ISCAM 3: CHAPTER 4 EXERCISES

1. 2015 Wimbledon
A tennis fan recorded data on a sample of 16 first-round men's singles matches from the 2015 Wimbledon Championships and also on a sample of 16 first-round women's matches. (The fan did not want to invest the time required to gather and record the data for all matches played in the tournament.) The matches were collected using *systematic sampling* from http://www.wimbledon.com/en_GB/scores/ where every 4th game as selected starting with the 1st, 2nd, 3rd, or 4th match for that sex on day 1 and day 2 of the tournament. Variables recorded include sex, number of sets played, number of games played (not including tie breakers), number of points won by the winner, number of points won by the loser, and length of match in minutes.

(a) Explain how the fan could have used *random sampling* to select the 32 first round matches (16 men's and 16 women's). Give lots of details. Do you think the results from the systematic sampling will be representative of the first round matches? Explain.
(b) Classify each of the recorded variables as categorical or quantitative.

The sorted data for the total number of points played in a match are given here:
Men: 140 159 164 164 166 171 186 187 199 201 220 237 248 318 334 339
Women: 95 102 103 108 111 113 115 118 124 162 165 200 213 238 221 256

(c) Determine (by hand) the five-number summary for each sex's distribution of the number of points played by each sex.
(d) For each sex, determine whether there are any outliers by the 1.5IQR criterion (Investigation 2.2).
(e) Construct a boxplot for each sex's distribution, placing them on the same scale. (Remember to label you axes and include scales.)
(f) Comment on what the numerical and graphical summaries reveal about the distributions of points between the two sexes.
(g) Did all of these men's matches play more points than all of the women's matches? Do men tend to play more points in their matches than women? Explain the difference in these two questions as you justify your answers.

2. 2015 Wimbledon (cont.)
Reconsider the tennis data from the 2015 Wimbledon Championships.

(a) Before turning to technology, make (educated) guesses for the values of the mean and standard deviation of the *number of points* played for each sex. Briefly explain your guesses.
(b) Create the "total number of points" variable (*winner points + loser points*) and use technology (Wimbledon2015.txt) to calculate these means and standard deviations. How were your guesses?
(c) The three largest values correspond to men's matches that took over 50 games (there is no tie break in the fifth set a Wimbledon). Make predictions for the effect that removing these three observations would have on the mean, median, standard deviation, and IQR of the points played by men.
(d) Remove these three observations and re-calculate these statistics. Which statistics were more affected by the removal of the observations? Explain why this makes sense.
(e) Is it reasonable to only remove these three observations because they are the largest? Explain.

3. 2015 Wimbledon (cont.)
Reconsider the 2015 Wimbledon tennis data again (Wimbledon2015.txt). Use technology to analyze the men's and women's distributions of the *sets*, *games*, and *time* variables. For each of these three variables, produce graphical and numerical summaries to compare the distributions between the two sexes, and write a paragraph comparing and contrasting them.

4. 2015 Wimbledon (cont.)
Reconsider the 2015 Wimbledon tennis data yet again (Wimbledon2015.txt).
(a) Use technology to create three new variables:

- Ratio of games to sets
- Ratio of points to games
- Ratio of time to points

Analyze these data to investigate whether men and women differ with regard to the distributions of these variables. For each of these three variables, produce graphical and numerical summaries to compare the distributions between the two sexes, and write a paragraph comparing and contrasting them.
(b) Use technology to create a new variable: *winner points – loser points* and examine the distribution of this point differential separately for males and females. Which sex tended to have smaller differentials? What does this imply? Which sex tended to have more variability in the differentials? What does this imply? How often did the winner score fewer points than the loser? Explain how this can happen in this context.

5. Feeling Motivated?
A psychology study investigated whether people display more creativity when they are thinking about intrinsic or extrinsic motivations. The subjects were 47 people with extensive experience with creative writing. They were randomly assigned to one of two groups: one group answered a survey about intrinsic motivations for writing (such as the pleasure of self-expression) and the other group answered a survey about extrinsic motivations (such as public recognition). Then all subjects were instructed to write a Haiku poem, and these poems were evaluated for creativity by a panel of judges. The researchers conjectured that subjects who were thinking about intrinsic motivations would display more creativity than subjects who were thinking about extrinsic motivations. The creativity scores from this study are in the file creativity.txt.
(a) Identify the explanatory and response variables. Also classify each as categorical or quantitative.
(b) Is this an observational study or a randomized experiment? Explain how you know.
(c) Examine the dotplots of the sample data produced by the Comparing Groups applet. Submit a screen capture of these graphs, and comment on what they reveal about the researchers' conjecture.
(d) Report the mean of the creativity scores for each group. Do these summary values indicate that the intrinsically motivated group did indeed display more creativity than the intrinsically motivated group?
(e) Carry out a randomization test using technology to determine whether the data provide statistically significant evidence that the type of motivation causes affects creativity score in the conjectured direction. Submit a screen capture of the resulting null distribution, and answer four questions:
 i) Describe the null model that underlies this simulation analysis.
 ii) Explain what variable is displayed in the null distribution.
 iii) Describe what the null distribution reveals.
 iv) Report the approximate p-value.
(f) Summarize your conclusion in the context of this study. Include an explanation of the reasoning process behind your conclusion. Be sure to address the issues of causation (i.e., is a cause-and-effect conclusion warranted?) and generalizability (i.e., how broadly can you legitimately generalize your conclusion?), as well as the issue of statistical significance.

6. Feeling Motivated? (cont.)
Reconsider the previous study.
(a) Suppose you thought the intrinsic motivation would, on average, add 10 points to the creativity scores. Specify the corresponding null and (two-sided) alternative hypotheses.

(b) Open the creativity.txt file. Are the data in stacked or unstacked format?
(c) Copy and paste the data into the flash-based Randomization Test applet. This applet lets you specify a hypothesized group 1 effect. Specify 10 as the hypothesized group 1 effect and generate 1000 repetitions. Explain why the resulting null distribution is centered where it is.
(d) Count the samples beyond the observed difference in sample means. Does 10 appear to be a plausible value for the difference in the underlying treatment means? Explain your reasoning.
(e) Use technology to compute a 95% confidence interval comparing the two treatments. Include your output and interpret the interval.
(f) Using the confidence interval, does 10 appear to be a plausible value for the difference in the underlying treatment means? Explain your reasoning.
(g) Use technology to carry out the two-sample *t*-test to obtain a p-value. How does the analysis compare?

7. Bumpus Data
In a famous 1898 lecture described in *The Statistical Sleuth*, a biologist named Bumpus presented data that he analyzed to study the process of natural selection. The data were obtained from adult male house sparrows, some of which had survived a particularly severe winter storm, and others of which had perished. Bumpus investigated whether those that survived had physical characteristics that may have helped them to withstand the storm. Data on the humerus (arm bone) lengths (in thousandths of an inch) follow and appear in Bumpus.txt:

Survived:
687 703 709 715 728 721 729 723 728 723 726 728 736 733 730 733 730 739 735
741 741 749 741 743 741 752 752 751 756 755 766 767 769 770 780

Perished:
659 689 703 702 709 713 720 729 726 726 720 737 739 731 738 736 738 744 745
743 754 752 752 765

(a) Is this an observational study or an experiment? Explain.
(b) Identify and classify the two variables represented in these data.
(c) Produce graphical and numerical summaries for comparing the distributions of humerus lengths between the two groups of sparrows. Write a paragraph addressing Bumpus' question of whether sparrows who survived tended to be physically superior (as measured by humerus length) to those who perished.
(d) Use technology to simulate a randomization test to investigate whether the difference in group means is significant. Use at least 1000 repetitions, and report the approximate p-value. Include your technology output and graphical display of the empirical randomization distribution.
(e) Summarize your conclusion and explain how it follows from your simulation analysis. Also address the issue of whether a cause-and-effect conclusion is warranted, paying attention to the design of the study.

