ISCAM 3: CHAPTER 5 HOMEWORK

1. Feeling Motivated?

The University of Pennsylvania's National Annenberg Election Survey of 2004 studied the humor of late-night comedians Jon Stewart, Jay Leno, and David Letterman. Between July 15 and September 16, 2004, they performed a content analysis of the jokes made by Jon Stewart during the "headlines" segment of *The Daily Show* and by Jay Leno and David Letterman during the monologue segments of their shows. The data that they gathered was:

Leno: 315 of the 1313 jokes were of a political nature
Letterman: 136 of the 648 jokes were of a political nature
Stewart: 83 of the 252 jokes were of a political nature

(a) Organize the data into a two-way table, with comedians in columns and the "political in nature or not" variable in rows. [*Hint*: Notice that the information above presents the number of "successes" and the sample size in each group, not the number of "failures."]
(b) For each comedian, calculate the conditional proportion of his jokes that were political in nature. Also display these proportions in a segmented bar graph. Comment on what your analysis reveals.
(c) For the three comedians combined, what proportion of the jokes were political in nature?
(d) Multiply your answer to (c) by each of the comedian's sample sizes to obtain the expected count of political jokes for each comedian. Then subtract these expected counts from each comedian's sample size to obtain the expected count of non-political jokes for each comedian.
(e) Which comedian came closest to the expected counts? Which told many more political jokes than expected? Which told far fewer political jokes than expected?
(f) Treat these as three independent random samples from the joke-producing processes of these comedians. Conduct a chi-square test to decide whether the proportions of political jokes differ significantly among these three comedians. (Like always, report the hypotheses, comment on the technical conditions, include a sketch of the sampling distribution of the test statistic, and calculate the test statistic and p-value.) Indicate whether the differences among the observed conditional proportions are statistically significant at the 0.10 level, and summarize your conclusions.

2. Regional Internet Users

The Pew Internet and American Life Project examined internet usage variations across 12 regions of the United States in 2001. (The data are based on telephone interviews conducted by Princeton Research Associates using a random digit sampling of the last two digits of telephone numbers.) The two-way table of counts of internet users by region is (we have switched the rows and columns to make the table fit onto the page more easily):

Region	Internet users	Non-internet users	Sample size
New England	541	338	879
Mid-Atlantic	1354	901	2255
National Capital	522	317	839
Southeast	1346	955	2301
South	1026	1005	2031
Industrial Midwest	1543	1114	2657

Upper Midwest	560	422	982
Lower Midwest	728	562	1290
Border States	1054	660	1714
Mountain States	499	289	788
Pacific Northwest	468	223	691
California	1238	706	1944

(a) For each region, determine the proportion of internet users. Which region has the largest proportion? Which has the smallest?
(b) State the null and alternative hypothesis for a chi-square test on these data. Is this a test of homogeneity of proportions or of independence? Explain.
(c) Determine whether the technical conditions for the chi-square test are satisfied. If they are, carry out the test. (Feel free to use technology.) Report the value of the test statistic and p-value, and summarize your conclusion.
(d) Identify the two or three cells of the table that contribute the most to the calculation of the test statistic. Is the observed count larger or smaller than the expected count in those cells? Comment on what this reveals.

3. Academy Award Life Expectancy

Redelmeier and Singh (2001) wanted to see whether the increase in status from wining an Academy Award is associated with long-term mortality among actors and actresses. All actors and actresses every nominated for an Academy Award in a leading or supporting role were identified (n = 762). For each, another cast member of the same sex who was in the same film and was born in the same era was identified (n = 887). Each person was categorized into one of three categories (based on highest achievement): winners (those who were nominated and won at least one Academy Award), nominees (nominated but never won), and controls (those who were never nominated). There were 235 winners, 527 nominees, and 887 controls. Of the winners, 99 had died by March 2000 compared to 221 nominees and 452 of the control group.

(a) Is this study and observational study or an experiment? If observational, is it a case-control, cohort, or cross-classified study? Is this design retrospective or prospective? Explain.
(b) Create a two-way table, segmented bar graph, and conditional proportions to compare the survival rates in these three groups. Comment on what this analysis reveals.
(c) Calculate the relative risk of death for the winners compared to the controls and then compared to the nominees. Comment on what this analysis reveals.
(d) Carry out a chi-square analysis to determine whether there is a statistically significant difference in the survival rate among these three groups. Summarize the conclusions you can draw from this study.

4. Low Birth Weights

To investigate whether different races experience different rates of low birth weight babies, we can examine 1992 data from the *National Vital Statistics Reports*. Reported there is the following information:

	White (non-Hispanic)	**Black (non-Hispanic)**	**Hispanic**
Number of live births	2,298,156	578,335	876,642
% low birth weight babies	6.9%	13.4%	6.5%

(a) Convert this table into a two-way table of *counts*, with race in columns and birth weight (low or not) in rows.
(b) Construct and comment on segmented bar graphs comparing the low birth weight percentages among these three racial groups.
(c) Consider these observations as a random sample from the birth process in the U.S. Conduct a chi-square test to assess whether the observed percentages of low birth weight babies differ more than would be expected by random variation alone. Report the hypotheses, validity of technical conditions, sketch of test statistic sampling distribution, test statistic, and p-value. (Provide the details of your calculations and/or relevant technology output.) Summarize your conclusion.
(d) Which 2-3 of the six cells in the table contribute the most to the calculation of the χ^2 test statistic? Is the observed count lower or higher than the expected count in those cells? Summarize what this reveals about the association between race and birth weights.

5. Racial Gestation

The *National Vital Statistics Reports* also provides data on gestation period for babies born in 2002. The following table classifies the births by the mother's race and by the duration of the pregnancy:

	White (non-Hispanic)	Black (non-Hispanic)	Hispanic
Pre-term (under 37 weeks)	251,132	101,423	99,510
Full term (37 - 42 weeks)	1,885,189	435,923	692,314
Post-term (over 42 weeks)	149,898	36,896	64,997

(*Note*: The totals do not add up to the same totals as in the previous table, because some of the gestation periods were not reported.)

(a) Identify the observational units and the variables represented in this table.
(b) Calculate conditional proportions and produce a segmented bar graph to compare the conditional proportions of gestation periods among the three races. Comment on what these proportions and this graph reveal.
(c) Consider these observations as a random sample from the birth process in the U.S. and conduct a chi-square test of whether these data suggest an association between race and length of gestation period. Report the hypotheses, validity of technical conditions, sketch of sampling distribution, test statistic, and p-value. (Provide the details of your calculations and/or relevant computer output.) Summarize your conclusion.
(d) Which 2-3 of the nine cells in the table contribute the most to the calculation of the χ^2 test statistic? Is the observed count lower or higher than the expected count in those cells? Summarize what this reveals about the association between race and length of gestation period.

6. U.S. Volunteerism

The 2003 study on volunteerism conducted by the Bureau of Labor Statistics reported the sample percentages who performed volunteer work, broken down by many other variables. For example, respondents were categorized by age. The following reports the percentage of sample respondents in each age group who had performed volunteer work in the previous year:

Age group	16–24 years	25–34 years	35–44 years	45–54 years	55–64 years	65 or more
% volunteer	21.9%	24.8%	34.1%	31.3%	27.5%	22.7%

(a) Is this information sufficient to construct a segmented bar graph for comparing the proportions of volunteers across the various age categories? If so, do so, and comment on what the graph reveals. If not, explain.
(b) Explain why this information is not sufficient to conduct a chi-square test of whether these sample proportions differ significantly across the age categories.

