CHAPTER 5

Inferences for Two Populations

CONTENTS

■ Example 5.1: Comparing Changes in Stock Prices

The year 2008 saw tremendous declines in stock market prices, but perhaps some industry categories saw greater declines than others. To examine this question, we randomly selected stock prices for 10 companies from the Standard & Poor Consumer Staples and Financial categories, respectively. The prices, as of 12/31/2008 and 12/31/2007, are shown in Table 5.1.

Figure 5.1 shows the box plots of the price changes, expressed both as a simple decrease (2007 price–2008 price) and as a percentage decrease. These plots suggest that stocks in the Financial category did see typically greater declines, together with greater variability. In fact, we will see that although there is strong evidence of mean decreases within each category, the evidence for a systematic difference between the Consumer Staples and Financial categories is surprisingly weak.

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00005-6

Table 5.1 Stock Prices

Consumer Staples			Financial		
Tkr	12/31/08	12/31/07	Tkr	12/31/08	12/31/07
MO	14.77	21.50	ALL	32.34	49.41
CPB	29.74	34.44	BAC	14.04	38.05
CL	67.67	75.22	CINF	28.55	36.96
CVS	28.58	39.25	EQR	29.09	33.82
HNZ	37.13	44.49	HCP	26.67	31.52
KFT	26.50	31.03	JPM	31.09	41.54
TAP	48.65	50.56	MMC	23.79	25.19
PG	60.86	70.62	NYX	26.90	83.81
SJM	43.03	49.69	RF	7.76	21.56
TSN	8.72	15.04	USB	24.93	29.90

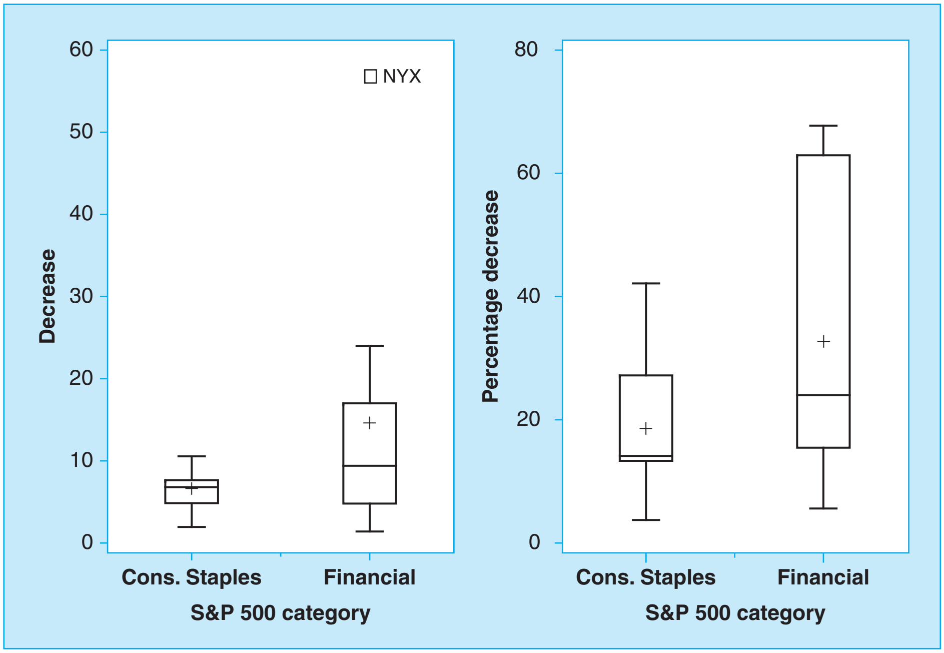

FIGURE 5.1

Changes in Stock Market Prices during 2008 for Example 5.1.

5.1 INTRODUCTION

In Chapter 4 we provided methods for inferences on parameters of a single population. A natural extension of these methods occurs when two populations are to be compared. In this chapter we provide the inferential methods for making comparisons on parameters of two populations. This leads to a natural extension, that of comparing more than two populations, which is presented in Chapter 6. So, why not go directly to comparing parameters of several populations and consider the case of two populations as a special case? There are several good answers to that question:

- Many interesting applications involve only two populations, for example, any comparisons involving differences between the two sexes, comparing a drug with a placebo, comparing old versus new, or before and after some event.
- Some of the concepts underlying comparing several populations are more easily introduced for the two-population case.
- The comparison of two populations results in a single easily understood statistic: the difference between sample means. As we shall see in Chapter 6, such a simple statistic is not available for comparing more than two populations. As a matter of fact, even when we have more than two populations, we will often want to make comparisons among specific pairs from the set of populations.

Populations that are to be compared arise in two distinct ways:

- The populations are actually different. For example, male and female students, two regions of a state or nation, or two different breeds of cattle. In Section 1.1 we referred to a study involving separate populations as an observational study.
- The populations are a result of an experiment where a single homogeneous population has been divided into two portions where each has been subjected to some sort of modification, for example, a sample of individuals given two different drugs to combat a disease, a field of an agricultural crop where two different fertilizer mixtures are applied to various portions, or a group of school children subjected to different teaching methods. In Section 1.1 this type of study was referred to as a designed experiment.

This latter situation constitutes the more common usage of statistical inference. In such experiments the different populations are usually referred to as “treatments” or “levels of a factor.” These terms will be discussed in greater detail in later chapters, especially Chapter 10.

There are also two distinct methods for collecting data on two populations, or equivalently, designing an experiment for comparing two populations. These are called (1) **independent samples** and (2) **dependent** or **paired samples**. We illustrate these two methods with a hypothetical experiment designed to compare the effectiveness of two migraine headache remedies. The response variable is a measure of headache relief reported by the subjects.

Independent Samples

A sample of migraine sufferers is randomly divided into two groups. The first group is given remedy A while the other is given remedy B, both to be taken at the onset of a migraine attack. The pills are not identified, so patients do not know which pill they are taking. Note that the individuals sampled for the two remedies are indeed independent of each other.

Dependent or Paired Samples

Each person in a group of migraine sufferers is given two pills, one of which is red and the other is green. The group is randomly split into two subgroups and one is told to take the green pill the first time a migraine attack occurs and the red pill for the next one. The other group is told to take the red pill first and the green pill next. Note that both pills are given to each patient so the responses of the two remedies are naturally paired for each patient.

These two methods of comparing the efficacy of the remedies dictate different inferential procedures. The comparison of means, variances, and proportions for independent samples are presented in Sections 5.2, 5.3, and 5.5, respectively, and the comparison of means and proportions for the dependent or paired sample case in Sections 5.4 and 5.5.

5.2 INFERENCES ON THE DIFFERENCE BETWEEN MEANS USING INDEPENDENT SAMPLES

We are interested in comparing two populations whose means are μ_1 and μ_2 and whose variances are σ_1^2 and σ_2^2, respectively. Comparisons may involve the means or the variances (standard deviations). In this section we consider the comparison of means.

For two populations we define the difference between the two means as

$$\delta = \mu_1 - \mu_2.$$

This single parameter δ provides a simple, tractable measure for comparing two population means, not only to see whether they are equal, but also to estimate the difference between the two. For example, testing the null hypothesis

$$H_0\colon \mu_1 = \mu_2$$

is the same as testing the null hypothesis

$$H_0\colon \delta = 0.$$

A sample of size n_1 is randomly selected from the first population and a sample of size n_2 is independently drawn from the second. The difference between the two sample means $(\bar{y}_1 - \bar{y}_2)$ provides the unbiased point estimate of the difference $(\mu_1 - \mu_2)$. However, as we have learned, before we can make any inferences about the difference between means, we must know the sampling distribution of $(\bar{y}_1 - \bar{y}_2)$.

5.2.1 Sampling Distribution of a Linear Function of Random Variables

The sampling distribution of the difference between two means from independently drawn samples is a special case of the sampling distribution of a **linear function of random variables**. Consider a set of n random variables $y_1, y_2, \ldots, y_n$, whose distributions have means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. A linear function of these random variables is defined as

$$L = \sum a_i y_i = a_1 y_1 + a_2 y_2 + \cdots + a_n y_n,$$

where the a_i are arbitrary constants. L is also a random variable and has mean

$$\mu_L = \sum a_i \mu_i = a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n.$$

If the variables are independent, then L has variance

$$\sigma_L^2 = \sum a_i^2 \sigma_i^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2.$$

Further, if the y_i are normally distributed, so is L.

5.2.2 The Sampling Distribution of the Difference between Two Means

Since sample means are random variables, the difference between two sample means is a linear function of two random variables. That is,

$$\bar{y}_1 - \bar{y}_2$$

can be written as

$$L = a_1 \bar{y}_1 + a_2 \bar{y}_2 = (1)\bar{y}_1 + (-1)\bar{y}_2.$$

In terms of the linear function specified above, $n = 2$, $a_1 = 1$, and $a_2 = -1$. Using these specifications, the sampling distribution of the difference between two means has a mean of $(\mu_1 - \mu_2)$.

Further, since the $\bar{y}_1$ and $\bar{y}_2$ are sample means, the variance of $\bar{y}_1$ is σ_1^2/n_1 and the variance of $\bar{y}_2$ is σ_2^2/n_2. Also, because we have made the assumption that the two samples are independently drawn from the two populations, the two sample means are independent random variables. Therefore, the variance of the difference $(\bar{y}_1 - \bar{y}_2)$ is

$$\sigma_L^2 = (+1)^2 \sigma_1^2/n_1 + (-1)^2 \sigma_2^2/n_2,$$

or simply

$$= \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

Note that for the special case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $n_1 = n_2 = n$, the variance of the difference is $2\sigma^2/n$.

Finally, the central limit theorem states that if the sample sizes are sufficiently large, $\bar{y}_1$ and $\bar{y}_2$ are normally distributed; hence for most applications L is also normally distributed.

Thus, if the variances σ_1^2 and σ_2^2 are known, we can determine the variance of the difference $(\bar{y}_1 - \bar{y}_2)$. As in the one-population case we first present inference procedures that assume that the population variances are known. Procedures using estimated variances are presented later in this section.