8. Bumpus Data (cont.)
Reconsider the previous question. Bumpus also recorded the weights (in grams) of each sparrow. One hypothesis is that heavier birds are bigger and stronger, therefore more likely to survive the storm. Another hypothesis is that heavier birds are less agile and less mobile, therefore less likely to survive the storm. A third possibility is that there is no association between a bird's weight and its capacity to survive the storm.
(a) Would you consider these data to be from independent random samples or a randomized experiment?
(b) Before you analyze the data, identify which of these three hypotheses you consider the most reasonable (intuitively). Explain briefly.

The data follow and appear in Bumpus.txt:
Survived:
24.5 26.9 26.9 24.3 24.1 26.5 24.6 24.2 23.6 26.2 26.2 24.8 25.4 23.7 25.7 25.7 26.3
26.7 23.9 24.7 28.0 27.9 25.9 25.7 26.6 23.2 25.7 26.3 24.3 26.7 24.9 23.8 25.6 27.0
24.7
Perished:
26.5 26.1 25.6 25.9 25.5 27.6 25.8 24.9 26.0 26.5 26.0 27.1 25.1 26.0 25.6 25.0 24.6
25.0 26.0 28.3 24.6 27.5 31.1 28.3
(c) Analyze these data with graphical and numerical summaries. Write a paragraph summarizing what your analysis reveals relevant to the competing hypotheses described above.
(d) Use technology to simulate a randomization test to investigate whether the difference in group means is statistically significant for the weight data. Use at least 1000 repetitions, and report the approximate p-value. Include your technology output and graphical display of the empirical randomization distribution.
(e) Summarize your conclusion and explain how it follows from your simulation analysis. Also address the issue of whether a cause-and-effect conclusion is warranted, paying attention to the design of the study.

9. Low Carb Diet

A study by Foster el al., reported in *The New England Journal of Medicine* (May, 2003), investigated the effectiveness of a popular "low-carb" diet. The researchers randomly assigned 63 obese men and women to either a low-carbohydrate, high-protein, high-fat (Atkins) diet or a low-calorie, high-carbohydrate, low-fat (conventional) diet. The mean amount of weight lost, as percent of body weight, after 3 months, 6 months and 12 months are shown in the table below.
(The baseline weight was carried forward in the case of missing values.)

Time	Diet	Sample size	Mean	SD
3 months	Low-carb	33	6.8	5.0
	Conventional	30	2.7	3.7
6 months	Low-carb	33	7.0	6.5
	Conventional	30	3.2	5.6
12 months	Low-carb	33	4.4	6.7
	Conventional	30	2.5	6.3

(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory and response variables.
(c) Report the relevant hypotheses (in symbols) for testing whether the mean weight losses differ significantly between the two diets.
(d) Calculate the *t*-test statistic for testing these hypotheses at the 3-month point. Also report the p-value and your test decision at the .05 significance level.
(e) Repeat (d) for comparing the weight losses between the two diets at the 6-month point and again at the 12-month point.
(f) Summarize your conclusions from these three tests. In particular, what do you notice about the trend in the p-value as time passes, and what does that reveal?
(g) Report the 95% confidence intervals for the difference in mean weight loss between the two diets at each time point. Comment on how these confidence intervals change across the three time points.

10. Marriage Ages

A student investigated whether husbands tend to be older than their wives. He gathered data on the ages of a sample of 24 couples, taken from marriage licenses filed in Cumberland County, Pennsylvania, in

June and July of 1993. These data can be accessed in a file MarriageAges.txt.

(a) For each couple, calculate the *difference* in ages (taking the husband's age minus the wife's age). Produce and comment on a dotplot of these differences, keeping in mind the research question of whether husbands tend to be older than their wives.
(b) State the null and alternative hypotheses (in symbols) for testing whether the sample data support the research conjecture that husbands tend to be older than their wives.
(c) Copy/paste the data into the Matched Pairs Randomization applet, and perform 1000 repetitions of the randomization. Submit a copy of the resulting dotplot of sample mean differences. Also use the simulation results to determine an empirical p-value.
(d) Describe what the empirical p-value in (c) represents (it's the probability of what?), and summarize the conclusion that you draw from it.
(e) Investigate and comment on whether the technical conditions of a paired *t*-test appear to be satisfied here.
(f) Calculate the paired *t*-test statistic and p-value. Would you reject the null hypothesis at the .05 significance level?
(g) Produce and interpret a 90% confidence interval for the population mean difference in ages between a husband and wife.
(h) Produce and interpret a 90% prediction interval for the difference in age between a husband and wife.

11. Cool Mice

Medical examiners can use the temperature of a dead body at a murder scene to estimate the time of death. But can a clever murderer disguise the time of death by reheating the victim's body? A scientist actually investigated this issue on mice. Hart (1951) used 19 mice as the experimental units. He sacrificed each mouse and then measured the cooling constant of its body. Then he reheated the mouse's body and measured its cooling constant in that reheated state. The results are in CoolMice.txt.

(a) Explain why these data call for a matched pairs analysis.
(b) Produce and comment on relevant graphical displays and numerical summaries for investigating the question of whether cooling constants for reheated mice are similar to those of freshly killed mice.
(c) Conduct a paired *t*-test or use the Matched Pairs Randomization applet to determine whether the data suggest a significant difference in average cooling constants between freshly killed and reheated mice. If you use the *t*-test, make sure comment on whether you believe the test procedure is valid and how you are decided.
(d) Construct and interpret a 95% confidence interval for estimating the population mean difference in cooling constants.
(e) Summarize the conclusions you would draw from this study. Make sure you comment on significance, confidence, generalizability, and causation.

12. Memorizing Letters

Students in a statistics course at Cal Poly were given 20 seconds to memorize as many letters as possible in a sequence of 30 letters. The letters and the sequence were exactly the same for all students, but the presentation of the letters differed. Twenty-seven students were randomly assigned to see letters arranged in recognizable three-letter chunks such as JFK-CIA-FBI and so on. For the other 26 students, the letters were in less recognizable chunks such as JFKC-IAF and so on. Students' "scores" were determined as the number of letters they memorized correctly in the sequence before their first mistake.

(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory and response variable. Identify each as categorical or quantitative.
(c) Which group would you expect to memorize more letters in general?