The sample sizes in each age group are not given in the report, but based on other information we can estimate them to be as follows:

Age group	16–24 years	25–34 years	35–44 years	45–54 years	55–64 years	65 or more
Sample size	9719	10613	12,070	10,959	7329	9310

(c) Use this information to produce a table of counts with age groups in columns and volunteer status (yes or no) in rows.
(d) Consider a chi-square test on the table that you produced in (c). Would this be a test of homogeneity of proportions or association between variables? Explain.
(e) Conduct the chi-square test. Report the hypotheses, check of technical conditions, sampling distribution, test statistic, and p-value. (Provide the details of your calculations and/or relevant computer output.) Summarize your conclusion.
(f) Construct a 2×6 table with the same row and column headings as in (c), but containing only + and – signs indicating whether the observed count is larger (+) or smaller (–) than expected in that cell. Does this table reveal a pattern? Explain what that pattern suggests about the relationship between age group and volunteerism.

7. U.S. Volunteerism (cont.)

Reconsider the previous question about volunteerism. Suppose that the sample sizes had all been smaller by a factor of 100 (so that the entire study included only about 600 subjects) but that the conditional proportions of volunteerism within each age group had all turned out the same.
(a) How (if at all) would you expect the segmented bar graph to change? Explain.
(b) How (if at all) would you expect the test statistic to change? Explain.
(c) How (if at all) would you expect the p-value to change? Explain.
(d) How (if at all) would you expect your conclusion to change? Explain.
(e) Repeat the chi-square analysis with this greatly reduced sample size (round the observed counts in the new table to the nearest integer). Confirm or correct your answers to (b)–(d) in light of this analysis.

8. U.S. Volunteerism

Suppose that a student wants to study blood types of male and female students on campus, so he/she takes a random sample of 100 males and an independent random sample of 150 females and classifies their blood types. The student organizes the resulting data into a two-way table as follows:

	Males	Females	Total
Type A	40	61	101
Type B	10	16	26
Type AB	6	9	15
Type O	44	64	108
Total	100	150	250

(a) Would a chi-square test applied to these data be a test of independence or a test of homogeneity of proportions? Explain.
(b) Conduct a chi-square test of these data. Report the expected counts, test statistic, and p-value. Are the technical conditions satisfied? What conclusion would you draw from the test?
(c) The test reveals something suspicious about these data. What is that? [*Hints*: Look at the p-value of this test, and think about what the distribution of p-values would look like if the null hypothesis were true. Also think about how often you would get such a large p-value due to random variation if the null hypothesis were true.]
(d) What does this unusual feature lead you to suspect about how the data were collected? Explain.

9. Designing Independence

Suppose that you take a random sample of 1000 college students and ask each to report his/her political inclination as liberal or conservative and also his/her preferred type of music between rock and classical.
(a) Identify the observational units and variables. For each variable, classify it as quantitative or categorical.
(b) Is this a cohort, case-control, or cross-classified design?
(c) If the two variables are (perfectly) independent in this sample, is it necessary for 500 students to be liberal and 500 to be conservative, and also for 500 students to prefer rock music and 500 to prefer classical? Or is just one of these conditions necessary for the variables to be independent? Or is neither necessary? Explain.
(d) Consider the following two-way table with marginal totals filled in:

	Liberal	Conservative	Total
Rock			800
Classical			200
Total	600	400	1000

Is it possible to fill in this table so that the two variables are (perfectly) independent? If so, do it. If not, explain.
(e) Explain how your selections in filling in the table relate to the idea of "expected counts."
(f) Now consider the following two-way table:

	Liberal	Conservative	Total
Rock	450		800
Classical			200
Total	600	400	1000

Is there any way to fill in the remainder of the table so that the two variables are (perfectly) independent? Explain.

(g) Without making the variables independent, is it possible to fill in the rest of the table in (f)? How many choices do you have for the three unfilled cells of the table?

(h) Explain how your answer to (g) relates to the idea of "degrees of freedom."

10. Designing Independence (cont.)

Reconsider the previous question. Now suppose that the sample of 1000 students are allowed to choose among three political viewpoints (liberal, moderate, conservative) but are still limited to choosing between two types of music (rock and classical). Consider the following 2×3 table with marginal totals filled in:

	Liberal	**Moderate**	**Conservative**	**Total**
Rock				800
Classical				200
Total	350	400	250	1000

(a) Is it possible to fill in this table so that the two variables are (perfectly) independent? If so, do it. If not, explain. [*Hint*: Make use of the "expected count" idea.]

(b) Without the stipulation that the two variables be independent, there are a huge number of ways to fill in the table, but once you have filled in some of the cells, you no longer have any choice about how to fill in the rest. How many cells are you free to manipulate before the rest become pre-determined? Explain.

11. Independence Properties

Suppose that a chi-square test of independence is to be applied to a 2×2 table containing $4n$ observations. Suppose that this table has the form:

	Group 1	**Group 2**
"Successes"	$n + c$	$n - c$
"Failures"	$n - c$	$n + c$

(a) Express the value of the chi-square test statistic χ^2 as a function of c.

(b) For a fixed value of n, is this an increasing or a decreasing function of c? Justify your answer, and explain why it makes sense.

(c) For what values of c (in terms of n) would the null hypothesis be rejected at the $\alpha = 0.01$ significance level? Explain.

12. Independence Properties (cont.)

Suppose that a chi-square test of independence is to be applied to a 2 × 2 table containing 2*n* observations. Suppose that for some value of *k* (with $0< k <1$), this table has the form:

	Group 1	**Group 2**
"Successes"	kn	$(1-k)n$
"Failures"	$(1-k)n$	kn

(a) Express the value of the chi-square test statistic χ^2 as a function of *k*.
(b) For a fixed value of *n*, is this an increasing of a decreasing function of *k*? Justify your answer, and explain why it makes sense.
(c) For what values of *k* (in terms of *n*) would the null hypothesis be rejected at the 0.05 significance level? Explain.

13. 2 × 2 Table Properties

Consider the hypothetical data in this 2 × 2 table:

	Group 1	**Group 2**
"Successes"	50	350
"Failures"	50	550

(a) Calculate the expected counts for each cell.
(b) Calculate the value of the χ^2 test statistic. Record the contribution that each cell of the table makes toward the calculation of this test statistic.
(c) Verify that the difference (in absolute value) between the observed and expected count is 10 for all four cells in the table.
(d) Do all four cells therefore contribute the same amount to calculation of the χ^2 test statistic? In fact, do any two of the cells contribute the same amount?
(e) Which cell contributes the most, and which contributes the least? Explain why this makes sense.

14. Blind Apples

Rosales, Yarbrough, Yarbrough, and Martella (1997) wanted to know whether the appearance of an apple affected a person's judgment of its taste. They set up a taste test with 20 subjects where each subject choose their favorite tasting apple, once when they could see the apples and once when they were blindfolded. During the blind taste test, Apple A was chosen as the favorite 2 times, Apple B 8 times, Apple C 7 times and apple D 3 times. Without the blindfold, they favorite apple was A:2, B:6, C:3, D:9. The ordering of the apples and whether the subjects were blind-folded first or second was randomly determined. Suppose you examine the following table:

	With sight	**Blindfolded**	**Total**
A favorite	2	2	4
B favorite	6	8	14
C favorite	3	7	10
D favorite	9	3	12
Total	20	20	40

(a) Is this an observational study or an experiment? Explain.
(b) Is this table an appropriate way to summarize the information? If not, explain why it is misleading.
(c) Suggest an alternative way of examining the results from this study.

15. Halloween Treat Choices

Recall from the Chapter1Exercises the study of whether children prefer fruit-flavored candy (high in sugar) or chocolate candy (high in sugar and fat) for Halloween treats, where 63 of 104 boys and 64 of 87 girls chose the chocolate candy.

(a) Reanalyze these data using the Chi-Square procedure. Compare the Chi-Square test statistic to that from the *two-sided* two-sample *z*-test. Can you find a mathematical relationship between the values of the F and *t* statistics? [*Hint*: See whether the relationship that you found in Investigation 5.3 also holds here.]
(b) In general, suggest a situation in which Chi-Square would be a more appropriate analysis than a two-sample *z* test.
(c) Suggest a situation in which a two-sample *z*-test would be a more appropriate analysis than a Chi-Square analysis.

