is limited to handling censoring for data problems like this one, in which a number of censored observations are tied for last.

4.7 EXERCISES

Conceptual Exercises

1. Cognitive Load. Suppose that there were two textbooks on coordinate geometry, one written with conventional worked problems and the other with modified worked problems. And suppose it is possible to identify a number of schools in which each of the textbooks is used. If you took random samples of size 14 from schools with each text and obtained exactly the same data as in the example in Section 4.1.2, would the analysis and conclusions be the same?

2. O-Ring Data. (a) Is it appropriate to use the two-sample t-test to compare the number of O-ring incidents in cold and warm launches? (b) Is it appropriate to use the rank-sum test? (c) Is it appropriate to use a permutation test based on the t-statistic?

3. O-Ring Data. When these data were analyzed prior to the *Challenger* disaster it was noticed that variability was greater in the group with the larger average, so a log transformation was used. Since the log of zero does not exist, all the zeros were deleted from the data set. Does this seem like a reasonable approach?

4. O-Ring Data. Explain why the two-sided p-value from the permutation test applied to the O-ring data is equal to the one-sided p-value (see Display 4.10).

5. O-Ring Data. If you looked at the source of the O-ring data and found that temperatures for each launch were recorded in degrees F rather than as over/under 65°F, what question would that raise? Would the answer affect your conclusions about the analysis?

6. Motivation and Creativity. In what way is the p-value for the motivation and creativity randomized experiment (Section 1.1.1) dependent on an assumed model?

7. Are there occasions when both the two-sample t-test and the rank-sum test are appropriate?

8. Can the rank-sum test be used for comparing populations with unequal variances?

9. Suppose that two drugs are both effective in prolonging length of life after a heart attack. Substantial statistical evidence indicates that the mean life length for those using drug A is greater than the mean life length for those using drug B, but the variation of life lengths for drug A is substantially greater as well. Explain why it is difficult to conclude that drug A is better even though the mean life length is greater.

10. In a certain problem, the randomization test produces an exact two-sided p-value of 0.053, while the t-distribution approximation produces 0.047. One might say that since the p-values are on opposite sides of 0.05, they lead to quite different conclusions and, therefore, the approximation is not adequate. Comment on this statement.

11. What is the difference between a permutation test and a randomization test?

12. Explain what is meant by the comment that there is no single test called a randomization test.

13. What confounding factors are possible in the O-ring failure problem?

Computational Exercises

14. O-Ring Study. Find the t-distribution approximation to the p-value associated with the observed t-statistic. Compare this approximation to the (correct) permutation test p-value.

15. Consider these artificial data:

Group 1:	1	5	
Group 2:	4	8	9

The difference in averages $\overline{Y}_1 - \overline{Y}_2$ is -4. What is a one-sided p-value from the permutation distribution of the *difference in averages*? (*Hint*: List the 10 possible groupings; compute the difference in averages for each of these groupings, then calculate the proportion of these less than or equal to -4.)

16. Consider these artificial data:

Group 1:	5	7	12
Group 2:	4	6	

Calculate a p-value for the hypothesis of no difference, using the permutation distribution of the difference in sample averages. (You do not need to calculate the t-statistic for each grouping, only the difference in averages.)

17. O-Ring Study. Suppose the O-ring data had actually turned out as shown in Display 4.13. These are the same 24 numbers as before, but with the 2 and 3 switched. What is the one-sided p-value from the permutation test applied to the t-statistic? (This can be answered by examining Display 4.10.)

DISPLAY 4.13 Hypothetical O-ring data

Launch temperature	Number of O-ring incidents
Below 65°F	1 1 1 2
Above 65°F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 3

18. Suppose that six persons have an illness. Three are randomly chosen to receive an experimental treatment, and the remaining three serve as a control group. After treatment, a physician examines all subjects and assigns ranks to the severity of their symptoms. The patient with the most severe condition has rank 1, the next most severe has rank 2, and so on up to 6. It turns out that the patients in the treatment group have ranks 3, 5, and 6. The patients in the control group have ranks 1, 2, and 4. Is there any evidence that the treatment has an effect on the severity of symptoms? Use the randomization distribution of the sum of ranks in the treatment group to obtain a p-value. (First find the sum of ranks in the treatment group. Then write down all 20 groupings of the 6 ranks; calculate the sum of ranks in the treatment group for each. What proportion of these give a rank-sum as large as or larger than the observed one?)