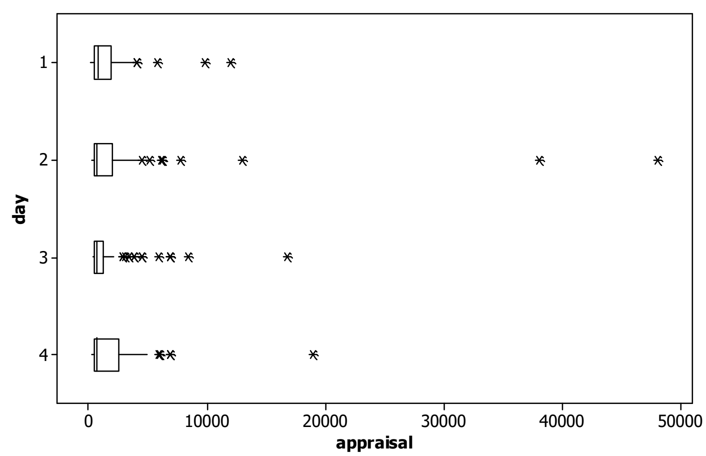
The resulting numbers of letters memorized successfully (MemoryLetters.txt) were:

JFK: 6, 6, 6, 8, 9, 9, 9, 9, 12, 15, 15, 15, 15, 18, 18, 18, 19, 21, 21, 21, 21, 21, 21, 21, 24, 27, 27
JFKC: 2, 3, 3, 3, 5, 6, 6, 6, 6, 8, 9, 9, 10, 13, 14, 14, 14, 14, 14, 15, 15, 15, 17, 18, 20, 24

(d) What proportion of the 27 scores in the JFK group are multiples of three? What about in the JFKC group of 26 scores? Explain why it makes sense that so many scores in the JFK group are multiples of three. (This aspect of a distribution, where the data are clustered at certain values, is called *granularity*.)
(e) Construct visual displays to compare the distributions of letters memorized correctly between the two groups. Report the five-number summary, as well as the mean and standard deviation, for each group. Write a paragraph comparing and contrasting the distributions. (Remember to comment on center, spread, shape, and outliers.)

13. Surfboard Lengths
A student collected data on surfers over several weeks at a local beach (Wood, 2004). The data are in the file surfer.txt. Two of the questions of interest are how the age distributions of men and women surfers compare, and how the lengths of surfboards used by men and women compare.
(a) Identify the observational units in this study.
(b) Classify each of these variables (age, sex, surfboard length) as categorical or quantitative.
(c) Produce graphical displays and numerical summaries to address the question of how the age distributions of men and women surfers compare. Write a paragraph summarizing your findings. Include well-labeled output as appropriate.
(d) Produce graphical displays and numerical summaries to address the question of how the surfboard length distributions of men and women surfers compare. Write a paragraph summarizing your findings. Include well-labeled output as appropriate.
(e) In exploration the data, are there any observations you would suggest removing from the dataset? Justify your answer.

14. Appraisal Prices
The following boxplots are the appraisal prices of pieces of art auctioned off over a four-day period in December of 2004:



(a) Comment on what these four distributions have in common.
(b) Would you expect the mean appraisal price to be larger than, smaller than, or close to the median appraisal price on these days? Explain.
(c) Day 2 has the smallest median appraisal price among these four days, but it has the largest mean. Explain, based on the boxplots, why this makes sense.

15. Appraisal Prices (cont.)
The auction data from the previous exercise appear in auction.txt, where the variables are day, appraisal price, starting price at the auction, and selling price at the auction.
(a) Create a new variable: ratio of starting price to appraisal price. How many and what proportion of the art pieces had a starting price of more than half their appraisal price? How many and what proportion of the art pieces had a starting price less than one-third their appraisal price?
(b) Produce graphical displays and numerical summaries to analyze the distribution of this "ratio" variable. Write a paragraph reporting your findings.
(c) Now compare the distribution of these ratios across the four days of the auction. Do the distributions appear to differ considerably across the days? Write a paragraph reporting your findings.

16. Roller Coaster Speeds (cont.)
Reconsider the data from the Chapter 2 exercises on 139 coasters in the United States, as downloaded from the www.rcdb.com site in November of 2003.
(a) Open the data file coasters.txt, which contains data on 145 roller coasters in the United States, as downloaded from the www.rcdb.com site in November of 2003. Use technology to produce boxplots of height (in feet) by type, length (in feet) by type, and drop (in feet) by type. Write a paragraph summarizing differences between wooden and steel coasters with regard to these variables.
(b) Another variable in the file is *age group* (column 13) which is coded as "1:older" for coasters opened in 1990 or earlier, coded as "2:middle" for coasters opened between 1991 and 1998 inclusive, and coded as "3:newer" for coasters opened in 1999 or later. Produce boxplots of height, length, and drop by this *age group* variable. Write a paragraph summarizing how roller coasters appear to have changed over time with respect to these variables.

17. Seeding Clouds (cont.)
Reconsider the cloud seeding data (CloudSeeding.txt). At the end of Investigation 4.7, you applied the log transformation to the rainfall amounts.
(a) Use technology to take the square root of the rainfall amounts. Produce graphical and numerical summaries for comparing the two groups on this transformed variable. Comment on what your analysis reveals.
(b) Repeat (a) for the reciprocal transformation.
(c) Which of the three transformations that you have tried thus far (log, square root, reciprocal) does the best job of making the distributions more symmetric? Justify your choice.

18. Memorizing Letters (cont.)
Reconsider the data from the memory experiment from Exercise 12 (MemoryLetters.txt)
(a) Use technology to simulate a randomization test to investigate whether the difference in group means is significant. Use at least 1000 repetitions, and report the approximate p-value. Include your technology output and graphical display of the empirical randomization distribution.
(b) Summarize your conclusion and explain how it follows from your simulation analysis. Also address the issue of whether a cause-and-effect conclusion is warranted, paying attention to the design of the study.
(c) Repeat this analysis on the group *medians*, and comment on whether your conclusion differs substantially.

19. Sleeping Student (cont.)
Reconsider the students' sleeping times from Chapters 0 and2 (SleepStudents.txt).
(a) Choose one of these three students, and conduct a simulation analysis to approximate a randomization test for comparing her school night *sleeping times* to her non-school night *sleeping times*. Submit a well-labeled histogram of your simulation results.
(b) Report the approximate p-value based on your simulation results. Does your analysis suggest that the difference in their mean *sleeping times* between school nights and non-school nights is unlikely to have occurred by chance?
(c) Is this a study for which the randomization in your simulation mirrors that in the design, or is the randomization hypothetical in this study? Explain.

20. Musical Dining
A study by North and Shilcock involved three weeks monitoring the effects of classical, pop music, and background silence on customers' spending in British restaurants. Each type of music was played for 6 nights (the order was randomly determined to guard against confounding). When classical music was played in the background, 120 diners spent an average of £24.13 per person on food and drinks. When pop music was played, the 142 diners spent an average of £21.92.
(a) What additional information from the two samples would you need in order to decide if the difference in spending between the classical and pop music was statistically significant?
(b) Sketch comparative boxplots for hypothetical spending distributions between these two groups, creating a situation where you think the difference would be statistically significant. Explain the reasoning behind your sketch.
(c) Sketch comparative boxplots for hypothetical spending distributions between these two groups, creating a situation where you think the difference would *not* be statistically significant. Explain the reasoning behind your sketch.