5.2.3 Variances Known

We first consider the situation in which both population variances are known. We want to make inferences on the difference

$$\delta = \mu_1 - \mu_2,$$

for which the point estimate is

$$\bar{y}_1 - \bar{y}_2.$$

This statistic has the normal distribution with mean $(\mu_1 - \mu_2)$ and variance $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$. Hence, the statistic

$$z = \frac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

has the standard normal distribution. Hypothesis tests and confidence intervals are obtained using the distribution of this statistic.

Hypothesis Testing

We want to test the hypotheses

$$H_0\colon \mu_1 - \mu_2 = \delta_0,$$
$$H_1\colon \mu_1 - \mu_2 \neq \delta_0,$$

where δ_0 represents the hypothesized difference between the population means. To perform this test, we use the test statistic

$$z = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}.$$

The most common application is to let $\delta_0 = 0$, which is, of course, the test for the equality of the two population means. The resulting value of z is used to calculate a

p value (using the standard normal table) or compared with a rejection region constructed for the desired level of significance. One- or two-sided alternative hypotheses may be used.

A confidence interval on the difference $(\mu_1 - \mu_2)$ is constructed using the sampling distribution of the difference presented above. The confidence interval takes the form

$$(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}.$$

■ Example 5.2

A production plant has two fabricating systems: one uses automated equipment, the other is manually operated. Since the automated system costs more to install, we want to know whether it provides increased production in terms of the mean number of finished products fabricated per day. Experience has shown that the daily production of the automated system has a standard deviation of $\sigma_1 = 10$, the manual system, $\sigma_2 = 20$.[1] Independent random samples of 100 days of production are obtained from company records for each system. The sample results are that the automated system had a sample mean production of $\bar{y}_1 = 254$, and the manual system a sample mean of $\bar{y}_2 = 248$. Is the automated system superior to the manual one?

Solution

To answer the question, we will test the hypothesis

$$H_0\colon \delta = \mu_1 - \mu_2 = 0 \text{ (or } \mu_1 = \mu_2),$$

where μ_1 is the average production of the automated system and μ_2 that of the manual system. The alternate hypothesis is

$$H_1\colon \delta = \mu_1 - \mu_2 > 0 \text{ (or } \mu_1 > \mu_2);$$

that is, the automated system has a higher production rate. Because of the cost of installing the automated system, $\alpha = 0.01$ is chosen to determine whether the manual system should be replaced by an automated system. The test statistic has a value of

$$z = \frac{(254 - 248) - 0}{\sqrt{(10^2/100) + (20^2/100)}} \\ = 2.68.$$

The p value associated with this test statistic is $p = 0.0037$. The null hypothesis is rejected for any significance level exceeding 0.0037; hence we can conclude that

[1] The fact that the automated system has a smaller variance is not of interest at this time.

average daily production will be increased by replacing the manual system with an automated one.

It is also of interest to estimate by what amount the average daily production will be increased. This can be determined by using a one-sided confidence interval similar to that discussed in Section 3.3. In particular, we determine the lower 0.99 confidence limit on the mean as

$$(254 - 248) - 2.326\sqrt{(10)^2/100 + (20)^2/100} = 0.80.$$

This means that the increase may be as low as one unit, which may not be sufficient to justify the expense of installing the new system, illustrating the principle that a statistically significant result does not necessarily imply practical significance as noted in Section 3.6. ■

5.2.4 Variances Unknown but Assumed Equal

The "obvious" methodology for comparing two means when the population variances are not known would seem to be to use the two variance estimates, s_1^2 and s_2^2, in the statistic described in the previous section and determine the significance level from the Student's t distribution. This approach will not work because the mathematical formulation of this distribution requires as its single parameter the degrees of freedom for a single variance estimate.

The solution to this problem is to assume that the two population variances are equal and find an estimate of that variance. The equal variance assumption is actually quite reasonable since in many studies, a focus on means implies that the populations are similar in many respects. Otherwise, it would not make sense to compare just the means (apples with oranges, etc.). If the assumption of equal variances cannot be made, then other methods must be employed, as discussed later in this section.

Assume that we have independent samples of size n_1 and n_2, respectively, from two normally distributed populations with equal variances. We want to make inferences on the difference $\delta = (\mu_1 - \mu_2)$. Again the point estimate of that difference is $(\bar{y}_1 - \bar{y}_2)$.

5.2.5 The Pooled Variance Estimate

The estimate of a common variance from two independent samples is obtained by "pooling," which is simply the weighted mean of the two individual variance estimates with the weights being the degrees of freedom for each variance. Thus the pooled variance, denoted by s_{p}^2, is

$$s_{\mathrm{p}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

We have emphasized that all estimates of a variance have the form

$$s^2 = \mathrm{SS}/\mathrm{df},$$

where, for example, $\mathrm{df} = (n-1)$ for a single sample, and consequently $\mathrm{SS} = (n-1)s^2$. Using the notation SS_1 and SS_2 for the sums of squares from the two samples, the pooled variance can be defined (and, incidentally, more easily calculated) as

$$s_{\mathrm{p}}^2 = \frac{\mathrm{SS}_1 + \mathrm{SS}_2}{n_1 + n_2 - 2}.$$

This form of the equation shows that the pooled variance is indeed of the form SS/df, where now $\mathrm{df} = (n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$. The pooled variance is now used in the t statistic, which has the t distribution with $(n_1 + n_2 - 2)$ degrees of freedom. We will see in Chapter 6 that the principle of pooling can be applied to any number of samples.

5.2.6 The "Pooled" *t* Test

To test the hypotheses

$$H_0\colon \mu_1 - \mu_2 = \delta_0,$$
$$H_1\colon \mu_1 - \mu_2 \neq \delta_0,$$

we use the test statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{\left(s_{\mathrm{p}}^2/n_1\right) + \left(s_{\mathrm{p}}^2/n_2\right)}},$$

or equivalently

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_{\mathrm{p}}^2(1/n_1 + 1/n_2)}}.$$

This statistic will have the t distribution and the degrees of freedom are $(n_1 + n_2 - 2)$ as provided by the denominator of the formula for s_{p}^2. This test statistic is often called the **pooled *t* statistic** since it uses the pooled variance estimate.

Similarly the confidence interval on $\mu_1 - \mu_2$ is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_{\mathrm{p}}^2(1/n_1 + 1/n_2)},$$

using values from the t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Example 5.3

Mesquite is a thorny bush whose presence reduces the quality of pastures in the Southwest United States. In a study of growth patterns of this plant, dimensions of samples of mesquite were taken in two similar areas (labeled A and M) of a ranch. In this example, we are interested in determining whether the average heights of the plants are the same in both areas. The data are given in Table 5.2.

Table 5.2 Heights of Mesquite

Location A ($n_A = 20$)		Location M ($n_m = 26$)		
1.70	2.00	1.30	0.90	1.50
3.00	1.30	1.35	1.35	1.50
1.70	1.45	2.16	1.40	1.20
1.60	2.20	1.80	1.00	0.70
1.40	0.70	1.55	1.70	1.20
1.90	1.90	1.20	1.50	0.80
1.10	1.80	1.00	0.65	
1.60	2.00	1.70	1.50	
2.00	2.20	0.80	1.70	
1.25	0.92	1.20	1.70	

Solution

As a first step in the analysis of the data, construction of a stem and leaf plot of the two samples (Table 5.3) is appropriate. The purpose of this exploratory procedure is to provide an overview of the data and look for potential problems, such as outliers or distributional anomalies. The plot appears to indicate somewhat larger mesquite bushes in location A. One bush in location A appears to be quite large;

Table 5.3 Stem and Leaf Plot for Mesquite Heights

Location A	Stem	Location M
0	3	
	2	
00022	2	2
6677789	1	5555677778
12344	1	0022223444
79	0	77889

however, we do not have sufficient evidence that this value represents an outlier or unusual observation that may affect the analysis.

We next perform the test for the hypotheses

$$H_0\colon \mu_A - \mu_M = 0 \text{ (or } \mu_A = \mu_M),$$
$$H_1\colon \mu_A - \mu_M \neq 0 \text{ (or } \mu_A \neq \mu_M).$$

The following preliminary calculations are required to obtain the desired value for the test statistic:

Location A	**Location** M
$n = 20$	$n = 26$
$\sum y = 33.72$	$\sum y = 34.36$
$\sum y^2 = 61.9014$	$\sum y^2 = 48.9256$
$\bar{y} = 1.6860$	$\bar{y} = 1.3215$
SS $= 5.0495$	SS $= 3.5175$
$s^2 = 0.2658$	$s^2 = 0.1407$

The computed t statistic is

$$\begin{aligned} t &= \frac{1.6860 - 1.3215}{\sqrt{\frac{5.0495 + 3.5175}{44}\left(\frac{1}{20} + \frac{1}{26}\right)}} \\ &= \frac{0.3645}{\sqrt{(0.1947)(0.08846)}} \\ &= \frac{0.3654}{0.1312} \\ &= 2.778. \end{aligned}$$

We have decided that a significance level of 0.01 would be appropriate. For this test we need the t distribution for $20 + 26 - 2 = 44$ degrees of freedom. Because Appendix Table A.2 does not have entries for 44 degrees of freedom, we use the next smallest degrees of freedom, which is 40. This provides for a more conservative test; that is, the true value of α will be somewhat less than the specified 0.01. It is possible to interpolate between 40 and 60 degrees of freedom to provide a more precise rejection region, but such a degree of precision is rarely needed. Using this approximation, we see that the rejection region consists of absolute values exceeding 2.7045.

The value of the test statistic exceeds 2.7045 so the null hypothesis is rejected, and we determine that the average heights of plants differ between the two locations. Using a computer program, the exact p value for the test statistic is 0.008.

The 0.99 confidence interval on the difference in population means, $(\mu_1 - \mu_2)$, is

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2}\sqrt{s_{\mathrm{p}}^2(1/n_1 + 1/n_2)},$$

which produces the values

$$0.3645 \pm 2.7045\ (0.1312) \quad \text{or} \quad 0.3645 \pm 0.3548,$$

which defines the interval from 0.0097 to 0.7193. The interval does not contain zero, which agrees with the results of the hypothesis test. ■

5.2.7 Variances Unknown but Not Equal

In Example 5.3 we saw that the variance of the heights from location *A* was almost twice that of location *M*. The difference between these variances probably is due to the rather large bush measured at location *A*. Since we cannot discount this observation, we may need to provide a method for comparing means that does not assume equal variances. (A test for equality of variances is presented in Section 5.3 and according to this test these two variances are not significantly different.)

