CHAPTER 14

Testing the Difference between Two Groups

Often a public or nonprofit administrator will have two samples and will want to know whether the values measured for one sample are different from those of the other sample on average. For example, a school administrator might want to know whether the reading levels of high school seniors improved after a special reading seminar was given: She wants a before-and-after comparison. A mental health counselor may want to know whether one type of treatment works better than another. A nonprofit executive may want to know whether agencies that contract with fund-raising firms net more donations than those that raise funds in-house. A librarian might want to know whether advertising affects circulation and thus might set up an experiment in which some branches advertise and others do not. In situations such as these, in which one wants to know whether two sample means are different or whether two sample proportions are different, the appropriate technique to use is a difference of means test.

Stating the Research and Null Hypotheses for Difference of Means Tests

In Chapter 11, we learned that when testing a hypothesis about a population using a single sample, the goal was to see whether a single sample with a particular mean could be drawn from a population with a known or hypothesized mean. The goal of testing hypotheses for two sample means is slightly different. When testing the difference between two sample means, the goal is to determine whether both sample means could have been drawn from the same population, or whether the two sample means are so different that they could not have been drawn from the same population.

For many management or policy evaluation issues, we expect the values for one sample to be different from those of the other sample. Research and null hypotheses are written to reflect this expectation. The general logic for the research hypothesis in a difference of means test is that one of the sample means is different (either smaller or larger) than the other sample mean. This statement is only a generic way of describing the underlying logic of a research hypothesis for a difference of means test. The actual hypotheses you write should be tailored to the specific research question at hand.

A thorough understanding of the research question helps clarify two important points necessary for carrying out a difference of means test: (1) the reason for hypothesizing a difference between the two groups (what makes one group different from the other?) and (2) the expected direction of the difference. For example, if we compare the academic performance of elementary school students who have participated in after-school learning programs to that of those who have not, it is too vague simply to hypothesize that the performance of the two groups will be different. Instead, we write the research hypothesis to reflect the expectation that the performance of program participants will be higher than that of nonparticipants.

Because we use samples to make inferences about larger populations, confirming the research hypothesis in a difference of means test indicates that the two population means are also different. In other words, there is a low probability that both samples could have been drawn from the same population. The general logic of the null hypothesis in a difference of means test is that the two sample means are not different. Failure to reject the null hypothesis indicates that the population means in question are not different.

To simplify, before examining the statistical issues involved in carrying out difference of means tests, we will illustrate how the process works using an example without data. Officials at the Bureau of Forms want to examine the effect that continuing education seminars have on employee performance. Half of the employees at the bureau have participated in continuing education seminars, whereas the other half have not. To see whether continuing education seminars are having an effect on performance, agency officials randomly select 50 employees who have participated in seminars and 50 who have not. After administering job skills tests to each of the samples, agency officials want to evaluate whether the test scores for the two groups of employees are different. The research and null hypotheses are as follows:

H_1: Employees who have taken continuing education seminars will have higher job skills scores.

H_0: There is no difference in job skills scores between employees who have taken continuing education seminars and those who have not.

In this case, if we were unable to reject the null hypothesis, the substantive conclusion would be that the mean test scores for the two populations of workers (those attending seminars and those not attending seminars) are not different. In other words, the population mean for both groups is the same, indicating that seminars are not leading to higher job skills scores.

If we were able to reject the null hypothesis, the substantive conclusion would be that the mean test scores for the two populations of workers are different. In other words, the scores for the population of workers who have participated in continuing education seminars are higher than the scores for workers who have not. Thus, the two population means are different, indicating that seminars are leading to higher job skills scores.

Now that you have been introduced to the general logic of difference of means tests, we present the statistical steps involved in the process.

Difference of Means Procedure

The best way to illustrate the difference of means test is with an example. The Ware County librarian wants to increase circulation from the Ware County bookmobiles. The librarian thinks that poster ads in areas where the bookmobiles stop will attract more browsers and increase circulation. To test this idea, the librarian sets up an experiment. Ten bookmobile routes are selected at random; on those routes, poster ads are posted with bookmobile information. Ten other bookmobile routes are randomly selected; on those routes, no advertising is done. In effect, the librarian has set up the following experiment:

Group	Treatment	Comparison
Experimental group	Place ads	Measure circulation
Control group	No ads	Measure circulation

After a week-long experiment, the information listed in Table 14.1 is available to the librarian.

The null hypothesis is that the mean circulation of the experimental group is not higher than the mean circulation of the control group. Testing the difference between two means tells us the probability that both groups could be drawn from the same population. More formally, the analyst wants to know if μ_e (the experimental mean) is greater than μ_c (the control mean) or, alternatively, to calculate

$$d = \mu_e - \mu_c$$

to test if $d = 0$. The procedure is as follows:

Step 1: Calculate the mean and standard deviation for each group. These calculations have already been performed and are shown in Table 14.1. We use the sample means and standard deviations as estimates of the population parameters.