21. Mirrors and Exercises
In a study reported in the journal *Health Psychology* (Ginis, Jung, and Gauvin, 2003), researchers investigated whether the presence or absence of mirrors during an exercise session would affect women's attitudes toward the session. The subjects were 58 sedentary women, who rode a stationary exercise bike for a 20-minute session. A week later the women returned for another 20-minute session, for which they were randomly assigned to exercise in front of either a mirrored or curtained wall. The first table in the research article describes characteristics of the sample, including the variables of age, body mass index, smoking status, and student status. Some of the statistics reported include:

	Mirror (n_m = 28)			**Curtain (n_c = 30)**		
Variable	**Proportion**	**Mean**	**Std. Dev.**	**Proportion**	**Mean**	**Std. Dev.**
Smoking	0.071			0.067		
Student	0.786			0.734		
Age		20.86	1.65		20.60	1.57
Body mass index		23.35	3.76		24.23	6.19

(a) Classify each of these four variables as categorical or quantitative.
(b) For each of these four variables, conduct a test of whether the two groups differ significantly on that variable. Report all of the test statistics and p-values. [*Hint*: Check technical conditions as much as possible for all four tests. For testing the proportions, if the technical conditions of the *z*-test are not satisfied, apply Fisher's Exact Test.]
(c) Why do you think the researchers collected and examined these data, performed these tests, and presented the results in the article?
(d) Do you think the researchers were pleased that none of these differences turned out to be statistically significant? Explain why.

22. Fish Oil

Researchers randomly assigned 14 male volunteers with high blood pressure to one of two diets for four weeks: a fish oil diet and regular oil diet. The subjects' diastolic blood pressure was measured at the beginning and end of the study, and the reduction was recorded for each subject (taken from Ramsey and Schafer (2002) based on a study by Knapp and Fitzgerald (1989)). Prior to conducting the study, researchers conjectured that those on the fish oil diet would tend to experience greater reductions in blood pressure than those on the regular oil diet. The resulting reductions in diastolic blood pressure, in millimeters of mercury were

Fish oil diet	8	12	10	14	2	0	0
Regular oil diet	–6	0	1	2	–3	–4	2

(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory variable and the response variable. Classify each as categorical or quantitative.
(c) State the hypotheses, in symbols and in words, for testing the researchers' conjecture about this study.
(d) Carry out a randomization test to determine whether the difference in group means is statistical significant at the 0.05 level.
(e) Is it appropriate and valid to carry out a pooled two-sample *t*-test here? Explain.
(f) Conduct a pooled two-sample *t*-test (whether you think it's valid to do so or not). Report the test statistic and p-value. How does the p-value from this pooled *t*-test compare to the p-value from the randomization test in (d)? Would you say that the pooled *t*-test provides a reasonably close approximation to the randomization test in this case? Explain.
(g) Use the pooled *t*-procedure to construct a 95% confidence interval for the treatment effect of the fish oil diet compared to the regular oil diet.

23. Fish Oil (cont.)

Reconsider the previous question about the fish oil study.
(a) Conduct a (non-pooled) two-sample *t*-test and confidence interval. Comment on how the results differ from those of the pooled test. Does the pooling appear to make much difference in this case?
(b) Explain why it would definitely not be appropriate to conduct a paired *t*-test on these data.

24. Fish Oil (cont.)

Reconsider the fish oil study again. Comment on how the p-value from the pooled *t*-test would change in the following situations. Provide an intuitive explanation for your reasoning in each case. Also provide an algebraic explanation based on the test statistic calculation in each case.
(a) What if the group means had been closer together (and everything else had been the same)?
(b) What if the group means had been further apart (and everything else had been the same)?
(c) What if the sample sizes had been larger (and everything else had been the same)?
(d) What if there had been more variability in each sample (and everything else had been the same)?

25. Fish Oil (cont.)

Reconsider the previous question. Comment on how the *width of a confidence interval* for the treatment effect would change in each of the four situations (group means closer together, group means further apart, large sample sizes, more variability in each sample). Again provide both an intuitive and an algebraic explanation for your reasoning in each case.

26. Backpack Weights (cont.)
Reconsider the backpack data from the Chapter 1 Exercises (backpack.txt). Analyze the data to examine whether the data suggest that male and female students differ significantly with regard to any of three variables: body weight, backpack weight, and ratio of backpack weight to body weight. Include both descriptive (graphical and numerical) and inferential (significance test and confidence interval) components to your analyses. For each variable, write a paragraph or two summarizing your findings.

27. Used Hondas
The UsedHondas.txt file contains data on a sample of 22 used Honda Civics and 23 used Honda Accords for sale on the web (Kelly Blue Book, kbb.com) August 17, 2015. Cars that were labeled as "new listings" were not included.
(a) Produce graphical displays and numerical summaries to compare the distributions of prices between the two models of cars. Comment on what this descriptive analysis reveals. (Like always, comment on shape, center, spread, and unusual observations.)
(b) Consider these for now to be random samples from the populations of all used Civics and all used Accords for sale on the web in August 2015. Conduct a two-sample *t*-test of the conjecture that Accords tend to cost more on average than Civics. Report your findings. (Include all components of the test in your report.)
(c) Estimate the difference in population means between the two car models with 90%, 95%, and 99% confidence intervals.
(d) Based on these intervals, how confident would you feel about concluding that used Accords cost more than more on average than used Civics? What about concluding that used Accords cost more than $850 more on average than used Civics? Explain.

28. Ideal Age
Social scientists have noted that American culture celebrates youth, and they have studied what American consider to be the ideal age. The Harris Poll asked a nationwide sample of 2306 adults on September 16-23, 2003 the following question: "If you could stop time and live forever in good health at a particular age, what age would you like to live at?" The mean response from men was 39 years, and the mean response from women was 43 years.
(a) Consider testing whether this difference in mean responses is statistically significant. What further information would you need to conduct a two-sample *t*-test?
(b) Suppose that the sample sizes were roughly the same for men and for women, so roughly 1153 in each group. With those sample sizes, does the distribution of "ideal ages" need to be normal in order for the *t*-procedures to be valid?
(c) Suppose that in each group, the standard deviation of the "ideal age" responses is 10 years. Sketch the sampling distribution of the test statistic and determine the observed test statistic and *p*-value of the two-sample *t*-test. Is the difference in mean responses significant at the 0.01 level?
(d) Repeat (c) if the standard deviation of the "ideal age" responses is 20 years for each group.
(e) How large would the standard deviation need to be in order for the sample results not to be statistically significant at the .01 level?
(f) Does this (your answer to part e) seem like a reasonable value for the standard deviation in this case? Explain.

29. Health Club Ages

A student collected data on ages of people who joined a local health club in August and September of 2004, also recording the sex of each person (Schmitt, 2004). The student took a systematic sample of every 5th male from a computerized list of males who joined each month and then again for females. The student wanted to test whether the ages of males and females differ significantly and whether the ages of new members in the two months differ significantly. The data are in the file GymMembership.txt.

