CHAPTER 3

A Closer Look at Assumptions

Although statistical computer programs faithfully supply confidence intervals and p-values whenever asked, the human data analyst must consider whether the assumptions on which use of the tools is based are met, at least approximately. In this regard, an important distinction exists between the mathematical assumptions on which use of t-tools is exactly justified and the broader conditions under which such tools work quite well.

The mathematical assumptions—such as those of the model for two independent samples from normal populations with the same standard deviation—are never strictly met in practice, nor do they have to be. The two-sample t-tools are often valid even if the population distributions are nonnormal or the standard deviations are unequal. An understanding of the broader conditions, provided by advanced statistical theory and computer simulation, is needed to evaluate the appropriateness of the tools for a particular problem. After checking the actual conditions with graphical displays of the data, the data analyst may decide to use the standard tools, use them but apply the label "approximate" to the inferences, or choose an alternative approach.

An effective alternative is to apply the standard tools after transforming the data. A transformation is useful if the tools are appropriate for the conditions of the transformed data and if the questions of interest are answerable on the new scale. A particularly important transformation is the logarithm, which permits a convenient description of a multiplicative effect.

3.1 CASE STUDIES

3.1.1 Cloud Seeding to Increase Rainfall—A Randomized Experiment

The data in Display 3.1 were collected in southern Florida between 1968 and 1972 to test a hypothesis that massive injection of silver iodide into cumulus clouds can lead to increased rainfall. (Data from J. Simpson, A. Olsen, and J. Eden, "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification," *Technometrics* 17 (1975): 161–66.)

DISPLAY 3.1 Rainfall (acre-feet) for days with and without cloud seeding

Rainfall from Unseeded Days ($n=26$)								
1,202.6	830.1	372.4	345.5	321.2	244.3	163.0	147.8	95.0
87.0	81.2	68.5	47.3	41.1	36.6	29.0	28.6	26.3
26.0	24.4	21.4	17.3	11.5	4.9	4.9	1.0	
Rainfall from Seeded Days ($n=26$)								
2,745.6	1,697.1	1,656.4	978.0	703.4	489.1	430.0	334.1	302.8
274.7	274.7	255.0	242.5	200.7	198.6	129.6	119.0	118.3
115.3	92.4	40.6	32.7	31.4	17.5	7.7	4.1	

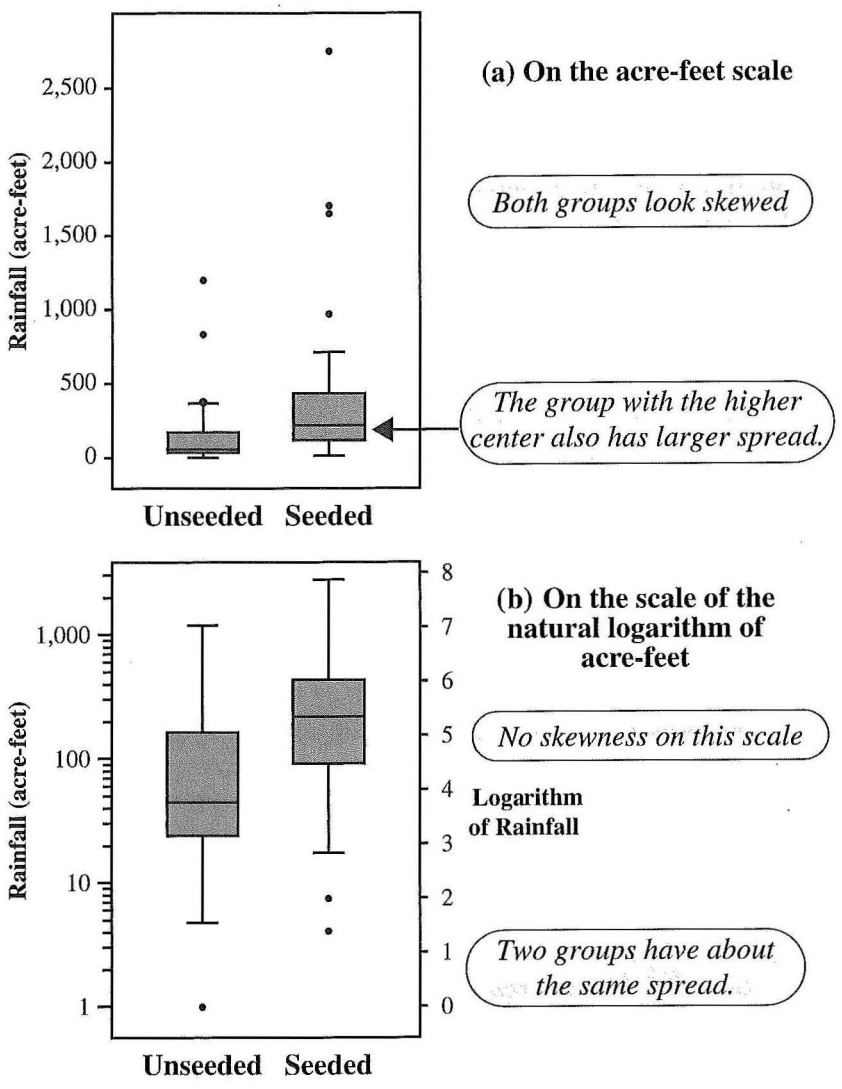
On each of 52 days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud on that day or to leave it unseeded as a control. An airplane flew through the cloud in both cases, since the experimenters and the pilot were themselves unaware of whether on any particular day the seeding mechanism in the plane was loaded or not (that is, they were *blind* to the treatment). Precipitation was measured as the total rain volume falling from the cloud base following the airplane seeding run, as measured by radar. Did cloud seeding have an effect on rainfall in this experiment? If so, how much?