Table 14.1 Librarian's Data

	Experimental Group	Control Group
Mean	526 books	475 books
Standard deviation	125	115

Step 2: Calculate the standard error of the mean estimate for each group.

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

Experimental group:

$$\text{s.e.} = \frac{125}{\sqrt{10}} = 39.5$$

Control group:

$$\text{s.e.} = \frac{115}{\sqrt{10}} = 36.4$$

Step 3: Calculate an overall or "pooled" standard error for both groups. The overall standard error is equal to the square root of the sum of the squared standard errors for each group. Symbolically, this can be expressed as

$$\text{s.e.}_d = \sqrt{\text{s.e.}_1^2 + \text{s.e.}_2^2}$$

For the present example, we have

$$\text{s.e.} = \sqrt{39.5^2 + 36.4^2} = \sqrt{1{,}560.25 + 1{,}324.96} = \sqrt{2{,}885.21} = 53.7$$

Step 4: Because we want to know the probability that the groups could have been drawn from the same population, and because we have a mean estimate and a standard error, we can calculate the following *t* score:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\text{s.e.}_d}$$

where $\overline{X}_1$ is the control group mean, $\overline{X}_2$ is the experimental group mean, and s.e. is the overall standard error. In the present example, we have

$$t = \frac{475 - 526}{53.7} = \frac{-51}{53.7} = -.95$$

Looking up a *t* score of –.95 in the *t* table presented as Table 3 (Statistical Tables) in the back of the book [degrees of freedom (df) is $n_1 + n_2 - 2$, or 18 in this case; the first column of Table 3 is labeled "df" for degrees of freedom], we find a probability of more than .1 (.18 if a computer is used). Statistically, we can say that there is more than 1 chance in 10 that the two samples could have been drawn from the same population (that is, there is no difference).

Step 5: If the research design shows no other possible causes, what can the librarian say managerially about the program?

Understanding the Three Major Difference of Means Tests

The preceding formula for a difference of means test is one of three such tests. It is the test that is used when the two samples are independent and the analyst is unwilling to assume that the two population variances are equal. There are two other tests: one for equal variances and one for dependent samples. To understand which test you should apply to a particular management question, you need to understand the difference between independent and dependent samples.

Independent samples are those in which cases across the two samples are not "paired" or matched in any way. The best procedure to obtain independent samples is through random sampling techniques. For example, if an analyst at the Internal Revenue Service (IRS) randomly selects two samples from a national database, each consisting of 250 tax returns, there is no reason to expect a one-to-one linkage or pairing between individual cases across the two samples. Case 1 from sample A might be a tax return filed by a male from Arkansas. Case 1 from sample B might be a tax return filed by a female from Nevada. The remaining 249 tax returns for each sample should be similarly diverse in terms of the background characteristics of filers. Because each sample is drawn randomly, we have no reason to expect paired relationships (such as each person in sample A being matched with a close relative in sample B) between individual cases across the two samples.

Dependent samples exist when each item in one sample is paired with an item in the second sample. A before-and-after test would generate dependent samples if the same cases were used both before and after. For example, an agency could select 50 employees with low performance scores and require them to attend mandatory performance workshops for a month. To see whether the workshops improve performance, the same 50 employees could be given performance exams after training has been completed. The scores for each employee are logically paired because each one has both a before and an after score.

The logic of dependent samples does not apply if the "before" and "after" cases are not paired. Let us assume that an agency with 1,000 employees has decided to select a random sample of 50 employees for drug testing. After obtaining the test results, agency officials undertake aggressive steps to reduce illegal drug usage among employees. Agency officials then draw a second random sample of 50 employees 3 months later to see whether the new policies are having an effect. The same 50 employees are not included in the before and after samples. Because cases for both samples were randomly selected, and there is no way to pair or connect the cases in the first sample with those in the second, the two samples are independent.

t Test Assuming Independent Samples with Unequal Variances

Of the three types of *t* test, the independent samples, unequal variances *t* test is the most conservative; that is, it is less likely than the other two *t* tests to reject the null hypothesis. Sampling error is one reason why samples often have unequal variances. The problem this circumstance poses for inference is that sampling error makes it harder to determine whether two sample means are truly different or different mainly because of different variances that result from sampling error. The *t* test for independent samples with unequal variances is conservative in the sense that the calculations for the standard errors are designed to take large differences in sample variances into account. This procedure helps to clarify whether the sample means themselves are truly different. The calculations for degrees of freedom when using the *t* test for independent samples with unequal variances are also more conservative than those for the other *t* tests, as explained in Box 14.1.

The independent samples, unequal variances *t* test is particularly useful when the number of cases in each sample is different or when the number of cases in one or both of the samples is small (less than 30 or so). For example, if one sample consists of 150 cases and the other of 20, the amount of sampling error for the smaller sample might be much larger than that for the larger sample. If this were the case, the variances for the two samples would also be different.

