3.6.2 Robustness and Transformation for Paired *t*-Tools

The one-sample t-test, of which the paired t-test is a special case, assumes that the observations are independent of one another and come from a normally distributed population. P-values and confidence intervals remain valid for moderate and large sample sizes for nonnormal distributions. For smaller sample sizes skewness can be a problem. When cluster or serial effects are present (see Section 3.2.4), the t-tools may give misleading results. When the observations within each pair are positive, either an apparent multiplicative treatment effect (in an experiment) or a tendency for larger differences in pairs with larger average values suggests the use of a log transformation. The transformation is applied before taking the difference, which is equivalent to forming a ratio within each pair and performing a one-sample analysis on the logarithms of the ratios. If there are n pairs, let $Z_i = \log(Y_{1i}) - \log(Y_{2i})$, which is the same as $\log(Y_{1i}/Y_{2i})$. In an observational study, $\exp(\overline{Z})$ is an estimate of the median of the ratios, Y_1/Y_2. (This is not the same as the ratio of the medians [see Exercise 20].) In a randomized, paired experiment, $\exp(\overline{Z})$ estimates a multiplicative treatment effect on the original scale. In both cases, the statistical work of testing and constructing a confidence interval is done on the log scale. The estimate and associated interval are transformed back to the original scale.

3.6.3 Example—Schizophrenia

In the schizophrenia example of Section 2.1.2, Z_i represents the logarithm of the left hippocampus volume of the unaffected twin divided by the left hippocampus volume of the affected twin in pair i. The average of the 15 log ratios is 0.1285. A one-sample analysis gives a p-value of 0.0065 for the test that the mean is zero and a 95% confidence interval from 0.0423 to 0.2147 for the mean itself. Taking antilogarithms of the estimate and the endpoints of the confidence interval yields the following conclusion: It is estimated that the median of the unaffected-to-affected volume ratios is 1.137. A 95% confidence interval for the median ratio is from 1.043 to 1.239.

3.7 SUMMARY

Cloud Seeding and Rainfall Study

The box plots of the rainfalls for seeded and unseeded days reveal that the two distributions of rainfall are skewed and that the distribution with the larger mean also has the larger variance. This is the situation where log-transformed data behave in accordance with the ideal model. A plot of the data after transformation confirms the adequacy of the transformation. The two-sample t-test can be used as an approximation to the randomization test, and the difference in averages (of log rainfall) can be back-transformed to provide a statement about a multiplicative treatment effect. In the example, it is estimated that the rainfall is 3.1 times as much when a cloud is seeded as when it is left unseeded.

Since randomization is used, the statistical conclusion implies that the seeding causes the increase in rainfall. Since the decision about whether to seed clouds is determined (in this case) by a random mechanism, and since the airplane crew is *blind* to which treatment they are administering, human bias can have had little influence on the result.

Agent Orange Study

Graphical analysis focuses attention on the possibly undue influence of two outliers, but analyses with and without the outliers reveal no such influence, so the t-tools are used on the entire data set. The form of the sampling from the populations of living Vietnam veterans and of other veterans is a major concern in accepting the reliability of the statistical analysis. Protocols for obtaining the samples have not been discussed here, except to note that random sampling is not being used. Conclusions based on the two-sample t-test are supplied, along with the caveat that there may be biases due to the lack of random sampling.

3.8 EXERCISES

Conceptual Exercises

1. Cloud Seeding. What is the experimental unit in the cloud seeding experiment?

2. Cloud Seeding. Randomization in the cloud seeding experiment was crucial in assessing the effect of cloud seeding on rainfall. Why?

3. Cloud Seeding. Why was it important that the airplane crew was unaware of whether seeding was conducted or not?

4. Cloud Seeding. Why would it be helpful to have the date of each observed rainfall?

5. Agent Orange. How would you respond to the comment that the box plots in Display 3.3 indicate that the dioxin levels in the Vietnam veterans tend to be larger since their values appear to be larger?