16. Seven-Game Series

An article in the May 24, 2004 issue of *Sports Illustrated* raised two separate questions about seven-game series in professional team sports. One question concerns the proportion of seven-game series that have gone to the full length of seven games. The article reported that through the year 2003, 44 of 131 (34%) series went to the full length in baseball, compared to 111 of 471 (24%) in hockey and 85 of 303 (28%) in basketball.

(a) Conduct a chi-square analysis of whether these percentages differ more than would be expected by random variation. Begin with graphical displays and numerical summaries, and then proceed to a chi-square test. Summarize your conclusions.
(b) Comment on whether these data come from random samples or from randomization to groups, or whether the randomness is hypothetical here.
(c) The other question posed by the article compares the proportion of "game sevens" that are won by the home team across these sports. The article reported that 23 of 44 (52%) were won by the home team in baseball, compared to 70 of 111 (63%) in hockey and 70 of 85 (82%) in basketball. Analyze these data to assess whether they provide evidence that the three proportions differ significantly, and write a paragraph or two summarizing your conclusions.

17. Feeling Motivated? (cont.)

Reconsider the University of Pennsylvania's National Annenberg Election Survey of 2004, which studied the humor of late-night comedians Jon Stewart, Jay Leno, and David Letterman. Interviewees were given a six-question test of their knowledge of presidential candidates' positions on the issues. The study reported that interviewees who do not regularly watch late-night television comedians got an average of 2.62 items correct, compared to 2.91 for regular

watchers of Letterman, 2.95 for regular watchers of Leno, and 3.59 for regular watchers of Stewart.
(a) What procedure would you use to test whether these four group means differ significantly?
(b) What further information do you need to apply this procedure?

18. Moving Running Times

The worksheet movies03RT.txt contains data on the running time of the movies analyzed in Investigation 5.9 (as reported by imdb.com). One question of interest is whether the average running times of movies differs across the rating categories (G, PG, PG-13, R) of the movies.
(a) Examine numerical and graphical summaries of these data to address this issue. Summarize what they reveal.
(b) Check and comment on the validity of the technical conditions for carrying out an Analysis of Variance to compare the average running time in these 4 categories.
(c) Identify the two movies that are extreme in their running times and remove them from the data set. Repeat your check on the technical conditions.
(d) Carry out the ANOVA (whether or not you believe it is valid) for the data set in (c). Remember to state the hypotheses, sketch the sampling distribution of the test statistic and report the observed test statistic and p-value. Summarize what this analysis reveals.
(e) What percentage of the variation in the running times is explained by the rating categories?
(f) Comment on an interesting aspect to the comparison between the 4 groups that is revealed by the numerical and graphical summaries but not by the ANOVA.

19. Cloud Seeding

Reconsider the cloud seeding data from Investigation 4.7 (CloudSeeding.txt)
(a) Would it be appropriate to conduct an ANOVA to compare the two group means? [*Hint*: See if the relationship that you found in Practice Problem 5.5B also holds here.] Explain.
(b) Conduct an ANOVA on the log-transformed rainfall amounts. Check the technical conditions, state the hypotheses, calculate the test statistic and p-value, and summarize your conclusions.
(c) Compare the ANOVA test statistic and p-value to that of a pooled, two-sample *t*-test. Can you find a mathematical relationship between the values of the *F* and *t* statistics?

20. Memorizing Words

Students conducted a class project in which subjects were asked to memorize as many words as possible in 30 seconds from a list of ten words. There were four different lists of words:

- one list had short, concrete words (car, rock, …)
- one list had long, concrete words (table, donkey, …)
- one list had short, abstract words (fear, sweet, …)
- one list had long abstract words (happy, laughter, …)

Each list was presented to 13 different people, chosen by convenience, among students on campus. Which list a subject received was determined randomly. The number of words successfully memorized (in any order) in 30 seconds was recorded. The data file WordMemory.txt contains the results.

(a) Is this an observational study or an experiment? Explain.
(b) Identify the observational units in this study.
(c) Report which is the explanatory variable and which is the response variable in this study. Classify each variable as categorical or quantitative.
(d) Analyze graphical and numerical summaries to investigate whether the distributions of numbers of words successfully memorized appear to differ among these four groups. (Comment on shape and spread as well as center.)
(e) Conduct an ANOVA to address the question of whether the group means differ significantly at the 0.05 significance level. State the hypotheses, check the technical conditions, sketch the sampling distribution of the test statistic, calculate the test statistic and p-value, and summarize your conclusions.
(f) Considering the design of this study, are you justified in concluding that there is a cause-and-effect relationship between word type and number of letters memorized? Explain.
(g) An alternative analysis is to examine the "length of word" and "concrete/abstract" variables separately (in the 8th and 9th columns). Which variable appears to have a stronger effect on the number of words memorized? Explain your conclusion based on this "two-way" ANOVA table.

21. Crash Tests

The National Transportation Safety Administration conducts crash tests on automobiles. The file crash.txt contains data on automobile crash tests in which stock automobiles are crashed into a wall at 35 miles per hour with dummies in the driver and front passenger seat (as reported by the Data and Story Library (DASL) web site, http://lib.stat.cmu.edu/DASL/Datafiles/Crash.html). Response variables are measurements of injury extent on head (column 5), chest (column 6), left leg (column 7), and right leg (column 8). Explanatory variables include whether the dummy was on the driver or passenger side (column 9), protective devices in the car (column 10), number of doors on the car (column 11: 2, 4, or other), year of make (column 12), and size of car (column 14).

(a) Produce boxplots of head injury measurements by number of doors. Write a paragraph comparing and contrasting the distributions of these measurements among the three groups. (Comment on shape, center, spread, unusual observations, and any other features of interest. Pay particular attention to the question of whether the extent of head injury seems to differ among the three groups.)
(b) In addition to the boxplots, produce normal probability plots of the head injury measurements for each of the "number of doors" categories. Do the data suggest that each of the population distributions of head injury measurements are normally distributed? Explain.
(c) Apply the log transformation to the head injury measurements. Then examine boxplots and normal probability plots of this transformed variable by number of doors. Do these distributions appear to be roughly normally distributed?
(d) Does the technical condition about equal population standard deviations appear to be satisfied on the transformed data? Explain.
(e) Conduct an ANOVA on these transformed data. Report the hypotheses along with the value of the *F* statistic and p-value. Summarize your conclusions about whether the data provide evidence that the extent of head injury varies among vehicles with different numbers of doors.

22. Crash Tests (cont.)
Reconsider the crash test data from the previous question. Investigate whether the data provide evidence that any of the other response variables differ significantly by the number of doors on the vehicle. Include both a descriptive (visual displays and numerical summaries) and inferential (ANOVA) analysis. Be sure to investigate whether a transformation is needed and, if so, whether the log transformation works well again. (If the log does not seem to work, try other power transformations such as square root, cube root, or reciprocal.) Also be sure to check the condition about equal standard deviations. Write a paragraph summarizing your findings.

23. Crash Tests (cont.)
Reconsider again the crash test data from the previous question. Now examine whether the data provide evidence that the extent of head injury varies across vehicles of different *years*. Write a paragraph or two summarizing your findings. Include both a descriptive (visual displays and numerical summaries) and inferential (ANOVA) analysis. Be sure to investigate whether a transformation is needed and, if so, whether the log transformation works well again. Also be sure to check the condition about equal standard deviations. Write a paragraph summarizing your findings.