19. Bumpus's Study. Use a statistical computer program to perform the rank-sum comparison of humerus lengths in the sparrows that survived and the sparrows that perished (Exercise 2.21). (a) What is the two-sided p-value? (b) Does the statistical computer package report the exact p-value or the one based on the normal approximation? (c) If it reports the one using the normal approximation, does it use a continuity correction to the Z-statistic? (d) How does the p-value from the rank-sum test compare to the one from the two-sample t-test (0.08) and the one from the two-sample t-test when the smallest observation is set aside (see Chapter 3, Exercise 28)? (e) Explain the relative merits of (i) the two-sample t-test using the strategy for dealing with outliers and (ii) the rank-sum test.

20. Trauma and Metabolic Expenditure. For the data in Exercise 18 in Chapter 3: (a) Determine the rank transformations for the data. (b) Calculate the rank-sum statistic by hand (taking the trauma

patients to be group 1.) (c) Mimic the procedures used in Display 4.5 and Display 4.7 to compute the Z-statistic. (d) Find the one-sided p-value as the proportion of a standard normal distribution larger than the observed Z-statistic.

21. Trauma and Metabolic Expenditure. Use a statistical computer package to verify the rank-sum and the Z-statistic obtained in Exercise 20. Is the p-value the same? (Does the statistical package use a continuity correction?)

22. Trauma and Metabolic Expenditure. Using the rank-sum procedure, find a 95% confidence interval for the difference in population medians: the median metabolic expenditure for the population of trauma patients minus the median metabolic expenditure for the population of nontrauma patients.

23. Motivation and Creativity. Use a statistical computer package to compute the randomization test's two-sided p-value for testing whether the treatment effect is zero for the data in Section 1.1.1 (file case0101). How does this compare to the results from the two-sample t-test (which is used as an approximation to the randomization test)?

24. Motivation and Creativity. Find a 95% confidence interval for the treatment effect (poem creativity score after intrinsic motivation questionnaire minus poem creativity score after extrinsic motivation questionnaire, from Section 1.1.1 (file case0101)) using the rank-sum procedure. (Use a statistical computer program.) How does this compare to the t-based confidence interval for the treatment effect?

25. Guinea Pig Lifetimes. Use the Welch t-tools to find a two-sided p-value and confidence interval for the effect of treatment on lifetimes of guinea pigs in Chapter 2, Exercise 11. Does the additive treatment effect seem like a sensible model for these data?

26. Schizophrenia Study. (a) Draw a histogram of the differences in hippocampus volumes in Display 4.12. Is there evidence that the population of differences is skewed? (b) Take the logarithms of the volumes for each of the 30 subjects, take the differences of the log volumes, and draw a histogram of these differences. Does it appear that the distribution of differences of log volumes is more nearly symmetric? (c) Carry out the paired t-test on the log-transformed volumes. How does the two-sided p-value compare with the one obtained on the untransformed data? (d) Find an estimate of and 95% confidence interval for the mean difference in log volumes. Back-transform these to get an estimate and confidence interval for the median of the population of ratios of volumes.

27. Schizophrenia Study. Find the two-sided p-value using the signed-rank test, as in Display 4.12, but after taking a log transformation of the hippocampus volumes. How does the p-value compare to the one from the untransformed data? Is it apparent from histograms that the assumptions behind the signed-rank test are more appropriate on one of these scales?

28. Darwin Data. Charles Darwin carried out an experiment to study whether seedlings from cross-fertilized plants tend to be superior to those from self-fertilized plants. He covered a number of plants with fine netting so that insects would be unable to fertilize them. He fertilized a number of flowers on each plant with their own pollen and he fertilized an equal number of flowers on the same plant with pollen from a distant plant. (He did not say how he decided which flowers received which treatments.) The seeds from the flowers were allowed to ripen and were set in wet sand to germinate. He placed two seedlings of the same age in a pot, one from a seed from a self-fertilized flower and one from a seed from a cross-fertilized flower. The data in Display 4.14 are the heights of the plants at certain points in time. (The fertilization experiments were described by Darwin in an 1878 book; these data were found in D. F. Andrews and A. M. Herzberg, *Data* (New York: Springer–Verlag, 1985), pp. 9–12.) (a) Draw a histogram of the differences. (b) Find a two-sided p-value for the hypothesis of no treatment effect, using the paired t-test. (c) Find a 95% confidence interval for the additive treatment effect. (d) Is there any indication from the plot in (a) that the paired

DISPLAY 4.14 Darwin's data: heights (inches) for 15 pairs of plants of the same age, one of which was grown from a seed from a cross-fertilized flower and the other of which was grown from a seed from a self-fertilized flower, first 5 of 15 rows

	Plant height (inches)	
Pair	**Cross-fertilized**	**Self-fertilized**
1	23.5	17.375
2	12	20.375
3	21	20
4	22	20
5	19.125	18.375

t-test may be inappropriate? (e) Find a two-sided p-value for the hypothesis of no treatment effect for the data in Display 4.14, using the signed-rank test.