Box plots of the data in Display 3.2(a) indicate that the rainfall tended to be larger on the seeded days. Both distributions were quite skewed, and more variability occurred in the seeded group than in the control group. The box plots in Display 3.2(b) are drawn from the same data, but on the scale of the natural logarithm of the rainfall measurements. On this scale, the measurements appear to have symmetric distributions, and the variation seems nearly the same.

Statistical Conclusion

It is estimated that the volume of rainfall on days when clouds were seeded was 3.1 times as large as when not seeded. A 95% confidence interval for this multiplicative effect is 1.3 times to 7.7 times. Since randomization was used to determine whether any particular suitable day was seeded or not, it is safe to interpret this as evidence that the seeding caused the larger rainfall amount.

DISPLAY 3.2 Box plots of rainfall amounts on original and transformed scales



3.1.2 Effects of Agent Orange on Troops in Vietnam—An Observational Study

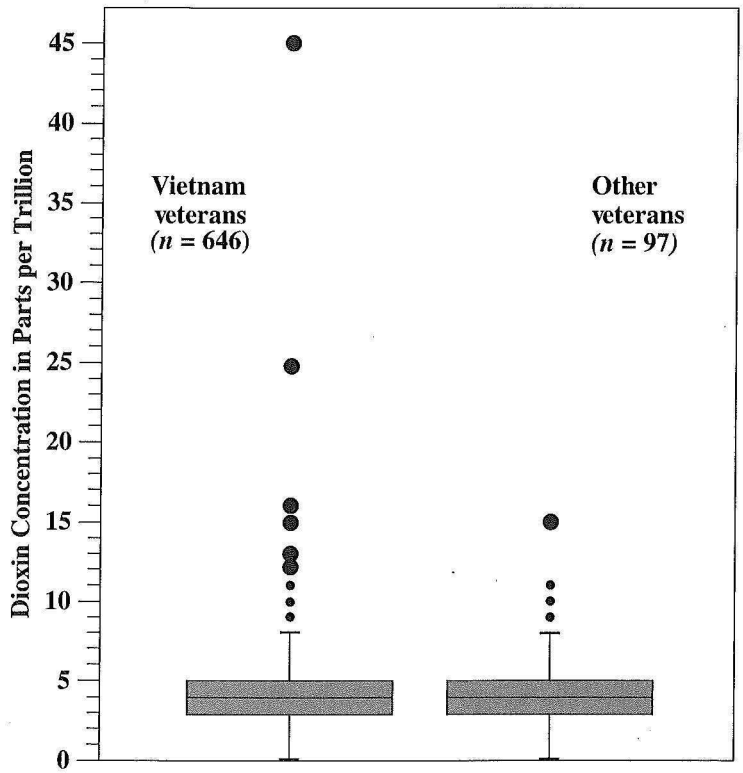
Many Vietnam veterans are concerned that their health may have been affected by exposure to Agent Orange, a herbicide sprayed in South Vietnam between 1962 and 1970. The particularly worrisome component of Agent Orange is a dioxin called TCDD, which in high doses is known to be associated with certain cancers. Studies have shown that high levels of this dioxin can be detected 20 or more years after

heavy exposure to Agent Orange. Consequently, as part of a series of studies, researchers from the Centers for Disease Control compared the current (1987) dioxin levels in living Vietnam veterans to the dioxin levels in veterans who did not serve in Vietnam.

The 646 Vietnam veterans in the study were a sample of U.S. Army combat personnel who served in Vietnam during 1967 and 1968, in the areas that were most heavily treated with Agent Orange. The 97 non–Vietnam veterans entered the Army between 1965 and 1971 and served only in the United States or Germany. Neither sample was randomly selected.

Blood samples from each veteran were analyzed for the presence of dioxin. Box plots of the observed levels are shown in Display 3.3. (Data from a graphical display in Centers for Disease Control Veterans Health Studies, "Serum 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin Levels in U.S. Army Vietnam-era Veterans," *Journal of the American Medical Association* 260 (September 2, 1988): 1249–54.) The question of interest is whether the distribution of dioxin levels tends to be higher for the Vietnam veterans than for the non–Vietnam veterans.

DISPLAY 3.3 Box plots of 1987 dioxin concentrations in 646 Vietnam veterans and 97 veterans who did not serve in Vietnam



Statistical Conclusion

These data provide no evidence that the mean dioxin level in surviving Vietnam combat troops is greater than that for non–Vietnam veterans (one-sided p-value = 0.40, from a two-sample t-test). A 95% confidence interval for the difference in population means is −0.48 to 0.63 parts per trillion.

Scope of Inference

Since the samples were not random, inference to the populations is speculative. Participating veterans may not be representative of their respective groups. For example, nonparticipating Vietnam veterans may have failed to participate because of dioxin-related illnesses. If so, statistical statements about the populations of interest could be seriously biased. It should also be noted that many Vietnam veterans are frustrated and insulted by the prevalence of weak Agent Orange studies, like this one, which appear to address the Agent Orange problem but which actually skirt the main health issues.

3.2 ROBUSTNESS OF THE TWO-SAMPLE *t*-TOOLS

3.2.1 The Meaning of Robustness

The two-sample t-tools were used in the analyses of the Agent Orange study and the cloud seeding study, even though the actual conditions did not seem to match the ideal models upon which the tools are based. In the cloud seeding study, the t-tools were applied after taking the logarithms of the rainfalls. The t-tools could be used for the Agent Orange study, despite a lack of normality in the populations, because of the robustness of the t-tools against nonnormality.

