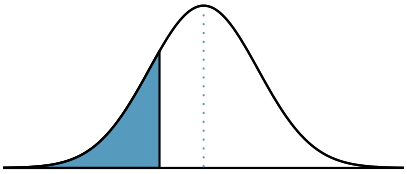
Find the SAT score for the 80^{th} percentile.

We look for the are to the value in the table closest to 0.8000. The closest value is 0.7995, which corresponds to $Z = 0.84$, where 0.8 comes from the row value and 0.04 comes from the column value. Next, we set up the equation for the Z-score and the unknown value x as follows, and then we solve for x:

$$Z = 0.84 = \frac{x - 1100}{200} \quad \rightarrow \quad x = 1268$$

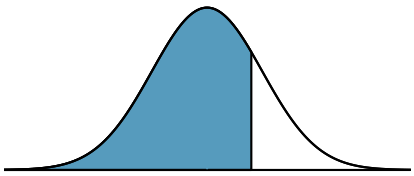
The College Board scales scores to increments of 10, so the 80^{th} percentile is 1270. (Reporting 1268 would have been perfectly okay for our purposes.)

For additional details about working with the normal distribution and the normal probability table, see Section 4.1, which starts on page 133.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.



Positive Z

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

*For $Z \geq 3.50$, the probability is greater than or equal to 0.9998.

C.2 *t*-Probability Table

A ***t*-probability table** may be used to find tail areas of a t-distribution using a T-score, or vice-versa. Such a table lists T-scores and the corresponding percentiles. A partial ***t*-table** is shown in Figure C.3, and the complete table starts on page 414. Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Figure C.3. If we want the value in this row that identifies the T-score (cutoff) for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

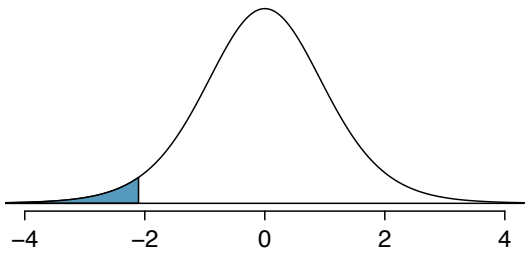
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Figure C.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

EXAMPLE C.4

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture and shade the area below -2.10:



To find this area, we first identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. That is, 2.5% of the distribution falls below -2.10.

In the next example we encounter a case where the exact T-score is not listed in the table.

EXAMPLE C.5

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure C.4. Estimate the proportion of the distribution falling above 1.65.

(E)

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

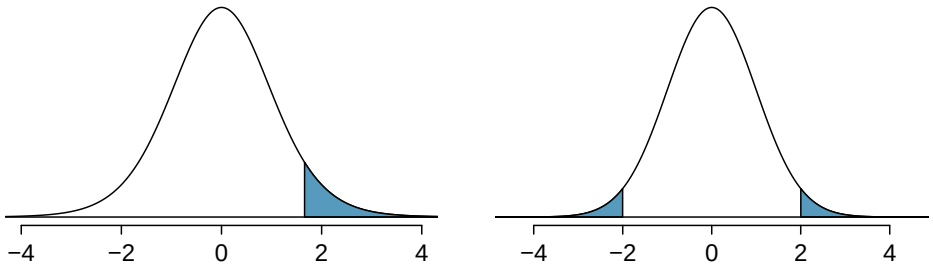


Figure C.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 475 degrees of freedom, with the area further than 2 units from 0 shaded.

EXAMPLE C.6

A t-distribution with 475 degrees of freedom is shown in the right panel of Figure C.4. Estimate the proportion of the distribution falling more than 2 units from the mean (above or below).

(E)

As before, first identify the appropriate row: $df = 475$. This row does not exist! When this happens, we use the next smaller row, which in this case is $df = 400$. Next, find the columns that capture 2.00; because $1.97 < 3 < 2.34$, we use the third and fourth columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.02 and 0.05. We use the two tail values because we are looking for two symmetric tails in the t-distribution.

(G)

GUIDED PRACTICE C.7

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[1]

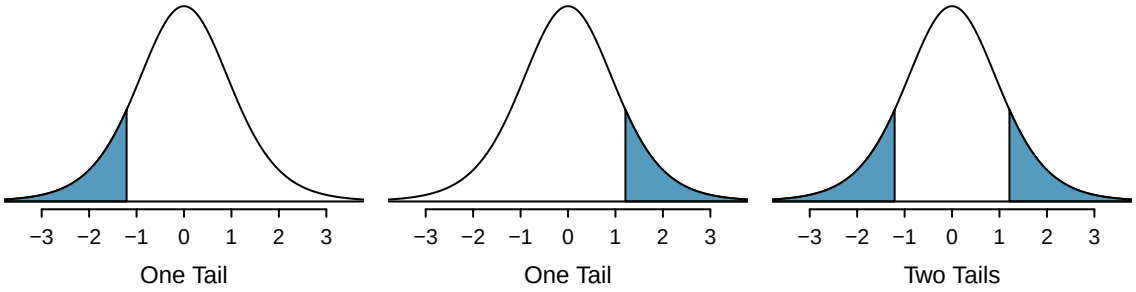
EXAMPLE C.8

Find the value of $t^\star_{18}$ using the t-table, where $t^\star_{18}$ is the cutoff for the t-distribution with 18 degrees of freedom where 95% of the distribution lies between $-t^\star_{18}$ and $+t^\star_{18}$.