(a) Start with the question of whether men's and women's ages differ significantly on average. Analyze the data to address this issue. Include both descriptive (graphical and numerical) and inferential (significance test and confidence interval) aspects to your analysis. Include all components (including a check of technical conditions), and summarize your findings.
(b) Repeat (a) for the question of whether mean ages of new members differed significantly between August and September.
(c) To what populations would you feel comfortable generalizing your findings? Explain.

30. Melting Chips

A study was carried out to see whether there is a difference in the melting times of semisweet chocolate chips and peanut butter chips. Twenty students in a statistics class were told to put a chip on their tongue, touch it to the roof of their mouth, and then time how long it was before the chip was completely melted, without any "encouragement" on their part. Each student repeated this with both types of chips, randomly determining which chip they would use first. The data are in the file ChipMelting.txt.

(a) Is this an observational study or an experiment? Identify the observational/experimental units and the variables of interest.
(b) Produce graphical displays and numerical summaries to analyze the differences in melting times between the two kinds of chips. Write a paragraph summarizing your findings.
(c) Conduct a two-sample *t*-test to determine whether there is a significant difference in the average melting time between these types of chips. Report the hypotheses, test statistic, and p-value.
(d) Explain why the analysis in (c) is not valid.
(e) Conduct a matched-pairs *t*-test of whether the data suggest that either type of chips tends to take longer to melt than the other. Report all components of the test, including graphical and numerical summaries of the *differences,* the check of technical conditions, and summarize your conclusions. Be sure to comment on whether a cause and effect conclusion can be drawn and the population that you are willing to generalize these results to.
(f) Construct and interpret a 90% confidence interval for the treatment effect on melting time of chocolate as opposed to peanut butter chips.
(g) Suppose that you had calculated the differences in melting times by subtracting in the opposite order. Describe specifically what effect this would have on the test statistic, the p-value, and the confidence interval.

31. Presidential Doctors?

Researchers examined the long-term survival of doctors graduating from one medical school over one century (Redelmeier and Kwong, 2004), comparing those who were presidents of their class to those who appeared alphabetically before or alphabetically after the president in the graduating class photograph. Statistics on long-term mortality were obtained from licensing authorities, medical obituaries, professional associations, alumni records, and national physician directories (follow-up 94%). They reported on 507 presidents and 1014 classmates.

(a) The researchers examined several base-line variables, including sex and whether or not the individual wore glasses. They found 93% of the presidents were male, compared to 85% of their classmates. They also found 9% of presidents were glasses, compare to 12% of their classmates. Are either of these differences statistically significant?

(b) As a measure of accomplishment after graduation, the researchers examined the number of announcements posted by each individual in the alumni notices. They found 21.9% of presidents reported professional accomplishments compared to 13.3% of their classmates. Is this difference statistically significant? (Include all steps of the test of significance and indicate which procedure you are using.)

(c) The overall-life expectancy for the presidents was 49.0 years compared to 51.4 years for their classmates. The two-sided p-value was reported to be 0.036. Assuming the standard deviations were similar in the two samples, use trial-and-error in some technology, or algebra to approximate the value of this standard deviation. What conclusion would you draw from this p-value?

(d) Write a paragraph summarizing your conclusions from these analyses.

32. Exam Performance

Suppose you want to compare student's performances on the first two exams in a course.

(a) Would it make more sense to design this study to use a paired *design* or an independent sample *design*? Explain.

(b) For the following summary data, calculate the paired *t*-statistic and *p*-value, and also the independent-samples *t*-statistics and p-value. Does pairing appear to have been useful in this situation? Explain.

Exam 1	$n_1 = 12$	$\bar{x}_1 = 86.4$	$s_1 = 9.5$
Exam 2	$n_2 = 12$	$\bar{x}_2 = 83.3$	$s_2 = 12.3$
Differences	$n_d = 12$	$\bar{x}_d = 3.2$	$s_d = 4.5$

(c) Repeat (b) for the summary data provided next. [*Hint*: If you pay close attention, you can avoid duplicating work.]

Exam 3	$n_3 = 12$	$\bar{x}_3 = 86.4$	$s_3 = 9.5$
Exam 4	$n_4 = 12$	$\bar{x}_4 = 83.3$	$s_4 = 12.3$
Differences	$n_d = 12$	$\bar{x}_d = 3.2$	$s_d = 18.0$

(d) Explain why pairing is so effective in one case and not in the other. You may want to speculate about what else might be true about how students' exam performance is related across the exams.

33. Laptop Fertility

A study published in the on-line journal *Human Reproduction* on December 9, 2004 suggested that using laptop computers could damage the fertility of males by increasing the scrotal temperature, which can affect the quality and quantity of men's sperm. Researchers studied a sample of 29 healthy males between the ages of 21 and 35 by measuring their scrotal temperature before and after using a computer on their laps. The article's abstract reports that the mean temperature was 2.1 degrees Centigrade higher with the computer resting on their laps even when it was not turned on. The mean temperature was 2.7 degrees Centigrade higher with the computer turned on. The abstract of the article did not report standard deviations but said that the p-values were less than 0.0001.

(a) Explain what makes this a matched-pairs design.

(b) State the hypotheses for the appropriate significance test, both in symbols and in words, for comparing temperatures with no laptop and with the laptop turned on.

(c) If the standard deviation of the temperature increases had been 2 degrees Centigrade, calculate the test statistic and p-value.

(d) Determine how large could the standard deviation have been and still produced a p-value of less than 0.0001.

(e) If you had access to the temperature data for each individual subject, what would you examine to assess whether the technical conditions for the paired *t*-test are satisfied? Explain.

34. Heart Transplant Mortality

Heart transplantation is no longer considered an experimental treatment but is used on thousands with end-stage organ disease. More than 22,000 Americans receive an organ transplant each year. Stanford University has long been recognized as the pioneering center for heart transplants and carried out the first adult heart transplantation in the U.S. in 1968. In the mid-1970s, data were examined to assess the effect of transplantation on survival. Crowley and Hu (1977) examined the data for the patients enrolled between September, 1967, and March, 1974. After a patient is enrolled, a donor heart, matched on blood type, is then sought. This wait could last anywhere from a few days to almost a year. Some patients die before a suitable heart is found. When a match is found, there could be as many as seven patients waiting. It is generally thought that no serious bias has been introduced by the selection of patients for the transplant operation. In fact, physicians believe that, if anything, less hardy patients tend to receive hearts preferentially over hardier ones. In the data set you will examine, 69 patients received a transplant and 34 did not. The researchers recorded the number of days they survived from the time of enrollment until death or until the study ended. These data can be found in transplants.txt

(a) Produce numerical and graphical summaries comparing the survival times of these two groups. What do you learn?
Technology hint: In R, we have to split the two groups out to look at the probability plots, for example:
```
> transplantsurvival=survival[group=="transplant"]
> controlsurvival=survival[group=="control"]
> qqnorm(controlsurvival)
> qqnorm(transplant survival)
```
(b) Would it be reasonable to use a two-sample *t*-test for these data? Explain.
(c) Transform the *survival times* by taking the natural log. Produce numerical and graphical summaries for comparing ln-survival time between transplant and non-transplant patients, including a normal probability plot for each group. Would it be reasonable to apply the *t*-procedures to these data?
(d) Calculate and interpret a 95% *t*-confidence interval for the difference in mean-ln(*survival times*) between these two groups.
(e) The confidence interval in (d) is for mean(ln(treatment)) – mean(ln(control)).