> *A statistical procedure is* ***robust to departures from a particular assumption*** *if it is valid even when the assumption is not met.*

Valid means that the uncertainty measures—the confidence levels and the p-values—are very nearly equal to the stated rates. For example, a procedure for obtaining a 95% confidence interval is valid if it is 95% successful in capturing the parameter. It is robust against nonnormality if it is roughly 95% successful with nonnormal populations.

Robustness of a tool must be evaluated separately for each assumption. The following sections detail the robustness of the two-sample t-tools against departures from the ideal assumptions of the normal, equal standard deviation model.

3.2.2 Robustness Against Departures from Normality

The *Central Limit Theorem* asserts that averages based on large samples have approximately normal sampling distributions, regardless of the shape of the population distribution. This suggests that underlying normality is not a serious issue, as long as sample sizes are reasonably large. The theorem provides only partial reassurance

of applicability with respect to t-tools. It states what the sampling distribution of an average should be, but it does not address the effects of estimating a population standard deviation. Many empirical investigations and related theory, however, confirm that the t-tools remain reasonably valid in large samples, with many nonnormal populations.

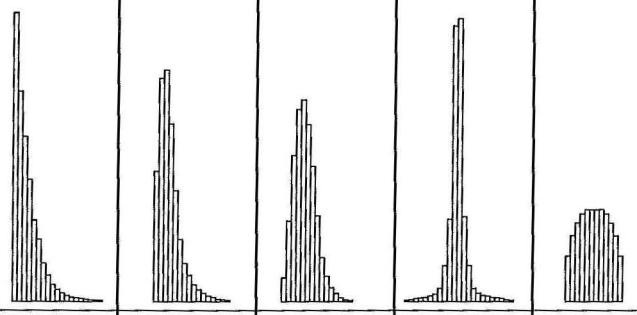
How large is large enough? That depends on how nonnormal the population distributions are. Because distributions can differ from the normal in infinitely many ways, the question of sample size is difficult to answer. Statistical theory does say something fairly general about the relative effects of skewness and long-tailedness (*kurtosis*):

1. If the two populations have the same standard deviations and approximately the same shapes, and if the sample sizes are about equal, then the validity of the t-tools is affected moderately by long-tailedness and very little by skewness.
2. If the two populations have the same standard deviations and approximately the same shapes, but if the sample sizes are not approximately the same, then the validity of the t-tools is affected moderately by long-tailedness and substantially by skewness. The adverse effects diminish, however, with increasingly large sample sizes.
3. If the skewness in the two populations differs considerably, the tools can be very misleading with small and moderate sample sizes.

Computer simulations can clarify the role of sample sizes. To further investigate the effect of nonnormality on 95% confidence intervals, a computer was instructed to generate samples from the nonnormal distributions shown in Display 3.4. For each pair of generated samples, it computed the 95% confidence interval for the difference in population means and recorded whether the interval actually captured the *true* difference in population means. The actual percentage of successful intervals from 1,000 simulations is shown in Display 3.4 for each set of conditions examined. The purpose of the simulation is to identify combinations of sample sizes and nonnormally shaped distributions for which the confidence interval procedure has a success rate of nearly 95%. An actual success rate less than 95% is undesirable because it means the intervals tend to be too short for the given confidence level and they therefore tend to exclude possible parameter values that shouldn't be excluded. The corresponding hypothesis test, in this case, tends to produce more false claims about statistical significance than expected. An actual success rate greater than 95% is also undesirable because it means the intervals tend to be too long and include possible parameter values that shouldn't be included. The corresponding hypothesis test, in this case, tends to miss statistically significant findings that it really should find.

Of the five distributions examined, only the long-tailed distribution appears to have success rates that are poor enough to cause potentially misleading statements—and even those are not too bad. This distribution can be recognized in practice by the presence of outliers. For the skewed distributions, however, the normality assumption does not appear to be a major concern even for small sample sizes, at least as long as the skewness is the same in the two populations and the sample sizes are roughly equal.

DISPLAY 3.4 Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but with same shape and SD, and equal sample sizes) (each percentage is based on 1,000 computer simulations)



Sample size	Strongly skewed	Moderately skewed	Mildly skewed	Long-tailed	Short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

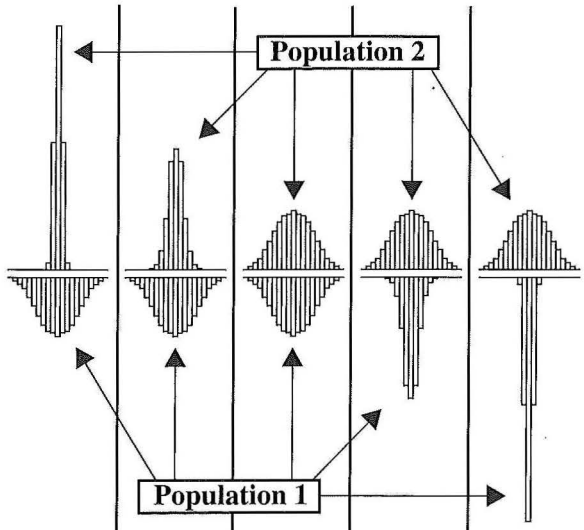
3.2.3 Robustness Against Differing Standard Deviations

More serious problems may arise when the standard deviations of the two populations are substantially unequal. In this case, the pooled estimate of standard deviation does not estimate any population parameter and the standard error formula, which uses the pooled estimate of standard deviation, no longer estimates the standard deviation of the difference between sample averages. As a result, the t-ratio does not have a t-distribution.