Although a *t* test that makes it harder to reject the null hypothesis might seem like a disadvantage, a more rigorous standard makes the occurrence of Type I errors less likely. A public or nonprofit manager could spend his or her entire life using this *t* test and make adequate decisions (similar to using the binomial probability distribution when the hypergeometric should be used). We illustrate the use of this test with an example.

Sharon Pebble, the city manager of Stone Creek, South Dakota, wants to determine whether her new personnel procedures are decreasing the time it takes to hire an employee. She takes a sample of 10 city bureaus and calculates the average time to hire an employee, in days, before and after implementation of the new procedure. She gets the following results:

Bureau	Before	After
A	36.4	32.2
B	49.2	45.2
C	26.8	31.3
D	32.2	27.1
E	41.9	33.4
F	29.8	29.0
G	36.7	24.1
H	39.2	38.2
I	42.3	38.0
J	41.9	37.2

Box 14.1 **Calculating Degrees of Freedom When Using the *t* Test for Independent Samples and Unequal Variances**

One reason the *t* test for independent samples and unequal variances is more conservative than other *t* tests is the way degrees of freedom are calculated. The formula* for calculating degrees of freedom when using the independent samples unequal variances *t* test is as follows:

$$\text{df} = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

where

s = sample variance

N = number of cases in sample

This formula generally produces lower df values than the formula that we have been using up to this point, df = $(n_1 + n_2 - 2)$. The lower the degrees of freedom, the larger the calculated *t* statistic must be when evaluating whether the null hypothesis can be rejected.

To simplify matters, we have used the less complex formula to calculate degrees of freedom for the problems at the end of the chapter. Although the degrees of freedom are sometimes the same using either formula (such as in the Stone Creek bureau example below), you should not assume that this situation will always be the case when you are analyzing your own data. Thus, you should familiarize yourself with the steps involved in calculating degrees of freedom for the independent samples, unequal variances *t* test. Statistical software packages (such as SPSS) and spreadsheet programs (such as Microsoft Excel) automatically calculate the correct degrees of freedom depending on the type of test (equal or unequal variances) that is selected.

*Formula for independent samples, unequal variances t test obtained from the National Institute of Standards and Technology Website: NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

The hypothesis, that personnel are being hired in less time than they were before the adoption of the new procedures, is the same for all three tests. The null hypothesis is also the same: There is no difference between the time it takes to hire new personnel before and after implementation of the new procedures.

First, we discuss the independent samples, unequal variances procedure.

Step 1: Estimate the mean and standard deviation for the period before the adoption of the new procedures and the period after:

Period	Mean	S
Before	37.64	6.71
After	33.57	6.23

Step 2: Calculate the standard error for each group:

$$\text{Before } \frac{6.71}{\sqrt{10}} = 2.12$$

$$\text{After } \frac{6.23}{\sqrt{10}} = 1.97$$

Step 3: Calculate the overall standard error:

$$\sqrt{2.12^2 + 1.97^2} = 2.89$$

Step 4: Calculate the *t* score for the difference of means:

$$\frac{37.64 - 33.57}{2.89} = \frac{4.07}{2.89} = 1.41$$

Step 5: Look up a *t* score of 1.41 with df (degrees of freedom) = 18 in Table 3 at the back of the book. That value is statistically significant at less than .1. So there is less than 1 chance in 10 that the samples could have been drawn from the same population and, thus, that the means are equal. The resulting t score is not large enough to safely reject the null hypothesis (assuming an alpha value of .05; review "How Sure Should a Person Be" in Chapter 12). Thus, Ms. Pebble could conclude that the new procedures had no impact.

t Test Assuming Independent Samples with Equal Variances

The *t* test for independent samples and equal variances is less conservative than the *t* test for independent samples and unequal variances, because the former generates smaller standard errors and larger *t* scores. There is nothing wrong with using the *t* test for independent samples and equal variances if you are certain that the two sample variances are equal. However, if you assume that the sample variances are equal and they really are not, the overall standard error produced using this test will usually be smaller than it should be. This result increases the likelihood of making a Type I error when testing a hypothesis (i.e., rejecting the null hypothesis when it should be retained).

You can formally evaluate whether two sample variances are equal by performing the Levene test, which is an option available in most statistical software packages (see Box 14.2 for an explanation of how to interpret the Levene test). An even simpler approach is always to use the independent samples, unequal variances *t* test unless you are absolutely certain that the two sample variances are equal.

Box 14.2 How to Interpret the Levene Test for Equality of Variances

When interpreting the Levene test, the null hypothesis is that the two sample variances are equal. The alternative (research) hypothesis is that the two sample variances are not equal. The test statistic in this case is an *F* statistic. If you use a statistical software package to perform a difference of means test, the program will calculate the exact probability that the null hypothesis is correct. The following table is a sample SPSS output for a difference of means test that includes results for the Levene test.