Before continuing, it should be noted that inferences on means may not be useful when variances are not equal. If, for example, the distributions of two populations look like those in Fig. 5.2, the fact that population 2 has a larger mean is only one factor in the difference between the two populations. In such cases it may be more useful to test other hypotheses about the distributions. Additional comments on this and other assumptions needed for the pooled t test are presented in Section 5.6 and also in Chapter 14.

Sometimes differences in variances are systematic or predictable. For some populations the magnitude of the variance or standard deviation may be proportional to the

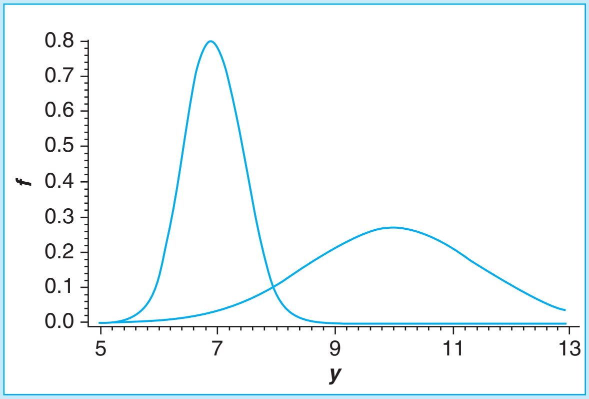

FIGURE 5.2
Distributions with Different Variances.

magnitude of the mean. For example, for many biological organisms, populations with larger means also have larger variances. This type of variance inequality may be handled by making "transformations" on the data, which employ the analysis of some function of the y's, such as log y, rather than the original values. The transformed data may have equal variances and the pooled t test can then be used. The use of transformations is more fully discussed in Section 6.4.

Not all problems with unequal variances are amenable to this type of analysis; hence we need alternate procedures for performing inferences on the means of two populations based on data from independent samples. For this situation we may use one of the following procedures with the choice depending on the sample sizes:

1. If both n_1 and n_2 are large (both over 30) we can assume a normal distribution and compute the test statistic

$$t' = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

   Since n_1 and n_2 are large, the central limit theorem will allow us to assume that the difference between the sample means will have approximately the normal distribution. Again, for the large sample case, we can replace σ_1 and σ_2 with s_1 and s_2 without serious loss of accuracy. Therefore, the statistic t' will have approximately the standard normal distribution.
2. If either sample size is not large, compute the statistic t' as in part (1). If the data come from approximately normally distributed populations, this statistic does have an approximate Student's t distribution, but the degrees of freedom cannot be precisely determined. A reasonable (and conservative) approximation is to use the degrees of freedom for the smaller sample. More precise but complex approximations are available. One such approximation, called Satterthwaite's approximation, is implemented in many statistical packages (see Steel and Torrie, 1980).

■ Example 5.4

In a study on attitudes among commuters, random samples of commuters were asked to score their feelings toward fellow passengers using a score ranging from 0 for "like" to 10 for "dislike." A sample of 10 city subway commuters (population 1) and an independent sample of 17 suburban rail commuters (population 2) were used for this study. The purpose of the study is to compare the mean attitude scores of the two types of commuters. It can be assumed that the data represent samples from normally distributed populations.

The data from the two samples are given in Table 5.4. Note that the data are presented in the form of frequency distributions; that is, a score of zero was given by three subway commuters and five rail commuters and so forth.

Table 5.4 Attitudes among Commuters

	SCORE										
Commuter Type	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Subway	3	1		2		1		1		2	
Rail	5	4	5	1	1	1					

Solution

Distributions of scores of this type typically have larger variances when the mean score is near the center (5) and smaller variances when the mean score is near either extreme (0 or 10). Thus, if there is a difference in means, there is also likely to be a difference in variances. We want to test the hypotheses

$$H_0\colon \mu_1 = \mu_2,$$
$$H_1\colon \mu_1 \neq \mu_2.$$

The t' statistic has a value of

$$t' = \frac{3.70 - 1.53}{\sqrt{(13.12/10) + (2.14/17)}} = 1.81.$$

The smaller sample has 10 observations; hence we use the t distribution with 9 degrees of freedom. The 0.05 critical value is ±2.262. The sample statistic does not lead to rejection at $\alpha = 0.05$; in fact, the p value is somewhat greater than 0.10. Therefore there is insufficient evidence that the attitudes of commuters differ.

Figure 5.3 shows the distributions of the two samples. The plot clearly shows the larger variation for the subway scores, but there does not appear to be much

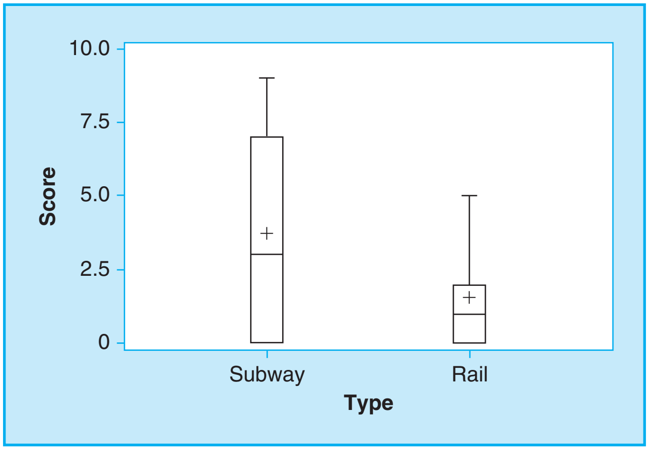

FIGURE 5.3
Box Plot of Commuters' Scores.

difference between the means. Even though the distributions appear to be skewed, Q–Q plots similar to those discussed in Section 4.5 (not shown here) do not indicate any serious deviations from normality.

If this data set had been analyzed using the pooled t test discussed earlier, the t value would be 2.21 with 25 degrees of freedom. The p value associated with this test statistic is about 0.04, which is sufficiently small to result in rejection of the hypothesis at the 0.05 significance level. Thus, if the test had been made under the assumption of equal variances (which in this case is not valid), an incorrect inference may have been made about the attitudes of commuters. ■

Actually the equal variance assumption is only one of several necessary to assure the validity of conclusions obtained by the pooled t test. A brief discussion of these issues and some ideas on remedial or alternate methods is presented in Section 5.6 and also in Chapter 14.

5.3 INFERENCES ON VARIANCES

In some applications it may be important to be able to determine whether the variances of two populations are equal. Such inferences are not only useful to determine whether a pooled variance may be used for inferences on the means, but also to answer more general questions about the variances of two populations. For example, in many quality control experiments, it is important to maintain consistency, and for such experiments inferences on variances are of prime importance, since the variance is a measure of consistency within a population.

In comparing the means of two populations, we are able to use the difference between the two sample means as the relevant point estimate and the sampling distribution of that difference to make inferences. However, the difference between two sample variances does not have a simple, usable distribution. On the other hand, the statistic based on the ratio s_1^2/s_2^2 is, as we saw in Section 2.6, related to the F distribution. Consequently, if we want to state that two variances are equal, we can express this relationship by stating that the ratio σ_1^2/σ_2^2 is unity. The general procedures for performing statistical inference remain the same.

Recall that the F distribution depends on two parameters, the degrees of freedom for the numerator and the denominator variance estimates. Also the F distribution is not symmetric. Therefore the inferential procedures are somewhat different from those for means, but more like those for the variance (Section 4.4).

To test the hypothesis that the variances from two populations are equal, based on independent samples of size n_1 and n_2 from normally distributed populations, use the following procedures:

1. The null hypothesis is

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{or} \quad H_0: \sigma_1^2/\sigma_2^2 = 1.$$

2. The alternative hypothesis is
$$H_1: \sigma_1^2 \neq \sigma_2^2 \quad \text{or} \quad H_1: \sigma_1^2/\sigma_2^2 \neq 1.$$
One-tailed alternatives are that the ratio is either greater or less than unity.
3. Independent samples of size n_1 and n_2 are taken from the two populations to provide the sample variances s_1^2 and s_2^2.
4. Compute the ratio $F = s_1^2/s_2^2$.
5. This value is compared with the appropriate value from the table of the F distribution, or a p value is computed from it. Note that since the F distribution is not symmetric, a two-tailed alternative hypothesis requires finding two separate critical values in the table.

As we discussed in Section 2.6 regarding the F distribution, most tables do not have the lower tail values. It was also shown that these values may be found by using the relationship
$$F_{(1-\alpha/2)}(\nu_1, \nu_2) = \frac{1}{F_{\alpha/2}(\nu_2, \nu_1)}.$$

An easier way of obtaining a rejection region for a two-tailed alternative is to always use the larger variance estimate for the numerator, in which case we need only the upper tail of the distribution, remembering to use $\alpha/2$ to find the critical value. In other words, if s_2^2 is larger than s_1^2, use the ratio $F = s_2^2/s_1^2$, and determine the F value for $\alpha/2$ with $(n_2 - 1)$ numerator and $(n_1 - 1)$ denominator degrees of freedom.

For a one-tailed alternative, simply label the populations such that the alternative hypothesis can be stated in terms of "greater than," which then requires the use of the tabled upper tail of the distribution.

Confidence intervals are also expressed in terms of the ratio σ_1^2/σ_2^2. The confidence limits for this ratio are as follows:

Lower limit:
$$\frac{(s_1^2/s_2^2)}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}.$$

Upper limit:
$$\frac{(s_1^2/s_2^2)}{F_{(1-\alpha/2)}(n_1 - 1, n_2 - 1)}.$$

In this case we must use the reciprocal relationship (Section 2.6) for the two tails of the distribution to compute the upper limit:
$$(s_1^2/s_2^2)F_{\alpha/2}(n_2 - 1, n_1 - 1).$$

Alternately, we can compute the lower limit for σ_2^2/σ_1^2, which is the reciprocal of the upper limit for σ_1^2/σ_2^2.