Theory shows that the t-tools remain fairly valid when the standard deviations are unequal, as long as the sample sizes are roughly the same. For clarification, a computer was again instructed to generate pairs of samples, this time from two normal populations with different standard deviations, as shown in Display 3.5. It computed 95% confidence intervals for each pair of samples and recorded whether the resulting interval successfully captured the true difference in population means. The actual percentages successful are displayed.

Notice that the success rates for the rows with equal sample sizes ($n_1 = n_2 = 10$ and $n_1 = n_2 = 100$) are very nearly 95%. Thus, as suggested by theory, unequal population standard deviations have little effect on validity if the sample sizes are equal. For substantially different σ's and different n's, however, the confidence intervals are unreliable. The worst situation is when the ratio of standard deviations is much different from 1 and the smaller sized sample is from the population with the larger standard deviation (as, for example, when $n_1 = 100$, $n_2 = 400$, and $\sigma_2/\sigma_1 = 1/4$).

DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)



n_1	n_2		$\sigma_2/\sigma_1 = 1/4$	$\sigma_2/\sigma_1 = 1/2$	$\sigma_2/\sigma_1 = 1$	$\sigma_2/\sigma_1 = 2$	$\sigma_2/\sigma_1 = 4$
10	10		95.2	94.2	94.7	95.2	94.5
10	20	Success	83.0	89.3	94.4	98.7	99.1
10	40	rates	71.0	82.6	95.2	99.5	99.9
100	100	for 95%	94.8	96.2	95.4	95.3	95.1
100	200	intervals	86.5	88.3	94.8	98.8	99.4
100	400		71.6	81.5	95.0	99.5	99.9

3.2.4 Robustness Against Departures from Independence

Cluster Effects and Serial Effects

Whenever knowledge that one observation is, say, above average allows an improved guess about whether another observation will be above average, *independence* is lacking. The methods of Chapter 2 may be misleading in this case. Two types of dependence (lack of independence) commonly arise in practical problems.

The first is a *cluster effect*, which sometimes occurs when the data have been collected in subgroups. For example, 50 experimental animals may have been collected from 10 litters and then randomly assigned to one of two treatment groups. Since animals from the same litter may tend to be more similar in their responses than animals from different litters, it is likely that independence is lacking.

The other type of dependence commonly encountered is caused by a *serial effect*, in which measurements are taken over time and observations close together in time tend to be more similar (or perhaps more different) than observations collected

at distant time points. This can also occur if measurements are made at different locations, and measurements physically close to each other tend to be more similar than those farther apart. In the latter case the dependence pattern is called *spatial correlation.*

The Effects of Lack of Independence on the Validity of the t-Tools

When the independence assumptions are violated, the standard error of the difference of averages is an inappropriate estimate of the standard deviation of the difference in averages. The t-ratio no longer has a t-distribution, and the t-tools may give misleading results. The seriousness of the consequences depends on the seriousness of the violation. It is generally unwise to use the t-tools directly if cluster or serial effects are suspected. Other methods that adjust for these effects are available (Chapters 9–15).

3.3 RESISTANCE OF THE TWO-SAMPLE *t*-TOOLS

Some practical suggestions will soon be provided for sizing up the actual conditions and choosing a course of action. The effect of outliers on the t-tools is discussed first, however, since decisions about how to deal with the outliers play an important role in the overall strategy.

3.3.1 Outliers and Resistance

An *outlier* is an observation judged to be far from its group average. The effect of outliers on the two-sample t-tools has partially been addressed in the discussion of robustness. In fact, it is evident from the theoretical results and the computer simulations in Display 3.4 that the p-values and confidence intervals may be unreliable if the population distributions are long-tailed. Since long-tailed distributions are characterized by the presence of outliers in the sample, outliers should cause some concern.

Long-tailed population distributions are not the only explanation for outliers, however. The populations of interest may be normal but the sample may be contaminated by one or more observations that do not come from the population of interest. Often it is philosophically difficult and practically irrelevant to distinguish between a natural long-tailed distribution and one that includes outliers that result from contamination, although in some cases the identification of clear contamination may dictate an obvious course of action. For example, if it is discovered that one member of a sample from a population of 25- to 35-year-old women is, in fact, over 50 years old, she should be removed from the sample.

It is useful to know how sensitive a statistical procedure may be to one or two outlying observations. The notion of resistance addresses this issue:

> *A statistical procedure is* ***resistant*** *if it does not change very much when a small part of the data changes, perhaps drastically.*

As an example, consider the hypothetical sample: 10, 20, 30, 50, 70. The sample average is 36, and the sample median is 30. Now change the 70 to 700, and what happens? The sample average becomes 162, but the sample median remains 30. The sample average is not a resistant statistic because it can be severely influenced by the change in a single observation. The median, however, is resistant.

Resistance is a desirable property. A resistant procedure is insensitive to outliers. A nonresistant one, on the other hand, may be greatly influenced by one or two outlying observations.

3.3.2 Resistance of *t*-Tools

Since t-tools are based on averages, they are not resistant. A small portion of the data can potentially have a major influence on the results. In particular, one or two outliers can affect a confidence interval or change a p-value enough to completely alter a conclusion.

If the outlier is due to contamination from another population, it can lead to false impressions about the population of interest. If the outlier does come from the population of interest, which happens to be long-tailed, the outcome is still undesirable for the following reason. In statistics, the goal is to describe *group* characteristics. An estimate of the center of a distribution should represent the typical value. The estimate is a good one if it represents the typical values possessed by the great majority of subjects; it is a bad one if it represents a feature unique to one or two subjects. Furthermore, a conclusion that hinges on one or two data points must be viewed as quite fragile.