Independent Samples Test

	Levene's Test for Equality of Variances		*t* Test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	*t*	df	Sig. (two-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	87.392	.000	–15.386	1,038	.000	–6.2228	.4044	–7.0164	–5.4292
Equal variances not assumed			–14.921	790.145	.000	–6.2228	.4171	–7.0415	–5.4041

In the present example, $F = 87.392$. With a level of significance of .000, this result indicates that the probability that the two sample variances are equal is extremely small. Thus, the null hypothesis should be rejected, and we should assume that the sample variances are not equal. Although a level of significance of .05 is commonly used to evaluate the test statistic, an analyst can also choose a more stringent threshold (such as .01) when testing the null hypothesis. For more information on how to calculate and interpret the Levene test, see Kurtz (1999, p. 185).

The *t* test for independent samples and equal variances operates as follows:

Step 1: Estimate the mean and standard deviation for the period before and the period after implementation of the new procedures, as shown in Step 1 in the prior example.

Step 2: Calculate an overall standard deviation by using the following formula:

$$s_d = \sqrt{\frac{[(n_1 - 1)s_1^2] + [(n_2 - 1)s_2^2]}{n_1 + n_2 - 2}}$$

The overall standard error is essentially a weighted average of the two standard deviations:

$$\sqrt{\frac{(9 \times 6.71^2) + (9 \times 6.23^2)}{10 + 10 - 2}} = 6.47$$

Step 3: Convert this overall standard deviation to a standard error with the following formula:

$$\text{s.e.} = s_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$6.47 \times \sqrt{\frac{1}{10} + \frac{1}{10}} = 2.89$$

In this case, the overall or pooled standard error is identical to the one in the preceding independent samples, unequal variances example because the sample sizes are equal. If the after sample had had 20 observations rather than 10, the final standard error would have been 1.99 for this method and 2.53 for the unequal variances method. The standard error for this method is always less than or equal to that of the other method; thus, the *t* score is always greater than or equal to the one for unequal variances. In the case of samples of 10 and 20, the *t* score for unequal variances would be 1.61 and for equal variances would be 2.05.

Step 4: Calculate the *t* score for the difference of means:

$$\frac{37.64 - 33.57}{2.89} = \frac{4.07}{2.89} = 1.41$$

Step 5: Look up a *t* score of 1.41 with df = 18. That value is statistically significant at less than .1. So there is less than 1 chance in 10 that the samples could have been drawn from the same population and, thus, that the means are equal.

t Test Assuming Dependent Samples

Finally, the *t*-test procedure for dependent samples is much different. In the present case, all of the items are paired because before and after data exist for each of the 10 bureaus. Because the items are paired, simply subtract one from the other to get *d*, the difference between the two items.

Step 1: Perform the pairwise subtractions to obtain the differences.

Bureau	Before	After	Difference
A	36.4	32.2	4.2
B	49.2	45.2	4.0
C	26.8	31.3	– 4.5
D	32.2	27.1	5.1
E	41.9	33.4	8.5
F	29.8	29.0	.8
G	36.7	24.1	12.6
H	39.2	38.2	1.0
I	42.3	38.0	4.3
J	41.9	37.2	4.7

The remaining steps are performed on the differences, rather than on the original data. The results are treated as a case of statistical inference on the differences.

Step 2: Calculate the mean and standard deviation of the differences. In this case, the mean is 4.07, and the standard deviation is 4.56.

Step 3: Calculate the standard error using the normal formula of dividing the standard deviation by the square root of the sample size. In this case, the standard error is

$$\text{s.e.} = \frac{4.56}{\sqrt{10}} = 1.44$$

Step 4: Calculate a t score with $n = 10$, or 9 df to see whether the mean is different from zero:

$$t = \frac{4.07 - 0}{1.44} = 2.83$$

Step 5: This t score is statistically significant at less than .01. The dependent samples t test produces the most significant results, but it can be used only when the samples are dependent. In this case, Ms. Pebble would likely conclude that there was a decrease in the time to hire new employees after the new procedures were implemented.

For the remainder of this chapter and the problems, we will assume independent samples and unequal variances unless otherwise specified.

Proportions

The t test is a technique that can be used both for the difference between two sample means and for the difference between two sample proportions. For example, the Morgan City parole board has been running an experimental program on one-third of its parolees. The parolees in the experimental program are placed in halfway houses run by nonprofit organizations that try to ease the parolees' readjustment to society. All other parolees are simply released and asked to check in with their parole officer once a month. The parole board wanted to evaluate the experimental program and decided that if the experimental program significantly reduced the recidivism rate of parolees, then the program would be declared a success. A random sample of 100 parolees who were placed in halfway houses is selected. Their names are traced through the Nationwide Criminal Data System (NCDS); 68 have been arrested again and convicted. Two hundred randomly selected parolees who were not assigned to halfway houses were also traced through the NCDS, and 148 of these people were in jail. Is the recidivism rate for parolees sent to halfway houses lower than the rate for the other parolees?