Data Problems

29. Salvage Logging. When wildfires ravage forests, the timber industry argues that logging the burned trees enhances forest recovery. The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2004—before any logging was done—and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. (Data from D.C. Donato et al., 2006. "Post-Wildfire Logging Hinders Regeneration and Increases Fire Risk," *Science*, 311: 352.) The numbers of seedlings in the logged (L) and unlogged (U) plots are shown in Display 4.15.

DISPLAY 4.15 Number of tree seedlings per transect in nine logged (L) and seven unlogged (U) plots affected by the Biscuit Fire, in 2004 and 2005, and the percentage of seedlings lost between 2004 and 2005; first 5 of 16 rows

Plot	Action	Seedlings2004	Seedlings2005	PercentLost
Plot 1	L	298	164	45.0
Plot 2	L	471	221	53.1
Plot 3	L	767	454	40.8
Plot 4	L	576	141	75.5
Plot 5	L	407	217	46.7

Analyze the data to see whether logging has any effect on the distribution of percentage of seedlings lost between 2004 and 2005, possibly using the following suggestions: (a) Use the rank-sum procedure to test for a difference between logged and unlogged plots. Also use the procedure to construct a 95% confidence interval on the difference in medians. (b) Use the t-tools to test for differences in mean percentages lost and to construct a 95% confidence interval. Compare the results with those in (a).

DISPLAY 4.16 Tolerance to sunlight (minutes) for 13 patients prior to treatment and after treatment with a sunscreen; first 5 of 13 rows

	Tolerance to sunlight (minutes)	
Patient	Pretreatment	During treatment
1	30	120
2	45	240
3	180	480
4	15	150
5	200	480

DISPLAY 4.17 Months of survival after beginning of study for 58 breast cancer patients

Control Patients ($n = 24$)
2, 6, 8, 10, 12, 12, 14, 14, 14, 16, 16, 16, 18, 18, 18, 20, 22, 22, 26, 34, 36, 38, 40, 48

Patients Given Group Therapy for One Year ($n = 34$)
2, 2, 4, 4, 4, 6, 6, 8, 10, 10, 12, 14, 16, 16, 16, 18, 20, 22, 32, 36, 46, 46, 48, 48, 58, 58, 66, 72, 72, 82, 122, 122*, 122*, 122*

*These three patients were still alive at the end of the 122-month study period.

30. Sunlight Protection Factor. A sunscreen sunlight protection factor (SPF) of 5 means that a person who can tolerate Y minutes of sunlight without the sunscreen can tolerate $5Y$ minutes of sunlight with the sunscreen. The data in Display 4.16 are the times in minutes that 13 patients could tolerate the sun (a) before receiving treatment and (b) after receiving a particular sunscreen treatment. (Data from R. M. Fusaro and J. A. Johnson, "Sunlight Protection for Erythropoietic Protoporphyria Patients," *Journal of the American Medical Association* 229(11) (1974): 1420.) Analyze the data to estimate and provide a confidence interval for the sunlight protection factor. Comment on whether there are any obvious potentially confounding variables in this study.

31. Effect of Group Therapy on Survival of Breast Cancer Patients. Researchers randomly assigned metastatic breast cancer patients to either a control group or a group that received weekly 90-minute sessions of group therapy and self-hypnosis, to see whether the latter treatment improved the patients' quality of life. The group therapy involved discussion and support for coping with the disease, but the patients were not led to believe that the therapy would affect the progression of their disease. Surprisingly, it was noticed in a follow-up 10 years later that the group therapy patients appeared to have lived longer. The data on number of months of survival after beginning of the study are shown in Display 4.17. (Data from a graph in D. Spiegel, J. R. Bloom, H. C. Kraemer, and E. Gottheil, "Effect of Psychosocial Treatment on Survival of Patients with Metastatic Breast Cancer," *Lancet* (October 14, 1989): 888–91.) Notice that three of the women in the treatment group were still alive at the time of the follow-up, so their survival times are only known to be larger than 122 months. Is there indeed evidence of an effect of the group therapy treatment on survival time and, if so, how much more time can a breast cancer patient expect to live if she receives this therapy? Analyze the data as best as possible and write a brief report of the findings.