3.4 PRACTICAL STRATEGIES FOR THE TWO-SAMPLE PROBLEM

Armed with information about the broad set of conditions under which the t-tools work well and the effect of outliers, the challenge to the data analyst is to size up the actual conditions using the available data and evaluate the appropriateness of the t-tools. This involves thinking about possible cluster and serial effects; evaluating the suitability of the t-tools by examining graphical displays; and considering alternatives.

In considering alternatives it is important to realize that even though the t-tools may still be valid when the ideal assumptions are not met, an alternative procedure that is more *efficient* (i.e., makes better use of the data) may be available. For example, another procedure may provide a narrower confidence interval.

Consider Serial and Cluster Effects

To detect lack of independence, carefully review the method by which the data were gathered. Were the subjects selected in distinct groups? Were different groups of subjects treated differently in a way that was unrelated to the primary treatment? Were different responses merely repeated measurements on the same subjects? Were observations taken at different but proximate times or locations? Affirmative answers to any of these questions suggest that independence may be lacking.

The principal remedy is to use a more sophisticated statistical tool. Identifiable clusters, which may be planned or unplanned, can be accounted for through analysis

of variance (Chapters 13 and 14) or possibly through regression analysis (Chapters 9–12). Serial effects require time series analysis, the topic of Chapter 15.

Evaluate the Suitability of the t-Tools

Side-by-side histograms or box plots of the two groups of data should be examined and departures from the ideal model should be considered in light of the robustness properties of the t-tools. It is important to realize that the conditions of interest, which are those of the populations, must be investigated through graphical displays of the samples.

If the conditions do not appear suitable for use of the t-tools, then some alternative is necessary. A transformation should be considered if the graphical displays of the transformed data appear to be closer to the ideal conditions. (See Section 3.5.) Alternative tools for analyzing two independent samples are the rank-sum procedure, which is resistant and does not depend on normality (Section 4.2); other permutation tests (Section 4.3.1); and the Welch procedure for comparing normal populations that have unequal standard deviations (Section 4.3.2).

A Strategy for Dealing with Outliers

If investigation reveals that an outlying observation was recorded improperly or was the result of contamination from another population, the solution is to correct it if the right value is known or to leave it out. Often, however, there is no way to know how the outliers arose. Two statistical approaches for dealing with this situation exist. One is to employ a resistant statistical tool, in which case there is no compelling reason to ponder whether the offending observations are natural, the result of contamination, or simply blunders. (The rank-sum procedure in Section 4.2 is resistant.) The other approach is to adopt the careful examination strategy shown in Display 3.6. An important aspect of adopting this procedure is that an outlier does not get swept under the rug simply because it is different from the other observations. To warrant its removal, an explanation for why it is different must be established.

Example—Agent Orange

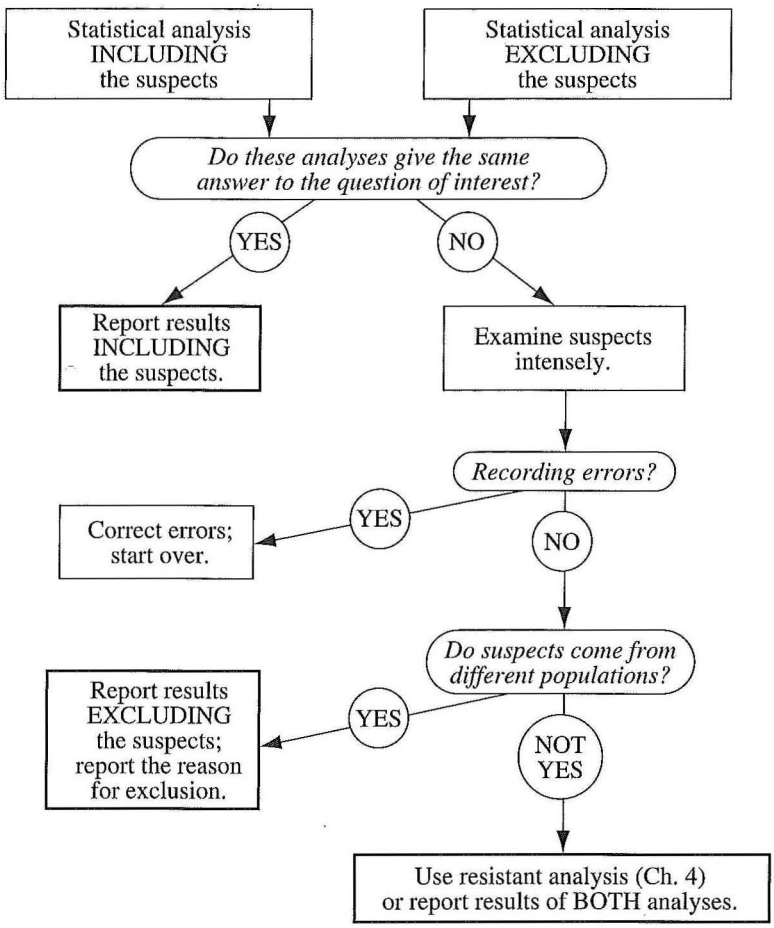
Box plots of dioxin levels in Vietnam and non–Vietnam veterans (Display 3.3) appear again in Display 3.7. The distributions have about the same shape and spread. Although the shape is not normal, the skewness is mild and unlikely to cause any problems with the t-test or the confidence interval. Two Vietnam veterans (#645 and #646) had considerably higher dioxin levels than the others.

From the results listed in Display 3.7 it is evident that the comparison of the two groups is changed very little by the removal of one or both of these outliers. Consequently, there is no need for further action. Even so, it is useful to see what else can be learned about these two, as indicated at the bottom of the display.