The process of analysis of variance for proportions is identical to that for sample means:

Step 1: Calculate the sample proportions (means) and standard deviations. For the experimental group, we have

$$p = \frac{68}{100} = .68$$

$$s = \sqrt{p(1-p)} = \sqrt{.68 \times .32} = .47$$

For the control group, we have

$$p = \frac{148}{200} = .74 \text{ and } s = \sqrt{.74 \times .26} = .44$$

Step 2: Calculate the standard error of the proportion estimate for each group.

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

Experimental group:

$$\text{s.e.} = \frac{.47}{\sqrt{100}} = .047$$

Control group:

$$\text{s.e.} = \frac{.44}{\sqrt{200}} = .031$$

Step 3: Calculate an overall standard error for both groups.

$$\text{s.e.}_d = \sqrt{\text{s.e.}_1^2 + \text{s.e.}_2^2} = \sqrt{.047^2 + .031^2} = \sqrt{.00317} = .056$$

Step 4: Convert the difference between the experimental and control groups into a t score with df = 298.

$$t = \frac{p_1 - p_2}{\text{s.e.}_d} = \frac{.74 - .68}{.056} = \frac{.06}{.056} = 1.07$$

Because the total number of cases far exceeds 30, we evaluate this t statistic by comparing it to the t values in the last row of Table 3 marked "∞" (indicating infinite degrees of freedom). The probability that the two samples could be drawn from the same population is greater than .10. We reach this conclusion because the value for the t statistic generated in step 4 ($t = 1.07$) does not exceed 1.282, the t score associated with the .10 level of significance in Table 3.

Step 5: Because we are dealing with crime, a manager should be more certain before acting. In this case a probability greater than .10 is not good enough to reject the null hypothesis; the result suggests that there is more than a .10 probability that the null hypothesis is correct. We thus conclude that the experimental program did not yield lower recidivism rates.

Let us look at one more example. Suppose the Morgan City parole board had a second experimental parole program in which parolees did community service work with local charities before they were granted parole. A sample of 100 of these parolees reveals 60 recidivists. Is the second experiment successfully reducing the recidivism rate in comparison to the control group? The calculations follow.

Experimental Group	Control Group
$n = 100$	$n = 200$
$p = .60$	$p = .74$
$s = .49$	$s = .44$
s.e. = .049	s.e. = .031

$$s.e._d = \sqrt{.049^2 + .031^2} = .058$$

$$t = \frac{.74 - .60}{058} = 2.41$$

This t score is statistically significant at less than .01, indicating that the probability of these samples being drawn from the same population is extremely small. What can you say from a research design perspective? From a management perspective? What decisions would you make?

Chapter Summary

Often a manager has samples from two groups (experimental and control, before and after, and so on) and wants to determine whether the two samples could be drawn from the same population and, hence, not differ significantly. This chapter illustrated the process of testing two sample means or two sample proportions to determine whether they could have been drawn from the same population. The procedure basically involves five steps. First, calculate the mean or proportion and standard deviation for each group. Second, calculate the standard error of the mean or proportion estimate for each group. Third, calculate an overall standard error for the groups. Fourth, calculate a *t* score and find its associated probability in the *t* table presented as Table 3 (Statistical Tables) at the back of the book. Finally, make an informed decision based on the analysis.

There are three major difference of means tests. The *t* test assuming independent samples with unequal variances is the most conservative because it consistently produces larger overall standard errors than the other two *t* tests. The *t* test assuming independent samples with equal variances can be used if an analyst is sure that the sample variances in question are equal. The Levene test is a formal test of the equality of sample variances and should be used to evaluate this assumption if an analyst intends to use this difference of means test. The *t* test assuming dependent samples is most appropriate when values for the same cases can be paired; for example, they occur at two different points in time (as in a before-and-after comparison) or in two different sets of scores for the same sample (as in scores on standardized tests of reading and mathematics). A thorough grasp of the difference between dependent and independent samples is needed to understand which of these tests should be used to examine a particular research or management question.

Problems

14.1 John Johnson, the local sheriff, suspects that many of his city's residents are operating motor vehicles without current inspection stickers. To determine whether this is true, John has his boys, John, H. R., and Charles, randomly stop 100 cars. Of these 100 cars, 43 do not have current inspection stickers. John decides to put some fear into drivers and launches a public relations campaign threatening to crack down. A month later, John wants to know whether the program worked. A random sample of 100 cars showed that 21 did not have valid inspection stickers. What can you tell John about the program? (Ask the statistical, research design, and management questions.)

14.2 The manager of the Houston Astros decides to see whether batting practice has any impact on the Astros' hitting. Twenty Astros take batting practice; they are randomly selected. Ten Astros, randomly selected, take no batting practice (the control group). After 25 games, the figures shown in the accompanying table are available. What can you tell the manager about his experiment, statistically and managerially?

