CHAPTER 10

Hypothesis Tests Concerning Two Populations

Statistics are like therapists—they will testify for both sides.

Fiorello La Guardia, former mayor of New York City

Numbers don't lie; and they don't forgive.

Harry Angstrom in *Rabbit Is Rich* by John Updike

CONTENTS

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00010-7

The importance of using a control in the testing of a new drug or a new procedure is discussed, and we see how this often results in comparisons between parameters of two different populations. We show how to test that two normal populations have the same population mean, both when the population variances are known and when they are unknown. We show how to test the equality of two population proportions.

10.1 INTRODUCTION

An ongoing debate of great importance turns on whether megadoses—of the order of 25,000 to 30,000 milligrams daily—of vitamin C can be effective in treating patients suffering from cancerous tumors. On one side of the controversy was the great U.S. chemist Linus Pauling, who was a strong advocate of vitamin C therapy, and a growing number of researchers, and on the other side are the majority of mainstream cancer therapists. While many experiments have been set up to test whether vitamin C is therapeutically effective, there has been controversy surrounding many of them. Some of these experiments, which reported negative results, have been attacked by vitamin C proponents as utilizing too small dosages of the vitamin. Others of the experiments, reported by Prof. Pauling and his associates, have been met with skepticism by some in the medical community. To settle all doubts, a definitive experiment was planned and carried out in recent years at the Mayo Clinic. In this famous study, part of a group of terminally ill cancer patients was given, in addition to the regular medication, large doses of vitamin C for three months. The remainder of the group received a placebo along with the regular medication. After the three-month period, the experiment was discontinued. These patients were then monitored until death to determine if the life span of those who had received vitamin C was longer than that of the control group. A summary statement was issued at the end of the experiment. This statement, which was widely disseminated by the news media, reported that there was no significant difference in the life span of those patients who had received the vitamin C treatment. This experiment, regarded by some in the medical community as being of seminal importance in discrediting vitamin C cancer therapy, was attacked by Pauling as being irrelevant to the claims of the proponents of vitamin C. According to the theory developed by Pauling and others, vitamin C would be expected to have protective value only while it was being taken and would not be expected to have any ongoing effect once it was discontinued. Indeed, according to earlier writings of Pauling, an immediate (as was done in the study) rather than a gradual stopping of vitamin C could have some potential negative effects. The controversy continues.

It is important to note that it would not have been sufficient for the Mayo Clinic to have given megadoses of vitamin C to all the volunteer patients. Even if there were significant increases in the additional life spans of these patients in comparison with the known life span distribution of patients suffering from this cancer, it would not be possible to attribute the cause of this increase to vitamin C. For one thing, the placebo effect—in which any type of "extra" treatment gives additional hope to a patient, and this in itself can have beneficial effects—could not be ruled out. For another, the additional life span could be due to factors totally unconnected to the experiment. Thus, to be able to draw a valid conclusion from the experiment, it was necessary to have a second group of volunteer patients, treated in all manners the same as the first, except that they did not receive additional vitamin C but rather only medication that looks and tastes like it. (Of course, to ensure that the two groups are as alike as possible, with the exception of their vitamin C intake, the group of volunteer patients was randomly divided into the two groups—the treated group, whose members received the vitamin C, and the control group, whose members received the placebo.) This resulted in two separate samples, and the resultant data were used to test the hypothesis that the mean additional lifetimes of these two groups are identical.

Indeed, in all situations in which one is trying to study the effect of a given factor, such as the administration of vitamin C, one wants to hold all other factors constant so that any change from the norm can be attributed solely to the factor under study. However, because this is often impossible to achieve outside experiments in the physical sciences, it is usually necessary to consider two samples—one of which is to receive the factor under study and the other of which is a control group that will not receive the factor—and then determine whether there is a statistically significant difference in the responses of these two samples. For this reason, tests concerning two sampled populations are of great importance in a variety of applications.