Notes

1. It is not useful to give a precise definition for an *outlier*. Subjective examination is the best policy. If there is any doubt about whether a particular observation deserves further examination, give it further examination.

DISPLAY 3.6 Examination strategy



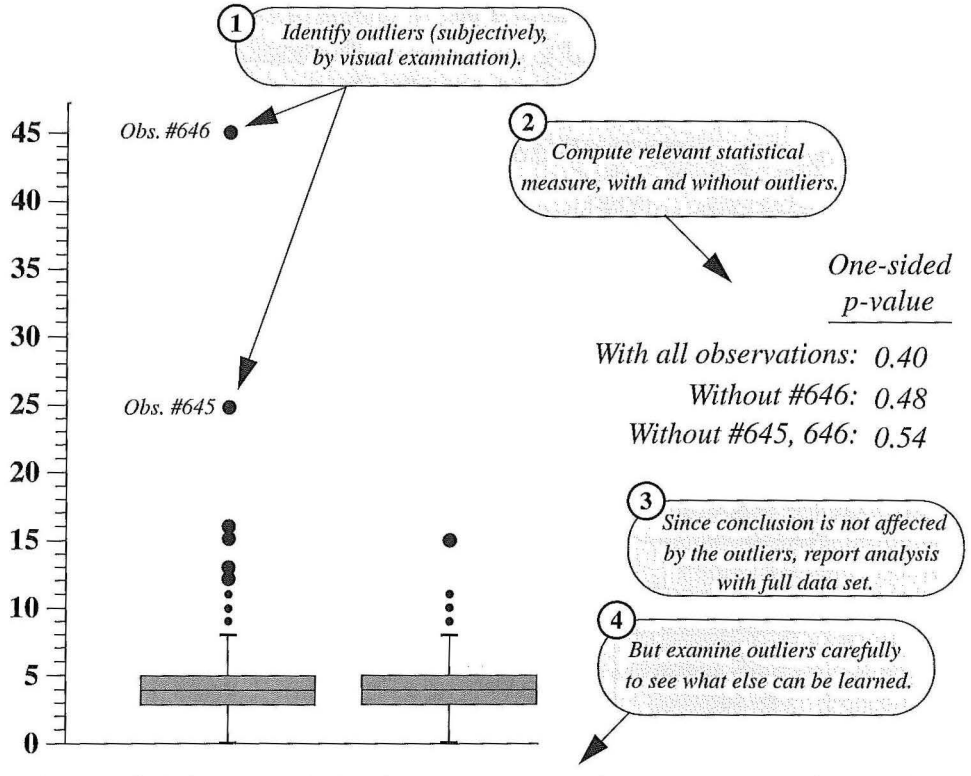
2. It is not surprising that the outliers in the Agent Orange example have little effect, since the sample sizes are so large.
3. The apparent difference in the box plots may be due to the difference in sample sizes. If the population distributions are identical, more observations will appear in the extreme tails from a sample of size 646 than from a sample of size 97.

3.5 TRANSFORMATIONS OF THE DATA

3.5.1 The Logarithmic Transformation

The most useful transformation is the *logarithm* (log) for positive data. The common scale for scientific work is the *natural* logarithm (ln), based on the number

DISPLAY 3.7 Outlier analysis for Agent Orange data: effect of outliers on the p-value, for equal population means



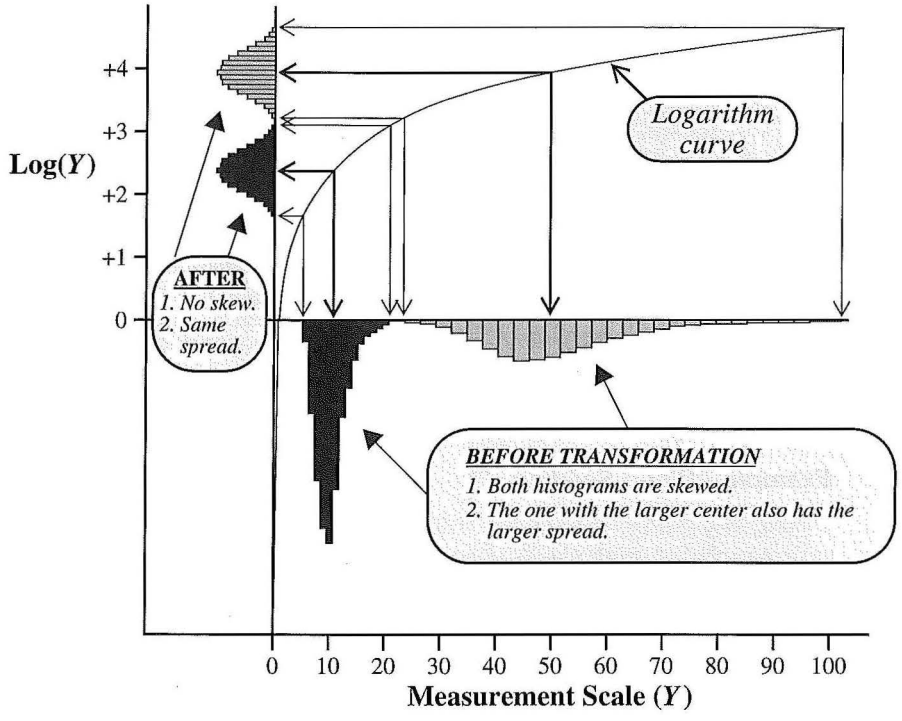
Veteran # 645: reported 180 days of indirect military exposure to herbicides.
Veteran # 646: reported no exposure (military or civilian) to herbicides.