6. Agent Orange. (a) What course of action would you propose for the statistical analysis if it was learned that Vietnam veteran #646 (the largest observation in Display 3.6) worked for several years, after Vietnam, handling herbicides with dioxin? (b) What would you propose if this was learned instead for Vietnam veteran #645?

7. Agent Orange. If the statistical analysis had shown convincing evidence that the mean dioxin levels differed in Vietnam veterans and other veterans, could one conclude that serving in Vietnam was responsible for the difference?

8. Schizophrenia. In the schizophrenia study in Section 2.1.2, the observations in the two groups (schizophrenic and nonschizophrenic) are not independent since each subject is matched with a twin in the other group. Did the researchers make a mistake?

9. True or false? A statistical computer package will only print out a p-value or confidence interval if the conditions for its validity are met.

10. True or false? A sample histogram will have a normal distribution if the sample size is large enough.

11. A woman who has just moved to a new job in a new town discovers two routes to drive from her home to work. The first Monday, she flips a coin, deciding to take route A if it comes up heads and to take route B if it is tails. The following Monday, she will take the other route. The first Tuesday, she flips the coin again with the same plan. And so on for the first week. At the end of two weeks, she has traveled both routes five times and can compare their average commuting times. Why should she not use the t-tools for two independent samples? What should she use?

12. In which ways are the t-tools more robust for larger sample sizes than for smaller ones (i.e., robust with respect to normality, equal SDs, and/or independence)?

13. Fish Oil. Why is a log transformation inappropriate for the fish oil data in Exercise 1.12?

14. Will an outlier from a contaminating population be more consequential in small samples or large samples?

15. What would you suggest as an alternative estimate of the standard deviation of the difference in sample averages when it is clear that the two populations have different SDs? (Check the formula for the standard deviation of the sampling distribution of the difference in averages, in Display 2.6.)

16. A researcher has taken tissue cultures from 25 subjects. Each culture is divided in half, and a treatment is applied to one of the halves chosen at random. The other half is used as a control. After determining the percent change in the sizes of all culture sections, the researcher calculates the standard error for the treatment-minus-control difference using both the paired t-analysis and the two independent sample (Chapter 2) t-analysis. Finding that the paired t-analysis gives a slightly larger standard error (and gives only half the degrees of freedom), the researcher decides to use the results from the unpaired analysis. Is this legitimate?

17. Respiratory breathing capacity of individuals in houses with low levels of nitrogen dioxide was compared to the capacity of individuals in houses with high levels of nitrogen dioxide. From a sample of 200 houses of each type, breathing capacity was measured on 600 individuals from houses with low nitrogen dioxide and on 800 individuals from houses with high nitrogen dioxide. (a) What problem do you foresee in applying t-tools to these data? (b) Would comparing the average *household* breathing capacities avoid the problem?

18. Trauma and Metabolic Expenditure. The following data are metabolic expenditures for eight patients admitted to a hospital for reasons other than trauma and for seven patients admitted for multiple fractures (trauma). (Data from C. L. Long, et al., "Contribution of Skeletal Muscle Protein in Elevated Rates of Whole Body Protein Catabolism in Trauma Patients," *American Journal of Clinical Nutrition* 34 (1981): 1087–93.)

Metabolic Expenditures (kcal/kg/day)								
Nontrauma patients:	20.1	22.9	18.8	20.9	20.9	22.7	21.4	20.0
Trauma patients:	38.5	25.8	22.0	23.0	37.6	30.0	24.5	

(a) Is the difference in averages resistant? (*Hint*: What happens if 20.0 is replaced by 200?)

(b) Replacing each value with its rank, from the lowest to highest, in the combined sample gives

Metabolic Expenditures (kcal/kg/day)								
Nontrauma patients:	3	9	1	4.5	4.5	8	6	2
Trauma patients:	15	12	7	10	14	13	11	

Consider the average of the ranks for the trauma group minus the average of the ranks for the nontrauma group. Is this statistic resistant?