In this chapter we will show how to test the hypothesis that two population means are equal when a sample from each population is available. In Sec. 10.2 we will suppose that the underlying population distributions are normal, with known variances. Although it is rarely the case that the population variances will be known, the analysis presented in this section will be useful in showing us how to handle the more important cases where this assumption is no longer made. In fact, we show in Sec. 10.3 how to test the hypothesis that the two population means are equal when the variances are unknown, provided that the sample sizes are large. The case where the sample sizes are not large is considered in Sec. 10.4. To be able to test the hypothesis in this case, it turns out to be necessary to assume that the unknown population variances are equal.

In Sec. 10.5 we consider situations in which the two samples are related because of a natural pairing between the elements of the two data sets. For instance, one of the data values in the first sample might refer to an individual's blood pressure before receiving any medication, whereas one of the data values in the second sample might refer to that same person's blood pressure after receiving medication.

In Sec. 10.6 we consider tests concerning the equality of two binomial proportions.

10.2 TESTING EQUALITY OF MEANS OF TWO NORMAL POPULATIONS: CASE OF KNOWN VARIANCES

Suppose that $X_1, \ldots, X_n$ are a sample from a normal population having mean μ_x and variance σ_x^2; and suppose that $Y_1, \ldots, Y_m$ are an independent sample from a normal population having mean μ_y and variance σ_y^2. Assuming that the population variances σ_x^2 and σ_y^2 are known, let us consider a test of the null hypothesis that the two population means are equal; that is, let us consider a test of

$$\mathrm{H}_0\colon \mu_x = \mu_y$$

against the alternative

$$\mathrm{H}_1\colon \mu_x \neq \mu_y$$

Since the estimators of μ_x and μ_y are the respective sample means

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{and} \quad \overline{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

it seems reasonable that H_0 should be rejected when $\overline{X}$ and $\overline{Y}$ are far apart. That is, for an appropriate constant c, it would seem that the test should be to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } |\overline{X} - \overline{Y}| \geq c \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

To specify the appropriate value of c, say, for a significance-level-α test, first we need to determine the probability distribution of $\overline{X} - \overline{Y}$. Now, $\overline{X}$ is normal with mean μ_x and variance σ_x^2/n. And similarly, $\overline{Y}$ is normal with mean μ_y and variance σ_y^2/m. Since the difference of independent normal random variables remains normally distributed, it follows that $\overline{X} - \overline{Y}$ is normal with mean

$$E\left[\overline{X} - \overline{Y}\right] = E\left[\overline{X}\right] - E\left[\overline{Y}\right] = \mu_x - \mu_y$$

and variance

$$\begin{aligned} \mathrm{Var}\left(\overline{X} - \overline{Y}\right) &= \mathrm{Var}\left(\overline{X}\right) + \mathrm{Var}\left(-\overline{Y}\right) \\ &= \mathrm{Var}\left(\overline{X}\right) + (-1)^2\,\mathrm{Var}\left(\overline{Y}\right) \\ &= \mathrm{Var}\left(\overline{X}\right) + \mathrm{Var}\left(\overline{Y}\right) \\ &= \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \end{aligned}$$

Hence, the standardized variable

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

has a standard normal distribution. Therefore, when the null hypothesis H_0: $\mu_x = \mu_y$ is true, the test statistic TS, given by

$$\mathrm{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \tag{10.1}$$

will have a standard normal distribution. Now, a standard normal random variable Z will, in absolute value, exceed $z_{\alpha/2}$ with probability α; that is,

$$P\{|Z| \geq z_{\alpha/2}\} = 2P\{Z \geq z_{\alpha/2}\} = \alpha$$

Thus, since we want to reject H_0 when |TS| is large, it follows that the appropriate significance-level-α test of

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\mathrm{TS}| \geq z_{\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

where the test statistic TS is given by Eq. (10.1).

An alternative way of carrying out this test is first to compute the value of the test statistic TS; say that the data yield the value v. The resulting p value for the test of H_0 versus H_1 is the probability that the absolute value of a standard normal random variable is at least as large as $|v|$. That is, if TS is v, then

$$p \text{ value} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

where Z is a standard normal random variable.

■ Example 10.1

Two new methods for producing a tire have been proposed. The manufacturer believes there will be no appreciable difference in the lifetimes of tires produced by these methods. To test the plausibility of such a hypothesis, a sample of 9 tires is produced by method 1 and a sample of 7 tires by method 2. The first sample of tires is to be road-tested at location A and the second at location B.