$e = 2.71828\ldots$. The logarithm of e is unity, denoted by $\log(e) = 1$. Also, the log of 1 is 0: $\log(1) = 0$. The general rule for using logarithms is that $\log(e^x) = x$. Another choice is the *common* logarithm based on the number 10, rather than e. Common logs are defined by $\log_{10}(10^x) = x$. Unless otherwise stated, *log* in this book refers to the natural logarithm.

Recognizing the Need for a Log Transformation

The data themselves usually suggest the need for a log transformation. If the ratio of the largest to the smallest measurement in a group is greater than 10, then the data are probably more conveniently expressed on the log scale. Also, if the graphical displays of the two samples show them both to be skewed and if the group with the larger average also has the larger spread (see Display 3.2), the log transformation is likely to be a good choice.

DISPLAY 3.8 The logarithmic transformation used to arrive at favorable conditions for the two-sample t-analysis



Display 3.8 illustrates the behavior of the log transformation. On the scale of measurement Y the two groups have skewed distributions with longer tails in the positive direction. The group with the larger center also has the larger spread. The measurements on the transformed scale have the same ordering, but small numbers get spread out more, while large numbers are squeezed more closely together. The overall result is that the two distributions on the transformed scale appear to be symmetric and have equal spread—just the right conditions for applying the t-tools.

3.5.2 Interpretation After a Log Transformation

For some measurements, the results of an analysis are appropriately presented on the transformed scale. Most users feel comfortable with the Richter scale for measuring earthquake strength, even though it is a logarithmic scale. Similarly, pH as a measure of acidity is the negative log of ion concentration. In other cases, however, it may be desirable to present the results on the original scale of measurement.

Randomized Experiment Model: Multiplicative Treatment Effect

If the randomized experiment model with additive treatment effect is thought to hold for the log-transformed data, then an experimental unit that would respond

to treatment 1 with a logged outcome of $\log(Y)$ would respond to treatment 2 with a logged outcome of $\log(Y) + \delta$. By taking antilogarithms of these two quantities, one finds that an experimental unit that would respond to treatment 1 with an outcome of Y would respond to treatment 2 with an outcome of Ye^{δ}. Thus, e^{δ} is the *multiplicative treatment effect* on the original scale of measurement. To test whether there is any treatment effect, one performs the usual t-test for the hypothesis that δ is zero with the log-transformed data. To describe the multiplicative treatment effect, one back-transforms the estimate of δ and the endpoints of the confidence interval for δ.

Interpretation After Log Transformation (Randomized Experiment)

Suppose $Z = \log(Y)$. It is estimated that the response of an experimental unit to treatment 2 will be $\exp(\overline{Z}_2 - \overline{Z}_1)$ times as large as its response to treatment 1.

Example—Cloud Seeding

Display 3.2 shows that the log-transformed rainfalls have distributions that appear satisfactory for using the t-tools; so in Display 3.9 a full analysis is carried out on the log scale. Tests and confidence intervals are constructed in the usual way but on the transformed data. The estimate of the additive treatment effect on log rainfall is back-transformed to an estimate of the multiplicative effect of cloud seeding on rainfall.

Population Model: Estimating the Ratio of Population Medians

The t-tools applied to log-transformed data provide inferences about the difference in means of the logged measurements, which may be represented as $\text{Mean}[\log(Y_2)] - \text{Mean}[\log(Y_1)]$, where $\text{Mean}[\log(Y_2)]$ symbolizes the mean of the logged values of population 2. A problem with interpretation on the original scale arises because the mean of the logged values is not the log of the mean. Taking the antilogarithm of the estimate of the mean on the log scale does *not* give an estimate of the mean on the original scale.

If, however, the log-transformed data have symmetric distributions, the following relationships hold:

$$\text{Mean}[\log(Y)] = \text{Median}[\log(Y)]$$

(and since the log preserves ordering)

$$\text{Median}[\log(Y)] = \log[\text{Median}(Y)],$$

where $\text{Median}(Y)$ represents the *population median* (the 50th percentile of the population). In other words, the 50th percentile of the logged values is the log of the 50th percentile of the untransformed values. Putting these two equalities together,

DISPLAY 3.9 Two-sample t analysis and statement of conclusions after logarithmic transformation—cloud seeding example

(1) Transform the data.

Unseeded		Seeded	
Y (acre-ft)	log (Y)	Y (acre-ft)	log (Y)
1202.6	7.092	2745.6	7.918
830.1	6.722	1697.8	7.437
372.4	5.920	1656.0	7.412
345.5	5.845	978.0	6.886
321.2	5.772	703.4	6.556
244.3	5.498	489.1	6.193
163.0	5.094	430.0	6.064
147.8	4.996	334.1	5.811
95.0	4.554	302.8	5.713
87.0	4.466	274.7	5.616
81.2	4.397	274.7	5.616
68.5	4.227	255.0	5.541
47.3	3.857	242.5	5.491
41.1	3.716	200.7	5.302
36.6	3.600	198.6	5.291
29.0	3.367	129.6	4.864
28.6	3.353	119.0	4.779
26.3	3.270	118.3	4.773
26.1	3.262	115.3	4.748
24.4	3.195	92.4	4.526
21.7	3.077	40.6	3.704
17.3	2.851	32.7	3.487
11.5	2.446	31.4	3.447
4.9	1.589	17.5	2.862
4.9	1.589	7.7	2.041
1.0	0.000	4.1	1.411

(2) Use the two-sample t-tools on the log rainfall.

Difference in averages = 1.1436 (SE = 0.4495).