19. In each of the following data problems there is some potential violation of one of the independence assumptions. State whether there is a cluster effect or serial correlation, and whether the questionable assumption is the independence within groups or the independence between groups.

(a) Researchers interested in learning the effects of speed limits on traffic accidents recorded the number of accidents per year for each of 10 consecutive years on roads in a state with speed limits of 90 km/h. They also recorded the number of accidents for the next 7 years on the same roads after the speed limit had been increased to 110 km/hr. The two groups of measurements are the number of accidents per year for those years under study. (Notice that there is also a potential confounding variable here!)

(b) Researchers collected intelligence test scores on twins, one of whom was raised by the natural parents and one of whom was raised by foster parents. The data set consists of test scores for the two groups, boys raised by their natural parents and boys raised by foster parents.

(c) Researchers interested in investigating the effect of indoor pollution on respiratory health randomly select houses in a particular city. Each house is monitored for nitrogen dioxide concentration and categorized as being either high or low on the nitrogen dioxide scale. Each member of the household is measured for respiratory health in terms of breathing capacity. The data set consists of these measures of respiratory health for all individuals from houses with low nitrogen dioxide levels and all individuals from houses with high levels.

Computational Exercises

20. Means, Medians, Logs, Ratios. Consider the following tuitions and their natural logs for five colleges:

College	In-State	Out-of-State	Out/In Ratio	Log(In-State)	Log(Out-of-State)
A	\$1,000	\$ 3,000	3	6.9078	8.0064
B	\$4,000	\$ 8,000	2	8.2941	8.9872
C	\$5,000	\$ 30,000	6	8.5172	10.3090
D	\$8,000	\$ 32,000	4	8.9872	10.3735
E	\$40,000	\$ 40,000	1	10.5966	10.5966

(a) Find the average In-State tuition. Find the average log(In-State). Confirm that the log of the average is *not* the same as the average of the logs. (b) Find the median In-State tuition and the median of the logs of In-State tuitions. Verify that the log of the median *is* the same as the median of the logs. (c) Compute the median of the ratios. Compute the differences of logged tuitions—log(Out-of-State) minus log(In-State) and compute the median of these differences. Verify that the median of the differences (of log tuitions) is equal to the natural log of the median of ratios (aside from some minor rounding error).

21. Umpire Life Lengths. When an umpire collapsed and died soon after the beginning of the 1990 U.S. major league baseball season, there was speculation that the stress associated with that job poses a health risk. Researchers subsequently collected historical and current data on umpires to investigate their life expectancies (Cohen et al., " Life Expectancy of Major League Baseball Umpires," *The Physician and Sportsmedicine*, 28(5) (2000): 83–89). From an original list of 441 umpires, data were found for 227 who had died or had retired and were still living. Of these, dates of birth and death were available for 195. Display 3.10 shows several rows of a generated data set based on the study.

DISPLAY 3.10 First 4 rows (of 227) from the umpire data set (Observed is the known lifetime for those umpires who had died by the time of the study [for whom Censored = 0] and the current age of those who had not yet died [for whom Censored = 1]; Expected is the expected life length—from actuarial life tables—for individuals who were alive at the time the person first became an umpire)

Umpire	Observed life length (yr)	Censored (0 if dead)	Expected life length (yr)
1	63	0	70
2	69	0	71
3	58	0	71
4	61	1	70
...			

(a) Use a t-test and confidence interval (possibly after transformation) to investigate whether umpires had smaller observed life lengths than expected, using only those with known life lengths (i.e., for whom *Censored* = 0)

(b) What are the potential consequences of ignoring those 214 of the 441 umpires on the original list for whom data was unavailable?

(c) What are the potential consequences of ignoring those 32 umpires in the data set who had not yet died at the time of the study? (See, for example, the survival analysis techniques in S. Anderson et al., *Statistical Methods for Comparative Studies*, New York: Wiley, 1980.)