Table 10.1 Tire Lives in Units of 1000 Kilometers

Tires tested at A	Tires tested at B	Tires tested at A	Tires tested at B
66.4	58.2	61.4	58.7
61.6	60.4	62.5	56.1
60.5	55.2	64.4	
59.1	62.0	60.7	
63.6	57.3		

It is known, from previous experience, that the lifetime of a tire tested at either of these locations is a normal random variable with a mean life due to the tire but with a variance that is due to the location. Specifically, it is known that the lifetimes of tires tested at location A are normal with a standard deviation equal to 3000 kilometers, whereas those tested at location B have lifetimes that are normal with a standard deviation of 4000 kilometers.

Should the data in Table 10.1 cause the manufacturer to reject the hypothesis that the mean lifetime is the same for both types of tires? Use a 5 percent level of significance.

Solution

Call the tires tested at location A the X sample and those tested at B the Y sample. To test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

we need to compute the value of the test statistic TS. Now, the sample means are given by

$$\overline{X} = 62.2444 \quad \overline{Y} = 58.2714$$

Since $n = 9$, $m = 7$, $\sigma_x = 3$, and $\sigma_y = 4$, we see that the value of the test statistic is

$$\text{TS} = \frac{62.2444 - 58.2714}{\sqrt{9/9 + 16/7}} = 2.192$$

Thus the p value is equal to

$$p \text{ value} = 2P\{Z \geq 2.192\} = 0.0284$$

and so the hypothesis of equal means is rejected at any significance level greater than or equal to 0.0284. In particular, it is rejected at the 5 percent ($\alpha = 0.05$) level of significance. ■

If we were interested in testing the null hypothesis

$$H_0: \mu_x \le \mu_y$$

against the one-sided alternative

$$H_1: \mu_x > \mu_y$$

then the null hypothesis will be rejected only when the test statistic TS is large. In this case, therefore, the significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS} \ge z_\alpha \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

where

$$\text{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

Equivalently, if the observed value of TS is v, then the p value is

$$p \text{ value} = P\{Z \ge v\}.$$

■ Example 10.2

Suppose the purpose of the experiment in Example 10.1 was to attempt to prove the hypothesis that the mean life of the first set of tires exceeded that of the second set by more than 1000 kilometers. Are the data strong enough to establish this at, say, the 5 percent level of significance?

Solution

Let Y_i denote the life of the ith tire of the second set, $i = 1, \ldots, 7$. If we set $W_i = Y_i + 1$, then we are interested in determining whether the data will enable us to conclude that $\mu_x > \mu_w$, where μ_x is the mean life of tires in the first set and μ_w is the mean of W_i. To decide this, we should take this conclusion to be the alternative hypothesis. That is, we should test

$$H_0: \mu_x \le \mu_w \quad \text{against} \quad H_1: \mu_x > \mu_w$$

In other words, a rejection of H_0 would be strong evidence for the validity of the hypothesis that the mean life of the first set of tires exceeds that of the second set by more than 1000 kilometers.

Table 10.2 Tests of Means of Two Normal Populations Having Known Variances when Samples are Independent

The sample mean of a sample of size n from a normal population having mean μ_x and known variance σ_x^2 is $\overline{X}$. The sample mean of a sample of size m from a second normal population having mean μ_y and known variance σ_y^2 is $\overline{Y}$. The two samples are independent.

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}$	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$ Do not reject otherwise	$2P\{Z \geq \lvert v \rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}$	Reject H_0 if TS $\geq z_\alpha$ Do not reject otherwise	$P\{Z \geq v\}$

To test this hypothesis, we compute the value of the test statistic TS, being careful to add 1 to the values given in Table 10.1 for tires tested at location B. This yields

$$\overline{X} = 62.2444 \quad \overline{W} = 59.2714$$

and

$$\text{TS} = \frac{62.2444 - 59.2714}{\sqrt{9/9 + 16/7}} = 1.640$$

Since we want to reject H_0 when TS is large, the p value is the probability that a standard normal will exceed 1.640. That is,

$$p \text{ value} = P\{Z \geq 1.640\} = 0.0505$$

Thus, even though the evidence is strongly in favor of the alternative hypothesis, it is not quite strong enough to cause us to reject the null hypothesis at the 5 percent level of significance. ■

Table 10.2 details both the two-sided test and the one-sided test presented in this section.