- Taking into account that the transformed variables are symmetric, how can we express this difference in terms of the medians of the transformed variables?
- Recall from Chapter 2 that the median of the ln-transformed data is equal to the ln of the median of the un-transformed data. Apply this relationship to this difference.
- Now apply a rule of logarithms to this expression to show this confidence interval relates to the ratio of the medians of the un-transformed data.

Exponentiate the endpoints of this interval to obtain a confidence interval for the *ratio* of the group medians.

(f) Carry out a randomization test comparing the **median** survival time between the two treatments (using the untransformed data). Be sure it's clear how you have done so (feel free to use the Comparing Groups (Quantitative) applet).

35. Body Mass Index (cont.)

Reconsider Chapter 2 Exercise 23, in which you analyzed body mass index measurements for samples of men and women (BodyMassIndex.txt).

(a) Estimate the difference in mean BMI values between men and women with a 95% confidence interval. Interpret this interval, and comment on whether the technical conditions appear to be satisfied.
(b) Select a transformation that makes the distributions of BMI values more symmetric for both sexes. Determine and interpret a 95% confidence interval for the difference in population means on this

transformed variable.

(c) Convert your interval in (b) back to the original scale by performing the inverse transformation on the endpoints of the interval. How does this interval compare to the original one in (a)?

36. Freshman Fifteen

Suppose that you want to design a study to investigate the common belief that college freshmen tend to gain fifteen pounds of weight (the so-called "freshman fifteen") during their first term away at college.

(a) Explain why a matched-pairs design would be preferable to a completely randomized design for this study.
(b) State the null and alternative hypotheses, in symbols and in words, for testing this common belief.
(c) Suppose that the mean weight gain during the first term on campus in a random sample of freshmen at one college is 13.6 pounds. What more do you need to know to conduct a matched-pairs *t*-test of your hypotheses?
(d) Describe a scenario in which the sample result in (c) would lead to rejecting the null hypothesis.
(e) Describe a scenario in which the sample result in (c) would lead to not rejecting the null hypothesis.

37. Improving SATs

Suppose that 5000 students are randomly assigned to either take an SAT coaching course or not, with the following results in their improvements in SAT scores:

	Sample Size	**Sample mean**	**Sample std. dev.**
Coaching group	2500	46.2	14.4
Control group	2500	44.4	15.3

(a) Conduct a test of whether the sample data provide evidence that SAT coaching is helpful (in increasing the mean improvement). State the hypotheses, and report the test statistic and p-value. Draw a conclusion in the context of this study.
(b) Produce a 99% CI for the treatment effect of the SAT coaching on improvements. Interpret this interval.
(c) Do the sample data provide very strong evidence that SAT coaching is helpful? Explain whether the p-value or the confidence interval helps you to decide.
(d) Do the sample data provide strong evidence that SAT coaching is very helpful? Explain whether the p-value or the confidence interval helps you to decide.

38. Melting Chips (cont.)

Recall the study of chocolate chip and peanut butter chip melting times from Exercise 30 (ChipMelting.txt).

(a) Determine whether the semi-sweet chocolate or peanut butter chip melted more quickly, for each student. Record the number of students for which the chocolate chip melted more quickly and the number for which the peanut butter chip melted more quickly. Also construct a bar graph to display these results. [*Minitab Hint*: You could use `MTB> let c3=(c1<c2)` and then `MTB> tally c3`.]
(b) Report the hypotheses, in words and in symbols, for a sign test of whether the data suggest that either type of chip tends to melt more quickly than the other.
(c) Conduct this sign test, report the p-value, and summarize your conclusion.

39. Memorizing Letters (cont.)

Reconsider Exercises 12 and 18 which you analyzed data from a memory experiment.

(a) Analyze these data with a one-sided, two-sample *t*-test. Summarize your findings, including all

aspects of a significance test.
(b) Compare your findings to those from an empirical randomization test.

40. Left-Handed Advantages?

Noroozian, Lotfi, Ghassemzadeh, Emami, and Mehrabi (2002) compared the acceptance rate of left-handers with that of right-handers in the College Entrance Examintion (CEE) for the national universities in Iran. About 1 million Iranian high school graduates take part each year in the CEE. An entrance exam score is obtained for each participant, which has a mean of 5000 and a standard deviation of 100. A comprehensive list of all participants between 1993–1997 was obtained, and 10,000 were chosen randomly from each year. Hand preference was exclusively defined as writing preference. The distribution of left-handers and the distribution of right-handers did not differ significantly with respect to sex. Of the 47,854 right-handers, the mean score on the CEE was 5020, with standard deviation 718. Of the 3,398 left-handers, the mean score on the CEE was 5060, with standard deviation 720.

(a) Is it appropriate to apply the two-sample *t*-procedures to these sample data, or do you not have enough information to decide? Explain.
(b) Is this a statistically significant difference in the mean CEE score between the population of right-handers and the population of left-handers?
(c) Compute a 95% confidence interval for the difference in mean score between the left-handed population and the right-handed population.
(d) Explain how this difference may be considered statistically significant but not practically significant. What is the cause for this?

41. Schizophrenic Twins

Recall the study of the volumes of the hippocampus brain regions of monozygotic twins who are discordant for schizophrenia from Practice Problem 4.9 (hippocampus.txt).

(a) Carry out a two-sample *t*-test using these data. What conclusion would you draw about whether the mean hippocampus volumes differ between those affected and those unaffected by schizophrenia?
(b) Explain why this test is inappropriate in light of the way the data were collected.
(c) Compare these results to the ones in Practice Problem 4.9. Does the pairing appear to have been effective? Explain.

42. Close Friends

One of the questions asked of a random sample of adult Americans on the 2004 General Social Survey was:

> *From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you? Just tell me their first names or initials.*

The interviewer then recorded how many names each person gave, with the person's sex.

(a) The relevant parameter for this study can be symbolized as μmen – μwomen. Describe what this parameter means in this context.
(b) State the appropriate null and alternative hypotheses (in symbols) for testing whether American men and women differ with regard to average number of close friends.

The survey responses are summarized in the following table (and in the data file CloseFriends.txt):

Number of close friends	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Number of men responses	196	135	108	100	42	40	33	654
Number of women responses	201	146	155	132	86	56	37	813

(c) Use technology to produce graphs for comparing the distribution of number of close friends between men and women. Comment on what the histograms reveal about the shapes of the distributions.
(d) Use technology to determine the sample mean and sample standard deviation of the number of close friends for each sex. Report these with appropriate symbols. Also show how to calculate the sample means by hand from the table above.
(e) Conduct a two-sample *t*-test of the hypotheses from (b). Report the test statistic and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.
(f) Produce a 95% confidence interval for the difference in population means (for the number of close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.
(g) Are the technical conditions for the two-sample *t*-test satisfied here? Explain.
(h) Now conduct a test of whether these sample data suggest that the *proportion* of Americans who say they have *zero* close friends differs between men and women. Report the hypotheses, test statistic, and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.
(i) Produce a 95% confidence interval for the difference in population proportions (who have zero close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.