22. Voltage and Insulating Fluid. Researchers examined the time in minutes before an insulating fluid lost its insulating property. The following data are the breakdown times for eight samples of the fluid, which had been randomly allocated to receive one of two voltages of electricity:

Times (min) at 26 kV: 5.79 1579.52 2323.70

Times (min) at 28 kV: 68.8 108.29 110.29 426.07 1067.60

(a) Form two new variables by taking the logarithms of the breakdown times: $Y_1 = \log$ breakdown time at 26 kV and $Y_2 = \log$ breakdown time at 28 kV.

(b) By hand, compute the difference in averages of the log-transformed data: $\overline{Y}_1 - \overline{Y}_2$.

(c) Take the antilogarithm of the estimate in (b): $\exp(\overline{Y}_1 - \overline{Y}_2)$. What does this estimate? (See the interpretation for the randomized experiment model in Section 3.5.2.)

(d) By hand, compute a 95% confidence interval for the difference in mean log breakdown times. Take the antilogarithms of the endpoints and express the result in a sentence.

23. Solar Radiation and Skin Cancer. The data in Display 3.11 are yearly skin cancer rates (cases per 100,000 people) in Connecticut, with a code identifying those years that came two years after higher than average sunspot activity and those years that came two years after lower than average sunspot activity. (Data from D. F. Andrews and A. M. Herzberg, *Data*, New York: Springer-Verlag, 1985.) (a) Is there any reason to suspect that using the two independent sample t-test to compare skin cancer rates in the two groups is inappropriate? (b) Draw scatterplots of skin cancer rates versus year, for each group separately. Are any problems indicated by this plot?

24. Sex Discrimination. With a statistical computer program, reanalyze the sex discrimination data in Display 1.3 but use the log transformation of the salaries. (a) Draw box plots. (b) Find a p-value for comparing the distributions of salaries. (c) Find a 95% confidence interval for the ratio of population medians. Write a sentence describing the finding.

DISPLAY 3.11 Partial listing of Connecticut skin cancer rates (per 100,000 people) from 1938 to 1972, with solar code (1 if there was higher than average sunspot activity and 2 if there was lower than average sunspot activity two years earlier)

Year	Rate	Code
1938	0.8	2
1939	1.3	1
1940	1.4	1
1941	1.2	1
...		
1972	4.8	1

DISPLAY 3.12 Proportions of pollen removed and visit durations (in seconds) by 35 bumblebee queens and 12 honeybee workers; partial listing.

Bee	Type	Removed	Duration
1	queen	0.07	2
2	queen	0.10	5
3	queen	0.11	7
4	queen	0.12	11
...			
45	worker	0.78	51
46	worker	0.74	64
47	worker	0.77	78

25. Agent Orange. With a statistical computer program, reanalyze the Agent Orange data of Display 3.3 with and without the two largest dioxin levels in the Vietnam veterans group. Verify the one-sided p-values in bubble 2 of Display 3.7.

26. Agent Orange. With a statistical computer package, reanalyze the Agent Orange data of Display 3.3 after taking a log transformation. Since the data set contains zeros—for which the log is undefined—try the transformation log(dioxin + .5). (a) Draw side-by-side box plots of the transformed variable. (b) Find a p-value from the t-test for comparing the two distributions. (c) Compute a 95% confidence interval for the difference in mean log measurements and interpret it on the original scale. (*Note*: Back-transforming does not provide an exact estimate of the ratio of medians since 0.5 was added to the dioxins, but it does provide an approximate one.)

27. Pollen Removal. As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily. (Data from L. D. Harder and J. D. Thompson, "Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants," *American Naturalist* 133 (1989): 323–44.) Their data appear in Display 3.12.

(a) (i) Draw side-by-side box plots (or histograms) of the proportion of pollen removed by queens and workers. (ii) When the measurement is the proportion P of some amount, one useful transformation is $\log[P/(1-P)]$. This is the log of the ratio of the proportion removed to the proportion not removed. Draw side-by-side box plots or histograms on

this transformed scale. (iii) Test whether the distribution of proportions removed is the same or different for the two groups, using the t-test on the transformed data.