	Batting Practice Group	No-Practice Group
Mean	.212	.193
Standard deviation	.026	.047

14.3 The police chief wants to know whether the city's African Americans feel that the police are doing a good job. In comparison to whites' evaluations, this information will tell the police whether they have a community relations problem in the African American community. A survey reveals the information in the accompanying table. What can you tell the police chief?

Opinion	African American	White
Feel police do good job	74	223
Do not feel police do good job	76	73

14.4 General Kleinherbst is concerned with the VD (venereal disease) epidemic among soldiers in Europe. At a nonroutine inspection of 100 troops, 31 were found to have VD. Kleinherbst requires all troops to view the award-winning film *VD: Just between Friends*. At another inspection 180 days later, Kleinherbst finds that 43 of the 200 troops inspected have VD. What can you say about the program statistically, managerially, and from a research design point of view?

14.5 Morgan City Fire Chief Sidney Pyro is concerned about the low efficiency scores that his firefighters receive at the state testing institute. Chief Pyro believes that these scores result because some firefighters are not in good physical condition. Pyro orders 75 randomly selected firefighters to participate in an hour of exercise per day. Another 200 firefighters have no required exercise. After 60 days, all firefighters are tested again by the state; the results are shown in the accompanying table. What can you tell the chief based on this information?

	Exercise Group	No-Exercise Group
Mean	74.5	70.6
Standard deviation	31.4	26.3

14.6 Two hundred people on the welfare rolls in Sunbelt County are randomly selected. One hundred are required to do public service work for the county; the other 100 continue as before. After 6 months, 63 of the public service workers are still on welfare, as are 76 of the control group. What can you say about the effectiveness of this program? What facts may explain these results?

14.7 Ashville City Maintenance Chief Leon Tightwad wants to reduce the costs of maintaining the city automobile fleet. Knowing that city cars are kept for only 1 year, Leon feels that the city's periodic maintenance schedule may cost more than it is worth. Leon randomly selects 75 cars out of 300 and performs no maintenance on these cars unless they break down. At the end of the year, he finds the results shown in the accompanying table. What can you tell Leon about this experiment?

	Maintained Cars	No Maintenance
Mean	$625	$575
Standard deviation	150	200

14.8 Refer to Problem 14.7. Charlie Hustle is in charge of selling Ashville's cars after they have been used 1 year. He believes that Leon's policy costs the city money, and he presents the figures on the cars' sales prices shown in the accompanying table. Does Charlie have an argument? On an overall basis, who will save the city the most money, Leon Tightwad or Charlie Hustle?

	Maintained Cars	No Maintenance
Mean	$14,456	$13,821
Standard deviation	1250	1200

14.9 Both the Brethren Charity and the Lost Souls Mission are operating marriage counseling programs. The Brethren program has a man–woman team to counsel people, whereas Lost Souls uses single counselors. Last year, 12 of 84 randomly selected couples receiving counseling at Brethren ended up divorced. Ten of the fifty-one randomly selected couples at Lost Souls were divorced. As a policy analyst, what can you say about the programs?

14.10 The William G. Harding School of Public Affairs would like to evaluate its affirmative action program for students. After extended discussion, the faculty decides that all students will take the state civil service exam, and the scores on this exam will be used as the criterion of success. Write a memo discussing the results shown in the accompanying table.

	Regular Students	Affirmative Action Students
Mean	86.4	84.1
Standard deviation	17.3	28.2
n	44	19

14.11 A professor thinks that the MPA students at the University of Arizona (UA) are brighter than those at the University of Georgia (UGA). To examine this hypothesis, he gives the same midterm to UA students that he gave to UGA students the previous year. He finds the following results:

	UA	UGA
Mean	83.1	88.7
Standard deviation	11.4	7.8
n	36	24

Present a testable hypothesis and a null hypothesis, and evaluate them. Present a conclusion in plain English.

14.12 The state personnel bureau wants to know whether people resign if they are not promoted during the year. Bureau researchers take a sample of 30 people who were promoted and find that 6 of them resigned; a sample of 45 people who were not promoted includes 15 who resigned. State a hypothesis and a null hypothesis, and test them. State your conclusion in plain English.

14.13 Iowa has decided to run a quasi-experiment in regard to its workfare program and the program's impact on incentives. Officials think that workfare increases the incentives to individuals to earn more money in addition to welfare. Two hundred recipients are selected; 120 are randomly assigned to a workfare program, and 80 are assigned to a control group. By follow-up interviews, the state finds out how much outside income per week is earned by each individual, with the following results:

	Workfare	Control
Mean	\$242.50	\$197.30
s	137	95

Present a hypothesis and a null hypothesis, and evaluate them. State a conclusion in plain English.