Test of the hypothesis of no effect of cloud seeding on log rainfall: one-sided p-value from two-sample t-test = 0.0070 (50 d.f.).

95% confidence interval for additive effect of cloud seeding on log rainfall: 0.2406 to 2.0467.

(3) Back-transform estimate and confidence interval.

Estimate = $e^{1.1436} = 3.1382$
Lower confidence limit = $e^{0.2406} = 1.2720$.
Upper confidence limit = $e^{2.0467} = 7.7425$.

(4) State the conclusions on the original scale.

Conclusion: *There is convincing evidence that seeding increased rainfall (one-sided p-value = 0.0070). The volume of rainfall produced by a seeded cloud is estimated to be 3.14 times as large as the volume that would have been produced in the absence of seeding (95% confidence: 1.27 to 7.74 times).*

it is evident that the antilogarithm of the mean of the log values is the median on the original scale of measurements.

If $\overline{Z}_1$ and $\overline{Z}_2$ are used to represent the averages of the logged values for samples 1 and 2, then $\overline{Z}_2 - \overline{Z}_1$ estimates $\log[Median(Y_2)] - \log[Median(Y_1)]$, and therefore

$$\overline{Z}_2 - \overline{Z}_1 \text{ estimates } \log\left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)}\right]$$

and, therefore,

$$\exp(\overline{Z}_2 - \overline{Z}_1) \text{ estimates } \left[\frac{\text{Median}(Y_2)}{\text{Median}(Y_1)}\right].$$

The point of this is that a very useful multiplicative interpretation emerges in terms of the ratio of population medians. This is doubly important because the median is a better measure of the center of a skewed distribution than the mean. The multiplicative nature of this relationship is captured with the following wording:

Interpretation After Log Transformation (Observational Study)

It is estimated that the median for population 2 is $\exp(\overline{Z}_2 - \overline{Z}_1)$ *times as large as the median for population 1.*

In addition, back-transforming the ends of a confidence interval constructed on the log scale produces a confidence interval for the ratio of medians.

Example (Sex Discrimination)

Although the analysis of the sex discrimination data of Section 1.1.2, was suitable on the original scale of the untransformed salaries, graphical displays of the log-transformed salaries indicate that analysis would also be suitable on the log scale. The average male log salary minus the average female log salary is 0.147. Since $e^{0.147} = 1.16$, it is estimated that the median salary for males is 1.16 times as large as the median salary for females. Equivalently, the median salary for males is estimated to be 16% more than the median salary for females. Since a 95% confidence interval for the difference in means on the log scale is 0.100 to 0.194, a 95% confidence interval for the ratio of population median salaries is 1.11 to 1.21 ($e^{0.100}$ to $e^{0.194}$). With 95% confidence, it is estimated that the median salary for males is between 11% and 21% greater than the median salary for females.

3.5.3 Other Transformations for Positive Measurements

There are other useful transformations for positive measurements with skewed distributions where the means and standard deviations differ between groups. The *square root* transformation $\sqrt{Y}$ applies to data that are counts—counts of bacteria clusters in a dish, counts of traffic accidents on a stretch of highway, counts of red giants in a region of space—and to data that are measurements of area. The *reciprocal* transformation $1/Y$ applies to data that are waiting times—times to failure of lightbulbs, times to recurrence for cancer patients treated with radiation, reaction times to visual stimuli, and so on. The reciprocal of a time measurement can often be interpreted directly as a rate or a speed. The *arcsine square root* transformation, $\text{arcsine}(\sqrt{Y})$, and the *logit* transformation, $\log[Y/(1-Y)]$, apply when the measurements are proportions between zero and one—proportions of trees infested by

a wood-boring insect in experimental plots, proportions of weight lost as a side effect of leukemia therapy, proportions of winning lottery tickets in clusters of a certain size, and so forth.

Only the log transformation, however, gives such ease in converting inferences back to the original scale of measurement. One may estimate the difference in means of $\sqrt{Y_2}$ and $\sqrt{Y_1}$, but the square of this difference does not make much sense on the original scale.

Choosing a Transformation

Formal statistical methods are available for selecting a transformation. Nevertheless, it is recommended here that a trial-and-error approach, with graphical analysis, be used instead. For positive data in need of a transformation, the logarithm should almost always be the first tried. If it is not satisfactory, the reciprocal or the square root transformations might be useful. Keep in mind that the primary goal is to establish a scale where the two groups have roughly the same spread. If several transformations are similar in their ability to accomplish this, think carefully about which one offers the most convenient interpretation.

Caveat About the Log Transformation

Situations arise where presenting results in terms of population medians is not sufficient. For example, the daily emissions of dioxin in the effluent from a paper mill have a very skewed distribution. An agency monitoring the emissions will be interested in estimating the total dioxin load released during, say, a year of operation. The total dioxin load would be the population mean times the population size, and therefore is estimated by the sample average times the population size. It cannot be estimated directly from the median, unless more specific assumptions are made.

3.6 RELATED ISSUES

3.6.1 Prefer Graphical Methods Over Formal Tests for Model Adequacy

Formal tests for judging the adequacy of various assumptions exist. Tests for normality and tests for equal standard deviation are available in most statistical computer programs, as are tests that determine whether an observation is an outlier. Despite their widespread availability and ease of use, these diagnostic tests are not very helpful for model checking. They reveal little about whether the data meet the broader conditions under which the tools work well. The fact that two populations are not exactly normal, for example, is irrelevant. Furthermore, the formal tests themselves are often not very robust against their own model assumptions. Graphical displays are more informative, if less formal. They provide a good indication of whether or not the data are amenable to t-analysis and, if not, they often suggest a remedy.

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.