24. Comparing Means
Suppose that instructors A, B, and C are each teaching three large sections of a course, and each instructor wants to study whether the mean exam scores differ significantly across his/her three sections. Suppose that each takes a random sample of ten students, and calculates the following descriptive statistics:

	A1	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**
Sample size	10	10	10	10	10	10	10	10	10
Sample mean	50	60	70	50	60	70	57	60	63
Sample std. dev.	24	24	24	5	5	5	5	5	5

(a) Based on these statistics, which instructor has the strongest evidence that the mean scores differ significantly across his/her three sections? Which has the least evidence? Explain your answers.
(b) Hypothetical data matching these statistics can be found in HypoAnova.txt. Perform an ANOVA for each instructor, and report the test statistics and p-values.
(c) Do the results in (b) confirm your answers to (a)? If so, explain. If not, explain how you would not change your answers to (a) and why.

25. ANOVA Power
Consider the following two sets of population means for three groups. In which of the two situations do you think an ANOVA F test would have greater power?
A: $\mu_1 = 95, \mu_2 = 100, \mu_3 = 105$ B: $\mu_1 = 90, \mu_2 = 100, \mu_3 = 110$
Explain your reasoning.

26. 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 data are in the file GymMembership.txt.

(a) Analyze the data with ANOVA for comparing the mean ages of men and women. Produce and comment on numerical and graphical summaries, state the hypotheses, check the technical conditions, sketch the relevant sampling distribution, calculate the test statistic and p-value, and summarize your conclusions.

(b) Analyze the data with ANOVA for comparing the mean ages of people who joined the club in August and in September. Produce and comment on numerical and graphical summaries, state the hypotheses, check the technical conditions, sketch the relevant sampling distribution, calculate the test statistic and p-value, and summarize your conclusions.

27. ANOVA vs. Two-Sample *t*-Tests

In Practice Problem 5.5B you saw that the pooled two-sample two-sided *t*-test is equivalent to an ANOVA.

(a) Suggest a situation in which ANOVA would be a more appropriate analysis than a two-sample *t* test.

(b) Suggest a situation in which a two-sample *t*-test would be a more appropriate analysis than an ANOVA.

28. Comparing Diets

Reconsider Example 5.4 and the *JAMA* study that randomly assigned subjects to one of four population diet programs. In that example, we included only subjects who completed the 12-month study in the analysis. The researchers also analyzed the data by including all 40 subjects in each group, with the assumption that all subjects who dropped out of the study had zero weight loss/gain. These data are in the data file (ComparingDiets.txt).

(a) Describe how including these subjects affects the descriptive statistics.

(b) Describe how including these subjects affects the boxplots.

(c) Produce boxplots of weight loss by whether the subject completed the study or not, for all four diet groups combined. Describe what these boxplots look like, and explain why they look like this.

(d) Repeat the ANOVA analysis for comparing weight loss among these four diet groups using these data. Provide the ANOVA table, and summarize your conclusion.

(e) Comment specifically on how including these additional subjects changes the ANOVA analysis, as compared to the ANOVA analysis based only on the subjects who completed the 12-month study.

29. Comparing Diets (cont.)

Reconsider the previous exercise and the diet comparison study. Now conduct ANOVA analyses on the weight losses after 2 months and after 6 months, both with all subjects and with only those subjects who completed the study. Present all of the ANOVA tables, and summarize

your findings. In particular, comment on whether any of these analyses reveal significant differences in weight loss among the four diet plans.

30. Appraisal Prices

The file auction.txt contains data on the appraisal prices of art auctioned off over a 4-day period in December of 2004. The variables included are *day*, *appraisal price*, *starting price* at the auction, and *selling price* at the auction. Test whether the data provide evidence that the mean price differs significantly on any of these four days compared to the others. Provide a full analysis, including graphical displays and a check of technical conditions, in addition to summarizing your conclusions.

31. Practicing Correlation Coefficients

Consider the following 4 data sets:

A: (1,3), (2,5), (3,6), (4,8)
B: (1,4), (2,7), (3,2), (4,4)
C: (1,8), (2,6), (3,2), (4,3)
D: (1,5), (2,3), (3,5), (4,2)

(a) Based on the changes in the x and y values, arrange these data sets in order from the most negative correlation to the most positive. Explain your reasoning.
(b) Enter these data values in your technology and confirm your ordering. If your ordering was wrong, provide an explanation for the correct ordering.

32. Correlation Properties

Suppose that you take a random sample of classes on campus and record the number of students enrolled in the class and the average evaluation score assigned by students to that teacher's effectiveness (on an A = 4, B = 3, … scale).

(a) If the correlation coefficient turned out to be very close to zero, would you conclude that larger classes tend to have lower teaching evaluation averages? Explain.
(b) Suppose that the correlation coefficient turned out to be $r = -0.5$. Would you expect this to be more statistically significant if the sample size were $n = 5$ or if the sample size were $n = 50$, or would you expect sample size not to matter? Explain.
(c) Suppose that you also record whether the teacher was male or female. Would it make sense to calculate the correlation coefficient between class size and sex? Explain.

33. Least Squares Coefficients

Re-consider the expressions for calculating least squares coefficients for the slope $\left(b_1 = r\frac{s_y}{s_x}\right)$ and intercept $\left(b_0 = \bar{y} - b_1\bar{x}\right)$ of a regression line. Use these formulas to explain what happens to the least squares line in the following situations. [*Hint*: You may find it helpful to draw sketches as well as analyze these situations algebraically.]

(a) The mean value of the response variable increases, and all else remains the same. [*Hint*: Report what happens to slope and to the intercept.]
(b) The mean value of the explanatory variable increases, and all else remains the same.
(c) The standard deviation of the values of the response variable increases, and all else remains the same.
(d) The standard deviation of the values of the explanatory variable increases, and all else remains the same.
(e) The correlation coefficient between the two variables moves closer to zero, and all else remains the same.

34. Least Squares Coefficients (cont.)

Re-consider the expressions for calculating least squares coefficients for the slope $\left(b_1 = r\frac{s_y}{s_x}\right)$ and intercept $\left(b_0 = \bar{y} - b_1\bar{x}\right)$ of a regression line. These indicate that the least squares regression line can be written as: $\hat{y} = \left(\bar{y} - r\frac{s_y}{s_x}\bar{x}\right) + r\frac{s_y}{s_x}x$.

(a) Solve this equation for x as a function of $\hat{y}$.
(b) Is this the same equation that you would get by reversing the roles of the explanatory (x) and response (y) variables in the least square equation? Explain.
(c) Under what conditions will these be the same line? Explain.

(*Note*: The fact that these are not (usually) the same reveals that with least squares regression, which variable occupies which role (explanatory or response) has an effect on the resulting line.)

(d) Re-consider the expression for calculating a correlation coefficient: $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$. Does reversing the roles of the explanatory (x) and response (y) variables affect the calculation of the correlation coefficient? Explain.

35. Height and Foot Size

Reconsider the height and foot length data of Investigation 5.8 (HeightFoot.txt). In that investigation you used foot length (in centimeters) to predict height (in inches). Recall that some summary statistics are:

	Mean	Std dev	Correlation
Height	67.75 in	5.004 in	0.711
Foot length	28.50 cm	3.445 cm	

Now suppose that you want to use these data to construct a model for predicting a student's foot length from his/her height. (After all, most people can tell you their height off the top of their heads, but few can tell you the length of their right foot in centimeters.)

(a) Use these summary statistics to determine the least squares regression line for predicting foot length from height.
(b) Interpret the slope coefficient of this line in context.
(c) Use the line to predict the foot length of a student who is 66 inches tall.
(d) Identify the units of measurement (e.g., inches, centimeters, no units) on the slope coefficient, the intercept coefficient, and the correlation coefficient.
(e) The least squares line for predicting height from foot length is $\hat{height} = 38.3 + 1.03\, ft$. Solve this equation for foot length as a linear function of height.
(f) Is the line in (e) the same as the line in (a)? Explain.

36. Influential Observations

In regression, an observation with an extreme value of the explanatory (x) variable has the *potential* to be influential. Why is it not necessarily influential?