43. Facebook Emotions

Kramer, Guillory, and Hancock (2014) examined data from Facebook on whether posters respond differently depending on the level of emotional content expressed in the News Feed of their friends. The Facebook News Feed filters content to reduce the amount of information presented at once. Facebook uses an algorithm that aims to identify the content that is most relevant and interesting. In this study, Facebook manipulated how much positive and how much negative content was shown in the feed (to people who viewed Facebook in English). In one part of the experiment, the exposure to positive emotional content was reduced, and in the other the exposure to negative emotional content was reduced. Both studies included a control condition in which a similar proportion of posts were omitted at random. The experiments took place for 1 week (January 11–18, 2012). Participants were randomly selected based on their User ID, resulting in a total of approximately155,000 participants per condition (689,003 participants overall) who posted at least one status update during the experimental period. One response variable was the percentage of all words produced by a person that was either positive or negative during the experimental period.

(a) Below are graphs from this study. Write a few sentences describing what these graphs reveal and whether they appear to support the theory that "affective states are contagious"?

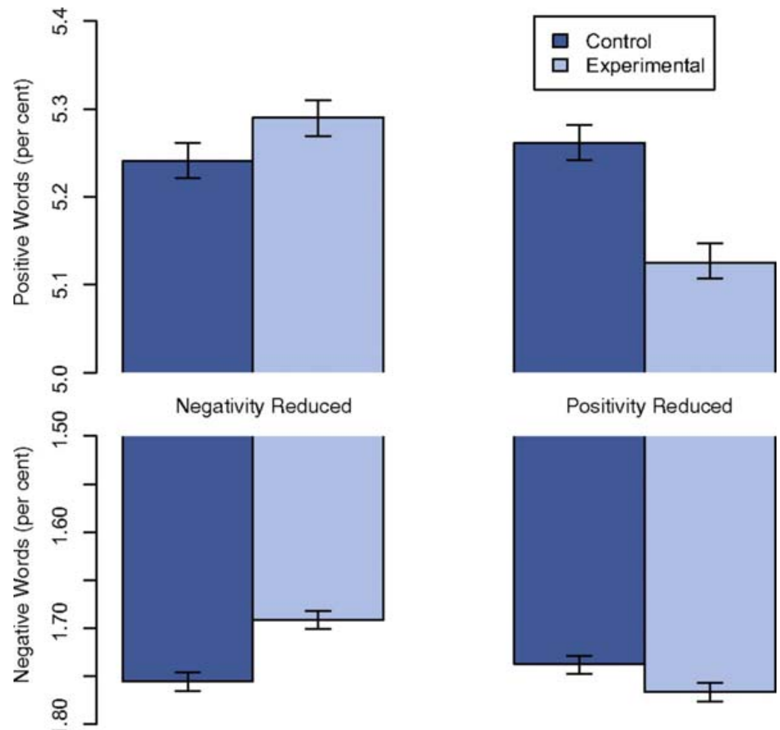


Fig. 1: Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

(b) Cite one disadvantage to these graphs compared to dotplots or histograms.

(c) Consider the mean percentage of positive words generated by people in the negatively reduced condition. Estimate the means and *standard errors* from the graph and estimate the p-value using a two-sample *t*-test. [*Hints*: Be clear what values you read off the top left graph. The bars in the graph represent standard errors, not standard deviations. You can either calculate the standardized statistic by hand using the standard errors, or, assuming 155,000 in each group, convert the standard errors to standard deviations and use technology. If by hand, because the sample sizes are so huge, you can use the normal distribution to convert the standardized statistic to a p-value.]

(d) Use your estimates to construct and interpret a 95% *t*-confidence interval. [*Hints*: If by hand, you can use the normal distribution to find the critical value rather than the t distribution.]

(e) Would you consider the results examined in (b) and (c) to be statistically significant? Would you consider them to be practically significant? Explain. [*Hint*: The second question is pretty subjective.]

(f) Do you expect the difference to be more or less statistically significant for the percentage of positive words generated by people in the positivity reduced condition? Explain your reasoning, discussing all relevant factors.

(g) Do you have any issues about the ethnical nature of this study? Explain. [*Note*: This question calls for opinion.]

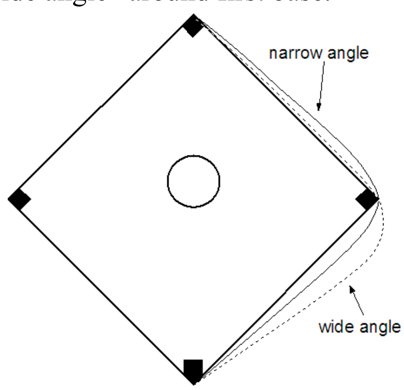
Important notes: It is important to note that a user's full content was always available by viewing a friend's content directly by going to that friend's "wall" or "timeline," rather than via the News Feed. Further, the omitted content may have appeared on prior or subsequent views of the News Feed. Finally, the experiment did not affect any direct messages sent from one user to another.

Posts were determined to be positive or negative if they contained at least one positive or negative word, as defined by Linguistic Inquiry and Word Count software (LIWC2007) word counting system, which correlates with self-reported and physiological measures of well-being, and has been used in prior research on emotional expression.

44. Fish Oil (cont.)
Reconsider the experiment on fish oil and blood pressure, described in previous exercises.
(a) Estimate the treatment effect with a 95% bootstrap interval (for the difference in group means).
(b) How does this interval compare to the *t*-intervals calculated in Exercises 23 and 24?
(c) Estimate the treatment effect with a 95% bootstrap interval (for the difference in group medians).

45. Academy Award Mortality
Redelmeier and Singh (2001) wanted to see whether the increase in status from winning an Academy Award is associated with long-term mortality among actors and actresses. They found 235 actors and actresses who had won at least one academy award and 527 who had been nominated but never win. At the time of the analysis, the average life expectancy for the winners was 79.7 years compared to 75.8 years for the nominees.
(a) Assuming a standard deviation of 17 years for each group, is this a statistically significant difference?
(b) In identifying the actors and actresses for this study, nominated actors/actresses were paired with a person of the same sex and similar age from the same film. Explain the advantages of this pairing.
(c) Summarize the conclusions you would draw from this study.

46. Rounding First Base
Imagine you are at the plate in baseball and have hit a hard line drive that you want to try to stretch from a single into a double. Does the path that you take to "round" first base make much of a difference? Have you seen what most professional baseball players do? Hollander and Wolfe (1999) report on a Master's Thesis by W. F. Woodward (1970) that investigated different base running strategies. For example, you could take a "narrow angle" or a "wide angle" around first base.