■ Example 5.5

In previous chapters we discussed a quality control example in which we were monitoring the amount of peanuts being put in jars. In situations such as this, consistency of weights is very important and therefore warrants considerable attention in quality control efforts. Suppose that the manufacturer of the machine proposes installation of a new control device that supposedly increases the consistency of the output from the machine. Before purchasing it, the device must be tested to ascertain whether it will indeed reduce variability. To test the device, a sample of 11 jars is examined from a machine without the device (population N), and a sample of 9 jars is examined from the production after the device is installed (population C). The data from the experiment are given in Table 5.5, and Fig. 5.4 shows side-by-side box plots for the weights of the samples. The sample from population C certainly appears to exhibit less variation. The question is, does the control device significantly reduce variation?

Table 5.5 Contents of Peanut Jars (oz.)

Population N without Control		Population C with Control	
8.06	8.39	7.99	8.03
8.64	8.46	8.12	8.14
7.97	8.28	8.34	8.14
7.81	8.02	8.17	7.87
7.93	8.39	8.11	
8.57			

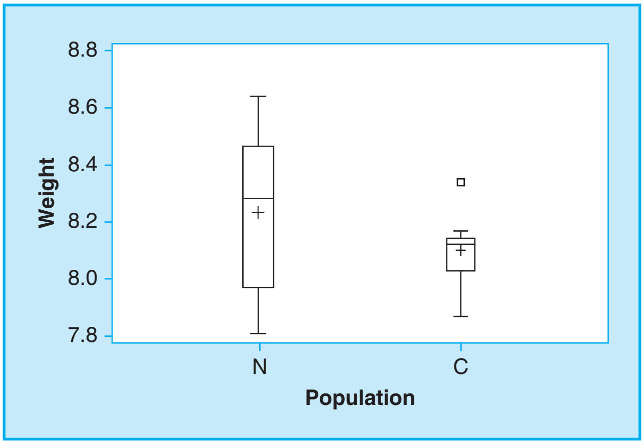

FIGURE 5.4
Box Plots of Weights.

Solution

We are interested in testing the hypotheses

$$H_0\colon \sigma_N^2 = \sigma_C^2 \ (\text{or} \quad \sigma_N^2/\sigma_C^2 = 1),$$
$$H_1\colon \sigma_N^2 > \sigma_C^2 \ (\text{or} \quad \sigma_N^2/\sigma_C^2 > 1).$$

The sample statistics are

$$s_N^2 = 0.07973 \quad \text{and} \quad s_C^2 = 0.01701.$$

Since we have a one-tailed alternative, we place the larger alternate hypothesis variance in the numerator; that is, the test statistic is s_N^2/s_C^2. The calculated test statistic has a value of $F = 0.07973/0.01701 = 4.687$. The rejection region for $\alpha = 0.05$ for the F distribution with 10 and 8 degrees of freedom consists of values exceeding 3.35. Hence the null hypothesis is rejected and the conclusion is that the device does in fact increase the consistency (reduce the variance).

A one-sided interval is appropriate for this example. The desired confidence limit is the lower limit for the ratio σ_N^2/σ_C^2, since we want to be, say, 0.95 confident that the variance of the machine without the control device is larger. The lower 0.95 confidence limit is

$$\frac{(s_N^2/s_C^2)}{F_{0.05}(10,8)}.$$

The value of $F_{0.05}(10,8)$ is 3.35; hence the limit is

$$4.687/3.35 = 1.40.$$

In other words we are 0.95 confident that the variance without the control device is at least 1.4 times as large as it is with the control device. As usual, the result agrees with the hypothesis test, which rejected the hypothesis of a unit ratio. ■

Count Five Rule The F test for equality of variances is sensitive to nonnormality of the data in the groups. A variety of tests that are less sensitive have been developed, and one of these will be introduced in Section 6.4. In the case where there are only two groups and the sample sizes are equal, McGrath and Yeh (2005) have proposed a simple rule called Count Five. Briefly, if you examine the absolute values of the deviations about each group mean, and the largest five all come from the same group, then you intuitively would believe that that group must have larger dispersion. In fact, the authors show that this is a test with surprisingly good properties. They also discuss the extension to unequal sample sizes.

CASE STUDY 5.1

Jerrold *et al.* (2009) compared typically developing children to young adults who had Down syndrome, with respect to a number of psychological measures thought to be related to the ability to learn new words. In Table 5.6, we present summary information on two of the measures:

1. Recall score, a measure of verbal short term memory.
2. Raven's CPM, scores on a task in which the participant must correctly identify an image that completes a central pattern.

The authors used the pooled t test to compare the typical scores in the two groups. For Raven's CPM, $t = -0.485$, p value $= 0.629$. For Recall Score, $t = -7.007$, p value < 0.0001. Hence, the two groups did not differ significantly with respect to mean Raven's CPM, but the Down syndrome group scored significantly differently (apparently lower) on Recall Score. Based on this and a number of other comparisons, the authors conclude that verbal short-term memory is a primary factor in the ability to learn new words.

The authors actually presented the results of the pooled t test (with 80 degrees of freedom) as an F test with 1 degree of freedom in the numerator and 80 in the denominator. The relation between these two test statistics will be explained in Chapter 6.

Table 5.6 Summary Statistics from Jerrold (2009)

	Down Syndrome Young Adults $n = 21$		**Typically Developing Children** $n = 61$	
	Mean	**S.D.**	**Mean**	**S.D.**
Raven's CPM	19.33	4.04	19.90	4.83
Recall Score	12.00	3.05	18.25	3.67

5.4 INFERENCES ON MEANS FOR DEPENDENT SAMPLES

In Section 5.2 we discussed the methods of inferential statistics as applied to two independent random samples obtained from separate populations. These methods are not appropriate for evaluating data from studies in which each observation in one sample is matched or paired with a particular observation in the other sample. For example, if we are studying the effect of a special diet on weight gains, it is not effective to randomly divide a sample of subjects into two groups and give the special diet to one of these groups and then compare the weights of the individuals from these two groups. Remember that for two independently drawn samples the estimate of the variance is based on the differences in weights among individuals in each sample, and these differences are probably larger than those induced by the special diet. A more logical data collection method is to weigh a random sample of individuals before they go on the diet and then weigh the same individuals after they have been subjected to the diet. The individuals' differences in weight before and after the special diet are then a more precise indicator of the effect of the diet. Of course, these two sets of weights are no longer independent, since the same individuals belong to both. The choice of data collection method (independent or dependent samples in this example) was briefly introduced in Section 5.1 and is an example of the use of a design of an experiment. (Experimental design is discussed briefly in Chapter 6 and more extensively in Chapter 10.)

For two populations, such samples are dependent and are called "paired samples" because our analysis will be based on the differences between pairs of observed values. For example, in evaluating the diet discussed above, the pairs are the weights obtained on individuals before and after the special diet and the analysis is based on the individual weight losses. This procedure can be used in almost any context in which the data can physically be paired.

For example, identical twins provide an excellent source of pairs for studying various medical and psychological hypotheses. Usually each of a pair of twins is given a different treatment, and the difference in response is the basis of the inference. In educational studies, a score on a pretest given to a student is paired with that student's post-test score to provide an evaluation of a new teaching method. Adjacent farm plots may be paired if they are of similar physical characteristics in order to study the effect of radiation on seeds, and so on. In fact, for any experiment where it is suspected that the difference between the two populations may be overshadowed by the variation within the two populations, the paired samples procedure should be appropriate.

Inferences on the difference in means of two populations based on paired samples use as data the simple differences between paired values. For example, in the diet study the observed value for each individual is obtained by subtracting the after weight from the before weight. The result becomes a single sample of differences, which can be analyzed in exactly the same way as any single sample experiment (Chapter 4). Thus the basic statistic is

$$t = \frac{\bar{d} - \delta_0}{\sqrt{s_d^2/n}},$$

where $\bar{d}$ is the mean of the sample differences, d_i; δ_0 is the population mean difference (usually zero); and s_d^2 is the estimated variance of the differences. When used in this way, the t statistic is usually called the "paired t statistic."

■ Example 5.6

For the first 60 years major league baseball consisted of 16 teams, eight each in the National and the American leagues. In 1961 the Los Angeles Angels and the Washington Senators became the first expansion teams in baseball history. It is conjectured that the main reason that the league allowed expansion teams was the fact that total attendance dropped from 20 million in 1960 to slightly over

17 million in 1961. Table 5.7 shows the total ticket sales for the 16 teams for the two years 1960 and 1961. Examination of the data (helped by Fig. 5.5) shows the reason that a paired t test would be appropriate to determine whether the average attendance did in fact drop significantly from 1960 to 1961. The variation among

Table 5.7 Baseball Attendance (Thousands)

Team	1960	1961	Diff.
1	809	673	−136
2	663	1123	460
3	2253	1813	−440
4	1497	1100	−397
5	862	584	−278
6	1705	1199	−506
7	1096	855	−241
8	1795	1391	−404
9	1187	951	−236
10	1129	850	−279
11	1644	1151	−493
12	950	735	−215
13	1167	1606	439
14	774	683	−91
15	1627	1747	120
16	743	597	−146

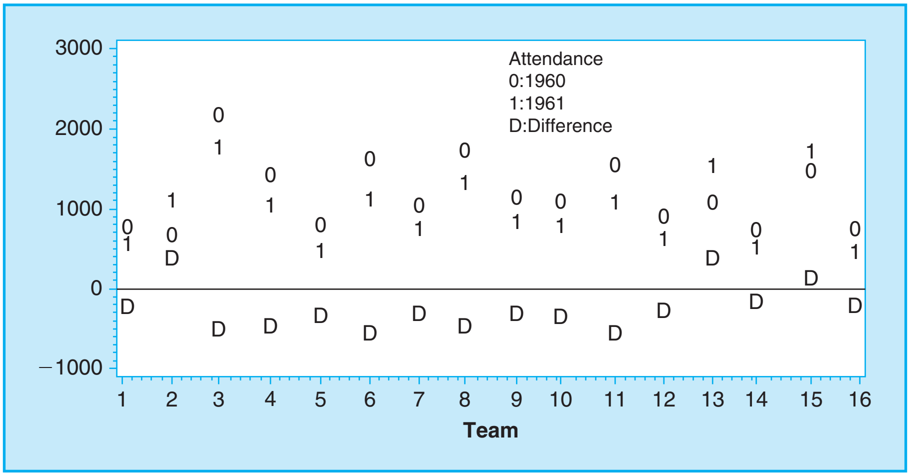

FIGURE 5.5
Baseball Attendance Data.

the attendance figures from team to team is extremely large—going from around 663,000 for team 2 to 2,253,000 for team 3 in 1960, for example. The variation between years by individual teams, on the other hand, is relative small—the largest being 506,000 by team 6.