(a) Draw a hypothetical scatterplot where one observation has an extreme x value but is not influential on the regression line. Then draw a hypothetical scatterplot where one observation has an extreme x value and is influential on the regression line. Explain the reasoning behind your drawings.
(b) Enter both sets of your hypothetical data into your technology. For each set of data, determine the least squares line and correlation coefficient both with and without the potentially influential observation. Report all four of these equations. Were you correct that the extreme observation was influential on the regression line in one dataset but not in the other?
(c) Comment on whether the extreme observations were influential on the correlation coefficient for each data set.

37. Resistant Lines

In Practice Problem 5.9, you made a prediction about whether the least squares line or the line that minimizes the sum of absolute errors would be more resistant to outliers. Investigate your prediction by using the applet that calculates the sum of squared errors and the sum of absolute errors from any line:

- Add an extreme (in the x-variable) point that does not fit with the linear pattern of the others.
- Determine the line that (roughly) minimizes the sum of absolute errors by moving the line until you have roughly minimized the sum of the absolute errors.
- Remove the extreme point, and re-determine the line that (roughly) minimizes the sum of absolute errors.
- Put the extreme point back and find the regression (least squares) line.
- Remove the point again and note the effect on the regression (least squares) line.

Report the equations of all four lines, and comment on how much the lines changed when the extreme point was included or excluded.

38. Surfboard Lengths

Do taller surfers tend to choose longer surfboards? Or do they tend to choose shorter ones? Or is there no association between a person's height and the length of his/her surfboard? Does the relationship between height and surfboard length differ between men and women? A student investigated this issue by collecting data over several weeks at a local beach (Wood, 2004). The data are in the file surfer.txt.

(a) Identify the observational units and the variables in this study. Classify each variable as categorical or quantitative.

(b) Construct a scatterplot with surfboard length as the response variable and height as the explanatory variable. Comment on what the graph reveals about a relationship between the two. (Be sure to comment on direction, strength, and linearity as well as any unusual observations.)

(c) Calculate the correlation coefficient between height and surfboard length. Comment on its direction and magnitude.

(d) Investigate and report on whether the technical conditions for the basic regression model are satisfied with these data.

(e) Report the hypotheses, sketch the relevant sampling distribution, and report the test statistic, and p-value for assessing whether there is an association between height and surfboard length. Is the sample slope statistically significant at the 0.10 level? Summarize your conclusion.

(f) Do you think it's reasonable to regard these sample data as representative of a larger population? What further information would you like to know?

39. Breaking Ice

Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on the Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file NenanaIceBreak.txt contains all of the data since 1917. Scientists have examined these data for evidence of global warming, which would suggest that the ice break day should be tending to occur earlier as time goes on.

(a) Examine a scatterplot of the day in which the ice broke (coded in column 7 with April 1 = 1) vs. year. Does it reveal any association between the two variables? In other words, is there any indication that the day on which spring begins is changing over time? Explain.

(b) Determine and report the regression line for predicting ice break day from year. Also calculate the correlation coefficient and the value of r^2. Comment on what these reveal, including an interpretation of the slope coefficient.

(c) Conduct a test for whether there is a linear association between ice break day and year. State the hypotheses, and report the test statistic and p-value. Check the technical conditions, and summarize your conclusions.

(d) Would you say that the p-value reveals evidence of a strong association, or strong evidence of an association? Explain.

(e) Do the data suggest that one can make better predictions by taking year into account, rather than simply using the average of the ice break days? Explain.
(f) What date would the regression model predict for the ice break-up in the year 2005? What about 2020? Explain why you should regard these predictions cautiously.

40. Cricket Thermometers

In the late 1890s, scientists first noted that the frequency of a cricket's chirps is related to air temperature. It is natural to ask whether *cricket chirp rates* can therefore be used to predict *temperature*. The data in crickets.txt come from an early study on snowy tree crickets by Bessey and Bessey (1898). For each of 30 crickets, the *frequency of chirping* (in chirps per minute) and the *air temperature* (in degrees Fahrenheit) were recorded.

(a) Which variable is explanatory and which is response?
(b) Produce a scatterplot, and comment on the direction, strength, and linearity of the association between chirp frequency and temperature.
(c) Calculate and comment on the correlation coefficient.
(d) Determine the least squares line for predicting temperature from chirp frequency. Interpret the value of the slope coefficient.
(e) Report and interpret the coefficient of determination.
(f) What temperature would the line predict when the cricket is chirping at a frequency of 120 chirps per minute?

41. Cricket Thermometers (cont.)

Reconsider the previous exercise about cricket thermometers. Now proceed to an inferential analysis.

(a) Examine residual plots to assess whether the technical conditions for inference appear to be satisfied. Summarize your findings.
(b) Report the hypotheses, sketch the relevant sampling distribution, and report the test statistic and p-value for testing whether the sample data provide strong evidence of a positive association between chirp frequency and temperature.
(c) Construct a 95% confidence interval for the population slope coefficient. Interpret this interval.
(d) Determine a 95% confidence interval for the average temperature on nights when the cricket is chirping at 120 chirps per minute.
(e) Determine a 95% prediction interval for the temperature on a particular night when the cricket is chirping at 120 chirps per minute.
(f) How do the intervals in (d) and (e) compare, both in midpoints and in widths? Explain why this makes sense.
(g) At what value of chirps per minute would a prediction interval for temperature be most narrow? Explain how you know.

42. Testing the Correlation Coefficient

A parallel test to the model utility test (with one explanatory variable) is based on the correlation coefficient rather than the least squares slope coefficient. Let r denote the sample correlation coefficient and ρ denote the population correlation coefficient. Then a test of whether $\rho = 0$ is based on the test statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, and the relevant reference distribution is a t-distribution with (n–2) degrees of freedom. Reconsider the data on predicting height from foot length in Investigation 5.8. Recall that the sample correlation coefficient is $r = 0.711$, the sample size is $n = 20$. Let ρ denote the correlation coefficient between height and foot length in the population from which this sample of students was selected.

(a) State the null and alternative hypotheses, in words and in symbols, for testing whether there is a positive association between height and foot length in the population.

(b) Use the above result to calculate the value of the test statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Then calculate the p-value of the test, and summarize your conclusion.

(c) Now use technology to fit this regression line and to report the test statistic for testing whether the population slope coefficient equals zero. Sketch the sampling distribution of the test statistic and report the test statistic and p-value. Summarize the conclusion that you would draw.

(d) Are the test statistics in (b) and (c) the same?

43. Testing the Correlation Coefficient (cont.)

Refer to the result stated in the previous question for conducting a test about a population correlation coefficient.

(a) Suppose that you take a random sample from a population and calculate the sample correlation coefficient between two quantitative variables to be $r = 0.5$. What further information would you need to decide if this provides strong evidence of an association between the variables in the population?

(b) Suppose that the sample size had been $n = 40$, and the sample correlation coefficient turned out to be $r = 0.5$. State the hypotheses, sketch the sampling distribution, calculate the test statistic, and determine the p-value. Is the sample correlation statistically significant at the 0.05 level?

(c) Repeat (b) with a sample size of $n = 15$.

(d) With the sample correlation coefficient of $r = 0.5$, express the test statistic as a function of the sample size n. Sketch a graph of this function for values of n from 3 through 50. (Feel free to use technology to evaluate the function and draw the graph.)

(e) Determine the smallest sample size for which a sample correlation coefficient of $r = 0.5$ is statistically significant at the 0.05 level. [*Hint*: Use your function/graph from (d) in conjunction with the inverse cdf function for the t-distribution.]

(f) If the sample correlation coefficient were $r = 0.3$, would the smallest sample size required for significance at the 0.05 level be greater, smaller, or the same as in (e)? Give an intuitive explanation.

(g) Determine the necessary sample size in (f). [*Hint*: First create a graph of the test statistic vs. sample size, as you did in (d) and (e).]