In Woodward's study, he used a stopwatch to time 22 different runners going from a spot 35 feet past home to a spot 15 feet before second. He had each runner use each method, with a rest period in between, randomizing which method they used first. The data in BaseRunning.txt shows the time (in seconds) for each running using the narrow angle and the wide angle.
(a) Explain why these data call for a matched pairs *analysis*.
(b) Produce, include, and comment on relevant graphical displays and numerical summaries for investigating the question of whether there is an advantage for taking wide angles or narrow angles. You should examine the angles individually as well as the differences. You should also comment on whether the pairing appeared to be useful in this study (be clear how you are deciding).
(c) Define the parameter of interest in this study and write the null and alternative hypotheses for testing whether there is an advantage for taking wide angles or narrow angles.
(d) Conduct a paired *t*-test or use the Matched Pairs applet to determine whether the data suggest a genuine difference in times for wide angles and narrow angles. If you use the *t*-test, make sure comment

on whether you believe the test procedure is valid and how you are deciding. (Remember to include your output.)
(e) Construct, include, and interpret a 95% confidence interval for estimating the population mean difference in base running times. (Be sure it's clear in your interpretation which method is faster.)
(f) Summarize the conclusions you would draw from this study. Make sure you comment on significance, confidence, generalizability, and causation.
(g) Reanalyze the data using a *sign test* (See Investigation 2.7). State the corresponding hypotheses (in symbols and in words), test statistic, and p-value. Compare your conclusions to those in question 2.

47. Tonsillectomies
Researchers investigated a possible link between having a tonsillectomy and developing Hodgkin's disease. They studied a sample of 85 Hodgkin's patients who had a sibling of the same sex within 5 years of age who was free of the disease (170 individuals total, 85 pairs). Taking into account the paired nature of these data produces the following table:

		Did Hodgkin's twin have tonsillectomy?		
		yes	**no**	**Total**
Did "control" twin have tonsillectomy?	**yes**	26	7	33
	no	15	37	52
	Total	41	44	85

Apply McNemar's Test to these data to determine whether there is evidence that the Hodgkins patients are the ones that are more likely to have had a tonsillectomy. Be sure to:

- State your hypotheses, including a definition of the parameter of interest.
- State the name of the appropriate probability distribution and also specify its parameter values.
- Report the p-value (include output) and summarize your conclusion in context.

NEW EXERCISES

48. 2015 Wimbledon (cont.)
Reconsider the tennis data from the 2015 Wimbledon Championships. In the day 1 matches, the "distance covered" (in meters) by the players was determined for 21 of the matches (WimbledonDistance2015.txt).
(a) Compare the distances covered by the winners (*DistanceCoveredWinner*) between the ladies' matches and the men's matches. Which sex tends to cover more distance? Which player covered the most distance? Can you explain why?
(b) Repeat (a) for the distances covered per point by the winners (*DistanceCovered/ptWinner*). Which comparison do you think is more reasonable? Explain.
(c) Who tends to cover more distance per point – the winners or the losers? Is this a statistically significant difference? (Show the details of a test of significance, justifying why your choice of analysis is appropriate.)

49. Longer Lives?
A group of Cal Poly students wanted to investigate whether men with children tend to live longer than men without children. They randomly sampled men from the obituaries page on the *San Luis Obispo Tribune's* website between June and November 2012. For each man selected, they noted the age at which the person died and whether or not the person had any children.

(a) State appropriate null and alternative hypotheses for testing whether the average lifespan is longer for men with children.
(b) Identify **and** classify the explanatory variable and the response variable in this study.
(c) Does this study involve random sampling or random assignment or both or neither?
(d) The data are in ChildrenandLifespan.txt. Create numerical and graphical summaries of the data comparing the two samples. Summarize what they reveal about the shapes, centers, and spreads of the two samples. Explain why the *shape* of the distribution of the response variable makes sense in this context.
(e) Do you consider the *t*-procedures valid for these data? Explain how you are deciding.
(f) Carry out a two-sample *t*-test to estimate p-value for this study. Include your output including a well-labeled graph of the null distribution with the p-value shaded. Would you reject or fail to reject the null hypothesis at the 5% level of significance?
(g) Calculate a 95% confidence interval for these data.
(h) Summarize the conclusions you would draw from this study including significance, estimation, causation, and generalizability. Provide a brief justification for each component.

50. NFL Referees

In the 2012 National Football League season, the first three weeks' games were played with replacement references because of a labor dispute between the NFL and its regular referees. Many fans and players were concerned with the quality of the replacement referees' performance. One variable of interest is the amount of time it took games to play (in minutes). The data in Referees.txt records this variable for these games and for the next three weeks' games that were played with regular referees.
(a) Create numerical and graphical summaries of the data comparing the two groups. Summarize what they reveal about the shapes, centers, and spreads of the two samples.
(b) Does this study involve random sampling or random assignment or both or neither?
(c) State null and alternative hypothesis for deciding whether games generally take less or more time to play with replacement referees than with regular referees.
(d) Would it be reasonable to use a two-sample *t*-test for these data? Explain.
(e) Transform the *game lengths* by taking the natural log and calculate a 95% *t*-confidence interval for the difference in mean-ln(*game length*) between these two "populations." [*Hint*: You may want to confirm that these distributions are a bit more normal looking.]
(f) The confidence interval in (e) is for mean(ln(replacement)) – mean(ln(regular)).

- Taking into account that the transformed variables are symmetric, how can we express this difference in terms of the medians of the transformed variables?
- Recall from Chapter 2 that the median of the ln-transformed data is equal to the ln of the median of the un-transformed data. Apply this relationship to this difference.
- Now apply a rule of logarithms to this expression to show this confidence interval relates to the ratio of the medians of the un-transformed data.

Exponentiate the endpoints of this interval to obtain a confidence interval for the *ratio* of the population medians. Interpret this interval in context.
(g) Based on the interval in (f), would you reject or fail to reject the null hypothesis is (c)? Can we conclude that the replacement referees caused games to last longer on average?

51. Video Game Aggression

To investigate an association between violent video games and aggressive behavior, British researchers Hollingdale and Greitemeyer (2014) randomly assigned 49 university students to play *Call to Duty: Modern Warfare* (a violent video game) and 52 students to play *LittleBigPlanet 2* (a non-violent/ neutral video game). After 30 minutes of playing the video games, the subjects were asked to complete a marketing survey investigating a new hot chili sauce recipe. They were told they were to prepare some

chili sauce for a taste tester and that the taste tester "couldn't stand hot chili sauce but was taking part due to good payment." They were then presented with what appeared to be a very hot chili sauce and asked to spoon what they thought would be an appropriate amount into a bowl for a new recipe. The amount of chili sauce was weighed in grams after the participant left the experiment. The *amount of chili sauce* was used as a measure of aggression, the more chili sauce, the greater the subject's aggression. The research question is: Do people tend to put more chili sauce into the recipe (and are thus more aggressive) after they play a violent video game as opposed to a non-violent one?
(a) State appropriate null and alternative hypotheses in words and/or symbols. Be sure to define any symbols you use.
(b) Put the data from ChiliSauceData.txt into the Comparing Groups (Quantitative) applet and carry out a randomization test using the *t* statistic to estimate the p-value. Make sure you include a copy of the null distribution indicating the p-value.
(c) Does the *t* distribution appear to be a reasonable mathematical model for this randomization distribution? Make sure you include a copy of the null distribution showing the overlaid *t* distribution. What degrees of freedom is being used? Do you consider the validity conditions met? (Explain.)
(d) Carry out a two-sample *t*-test and compare this p-value to that you found with the randomization test. Which do you consider more reliable? Do they lead to the same conclusion?
(e) Use the *t*-procedure to find a 95% confidence interval for the parameter of interest.
(f) Summarize your conclusions from this study, including comments on significance, estimation, causation, and generalizability.