(b) Draw side-by-side box plots of duration of visit on (i) the natural scale, (ii) the logarithmic scale, and (iii) the reciprocal scale. (iv) Which of the three scales seems most appropriate for use of the t-tools? (v) Compute a 95% confidence interval to describe the difference in means on the chosen scale. (vi) What are relative advantages of the three scales as far as interpretation goes? (vii) Based on your experience with this problem, comment on the difficulty in assessing equality of population standard deviations from small samples.

28. Bumpus's Data. Obtain p-values from the t-test to compare humerus lengths for sparrows that survived and those that perished (Exercise 2.21), with and without the smallest length in the perished group (length = 0.659 inch). Do the conclusions depend on this one observation? What action should be taken if they do?

29. Cloud Seeding—Multiplicative vs. Additive Effects. On the computer, create a variable containing the rainfall amounts for only the unseeded days. (a) Create four new variables by adding 100, 200, 300, and 400 to each of the unseeded day rainfall amounts. Display a set of five box plots to illustrate what one might expect if the effect of seeding were additive. (b) Create four additional variables by multiplying each of the unseeded day rainfall amounts by 2, by 3, by 4, and by 5. Display a set of five box plots to illustrate what could be expected if the effect of seeding were multiplicative. (c) Which set of plots more closely resembles the actual data?

Data Problems

30. Education and Future Income. Display 3.13 shows the first five rows of a data set with annual incomes in 2005 of the subset of National Longitudinal Survey of Youth (NLSY79) subjects (described in Exercise 2.22) who had paying jobs in 2005 and who had completed either 12 or 16 years of education by the time of their interview in 2006. All the subjects in this sample were between 41 and 49 years of age in 2006. Analyze the data to describe the amount (or percent) by which the population distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education. (*Note*: The NLSY79 data set codes all incomes above $150,000 as $279,816. To make an exercise version that better matches the actual income distribution, those values have been replaced in the data set by computer-simulated values from a realistic distribution of incomes greater than $150,000.)

DISPLAY 3.13 Annual incomes in 2005 (in U.S. dollars) of 1,020 Americans who had 12 years of education and 406 who had 16 years of education by the time of their interview in 2006. "Subject" is a subject identification number, first 5 of 1,426 rows

Subject	Educ	Income2005
2	12	5,500
6	16	65,000
7	12	19,000
13	16	8,000
21	16	253,043

31. Education and Future Income II. The data file ex0331 contains a subset of the NLSY79 data set (see Exercise 30) with annual incomes of subjects with either 16 or more than 16 years of education. Analyze the data to describe the amount (or percent) by which the population distribution of incomes for those with more than 16 years of education exceeds the distribution for those with 16 years of education.

32. College Tuition. Display 3.14 shows the first five rows of a data set with 2011–2012 in-state and out-of-state tuitions for random samples of 25 private and 25 public, four-year colleges and universities in the United States. Analyze the data to describe (a) the extent to which out-of-state tuition is more expensive than in-state tuition in the population of public schools, (b) the extent to which private school in-state tuition is more expensive than public school in-state tuition, and (c) the extent to which private school out-of-state tuition is more expensive than public school out-of-state tuition. (Data sampled from College Board: http://www.collegeboard.com/student/ (11 July 2011).)

DISPLAY 3.14 In-state and out-of-state tuitions for 25 public and 25 private colleges and universities in the United States, first 5 of 50 rows

College	Type	InState	OutofState
Albany State University	public	$5,434	$17,048
Appalachian State University	public	$5,175	$16,487
Argosy University: Nashville	private	$19,596	$19,596
Brescia University	private	$18,140	$18,140
Central Connecticut State University	public	$8,055	$18,679

33. Brain Size and Litter Size. Display 3.15 shows relative brain weights (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2 and for 45 species of mammal whose average litter size is greater than or equal to 2. (These are part of a larger data set considered in Section 9.1.2.) What evidence is there that brain sizes tend to be different for the two groups? How big of a difference is indicated? Include the appropriate statistical measures of uncertainty in carefully worded sentences to answer these questions.