44. Airfares

The following table reports the cheapest available airfare (as reported by the Sunday newspaper) and distance (in miles) to various destinations from Baltimore (also found in airfare.txt):

Destination	Distance	Airfare
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258

Destination	Distance	Airfare
Miami	946	198
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

(a) Which is the explanatory variable and which is the response?
(b) Produce and examine a scatterplot with the response variable on the vertical axis. Comment on the direction, strength, and linearity of the relationship between these two variables.
(c) Determine the correlation coefficient between these variables, and also report the p-value for assessing whether the relationship is statistically significant. Interpret both the correlation coefficient and the p-value, and summarize what they reveal.
(d) Produce a fitted line plot, and report the regression equation for predicting airfare from distance, along with the value of r^2.
(e) Interpret what the value of the slope coefficient represents in this context.
(f) What percentage of the variability in airfares is explained by the linear relationship with distance?
(g) Identify the destination with the largest residual (in absolute value). Also report the value of its residual, and explain how this would be calculated by hand.
(h) Which destination would you expect to have the most influence on the line? Explain.
(i) What airfare would the line predict for a destination that is 500 miles away?

45. Airfares (cont.)

Reconsider the airfare data (airfare.txt).

(a) Run the regression analysis and store the residuals in their own column.

(b) Use technology to calculate the variance in the airfare values, $\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$

(c) Use technology to calculate the variance in the residuals, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-1)$

(d) Compute the ratio of the variance in the residuals to the variance in the airfares. How is this value related to the coefficient of determination (R-Sq) in the technology output?

(*Notes*: The numerator of your calculation in (b) is reported in the Total row of the SS column. The numerator of your calculation in (c) is reported in the Residual Error row of the SS column.)

46. Used Hondas (cont.)

The data in UsedHondas.txt come from a sample of 45 used Honda Civics and Accords listed for sale on the web in August 2015. In addition to model of car, other variables recorded are age in years, mileage, and price.

(a) Examine scatterplots and correlation coefficients to determine whether age or mileage would be a better predictor of price. Choose one of these explanatory variables, and explain your choice.
(b) Determine the least squares line for predicting price from whichever explanatory variable you selected.
(c) Conduct a test of whether the sample data provide strong evidence of a linear relationship in the population between price and your explanatory variable. Include a check of the technical conditions with your analysis as well as the other steps in a test of significance.
(d) Determine a 95% confidence interval for the population slope coefficient, and interpret what this parameter and interval represent.
(e) Choose a particular value of your explanatory variable, and produce and interpret a 95% prediction interval for such a car.

47. Used Hondas (cont.)

Reconsider the data on used Hondas from the previous question, and continue to focus on the issue of finding a model to predict the price of a used Honda. We can include both age and mileage in the same model.

Minitab: Enter both age and mileage in the Continuous predictors box.
R: > summary(lm(price~age+mileage))
JMP: Analyze > Fit Model

(a) Run a regression to predict price, including both of these variables as predictors. How much of the variability in price is explained by the linear regression on these two variables combined? Is this larger than when you used just one variable in the previous problem? Explain why this makes sense.
(b) Use this equation to predict the price of the following cars:
3 years old, 30,000 miles
4 years old, 30,000 miles
3 years old, 60,000 miles
4 years old, 60,000 miles
4 years old, 60,100 miles
Based on these results, provide an interpretation of the slope coefficient of each predictor variable in this study.
(c) The overall *F* statistic now tests the null hypotheses that both population slope coefficients are zero (H_0: $\beta_1 = \beta_2 = 0$). What conclusion would you come to from this p-value?
(d) The individual *t* statistics still test the individual slope coefficients but conditional on the other variable being in the model. Does *age* appear to be a useful predictor even if we know the mileage of the car? Does *mileage* appear to be a useful predictor even if we know the age of the car? Justify your responses.

48. Used Hondas (cont.)

Reconsider the previous questions. It is also easy to add binary variables into a regression model. Rerun the regression model using *age*, *mileage,* and *model* as predictor variables.

(a) Report the regression equation for each model. [*Hints*: In R and Minitab, the intercept represents the intercept for Accord and the coefficient of *model* represents the change in price for a Civic compared to an Accord.]
(b) Report the value of r^2 and comment on how it has changed.
(c) Provide an interpretation for the coefficient of model type in this regression equation. What does this say about the difference in average price between Civics and Accords, after adjusting for age and mileage? Support your statements with labeled scatterplots (coded by model type) of *price* vs. *age* and *price* vs. *mileage*.
(d) Does model *type* appear to be a useful predictor variable even after we know the mileage and age of the car? Explain.

49. Surfboard Lengths (cont).

Reconsider Exercise 38 about surfers (surfer.txt).
(a) Create a scatterplot of surfboard length vs. height using separate symbols for the two sexes. Comment on what this graph reveals that the earlier scatterplot did not.
(b) Separate the data by sex ("unstack" the columns). Carry out a separate regression analysis for each sex (scatterplot, correlation coefficient, inference for regression). Report your findings including how the direction, strength, and linearity differs between the two groups.
(c) Run a regression analysis using both height and sex (with sex coded as 0 = male and 1 = female in the second column) as predictor variables. Provide an interpretation of the coefficient of sex in this model. [*Hint*: Follow our earlier method, if the explanatory variable changes by 1…].
(d) Do either sex or height appear to be statistically significant predictors of surfboard length in this model, assuming that the other variable is already in the model? Explain. [*Hint*: Examine the p-values for testing whether the variable's coefficient differs from zero.]
(e) Suggest an advantage to running a regression with two predictor variables instead of looking at the regression model for each sex separately.

50. Backpack Weights (cont.)

Reconsider the data from the Chapter 2 exercises on weights of backpacks (backpack.txt). Consider predicting the weight of a student's backpack from his/her body weight.
(a) Examine a scatterplot of backpack weight vs. body weight. Fit a regression line, and examine residual plots. Do these suggest that a transformation would be appropriate? Explain.
(b) Transform both weight variables by taking the logarithm base 10. Fit a regression line, and examine residual plots. Do these residual plots look better now? Explain.
(c) Conduct a test of whether the sample data provide strong evidence of a linear relationship in the population between log(backpack weight) and log(body weight).
(d) Predict the backpack weight for as student weighing 150 pounds. [*Hint*: Remember to "back-transform" so that your prediction is in pounds, not in log(pounds).]
(e) Determine a 90% prediction interval for your prediction in (d).

51. Backpack Weights (cont.)
Reconsider the previous exercise involving a model for predicting a student's backpack weight from his/her body weight.
(a) Repeat the analysis, but take the logarithms with natural log (base *e*) rather than base 10.
(b) How does this change the regression equation?
(c) How does this change r^2?
(d) How does this change the prediction for the backpack weight of a 150-pound student?
(e) How does this change the 90% prediction interval for the backpack weight of a 150-pound student?

52. Coefficient of Determination
Consider the following statements about the coefficient of determination r^2. Explain why each one is wrong.
(a) It is the percentage of points that fall on the line.
(b) It is the percentage of correct predictions made by the line.
(c) It can be negative, just not below –1.
(d) If it is close to 1, then a linear model is the most appropriate model.
(e) If it is close to 1, then there is most likely a cause-and-effect relationship between the variables.

53. Residual Properties
Consider the sum of residuals from a least squares line.
(a) Show algebraically that this sum always equals 0. [*Hints*: Remember that a residual is $y_i - \hat{y}_i$ and that $\hat{y}_i = b_0 + b_1 x_i$, so you want to analyze $\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]$. More important, remember that with least squares lines, we know that $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$.]
(b) Does it follow from this result that the *mean* of the residuals from a least squares line must equal zero? Explain.
(c) Does it follow from this result that the *median* of the residuals from a least squares line must equal 0? Explain.