PROBLEMS

1. An experiment is performed to test the difference in effectiveness of two methods of cultivating wheat. A total of 12 patches of ground are treated with shallow plowing and 14 with deep plowing. The average yield per ground area of the first group is 45.2 bushels, and the average

yield for the second group is 48.6 bushels. Suppose it is known that shallow plowing results in a ground yield having a standard deviation of 0.8 bushels, while deep plowing results in a standard deviation of 1.0 bushels.

(a) Are the given data consistent, at the 5 percent level of significance, with the hypothesis that the mean yield is the same for both methods?

(b) What is the p value for this hypothesis test?

2. A method for measuring the pH level of a solution yields a measurement value that is normally distributed with a mean equal to the actual pH of the solution and with a standard deviation equal to 0.05. An environmental pollution scientist claims that two different solutions come from the same source. If this is so, then the pH level of the solutions will be equal. To test the plausibility of this claim, 10 independent measurements were made of the pH level for both solutions, with the following data resulting:

Measurements of solution A	Measurements of solution B	Measurements of solution A	Measurements of solution B
6.24	6.27	6.26	6.31
6.31	6.25	6.24	6.28
6.28	6.33	6.29	6.29
6.30	6.27	6.22	6.34
6.25	6.24	6.28	6.27

(a) Do these data disprove the scientist's claim? Use the 5 percent level of significance.

(b) What is the p value?

3. Two machines used for cutting steel are calibrated to cut exactly the same lengths. To test this hypothesis, each machine is used to cut 10 pieces of steel. These pieces are then measured (with negligible measuring error). Suppose the resulting data are as follows:

Machine 1	Machine 2	Machine 1	Machine 2
122.40	122.36	121.76	122.40
123.12	121.88	122.31	122.12
122.51	122.20	123.20	121.78
123.12	122.88	122.48	122.85
122.55	123.43	121.96	123.04

Assume that it is known that the standard deviation of the length of a cut (made by either machine) is equal to 0.50.

(a) Test the hypothesis that the machines are set at the same value, that is, that the mean lengths of their cuttings are equal. Use the 5 percent level of significance.
(b) Find the p value.

4. The following are the values of independent samples from two different populations.

Sample 1: 122, 114, 130, 165, 144, 133, 139, 142, 150

Sample 2: 108, 125, 122, 140, 132, 120, 137, 128, 138

Let μ_1 and μ_2 be the respective means of the two populations. Find the p value of the test of the null hypothesis

$$H_0: \mu_1 \leq \mu_2$$

against the alternative

$$H_1: \mu_1 > \mu_2$$

when the population standard deviations are $\sigma_1 = 10$ and
(a) $\sigma_2 = 5$
(b) $\sigma_2 = 10$
(c) $\sigma_2 = 20$

5. In this section, we presented the test of

$$H_0: \mu_x \leq \mu_y \quad \text{against} \quad H_1: \mu_x > \mu_y$$

Explain why it was not necessary to separately present the test of

$$H_0: \mu_x \geq \mu_y \quad \text{against} \quad H_1: \mu_x < \mu_y$$

6. The device used by astronomers to measure distances results in measurements that have a mean value equal to the actual distance of the object being surveyed and a standard deviation of 0.5 light-years. An astronomer is interested in testing the widely held hypothesis that asteroid A is at least as close to the earth as is asteroid B. To test this hypothesis, the astronomer made 8 independent measurements on asteroid A and 12 on asteroid B. If the average of the measurements for asteroid A was 22.4 light-years and the average of those for asteroid B was 21.3, will the hypothesis be rejected at the 5 percent level of significance? What is the p value?

7. The value received at a certain message-receiving station is equal to the value sent plus a random error that is normal, with mean 0 and standard deviation 2. Two messages, each consisting of a single value, are to be

sent. Because of the random error, each message will be sent 9 times. Before reception, the receiver is fairly certain that the first message value will be less than or equal to the second. Should this hypothesis be rejected if the average of the values relating to message 1 is 5.6 whereas the average of those relating to message 2 is 4.1? Use the 1 percent level of significance.