Solution

The attendance data for the 16 major league teams for 1960 and 1961 are given in Table 5.7. The individual differences $d = y_{1961} - y_{1960}$ are used for the analysis. Positive differences indicate increased attendance while negative numbers that predominate here indicate decreased attendance. The hypotheses are

$$H_0\colon \delta_0 = 0,$$
$$H_1\colon \delta_0 < 0,$$

where δ_0 is the mean of the population differences. Note that we started out with 32 observations and ended up with only 16 pairs. Thus the mean and variance used to compute the test statistic are based on only 16 observations. This means that the estimate of the variance has 15 degrees of freedom and thus the t distribution for this statistic also has 15 degrees of freedom.

The test statistic is computed from the differences, d_i, using the computations

$$n = 16, \quad \sum d_i = -2843, \quad \sum d_i^2 = 1{,}795{,}451,$$
$$\bar{d} = -177.69, \quad \text{SS}_d = 1{,}290{,}285, \quad s_d^2 = 86{,}019,$$

and the test statistic t has the value

$$t = (-177.69)/\sqrt{(86{,}019/16)} = -2.423.$$

The (one-tailed) 0.05 rejection region for the Student's t distribution with 15 degrees of freedom is -1.7531; hence we reject the null hypothesis and conclude that average attendance has decreased. The p value for this test statistic (from a computer program) is $p = 0.0150$.

A confidence interval on the mean difference is obtained using the t distribution in the same manner as was done in Chapter 4. We will need the upper confidence limit on the increase (equivalent to lower limit for decrease) from 1960 to 1961. The upper limit is

$$\bar{d} + t_\alpha\sqrt{s_d^2/n},$$

which results in

$$-177.69 + (1.753)\sqrt{(86{,}019/16)} = -49.16;$$

hence, we are 0.95 confident that the true mean decrease is at least 49.16 (thousand).

The benefit of pairing Example 5.6 can be seen by pretending that the data resulted from independent samples. The resulting pooled t statistic would have the value $t = -1.164$ with 30 degrees of freedom. This value would not be significant at the 0.05 level and the test would result in a different conclusion. The reason for this result is seen by examining the variance estimates. The pooled variance estimate is quite large and reflects variation among teams that is irrelevant for studying year-to-year attendance changes. As a result, the paired t statistic will detect smaller differences, thereby providing more power, that is, a greater probability of correctly rejecting the null hypothesis (or equivalently give a narrower confidence interval). ■

It is important to note that while we performed both tests for this example, it was for demonstration purposes only! In a practical application, only procedures appropriate for the design employed in the study may be performed. That is, in this example only the paired t statistic may be used because the data resulted from paired samples.

The question may be asked: "Why not pair all two-population studies?" The answer is that not all experimental situations lend themselves to pairing. In some instances it is impossible to pair the data. In other cases there is not a sufficient physical relationship for the pairing to be effective. In such cases pairing will be detrimental to the outcome because in the act of pairing we "sacrifice" degrees of freedom for the test statistic. That is, assuming equal sample sizes, we go from $2(n-1)$ degrees of freedom in the independent sample case to $(n-1)$ in the paired case. An examination of the t table illustrates the fact that for smaller degrees of freedom the critical value are larger in magnitude, thereby requiring a larger value of the test statistic. Since pairing does not affect the mean difference, it is effective only if the variances of the two populations are definitely larger than the variances among paired differences. Fortunately, the desired condition for pairing often occurs if a physical reason exists for pairing.

■ Example 5.7

Two measures of blood pressure are known as systolic and diastolic. Now everyone knows that high blood pressure is bad news. However, a small difference between the two measures is also of concern. The estimation of this difference is a natural application of paired samples since both measurements are always taken together for any individual. In Table 5.8 are systolic (RSBP) and diastolic (RDBP) pressures of 15 males aged 40 and over participating in a health study. Also given is the difference (DIFF). What we want to do is to construct a confidence interval on the true mean difference between the two pressures.

Table 5.8 Blood Pressures of Males

OBS	RSBP	RDBP	DIFF
1	100	75	25
2	135	85	50
3	110	78	32
4	110	75	35
5	142	96	46
6	120	74	46
7	140	90	50
8	110	76	34
9	122	80	42
10	140	90	50
11	150	110	40
12	120	78	42
13	132	88	44
14	112	72	40
15	120	80	40

Solution

Using the differences, we obtain $\bar{d} = 41.0667$ and $s_d^2 = 52.067$, and the standard error of the difference is

$$\sqrt{\frac{52.067}{15}} = 1.863.$$

The 0.95 two-tailed value of the t distribution for 14 degrees of freedom is 2.148. The confidence interval is computed

$$41.0667 \pm (2.1448)(1.863),$$

which produces the interval 37.071 to 45.062.

If we had assumed that these data represented independent samples of 15 systolic and 15 diastolic readings, the standard error of mean difference would be 4.644, resulting in a 0.95 confidence interval from 31.557 to 50.577, which is quite a bit wider. As noted, pairing here is obvious, and it is unlikely that anyone would consider independent samples. ■

5.5 INFERENCES ON PROPORTIONS

In Chapter 2 we presented the concept of a binomial distribution, and in Chapter 4 we used this distribution for making inferences on the proportion of "successes" in a binomial population. In this section we present procedures for inferences on differences in the proportions of successes using independent as well as dependent samples from two binomial populations.

5.5.1 Comparing Proportions Using Independent Samples

Assume we have two binomial populations for which the probability of success in population 1 is p_1 and in population 2 is p_2. Based on independent samples of size n_1 and n_2 we want to make inferences on the difference between p_1 and p_2, that is, $(p_1 - p_2)$. The estimate of p_1 is $\hat{p}_1 = y_1/n_1$, where y_1 is the number of successes in sample 1, and likewise the estimate of p_2 is $\hat{p}_2 = y_2/n_2$. Assuming sufficiently large sample sizes (see Section 4.3), the difference $(\hat{p}_1 - \hat{p}_2)$ is normally distributed with mean

$$p_1 - p_2$$

and variance

$$p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2.$$

Therefore the appropriate statistic for inferences on $(p_1 - p_2)$ is

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}},$$

which has the standard normal distribution.

Note that the expression for the variance of the difference contains the unknown parameters p_1 and p_2. In the single-population case, the null hypothesis value for the population parameter p was used in calculating the variance. In the two-population case the null hypothesis is for equal proportions and we therefore use an estimate of this common proportion for the variance formula. Letting $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions for samples 1 and 2, respectively, the estimate of the common proportion p is a weighted mean of the two-sample proportions,

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2},$$

or, in terms of the observed frequencies,

$$\bar{p} = \frac{y_1 + y_2}{n_1 + n_2}.$$

The test statistic is now computed:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}}.$$

In construction of a confidence interval for the difference in proportions, we can not assume a common proportion, hence we use the individual estimates $\hat{p}_1$ and $\hat{p}_2$ in the variance estimate. The $(1-\alpha)$ confidence interval on the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{(\hat{p}_1(1-\hat{p}_1)/n_1) + (\hat{p}_2(1-\hat{p}_2)/n_2)}.$$

As in the one-population case the use of the t distribution is not appropriate since the variance is not calculated as a sum of squares divided by degrees of freedom. However, samples must be reasonably large in order to use the normal approximation.

■ Example 5.8

A candidate for political office wants to determine whether there is a difference in his popularity between men and women. To establish the existence of this difference, he conducts a sample survey of voters. The sample contains 250 men and 250 women, of which 42% of the men and 51.2% of the women favor his candidacy. Do these values indicate a difference in popularity?

Solution

Let p_1 denote the proportion of men and p_2 the proportion of women favoring the candidate, then the appropriate hypotheses are

$$H_0{:}\, p_1 = p_2,$$
$$H_1{:}\, p_1 \neq p_2.$$

The estimate of the common proportion is computed using the frequencies of successes:

$$\bar{p} = (105 + 128)/(250 + 250) = 0.466.$$

The test statistic then has the value

$$\begin{aligned} z &= (0.42 - 0.512)/\sqrt{[(0.466)(0.534)(1/250 + 1/250)]} \\ &= -0.092/0.0446 = -2.06. \end{aligned}$$

The two-tailed p value for this test statistic (obtained from the standard normal table) is $p = 0.0392$. Thus the hypothesis is rejected at the 0.05 level, indicating that there is a difference between the sexes in the degree of support for the candidate.

We can construct a 0.95 confidence interval on the difference $(p_1 - p_2)$ as

$$(0.42 - 0.512) \pm (1.96)\sqrt{[(0.42)(0.58)/250] + [(0.512)(0.488)/250]},$$

or

$$-0.09 \pm (1.96)(0.0444).$$

Thus we are 95% confident that the true difference in preference by sex is between 0.005 and 0.179. ■

An Alternate Approximation for the Confidence Interval

In Section 4.3 we gave an alternative approximation for the confidence interval on a single proportion. In Agresti and Caffo (2000), it is pointed out that the method of obtaining a confidence interval on the difference between p_1 and p_2 presented previously also tends to result in an interval that does not actually provide the specified level of confidence.