14.14 Wisconsin contracts with private organizations to operate its job placement program. The state needs to evaluate the quality of the program offered by one of its vendors, the Beaver Dam Job Placement Center. One hundred unemployed individuals are selected at random. Sixty of these are run through the Beaver Dam program; the others serve as a control group. Sixty percent of the Beaver Dam program group get jobs; the average salary of those jobs is \$29,847 (with a standard deviation of \$1,800). Of the control group, 30% get jobs; the average salary of those jobs is \$27,567 (standard deviation \$3,600). This program can be evaluated by two different criteria. Perform the calculations for both criteria, and present your conclusions.

14.15 Enormous State University has an MPA program. The MPA director is concerned with the small number of MPA students who are being awarded Presidential Management Internships (PMIs). She thinks that this might be because MPA students lack interviewing skills. To experiment with this notion, 10 of the 20 PMI nominees are sent to a special interviewing workshop; the other 10 do not attend the workshop. Seven of the ten attending the workshop receive PMIs, and three of those not attending the workshop receive PMIs. Present a hypothesis and a null hypothesis, and evaluate them. State your conclusion in plain English.

14.16 The Department of Human Services (DHS) has contracted with the Institute for Research on Poverty to run an experimental job training program. A group of 200 individuals are randomly selected from among the hard-core unemployed. A control group of 50 is selected at the same time. The 200 individuals in the experimental group are assigned to a program that attempts to place them in jobs. DHS has defined placement of the individual in a job for 6 months as a success. Of this group, 38 are still employed after 6 months. Of the control group, 11 are employed after 6 months. Present a hypothesis and a null hypothesis, and test them. Present a conclusion in plain English.

14.17 As a National Institutes of Health administrator, you wish to evaluate an experiment at the University of Illinois concerning the impact of exercise on individuals with high-cholesterol diets. The Illinois researchers take 25 pigs that have high-cholesterol diets; 10 of these are randomly selected and made to jog on a treadmill for 2 miles a day. The other 15 pigs do not jog (although they might play golf or get exercise in other ways). After 6 months, each pig is tested for cholesterol in the bloodstream (measured in parts per million) with the following results:

	Exercise Group	Others
Mean	160	210
Standard deviation	40	60

Present a hypothesis and a null hypothesis, and evaluate them. Present a statistical conclusion in plain English.

14.18 The Austin Independent School District wants to know whether the LBJ Magnet School for the Sciences is improving student performance. One hundred students were admitted as sophomores last year to the LBJ school. These students scored a mean of 14.7 on the junior year math achievement test (14 years, 7 months, or about a college sophomore level) with a standard deviation of 1.1. Twenty-three of these students play football. Education researcher Lana "Ein" Stein selects a control group of students who, in their sophomore year, performed comparably to the LBJ students in their sophomore year. These 144 students did not attend a magnet school. Their junior math achievement test produced a mean of 13.6 and a standard deviation of 2.9. Their mean IQ score was 117. Present a hypothesis, test this hypothesis, and present a conclusion in plain English regarding the magnet school students.

14.19 The Wisconsin legislature is considering a mandatory motorcycle helmet law. What legislators don't know is whether the law would encourage more people to use helmets. Senator I. C. Probability tells you that Minnesota has a law similar to the one that Wisconsin is considering. He would like you to compare the use of motorcycle helmets in Minnesota and Wisconsin. A survey is taken in both states, resulting in the statistics presented below. Present a hypothesis and a null hypothesis, and test them. Present your conclusion in plain English.

	Minnesota	Wisconsin
n	75	110
Number using helmets	37	28

14.20 Madonna Lewis's job in the Department of Sanitary Engineering is to determine whether new refuse collection procedures have improved the public's perception of the department. Public opinion surveys were taken both before and after the new procedures were implemented. The results are as follows:

	Before	After
The department is doing a good job	23	47
The department is doing a poor job	79	73

Present a hypothesis and a null hypothesis, and evaluate the hypotheses.

14.21 Edinburg attorney J. L. "Bubba" Pollinard is collecting data for a discrimination suit. He asks 500 Latino people whether they believe that the city is biased against them; 354 say it is. Bubba asks 300 Anglo residents the same question, and 104 state that the city is biased against them. Present a hypothesis and a null hypothesis, test the hypotheses, and present a conclusion in plain English.

14.22 Internal auditors for the city of Austin, Texas, periodically analyze patterns in parking meter collections. Specifically, the auditors focus on whether daily collection totals for the city's two collection teams are dramatically different. The auditors feel that the average daily receipt figures for each team should not be dramatically different; large differences between teams could indicate employee thefts or misreporting of receipts. The auditors ask you, as the chief statistician, to run a difference of means test on 200 randomly selected days of receipts (100 days for each team). Present a hypothesis and a null hypothesis, and conduct a difference of means test. Based on your analysis, what can you tell the auditors? (*Note: The data set for this problem is available on the book's Companion Website*.)