DISPLAY 3.15 Relative brain sizes, 1,000 × (Brain weight/Body weight), for 96 species of mammals

1,000 × (Brain weight/Body weight) for 51 species with average litter size < 2

0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63	1.73	2.17	2.42
2.48	2.74	2.74	2.79	2.90	3.12	3.18	3.27	3.30	3.61	3.63	4.13
4.40	5.00	5.20	5.59	7.04	7.15	7.25	7.75	8.00	8.84	9.30	9.68
10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15	14.27	14.56	15.84
18.55	19.73	20.00									

1,000 × (Brain weight/Body weight) for 45 species with average litter size ≥ 2

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56	2.58	3.24	3.39
3.53	3.77	4.36	4.41	4.60	4.67	5.39	6.25	7.02	7.89	7.97	8.00
8.28	8.83	8.91	8.96	9.92	11.36	12.15	14.40	16.00	18.61	18.75	19.05
21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	36.35			

Answers to Conceptual Exercises

1. The target clouds on a day that was deemed suitable for seeding.

2. Uncontrollable confounding factors probably explain the variability in rainfall from clouds treated the same way. Randomization is needed to ensure that the confounding factors do not tend to be unevenly distributed in the two groups.

3. Blinding prevents the intentional or unintentional biases of the human investigators from having a chance to make a difference in the results.

4. There may be serial correlation. A plot of rainfall versus date could be used to check.

5. Larger values are to be expected by chance if the populations are the same, since the sample of Vietnam veterans is so much larger than the sample of non–Vietnam veterans.

6. (a) He would not be representative of the target population and should be removed from the data set for analysis. (b) Same thing.

7. No, not from the statistics alone since this is an observational study. It could be said, however, that the data are consistent with that theory.

8. No. The dependence is the result of matching and is desirable. The two-sample t-tools are not appropriate (but the paired t-tools are).

9. False.

10. False. An *average* from a sample will have a sampling distribution that will tend toward normal with large sample sizes, but the sample histogram should mirror the population distribution. As the sample size gets larger, the sample histogram should become a better approximation to the population histogram.

11. There is a cluster effect: the particular day of the week. She should use a paired-t analysis, as will be discussed in Chapter 4.

12. The t-test is robust in validity to departures from normality, especially as the sample size gets large. The robustness with respect to equal standard deviations does not depend much on what the sample sizes are, so long as they are reasonably equal. Sample size does not affect robustness with respect to independence.

13. You cannot take logarithms of negative numbers.

14. It will be more consequential in smaller samples; its effect gets washed out in large ones.

15. Replace the population SDs in the formula (Section 2.2.2) by individual *sample* SDs.

16. No. The paired analysis must be used, even though the inferences may not appear to be as precise. The unpaired analysis is inappropriate.

17. (a) Dependence of measurements on individuals in the same household (cluster effect). (b) Maybe. Getting a single measure for each household may be an easy way out of the dependence problem, but care should be used as these groups also tend to differ in the average number of persons per household.

18. (a) No. (b) Yes.

19. (a) Serial correlation both within and between groups. (Confounding variable is the time at which observations were made.) (b) Cluster effect between groups. (c) Cluster effect (members of the same household should be similar) within groups.

CHAPTER 4

Alternatives to the *t*-Tools

The t-tools have an extremely broad range of application, extending well beyond the strict confines of the ideal model because of robustness. They extend even further when the possibilities of transforming the data and dealing with outliers are taken into account.

Nevertheless, situations arise where the t-tools cannot be applied, because the model assumptions of the t-test are grossly violated. For these situations, a host of other methods, based on different models, may be used. Some are presented in this chapter. Most notable are two distribution-free methods, based on models that do not specify any particular population distributions. The rank-sum test for two independent samples and the signed-rank test for a sample of pairs are useful alternatives, particularly when outliers may be present or when the sample sizes are too small to permit the assessment of distributional assumptions.