The solution, as proposed by Agresti and Caffo, is to add one success and one failure to each sample, and then use the standard formula to calculate the confidence interval. This adjustment results in much better performance of the confidence interval, even with relative small samples. Using this adjustment, the interval is based on new estimates of $p_1, \tilde{p}_1 = (y_1 + 1)/(n_1 + 2)$ and $p_2, \tilde{p}_2 = (y_2 + 1)/(n_2 + 2)$. For Example 5.8, the interval would be based on $\tilde{p}_1 = 106/252 = 0.417$ and $\tilde{p}_2 = 129/252 = 0.512$. The resulting confidence interval would be

$$0.417 - 0.512 \pm (1.96)\sqrt{\frac{(0.417)(0.583)}{252} + \frac{(0.512)(0.488)}{252}}$$

or

$$-0.095 \pm 0.087, \text{ or}$$

the interval would be from −0.182 to −0.008. As in Chapter 4, this interval is not much different from the one constructed without the adjustment, mainly because the sample sizes are quite large and both sample proportions are close to 0.5. If the sample sizes were small, this approximation would result in a more reliable confidence interval.

5.5.2 Comparing Proportions Using Paired Samples

A binomial response may occur in paired samples and, as is the case for inferences on means, a different analysis procedure that is most easily presented with an example must be used.

CASE STUDY 5.2

Butler *et al.* (2004) studied audit conclusions available from Compustat. During the period after the institution of the SAS 58 reporting protocols, Big 5 accounting firms issued 4911 unqualified (favorable) opinions out of 6638 reports. Non-Big 5 accounting firms issued 912 unqualified opinions out of 1397 reports.

We can use the independent samples z test for proportions to compare the probability of receiving an unqualified (favorable) opinion from the two types of accounting firms, $z = 6.62$, p value < 0.0001. The two types of firms have substantially different probabilities of issuing an unqualified opinion.

In interpreting this result, it is important to remember that this is observational data rather than experimental. The researchers did not randomly assign companies to accounting firms. Hence, the difference we have seen may not be because of the accounting firms' practices or skill, but because of the types of companies selecting the firms. For example, smaller or financially less stable companies may tend to choose non-Big 5 accounting firms.

■ Example 5.9

In an experiment for evaluating a new headache remedy, 80 chronic headache sufferers are given a standard remedy and a new drug on different days, and the response is whether their headache was relieved. In the experiment 56, or 70%, were relieved by the standard remedy and 64, or 80%, by the new drug. Do the data indicate a difference in the proportion of headaches relieved?

Solution

The usual binomial test is not correct for this situation because it is based on a total of 160 observations, while there are only 80 experimental units (patients). Instead, a different procedure, called McNemar's test, must be used. For this test, the presentation of results is shown in Table 5.9. In this table the 10 individuals helped by neither drug and the 50 who were helped by both are called **concordant pairs**, and do not provide information on the relative merits of the two preparations. Those whose responses differ for the two drugs are called **discordant pairs**. Among these, the 14 who were not helped by the standard but were helped by the new can be called "successes," while the 6 who were helped by the old and not the new can be called "failures." If both drugs are equally effective, the proportion of successes among the discordant pairs should be 0.5, while if the new drug is

Table 5.9 Data on Headache Remedy

	STANDARD REMEDY		
	Headache	**No Headache**	**Totals**
New drug			
Headache	10	6	16
No Headache	14	50	64
Totals	24	56	80

more effective, the proportion of successes should be greater than 0.5. The test for ascertaining the effectiveness of the new drug, then, is to determine whether the sample proportion of successes, 14/20 = 0.7, provides evidence to reject the null hypothesis that the true proportion is 0.5. This is a simple application of the one-sample binomial test (Section 4.3) for which the test statistic is

$$z = \frac{0.7 - 0.5}{\sqrt{[(0.5)(0.5)]/20}} = 1.789.$$

Since this is a one-tailed test, the critical value is 1.64485, and we may reject the hypothesis of no effect. ■

5.6 ASSUMPTIONS AND REMEDIAL METHODS

This chapter has been largely concerned with the comparison of means and variances of two populations. Yet we noted in Chapter 1 that means and variances are not necessarily good descriptors for populations with highly skewed distributions. This consideration leads to a discussion of assumptions underlying the proper use of the methods presented in this chapter. These assumptions can be summarized as follows.

1. *The pooled t statistic:*
   (a) The two samples are independent.
   (b) The distributions of the two populations are normal or of such a size that the central limit theorem is applicable.
   (c) The variances of the two populations are equal.
2. *The paired t statistic:*
   (a) The observations are paired.
   (b) The distribution of the differences is normal or of such a size that the central limit theorem is applicable.
3. *Inferences on binomial populations:*
   (a) Observations are independent (for McNemar's test *pairs* are independent).
   (b) The probability of success is constant for all observations.
   (c) Sample sizes are adequate for the normal approximation.
4. *Inferences on variances:*
   (a) The samples are independent.
   (b) The distributions of the two populations are approximately normal.

When assumptions are not fulfilled, the analysis is not appropriate and/or the significance levels (p values) are not as advertised. In other words, conclusions that arise from the inferences may be misleading, which means any recommendations or actions that follow may not have the expected results.

Most of the assumptions are relatively straightforward and violations easily detected by simply examining the data collection procedure. Major problems arise from (1) distributions that are distinctly nonnormal so that the means and variances are not useful measures of location and dispersion and/or the central limit theorem does not work, and, of course, (2) scenarios where the equal variance assumption does not hold.

Violation of distributional assumptions may be detected by the exploratory data analysis methods described in Chapter 1, which should be routinely applied to all data. The F test for equal variances may be used to detect violation of the equal variance assumption.[2]

What to do when assumptions are not fulfilled is not clear-cut. For the t statistics, minor violations are not particularly serious because these statistics are relatively robust; that is, they do not lose validity for modest departures from the assumptions. The inferences on variances are not quite so robust, because if a distribution is distinctly nonnormal, the variance may not be a good measure of dispersion. Therefore, for cases in which the robustness of the t statistics fails as well as for other cases of violated assumptions, it will be necessary to investigate other analysis strategies. In Section 4.5 we used a test on the median in a situation where the use of the mean was not appropriate. The procedure for comparing two medians is illustrated below.

Comparing medians is, however, not always appropriate. For example, population distributions may have different shapes and then neither means nor variances nor medians may provide the proper comparative measures. A wide variety of analysis procedures, called **nonparametric methods**, are available for such situations and a selection of such methods is presented in Chapter 14, where Section 14.3 is devoted to a two-sample comparison.

■ Example 1.4: Revisited

In Example 4.7 we noted that the existence of extreme observations may compromise the usefulness of inferences on a mean and that an inference on the median may be more useful. The same principle can be applied to inferences for two populations. One purpose of collecting the data for Example 1.4 was to determine whether Cytosol levels are a good indicator of cancer. We noted that the distribution of Cytosol levels (Table 1.11 and Fig. 1.11) is highly skewed and dominated by a few extreme values. For comparing Cytosol levels for patients diagnosed as having or not having cancer, the side-by-side box plots in Fig. 5.6 also show that the variances of the two samples are very different. How can the comparison be made?

Solution

Since we can see that using the t test to compare means is not going to be appropriate, it may be more useful to test the null hypothesis that the two populations have the same median. The test is performed as follows:

1. Find the overall median, which is 25.5.

[2] Some will argue that one should not test for violation of assumptions. We will not attempt to answer that argument.

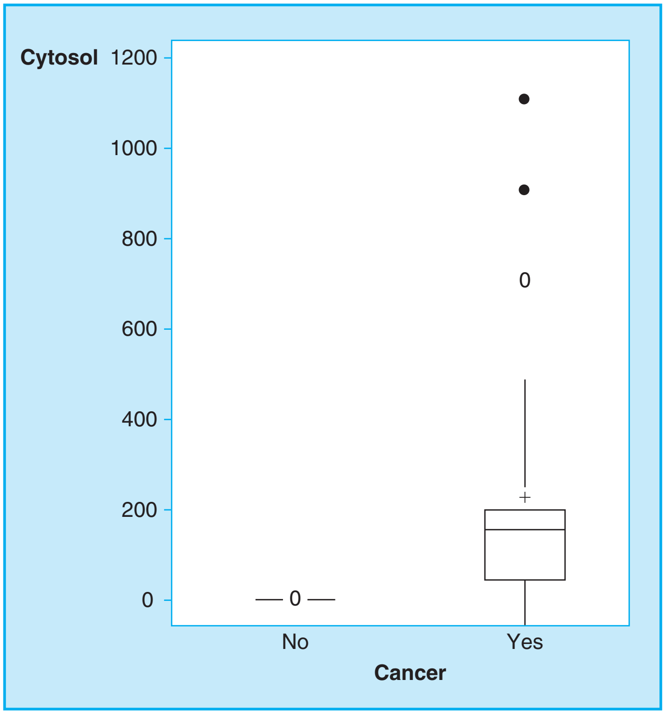

FIGURE 5.6
Box Plot of CYTOSOL.

2. Obtain the proportion of observations above the median for each of the two samples. These are $0/17 = 0.0$ for the no cancer patients and $21/25 = 0.84$ for the cancer patients.
3. Test the hypothesis that the proportion of patients above the median is the same for both populations, using the test for equality of two proportions. The overall proportion is 0.5; hence the test statistic is

$$\begin{aligned} z &= \frac{0.0 - 0.84}{\sqrt{(0.5)(0.5)(1/17 + 1/25)}} \\ &= \frac{-0.84}{0.157} \\ &= -5.35, \end{aligned}$$

which easily leads to rejection.

In this example the difference between the two samples is so large that any test will declare a significant difference. However, the median test has a useful interpretation in that if the median were to be used as a cancer diagnostic, none of the no-cancer patients and only four of the cancer patients would be misdiagnosed. ■

Example 5.4: Revisited

This example had unequal variances and was analyzed using the unequal variance procedure, which resulted in finding inadequate evidence of different mean attitude scores for the two populations of commuters. Can we use the procedure above to perform the same analysis? What are the results?

Solution

Using the test for equality of medians, we find that the overall median is 2 and the proportions of observations above the median are 0.6 for the subway and 0.38 for the rail commuters. The binomial test, for which sample sizes are barely adequate, results in a z statistic of 1.10. There is no significant evidence that the median scores differ.