54. Money Making Movies
Reconsider Investigation 5.9 in which you analyzed box office revenues and critics' rating scores of movies (movies03.txt).
(a) If we utilize the data set that does not have the six highest-grossing movies (columns 8–13), are the conditions to do inference for regression met? If not, which condition(s) appear violated? Does the large sample size allay any of our concerns?
(b) Decide whether the relationship between critics' ratings and revenue is statistically significant. (State the hypotheses, sketch the relevant sampling distributions, and report the test statistic and p-value as well as your conclusion in context.)

55. Coefficient of Determination (cont.)

Recall from an earlier exercise, that the coefficient of determination can be calculated as

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-1)s_y^2}$$

where $\hat{y}_i = b_0 + b_1 x_i$ and $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$. Show algebraically that this expression is equal to the correlation coefficient squared. [*Hint*: Substitute the expressions for $\hat{y}_i$ and b_0 and *complete the square*:

$$\sum_{i=1}^{n}(y_i - \bar{y} + b_1\bar{x} - b_1 x_i)^2 = \sum_{i=1}^{n}\left[(y_i - \bar{y})^2 + 2b_1(y_i - \bar{y})(x_i - \bar{x}) + b_1^2(x_i - \bar{x})^2\right].)$$

56. Anscombe's Data

A classic (hypothetical) data set was created by statistician Frank Anscombe (1973). Consider the following four sets of (*x*,*y*) data, also found in anscombe.txt:

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	7.46	10	9.14	8	6.58
8	6.95	8	6.77	8	8.14	8	5.76
13	7.58	13	12.74	13	8.74	8	7.71
9	8.81	9	7.11	9	8.77	8	8.47
11	8.33	11	7.81	11	9.26	8	7.04
14	9.96	14	8.84	14	8.10	8	5.25
6	7.24	6	6.08	6	6.13	8	5.56
4	4.26	4	5.39	4	3.10	8	7.91
12	10.84	12	8.15	12	9.13	8	6.89
7	4.82	7	6.42	7	7.26	8	8.84
5	5.68	5	5.73	5	4.74	19	12.50

(a) For each of the four data sets, calculate the mean of each variable, the standard deviation of each variable, and the correlation coefficient between them.

(b) What do you notice about these summary statistics? Based on this realization, what can you say about the least squares regression lines for predicting *y* from *x* in each of these four data sets? Explain.

(c) Use these statistics to determine the least square regression line for predicting *y* from *x* in each of the four data sets. Also calculate the coefficient of determination for each data set. How do they compare?

(d) For each data set, create a scatterplot of *y* vs. *x* with the regression line superimposed. Also, for each data set, create a scatterplot of residuals vs. *x*.

(e) Would you say that the regression line fit the data equally well in all four data sets? Would you say that the regression model is equally valid for all four data sets? If not, for which data set(s) does the regression model seem to be most appropriate? Explain.

57. Comparing Diets (cont.)

Reconsider Example 5.4 and the *JAMA* study comparing popular diet plans (ComparingDietsFull.txt).

(a) Report the equation of the regression line for predicting weight loss based on adherence level.
(b) What weight loss would this model predict for an individual with an adherence level of 5.5?
(c) Repeat the prediction in (b) for individuals with adherence levels of 1.5 and 7.5.
(d) Determine 95% prediction intervals for the weight loss of the three individuals in (b) and (c).
(e) Which of these three prediction intervals is widest? Which is narrowest? Offer an explanation for why this ordering makes sense.

58. Comparing Diets (cont.)

Once again reconsider Example 5.4 and the *JAMA* study comparing popular diet plans (ComparingDiets.txt).

(a) Fit separate regression models for predicting weight loss from adherence level for the four different diet plans. Comment on how the intercept and slope coefficients of these models compare.
(b) Use each of these four regression models to predict the weight loss of an individual with an adherence level of 5.5.
(c) Does one of these four diets always produce a higher predicted weight loss than the others, for every adherence level between 1 and 10? Justify your answer. [*Hint*: Drawing the four lines might be helpful.]

59. Dr. Spock's Trial

Reconsider Investigation 5.1. Another way to analyze the data for the trial of Dr. Spock is to look at the percentage of women on the different venires and determine whether the mean percentage of women is equal across the seven judges. Following are the percentages of women on the venires for a contemporary sample from each of the judges. These data are also in SpockPers.txt.

Judge 1	Judge 2	Judge 3	Judge 4	Judge 45	Judge 6	Judge 7
16.8	27.0	21.0	24.3	17.7	16.5	6.4
30.8	28.9	23.4	29.7	19.7	20.7	8.7
33.6	32.0	27.5	21.5	23.5		13.3
40.5	32.7	27.5	27.9	26.4		13.6
48.9	35.5	30.5	34.8	26.7		15.0
45.6	31.9	40.2	29.5			17.7
32.5	29.8					18.6
33.8	31.9					23.1
33.8	36.2					15.2

(a) Explain what information we learn from analyzing the data this way that we did not see when we carried out the chi-square test on the overall proportion of women for each judge. Why might this information be useful?
(b) Produce numerical and graphical summaries to compare the percentages across the seven judges.
(c) Carry out an ANOVA to test if at least one judge has a different mean percentage than the others. Did you state the null and alternative hypotheses in terms of population parameters or in terms of treatment effects?
(d) Comment on whether you believe the technical conditions for this procedure are met.

60. Draft Lottery

In 1970 the United States conducted a lottery to determine which young men born in 1952 (a leap year) would be drafted to serve in the armed forces. Each of the 366 birthdates of the year was assigned a draft number; the lower the draft number, the earlier someone born on that date would be drafted. The draft numbers of the 366 birthdates, numbered sequentially from 1–366, are in the file DraftLottery.txt.

(a) Construct a scatterplot of draft number vs. sequential date. Does it appear to reveal a random scattering of points? Is this what should result from a fair lottery? Explain.
(b) Calculate the correlation coefficient between draft number and sequential date, along with its p-value.
(c) Does the p-value reveal evidence of a strong association, or does it reveal strong evidence of an association? (State the hypotheses that are being tested.) Explain.
(d) What is the probability that a fair, random lottery would produce a correlation coefficient as large as this lottery did by chance? What would you conclude about the hypothesis that this was a fair, random lottery?
(e) Calculate the mean draft number for each birth month. Produce a scatterplot of these means vs. month number. Comment on what this reveals that the original scatterplot did not.

61. Draft Lottery (cont.)

Reconsider the previous question about the draft lottery. An alternative analysis could categorize both variables. The birthdate variable could be classified into three groups: January-April, May-August, and September-December. The draft number variable could simply be classified as top third (1–122), middle third (123-244), and bottom third. The resulting two-way table of counts is:

	January – April	**May – August**	**September – December**	**Total**
1–122	29	45	48	122
123–244	42	28	52	122
245–366	50	50	22	122
Total	121	123	122	366

(a) Calculate conditional proportions and construct a segmented bar graph to display the conditional distribution of draft number in each birth month category. Summarize what the graph reveals about a possible association.
(b) Conduct a chi-square analysis of the table. State hypotheses, sketch the sampling distribution, check technical conditions, and calculate the test statistic and p-value. (Feel free

to use technology.) Interpret what the p-value reveals about how often a fair lottery would produce such extreme results by chance.

(c) Which of the nine cells in the table contributed the most to the calculation of the test statistic value? Were the observed counts higher or lower than expected in those cells? Interpret what this reveals about the association between birth month and draft number.

(d) Are the results of this chi-square analysis consistent with the results of the correlation analysis in the earlier question? Explain.

62. Draft Lottery (cont.)

Reconsider again the exercises about the 1970 draft lottery. Consider yet another way to analyze these data: by assessing whether the mean draft numbers in the 12 months differ more than would be expected by chance from a fair, random lottery.

(a) What procedure conducts such an assessment (of whether group means differ significantly)?

(b) Use technology (DraftLottery.txt) to conduct an ANOVA to address this question. Report the ANOVA table, *F* statistic, and p-value. Interpret the p-value and summarize your conclusion about whether the mean draft numbers differ significantly across the 12 months.