14.23 The director of the Wisconsin Department of Business Licensing is looking for ways to improve employee productivity. Specifically, she would like to see an improvement in the percentage of applications that employees process correctly. The director randomly selects 50 employees and gathers data on the percentage of applications each one correctly processed last month. On the recommendation of a consultant, the director has these 50 employees complete a 3-day workshop in Proactive Synergy Restructuring Techniques (PSRTs). At the end of the month following the PSRT training, the director collects the application processing data for the same 50 employees. Help the director analyze these data. From a statistical standpoint, what can you tell the director? (*Note: The data set for this problem is available on the book's Companion Website*.)

14.24 Dan Stout, a researcher at the Wisconsin Department of Public Health, has begun work on a study examining body mass index (BMI) values for Wisconsin residents. Mr. Stout is particularly interested in BMI variations across the state's two largest cities: Milwaukee and Madison. Mr. Stout believes that citizens in Madison are more physically fit and should thus have lower average BMI scores than citizens in Milwaukee. As a trial run before conducting the larger study, Mr. Stout has obtained two random samples of BMI data for 120 residents in each city (Milwaukee = BMIMIL, Madison = BMIMAD). What can Mr. Stout conclude from these data? (*Note: The data set for this problem is available on the book's Companion Website.*)

14.25 Dr. Sheila Roberts, head of the Department of Public Administration at Eastern Seaboard State University, is concerned about whether student performance in online courses is worse than student performance in traditional courses. Specifically, she believes that the lack of face-to-face student–instructor interaction in online courses may be an impediment to learning. To test this hypothesis, Dr. Roberts randomly selects the final grades from 240 students enrolled in Principles of Public Administration over the past year (120 students from online sections and 120 from traditional sections). What can Dr. Roberts conclude about her hypothesis? (*Note: The data set for this problem is available on the book's Companion Website.*)

14.26 The Department of Service Financing in the city of Belmont, New York, has been experimenting with having city units provide services in-house versus having private contractors provide the same services. In the city Grounds Department, half of the landscaping work is performed by city crews, whereas the other half is performed by a private landscaping firm. The city manager has collected random samples of weekly expense report data for both providers. He asks you to conduct a difference of means test. What can the city manager conclude about the difference between in-house and private service provision? (*Note: The data set for this problem is available on the book's Companion Website.*)

14.27 John P. Smith, Director of the Texas Nonprofits Working Group, is interested in the emerging trend of small nonprofits (defined as those with budgets less than $2 million per year) collaborating with each other for the purpose of sharing administrative services (e.g., accounting, human resources, and information technology services). Mr. Smith hypothesizes that nonprofits engaged in collaborative relationships should be able to spend less on administrative services than nonprofits not engaged in collaborative cost-sharing agreements. To test this hypothesis, Mr. Smith collects data for a random sample of 150 small nonprofits in the state. Specifically, the variable of interest is the percent of each organization's annual budget spent on administrative services. In the sample, 75 of these organizations are engaged in collaborative cost-sharing agreements (SHARE); the other 75 are not (NOSHARE). Upon running a difference of means test, what can Mr. Smith conclude about his hypothesis? (*Note: The data set for this problem is available on the book's Companion Website.*)

14.28 A local foundation in Milwaukee has provided the Technical College of Milwaukee (TCM) with funding for a pilot program aimed at improving the academic performance of students who are single mothers. A number of these students have told school officials that laptop computers would help them immensely, due to unpredictable work and childcare scheduling issues that make it difficult for them to use the computer labs on campus. The foundation would like to see evidence on the program's effectiveness before making a larger financial commitment. Accordingly, administrators at TCM randomly assign the available laptops to 100 students from the group of students who requested laptops (COMSTUD). Another 100 of the students who requested (but did not receive) laptops were selected as a control group (NSTUD). At the end of the semester, the administrators conduct a difference of means test on the semester grade point averages (GPAs) for the two groups (the average GPAs were roughly the same for both groups at the start of the semester). What can they tell the foundation based on these results? (*Note: The data set for this problem is available on the book's Companion Website*.)

14.29 The federal government has asked officials in Milwaukee County to collect data on the Supplemental Nutrition Program for Women, Infants, and Children (WIC) in Milwaukee. Federal officials are concerned about whether children born to non-English-speaking parents are as healthy as those born to native speakers. County officials decide to conduct a pilot study to examine whether there are any differences between these groups. Specifically, officials will examine birth weight data for the two groups (measured in pounds and ounces). Two randomly selected samples of WIC participants are included in the study (each sample includes 100 individuals). The first sample of birth weight data is for babies born to non-English-speaking (NENGLISH) parents. The second sample is birth weight data for babies born to native speakers (ENGLISH). State the null and alternative hypotheses. Test the hypotheses using a difference of means test. (*Note: The data set for this problem is available on the book's Companion Website.*)