5.7 CHAPTER SUMMARY

Solution to Example 5.1

In the introduction to this chapter, we posed the question of whether the typical declines in stock prices differed for the Consumer Staples and Financial categories. A glance at Figure 5.1 shows that although the relative decreases are still right-skewed, they are less drastically nonnormal than the decreases. Hence, our analysis will use the relative decreases, shown in Table 5.10. Table 5.11 presents the summary statistics and a variety of test statistics. Each t test statistic has its degrees of freedom noted within parentheses.

Table 5.10 Relative Decreases in Stock Prices during 2008

Consumer Staples	31.30	13.65	10.04	27.18	16.54	14.60	3.78	13.82	13.40	42.02
Financial	34.55	63.10	22.75	13.99	15.39	25.16	5.56	67.90	64.01	16.62

Table 5.11 Summary Values for Relative Decreases in Stock Prices during 2008

	n	Sample Mean	Sample S.D.	Paired t Test
Consumer Staples	10	18.634	11.408	$t(9) = 5.17$ p value $= 0.0006$
Financial	10	32.902	23.447	$t(9) = 4.44$ p value $= 0.0016$
Independent samples t unequal variance		$t(13) = -1.73$ p value $= 0.1071$		

The paired t tests are for the null hypotheses that the mean relative decrease within each category is zero. Each test statistic is clearly significant, but how do we interpret these results? The easiest interpretation is that we have strong evidence that if we had written down the price decreases for *all* the stocks in, say, the Consumer Staples category, that mean decrease would be nonzero. In fact, since the collection of all stocks in this category is only four times the size of our sample, we have somewhat underestimated the strength of evidence for this interpretation. This conclusion is not very interesting; after all, with a little more work, we could have written down the price decreases in the complete collection. If we had done so, would this have meant that we would not need test statistics? We would still do so, if we regard the actual observations on the stocks as a random sample from a larger conceptual population of all possible relative decreases. The values in this hypothetical population follow a probability distribution influenced by the market's conditions. To generalize our results to this population of "what might have been" requires the kinds of hypothesis tests presented here.

The independent samples t test is for the null hypothesis that the means in these two underlying hypothetical distributions are equal. We have used the unequal variance version since the box plots have apparently different spreads. The 13 degrees of freedom was computed by SAS using Satterthwaite's approximation. The apparent difference in the means is not significant (p value $= 0.1071$)! That is, there is no significant evidence of a difference between mean changes in the Consumer Staples and Financial categories. The conclusion refers to the means of the hypothetical underlying distributions.

Dispersion is also of interest when comparing groups of stock prices. An F test for the null hypothesis of equal variances showed modest evidence of a difference ($F(9,9) = 4.22, p$ value $= 0.0431$). The Count Five Rule did not establish a difference, as only four of the five largest absolute deviations were from the Financial category.

Since the sample sizes are small and the sample distributions are still skewed, we might prefer a comparison of medians. The overall median for the combined sample was 16.58. There were three observations in Consumer Staples that exceeded this amount, and seven in Financial. Using the independent samples z test for proportions, these proportions are not significantly different ($z = 1.79$, p value $= 0.074$). Hence, there is no evidence that the medians differ. Since the sample sizes within each group are small, the z test for proportions may not be appropriate. Another test, called Fisher's exact test (see Section 12.4) would be used instead. Calculation of its p value is best done by statistical software. In this case, SAS gives Fisher's exact test p value $= 0.1789$, again leading to the conclusion that the medians do not differ more than could be attributed to chance. ■

This chapter provides the methodology for making inferences on differences between two populations. The focus is on differences in means, variances, and proportions. In performing two-sample inferences it is important to know whether the two samples are independent or dependent (paired). The following specific inference procedures were presented in this chapter:

- Inferences on means based on independent samples where the variances are assumed known use the variance of a linear function of random variables to generate a test statistic having the standard normal distribution. This method has little direct practical application but provides the principles to be used for the methods that follow.
- Inferences on means based on independent samples where the variances can be assumed equal use a single (pooled) estimate of the common variance in a test statistic having the Student's t distribution.
- Inferences on means based on independent samples where the variances cannot be assumed equal use the estimated variances as if they were the known population variances for large samples. For small samples an approximation must be used.
- Inferences on means based on dependent (paired) samples use differences between the pairs as the variable to be analyzed.
- Inferences on variances use the F distribution, which describes the sampling distribution on the ratio of two estimated variances.
- Inferences on proportions from independent samples use the normal approximation of the binomial to compute a statistic similar to that for inferences on means when variances are assumed known.
- Inferences on proportions from dependent samples use a statistic based on information only on pairs whose responses differ between the two groups.
- Inferences on medians are performed by adapting the method used for inferences on proportions.
- A final section discusses assumptions underlying the various procedures for comparing two populations and includes a brief discussion of detection of violations and some alternative methods.

5.8 CHAPTER EXERCISES

Concept Questions

Indicate true or false for the following statements. If false, specify what change will make the statement true.

1. ______________One of the assumptions underlying the use of the (pooled) two-sample test is that the samples are drawn from populations having equal means.

2. ______________In the two-sample t test, the number of degrees of freedom for the test statistic increases as sample sizes increase.

3. ______________A two-sample test is twice as powerful as a one-sample test.

4. ______________If every observation is multiplied by 2, then the t statistic is multiplied by 2.

5. ______________ When the means of two independent samples are used to compare two population means, we are dealing with dependent (paired) samples.

6. ______________ The use of paired samples allows for the control of variation because each pair is subject to the same common sources of variability.

7. ______________ The χ^2 distribution is used for making inferences about two population variances.

8. ______________ The F distribution is used for testing differences between means of paired samples.

9. ______________ The standard normal (z) score may be used for inferences concerning population proportions.

10. ______________ The F distribution is symmetric and has a mean of 0.

11. ______________ The F distribution is skewed and its mean is close to 1.

12. ______________The pooled variance estimate is used when comparing means of two populations using independent samples.

13. ______________ It is not necessary to have equal sample sizes for the paired t test.

14. ______________ If the calculated value of the t statistic is negative, then there is strong evidence that the null hypothesis is false.

Practice Exercises

The following exercises are designed to give the reader practice in doing statistical inferences on two populations through the use of sample examples with small data sets. The solutions are given in the back of the text.

1. An engineer was comparing the output from two different processes by independently sampling each one. From process A she took a sample of $n_1 = 64$, which yielded a sample mean of $\bar{y}_1 = 12.5$. Process A has a known standard deviation, $\sigma = 2.1$. From process B she took a sample of $n_2 = 100$, which yielded a sample mean of $\bar{y}_2 = 11.9$. Process B has a known standard deviation of $\sigma = 2.2$. At $\alpha = 0.05$ would the engineer conclude that both processes had the same average output?

2. The results of two independent samples from two populations are listed below:

Sample 1: 17, 19, 10, 29, 27, 21, 17, 17, 14, 20
Sample 2: 26, 24, 26, 29, 15, 29, 31, 25, 18, 26

Use the 0.05 level of significance and test the hypothesis that the two populations have equal means. Assume the two samples come from populations whose standard deviations are equal.

3. Using the data in Exercise 2, compute the 0.90 confidence interval on the difference between the two population means, $\mu_1 - \mu_2$.

4. The following weights in ounces resulted from a sample of laboratory rats on a particular diet. Use $\alpha = 0.05$ and test whether the diet was effective in reducing weight.

Rat	1	2	3	4	5	6	7	8	9	10
Before	14	27	19	17	19	12	15	15	21	19
After	16	18	17	16	16	11	15	12	21	18

5. In a test of a new medication, 65 out of 98 males and 45 out of 85 females responded positively. At the 0.05 level of significance, can we say that the drug is more effective for males?

Exercises

1. Two sections of a class in statistics were taught by two different methods. Students' scores on a standardized test are shown in Table 5.12. Do the results present evidence of a difference in the effectiveness of the two methods? (Use $\alpha = 0.05$.)

Table 5.12 Data for Exercise 1

Class A		Class B	
74	76	78	79
97	75	92	76
79	82	94	93
88	86	78	82
78	100	71	69
93	94	85	84
		70	

2. Construct a 95% confidence interval on the mean difference in the scores for the two classes in Exercise 1.

3. Table 5.13 shows the observed pollution indexes of air samples in two areas of a city. Test the hypothesis that the mean pollution indexes are the same for the two areas. (Use $\alpha = 0.05$.)

4. A closer examination of the records of the air samples in Exercise 3 reveals that each line of the data actually represents readings on the same day: 2.92 and 1.84 are from day 1, and so forth. Does this affect the validity of the results obtained in Exercise 3? If so, reanalyze.

5. To assess the effectiveness of a new diet formulation, a sample of 8 steers is fed a regular diet and another sample of 10 steers is fed a new diet. The weights of the steers at 1 year are given in Table 5.14. Do these results imply that the new diet results in higher weights? (Use $\alpha = 0.05$.)

Table 5.13 Data for Exercise 3

Area A	Area B
2.92	1.84
1.88	0.95
5.35	4.26
3.81	3.18
4.69	3.44
4.86	3.69
5.81	4.95
5.55	4.47

Table 5.14 Data for Exercise 5

Regular Diet	New Diet
831	870
858	882
833	896
860	925
922	842
875	908
797	944
788	927
	965
	887

6. Assume that in Exercise 5 the new diet costs more than the old one. The cost is approximately equal to the value of 25 lb. of additional weight. Does this affect the results obtained in Exercise 5? Redo the problem if necessary.

7. In a test of the reliability of products produced by two machines, machine A produced 7 defective parts in a run of 140, while machine B produced 10 defective parts in a run of 200. Do these results imply a difference in the reliability of these two machines?

8. In a test of the effectiveness of a device that is supposed to increase gasoline mileage in automobiles, 12 cars were run, in random order, over a prescribed course both with and without the device in random order. The mileages (mpg) are given in Table 5.15. Is there evidence that the device is effective?