32. Therapeutic Marijuana. Nausea and vomiting are frequent side effects of cancer chemotherapy, which can contribute to the decreased ability of patients to undergo long-term chemotherapy

DISPLAY 4.18 Number of vomiting and retching episodes for 15 chemotherapy-receiving cancer patients, under placebo and marijuana treatments; first 5 of 15 rows

	Total number of vomiting and retching episodes	
Subject number	Marijuana	Placebo
1	15	23
2	25	50
3	0	0
4	0	99
5	4	31

schedules. To investigate the capacity of marijuana to reduce these side effects, researchers performed a double-blind, randomized, crossover trial. Fifteen cancer patients on chemotherapy schedules were randomly assigned to receive either a marijuana treatment or a placebo treatment after their first three chemotherapy sessions, and then "crossed over" to the opposite treatment after their next three sessions. The treatments, which involved both cigarettes and pills, were made to appear the same whether in active or placebo form. Shown in Display 4.18 are the number of vomiting and retching episodes for the 15 subjects. Does marijuana treatment reduce the frequency of episodes? By how much? Analyze the data and write a statistical summary of conclusions. (Data from A. E. Chang et al., "Delta-9-Tetrahydrocannabinol as an Antiemetic in Cancer Patients Receiving High-Dose Methotrexate," *Annals of Internal Medicine*, Dec. 1979. The order of the treatments is unavailable.)

Answers to Conceptual Exercises

1. The analysis would be the same. The conclusions would be very different. You could infer a real difference in the solution times of the two groups, but you could not attribute it to the different text types because of the possibility of a host of confounding factors.

2. (a) No, the extent of the nonnormality in these small and unequally sized samples is more than can be tolerated by the two-sample t-test. (b) Probably not. The spreads apparently are not equal. (c) Yes. The permutation test for significance requires no model or assumptions.

3. No! Observations cannot be deleted simply because the transformation does not work on them. In this case, a major portion of the data was deleted, leaving a very misleading picture.

4. There are 105 groupings that lead to t-statistics greater than or equal to 3.888 and no groupings that lead to t-statistics less than or equal to -3.888.

5. Was the 65°F cutoff chosen to maximize the apparent difference in the two groups? If so, the p-value would not be correct. Why? Because the p-value represents the chance of getting evidence as damaging to the hypothesis when there is no difference. The "chance" is the frequency of occurrence in replications of the study, *using the same statistical analysis*. The p-value calculation assumes that the 65°F cutoff will always define the groups. If the choice of cutoff was part of the statistical analysis used on this data set, the calculation was not correct. To get a correct p-value would require that you allow for a different cutoff to be chosen in each replication. Further discussion of data snooping is given in Chapter 6.

6. The p-value is based on the two-sample t-test but it is now understood that this p-value serves as an approximation to the p-value from the exact randomization test. For this approximation to be valid, the histograms of creativity scores should be reasonably normal (which they are).

7. Yes. Since the population model for the t-tools requires that the populations be normal with equal spread they will necessarily have the same shape and spread. Therefore, the assumptions for the rank-sum test are also satisfied.

8. Yes, but the meaning may be unclear if the variances are substantially unequal.

9. Generally, it is hard to make use of the difference in the centers of two distributions when the spreads are quite different. Specifically, the mean life length for drug A may be longer, but more people who use it may die sooner than for drug B.

10. The statement takes the rejection region approach too literally. There is very little difference in the degree of evidence against the null hypothesis in p-values of 0.053 and 0.047, so the approximation is pretty good.

11. A randomization test is a permutation test applied to data from a randomized experiment.

12. There is a different permutation distribution for each statistic that can be calculated from the data.

13. Perhaps workers tended to make more mistakes in cold weather or wind stress was greater on days with cold weather.

CHAPTER 5

Comparisons Among Several Samples

The issues and tools associated with the analysis of three or more independent samples (or treatment groups in a randomized experiment) are very similar to those for comparing two samples. Notable differences, however, stem from the particular kinds of questions and the greater number of them that may be asked.

An initial question, often asked in a preliminary stage of the analysis, is whether all of the means are equal. An easy-to-use F-test is available for investigating this. A typical analysis, however, goes beyond the F-test and explores linear combinations of the means to address particular questions of interest. A simple example of a linear combination of means is $\mu_3 - \mu_1$, the difference between means in populations 3 and 1. Inferences about this parameter can be made with t-tools just as for the two-independent-sample problem, with the important difference that the pooled estimate of standard deviation is from all groups, not just from those being compared.

This chapter discusses the use of the pooled estimate of variance for specific linear combinations of means and the one-way analysis of variance F-test for testing the equality of several means. The next chapter looks at linear combinations more generally and the problem of compounded uncertainty from the multiple, simultaneous comparisons of means.