(c) How do the p-values from these three tests (correlation, chi-square, ANOVA) compare? Do they all lead to similar conclusions? Explain.

63. Transforming the Regression Slope

In this Exercise, you will consider the interpretation of the regression slope coefficient when you take log transformations. Suppose we have a regression equation of the form:

$$\log\hat{g}_{10}(y) = b_0 + b_1 x.$$

The back-transformation would imply $y = 10^{b_0}10^{b_1 x}$.

(a) Write out the expressions for $\hat{y}$ at x and at $x + 1$. What is the relationship between these two values? [*Hint:* Think ratio.]

(b) Use your answer to (a) to suggest an interpretation for the slope coefficient b_1 in this scenario.

In making this interpretation, we need to remember that the linear regression model is really making statements about $E(Y$ at each $x)$. So the left-hand side of our first equation is the prediction for $E(\log(Y)$ at each $x)$.

(c) In Investigation 2.8 what did you find about the equality of the mean of the log(variable) versus the log of the mean of the variable? [*Hint*: The same properties apply to log base 10 and to natural logs.]

(d) In Investigation 2.8, what did you find about the equality of the median of the log(variable) versus the log of the median of the variable?

(e) Use your answers to (c) and (d) to suggest why the interpretation of the slope in this case should be about the predicted change in the median response value instead of the mean response value.

(f) Suppose we have a regression equation for the form $\log\hat{g}_{10}(y) = b_0 + b_1 \log_{10}(x)$. Suggest an appropriate interpretation of the coefficient b_1. [*Hint*: Instead of considering an increase in x of one unit, what happens if our new value of x is 10 times the old value of x?]

64. The 2000 Presidential Election

The 2000 U.S. Presidential election is infamous for the close outcome and the confusing "butterfly ballot" which some voters claimed led them to inadvertently select a candidate for whom they did not intend to vote (Pat Buchanan instead of Al Gore). In particular, Palm Beach County recorded an unusually large number of Buchanan votes. One way to examine evidence for this claim is to look at the number of votes given to the candidates in different counties.

(a) Open the 2000pres.txt worksheet. Produce and interpret graphical and numerical summaries of the relationship between the number of votes for Buchanan in the different counties and the number of votes for Bush in these same counties.
(b) Use technology to determine the least-squares line for predicting the number of Buchanan votes from the number of Bush votes. Provide an interpretation of each coefficient in this context. Also report and interpret the value of r^2.
(c) Is there preliminary evidence that Buchanan received more votes than expected in Palm Beach County? Explain. Does Palm Beach County have the largest *residual* value?
(d) Remove Palm Beach County from the data set and recalculate the correlation coefficient, the least squares line and r^2. Do these measures appear to be *resistant*? Explain.
(e) Find the observation for Dade County. In what way is Dade County unusual in this data set?
(f) Remove Dade County from the data set and recalculate the correlation coefficient, the least squares line and r^2. Would you say that Dade County was *influential* in the calculations of these values? Explain.
(g) Now remove Sarasota County and recalculate the correlation coefficient, the least squares line and r^2. Would you say that Sarasota County was an influential observation?

65. Televisions and Life Expectancy

The file tvlife.txt contains data on the forty largest countries in the world (according to 1990 population figures) from *The World Almanac and Book of Facts 1993*. Column 3 is the country's life expectancy at birth and Column 4 is the number of people per television set.

(a) Fit the regression model for predicting the life expectancy of a country from the number of people per television. How much of the variation in life expectancies is explained by this regression on number of people per television?
(b) Examine the residual plots, is the basic regression model appropriate for these data?
(c) Transform the people per television variable by taking the log base 10 (remember to name the column). Does this relationship between life expectancy and $\log_{10}$(people per television) appear to be linear?
(d) Fit the regression model for predicting the life expectancy of a country from the log(people per television). Use the model to predict the life expectancy for a country with 10 people per television and for 100 people per television. What is the difference between these two predicted values?
(e) Is the basic regression model appropriate for these data? Is the relationship statistically significant?
(f) Do your answers to (e) imply that sending televisions to countries such as Angola and Haiti will improve the life expectancy in those countries? Explain. If not, suggest a potential confounding variable.

66. Income and Back Pain

A recent study (Center on an Aging Society, March 2003) examined the relationship between median annual earnings of adults and work limitations due to back pain. The following two-way table displays the results for the 18–44 years in the study.

	Difficulties at work due to back pain	Have not experienced difficulties at work
<5k	21	688
5K-10K	21	1047
10K-15K	19	1303
15K-20K	16	1822
20K-25K	10	1800
25K-35K	23	3436
35K-50K	17	3314
50K+	9	2585

(a) Produce numerical and graphical summaries to describe the relationship in this sample.
(b) Is there evidence of a significant association between annual earnings and whether or not 18-44 year-olds experience difficulties at work due to back pain? Explain.

67. Teaching Morals (cont.)

Recall the study from the Chapter 3 Exercises on which children's story was more effective in teaching Canadian children aged 3 – 7 years to tell the truth if they committed an infraction (peeking at a toy when the researcher left the room). Below are the results for the how often children peeked for each age group.

Table 1. Percentage of Children Who Peeked at the Target Toy in Experiments 1 and 2

Experiment and age group	Percentage of children
Experiment 1	
3-year-olds	88 (42 out of 48)
4-year-olds	81 (50 out of 62)
5-year-olds	70 (38 out of 54)
6-year-olds	62 (29 out of 47)
7-year-olds	68 (39 out of 57)

(a) Is there convincing evidence that the likelihood of peeking differs across the five age groups? [*Hints*: Use both a simulation-based analysis and a theory-based test and compare the results.]
(b) Are the validity conditions for the theory-based test met here? Justify your answer.
(c) What feature of the association do you see in the data that is not captured by your analysis?

NEW EXERCISES

68. 2015 Wimbledon (cont.)
Reconsider the data from the 2015 Wimbledon Championships on distance covered for 21 matches on Day 1 of the tournament (WimbledonDistance2015.txt).
(a) Does there appear to be a relationship between the difference in the distance covered by the winners and losers and the duration (in minutes) of the match? Justify your answer.
(b) Identify and remove the outlier and repeat your analysis from (a). Does the observation appear to be *influential*? Explain.

69. Keeping the weight off?
An *Associated Press* article (October 2007) describes an experiment in which 291 people who had lost at least 10% of their body weight in a medical weight loss program were assigned at random to one of three groups for follow-up. After 18 months, participants in each group were classified according to whether or not they had regained more than 5 pounds.

		Met monthly in person	Met online monthly in a chat room	Received a monthly newsletter by mail	Total
Regained > 5 lbs?	**Yes**	45	53	70	168
	No	52	44	27	123
	Total	97	97	97	291

(a) Identify the observational units.
(b) Identify the explanatory and response variables, and their type (categorical or quantitative).
(c) Is this an observational study or an experiment? How are you deciding?
(d) State the relevant hypotheses, in words, in the context of the study.
(e) Enter the two-way table (with appropriate one-word row and column labels) into the Analyzing Two-way Tables applet. Using the MAD as the statistic, carry out a randomization test to estimate a p-value for this test. Include all relevant output.
(f) Using the Chi-square statistic as the statistic, carry out a randomization test to estimate a p-value. Include all relevant output.
(g) Overlay the chi-square distribution of your simulated null distribution (include screen capture). Does the model appear appropriate? Is this what you would expect based on the two-way table? Explain.
(h) Use technology to carry out a chi-square test. Include your output, including the degrees of freedom, test statistic, and p-value.
(i) State your conclusion in the context of the study. Be sure to comment on statistical significance, causation, and generalizability, and how you are deciding on each of those aspects.
(j) Is there a particular follow-up group type that you would recommend to increase the likelihood of keeping weight off after losing weight in a weight loss program? Explain how you are deciding, and support your explanation with appropriate numbers.