Table 5.15 Data for Exercise 8

Car No.	Without Device	With Device
1	21.0	20.6
2	30.0	29.9
3	29.8	30.7
4	27.3	26.5
5	27.7	26.7
6	33.1	32.8
7	18.8	21.7
8	26.2	28.2
9	28.0	28.9
10	18.9	19.9
11	29.3	32.4
12	21.0	22.0

9. A new method of teaching children to read promises more consistent improvement in reading ability across students. The new method is implemented in one randomly chosen class, while another class is randomly chosen to represent the standard method. Improvement in reading ability using a standardized test is given for the students in each class in Table 5.16. Use the appropriate test to see whether the claim can be substantiated.

Table 5.16 Data for Exercise 9

New Method		Standard Method	
13.0	16.7	20.1	27.0
15.1	16.7	16.7	19.2
16.5	18.4	25.6	19.3
19.0	16.6	25.4	26.7
20.2	19.4	22.0	14.7
19.9	23.6	16.8	16.9
23.3	16.5	23.8	23.7
17.3	24.5	23.6	21.7

10. The manager of a large office building needs to buy a large shipment of light bulbs. After reviewing specifications and prices from a number of suppliers, the choice is narrowed to two brands whose specifications with respect to price and quality appear identical. He purchases 40 bulbs of each brand and subjects them to an accelerated life test, recording hours to burnout, as shown in Table 5.17.
    (a) The manager intends to buy the bulbs with a longer mean life. Do the data provide sufficient evidence to make a choice?

(b) To save labor expense, the owners have decided that all bulbs will be replaced when 10% have burned out. Is the decision in part (a) still valid? Is an alternate test possibly more useful? (Suggest the test only; do not perform.)

Table 5.17 Data for Exercise 10

Brand *A* Life (Hours)				**Brand *B* Life (Hours)**			
915	992	1034	1080	1235	1238	1248	1273
1137	1211	1211	1218	1275	1282	1298	1303
1260	1276	1289	1306	1307	1335	1337	1339
1319	1336	1360	1387	1360	1383	1384	1384
1400	1405	1419	1437	1388	1390	1390	1390
1488	1543	1581	1603	1394	1394	1403	1410
1606	1614	1635	1669	1417	1419	1423	1426
1683	1746	1752	1776	1430	1442	1448	1469
1881	1928	1940	1960	1478	1485	1486	1501
2029	2053	2063	2737	1508	1514	1515	1517

11. Chlorinated hydrocarbons (mg/kg) found in samples of two species of fish in a lake are as follows:

Species 1:	34	1	167	20		
Species 2:	45	86	82	70	160	170

Perform a hypothesis test to determine whether there is a difference in the mean level of hydrocarbons between the two species. Check assumptions.

12. Eight samples of effluent from a pulp mill were each divided into 10 batches. From each sample, 5 randomly selected batches were subjected to a treatment process intended to remove toxic substances. Five fish of the same species were placed in each batch, and the mean number surviving in the 5 treated and untreated portions of each effluent sample after 5 days were recorded and are given in Table 5.18. Test to see whether the treatment increased the mean number of surviving fish.

Table 5.18 Data for Exercise 12

	MEAN NUMBER SURVIVING							
Sample No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Untreated	5	1	1.8	1	3.6	5	2.6	1
Treated	5	5	1.2	4.8	5	5	4.4	2

13. In Exercise 13 of Chapter 1, the half-life of aminoglycosides from a sample of 43 patients was recorded. The data are reproduced in Table 5.19. Use these data to see whether there is a significant difference in the mean half-life of Amikacin and Gentamicin. (Use $\alpha = 0.10$.)

Table 5.19 Half-Life of Aminoglycosides by Drug Type

Pat	Drug	Half-Life	Pat	Drug	Half-Life	Pat	Drug	Half-Life
1	G	1.60	16	A	1.00	31	G	1.80
2	A	2.50	17	G	2.86	32	G	1.70
3	G	1.90	18	A	1.50	33	G	1.60
4	G	2.30	19	A	3.15	34	G	2.20
5	A	2.20	20	A	1.44	35	G	2.20
6	A	1.60	21	A	1.26	36	G	2.40
7	A	1.30	22	A	1.98	37	G	1.70
8	A	1.20	23	A	1.98	38	G	2.00
9	G	1.80	24	A	1.87	39	G	1.40
10	G	2.50	25	G	2.89	40	G	1.90
11	A	1.60	26	A	2.31	41	G	2.00
12	A	2.20	27	A	1.40	42	A	2.80
13	A	2.20	28	A	2.48	43	A	0.69
14	G	1.70	29	G	1.98			
15	A	2.60	30	G	1.93			

14. Draw a stem and leaf plot of half-life for each drug in Exercise 13. Do the assumptions necessary for the test in Exercise 13 seem to be satisfied by the data? Explain.

15. In Exercise 12 of Chapter 1 a study of characteristics of successful salespersons indicated that 44 of 120 sales managers rated reliability as the most important characteristic in salespersons. A study of a different industry showed that 60 of 150 sales managers rated reliability as the most important characteristic of a successful salesperson.

(a) At the 0.05 level of significance, do these opinions differ from one industry to the other?

(b) Construct the power curve for this test. (*Hint:* The horizontal axis will be the difference between the proportions.)

16. Elevated levels of blood urea nitrogen (BUN) denote poor kidney function. Ten elderly cats showing early signs of renal failure are randomly divided into two groups. Group 1 (control group) is placed on a standard high-protein diet. Group 2 (intervention group) is placed on a low-phosphorus high-protein diet. Their BUN is measured both initially and three months later. The data is shown in Table 5.20. Use $\alpha = 0.05$ in all parts of this problem.

(a) Was there a significant increase in mean BUN for Group 1?
(b) Was there a significant increase in mean BUN for Group 2?
(c) Did the two groups differ in their mean change in BUN? If so, which appeared to have the least increase?

Table 5.20 Data for Exercise 16

	Group 1 (control)			Group 2 (intervention)	
Cat	Initial BUN	Final BUN	Cat	Initial BUN	Final BUN
1	52	58	6	55	53
2	41	41	7	61	64
3	49	58	8	48	50
4	62	75	9	40	42
5	39	44	10	54	52

17. Researchers at Wolfson Children's Hospital, Jacksonville, FL tested new technology meant to reduce the number of attempts needed to draw blood from children. They collected data on the number of successes on the first attempt using the new technology and on a historical comparison group using standard technology. This data is summarized in Table 5.21.

Table 5.21 Data for Exercise 17

	Standard Technology	New Technology
Successful on 1st	74	73
Unsuccessful on 1st	76	18
Total	150	91

(a) Is there evidence that the new technology changes the probability of a success on the first attempt? Is the change for the better or for the worse?
(b) The researchers also recorded the ages of the children. In the standard technology group, the 150 children had a mean age of 5.73 and a standard deviation of 6.15. In the new technology group, the mean age was 9.02 with a standard deviation of 6.10. Does the mean age of the children in the two groups differ significantly?
(c) How do the results of part (b) complicate the interpretation of part (a)? (Private communication, H. Hess, Wolfson Children's Hospital, 2009.)

18. Garcia and Ybarra (2007) describe an experiment in which 174 undergraduates were randomly divided into a people-accounting condition (describing a numerical imbalance in an award) and a control condition. A situation was described to them, and they made a choice that could either add to or detract from the

imbalance. Of the undergraduates in the control condition, 34% made a choice that detracted from the imbalance. Of those in the people-accounting condition, 55% made a choice that detracted from the imbalance. Is the difference in the two groups' proportions greater than can be attributed to Chance? (Assume there were 88 people in the control condition, and 86 in the people-accounting condition, and use $\alpha = 0.01$.)

19. In an experiment in which infants interacted with objects, Sommerville *et al.* (2005) randomly divided 30 infants into a reach-first versus watch-first condition. The authors' state,

> *Whereas 11 of 15 infants in the reach-first condition looked longer at the new goal events than the new path events, only 4 of 15 infants in the watch-first condition showed this looking time preference.*

Is the difference observed between the two groups greater than can be attributed to chance if you use $\alpha = 0.05$? What if you use $\alpha = 0.01$?

20. Martinussen *et al.* (2007) compared "burnout" among a sample of Norwegian police officers to a comparison group of air traffic controllers, journalists, and building constructors. Burnout was measured on three scales: exhaustion, cynicism, and efficacy. The data is summarized in Table 5.22. The authors state,

> *The overall level of burnout was not high among police compared to other occupational groups sampled from Norway. In fact, police scored significantly lower on exhaustion and cynicism than the comparison group, and the difference between groups was largest for exhaustion.*

Substantiate the authors' claims.

Table 5.22 Summary Statistics for Exercise 20

	Police, $n = 222$		Comparison Group, $n = 473$	
	Mean	**S.D.**	**Mean**	**S.D.**
Exhaustion	1.38	1.14	2.20	1.46
Cynicism	1.50	1.33	1.75	1.34
Efficacy	4.72	0.97	4.69	0.89

Projects

1. **Lake Data Set.** The Florida Lakewatch data set is described in Appendix C.1. It contains water quality information on a sample of lakes in North Central Florida taken during 2005. For most of the lakes, total phosphorus level is reported for a summer month and also for a winter month. Does the typical total phosphorus level appear to differ in the two months?
   (a) The summer versus winter difference can be expressed as
      (1) the simple difference WTRTP – SMRTP,

(2) the ratio WTRTP/SMRTP, or

(3) the logarithm of the ratio.

Using a graphical display of each of these variables, decide which could best be analyzed using a t test. (Only those lakes where both a summer and winter value are present can be used.)

(b) For the variable you chose in part (a), how would you express the null hypothesis of "no difference between the winter and summer typical phosphorous levels"?

(c) Carry out the hypothesis test for part (b), and interpret the results.

2. **NADP Data Set.** The NADP data set (see Appendix C.3) contains data on the water chemistry of precipitation (rain, snow, and sleet) at a large number of sites throughout the United States. Rainfall in the United States tends to be acidic, especially in the Northeast. A major focus of air quality rules in the last 30 years has been to reduce pollutants that contribute to this acid rain phenomenon. Using the pH values (PHLAB90 and PHLAB20) for the continental United States (MISSIS = 'E' or 'W'), is there evidence that pH values are increasing? (Recall that low pH values indicate more acidity.) Is the distribution of changes in pH different east and west of the Mississippi?

This page intentionally left blank