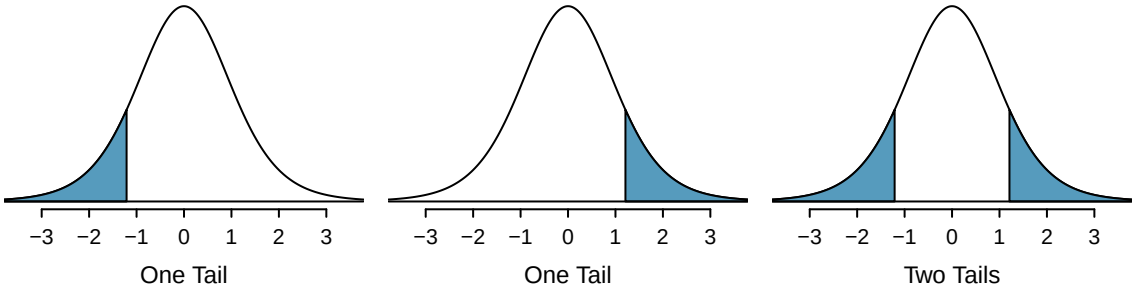
(E)

For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between $-t^\star_{18}$ and $t^\star_{18}$; this is the same as where the two tails have a total area of 0.05. We look in the t-table on page 412, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

[1] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77
	28	1.31	1.70	2.05	2.47	2.76
	29	1.31	1.70	2.05	2.46	2.76
	30	1.31	1.70	2.04	2.46	2.75



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65
	80	1.29	1.66	1.99	2.37	2.64
	90	1.29	1.66	1.99	2.37	2.63
	100	1.29	1.66	1.98	2.36	2.63
	150	1.29	1.66	1.98	2.35	2.61
	200	1.29	1.65	1.97	2.35	2.60
	300	1.28	1.65	1.97	2.34	2.59
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.645	1.96	2.33	2.58

C.3 Chi-Square Probability Table

A **chi-square probability table** may be used to find tail areas of a chi-square distribution. The **chi-square table** is partially shown in Figure C.5, and the complete table may be found on page 417. When using a chi-square table, we examine a particular row for distributions with different degrees of freedom, and we identify a range for the area (e.g. 0.025 to 0.05). Note that the chi-square table provides upper tail values, which is different than the normal and t-distribution tables.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure C.5: A section of the chi-square table. A complete table is in Appendix C.3.

EXAMPLE C.9

Figure C.6(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Figure C.5 to estimate the shaded area.

(E)

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure C.6(a) has area 0.1.

This example was unusual, in that we observed the *exact* value in the table. In the next examples, we encounter situations where we cannot precisely estimate the tail area and must instead provide a range of values.

EXAMPLE C.10

Figure C.6(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The area above value 4.3 has been shaded; find this tail area.

(E)

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure C.6(b) is between 0.1 and 0.2.

EXAMPLE C.11

Figure C.6(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

(E)

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than* 0.3.

EXAMPLE C.12

Figure C.6(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.

(E)

The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.

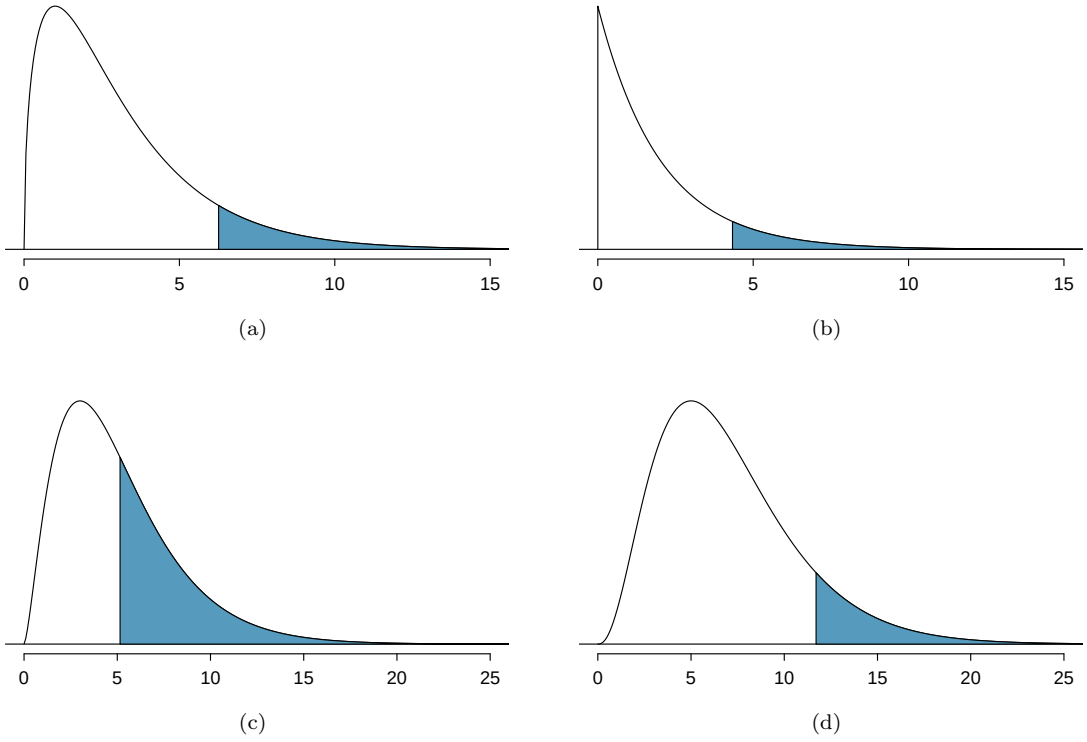


Figure C.6: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

Index