8. A large industrial firm has its manufacturing operations at one end of a large river. A public health official thinks that the firm is increasing the polychlorinated biphenyl (PCB) level of the river by dumping toxic waste. To gain information, the official took 12 readings of water from the part of the river situated by the firm and 14 readings near the other end of the river. The sample mean of the 12 readings of water near the firm was 32 parts per billion, and the sample mean of the other set of 14 readings was 22 parts per billion. Assume that the value of each reading of water is equal to the actual PCB level at that end of the river where the water is collected plus a random error due to the measuring device that is normal, with mean 0 and standard deviation 8 parts per billion.
   (a) Using the given data and the 5 percent level of significance, can we reject the hypothesis that the PCB level at the firm's end of the river is no greater than the PCB level at the other end?
   (b) What is the p value?

10.3 TESTING EQUALITY OF MEANS: UNKNOWN VARIANCES AND LARGE SAMPLE SIZES

In the previous section we supposed that the population variances were known to the experimenter. However, it is far more common that these parameters are unknown. That is, if the mean of a population is unknown, then it is likely that the variance will also be unknown.

Let us again suppose that we have two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ and are interested in testing a hypothesis concerning their means μ_x and μ_y. Although we do not assume that the population variances σ_x^2 and σ_y^2 are known, we will suppose that the sample sizes n and m are large.

To determine the appropriate test in this situation, we will make use of the fact that for large sample sizes the sample variances will approximately equal the population variances. Thus, it seems reasonable that we can substitute the sample variances S_x^2 and S_y^2 for the population variances and make use of the analysis developed in the previous section. That is, analogous with the result that

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

has a standard normal distribution, it would seem that for large values of n and m, the random variable

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2/n + S_y^2/m}}$$

will have an approximately standard normal distribution. Since this result is indeed true, it follows that we can utilize the same tests developed in Sec. 10.2 except that the sample variances are now utilized in place of the population variances. For instance, the significance-level-α test of

$$\mathrm{H}_0\colon \mu_x = \mu_y$$

against

$$\mathrm{H}_1\colon \mu_x \neq \mu_y$$

is to reject when $|\mathrm{TS}| \geq z_{\alpha/2}$, where the test statistic TS is now given by

$$\mathrm{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$$

An equivalent way of determining the outcome is first to determine the value of the test statistic TS, say it is v, and then to calculate the p value, given by

$$p \text{ value} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

Also, if we want to test the one-sided hypothesis

$$\mathrm{H}_0\colon \mu_x \leq \mu_y$$

against

$$\mathrm{H}_1\colon \mu_x > \mu_y$$

then we use the same test statistic as before. The test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } \mathrm{TS} \geq z_\alpha \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

Equivalently, if the observed value of TS is v, then the p value is

$$p \text{ value} = P\{Z \geq v\}$$

Remarks *We have not yet specified how large n and m should be for the preceding to be valid. A general rule of thumb is for both sample sizes to be at least 30, although values of 20 or more will usually suffice.*

Even when the underlying population distributions are themselves not normal, the central limit theorem implies that the sample means $\overline{X}$ and $\overline{Y}$ will be approximately normal. For this reason the preceding tests of population means can be used for arbitrary underlying distributions provided that the sample sizes are large. (Again, sample sizes of at least 20 should suffice.)

■ Example 10.3

To test the effectiveness of a new cholesterol-lowering medication, 100 volunteers were randomly divided into two groups of size 50 each. Members of the first group were given pills containing the new medication, while members of the second, or *control*, group were given pills containing lovastatin, one of the standard medications for lowering blood cholesterol. All the volunteers were instructed to take a pill every 12 hours for the next 3 months. None of the volunteers knew which group they were in.

Suppose that the result of this experiment was an average reduction of 8.2 with a sample variance of 5.4 in the blood cholesterol levels of those taking the old medication, and an average reduction of 8.8 with a sample variance of 4.5 of those taking the newer medication. Do these results prove, at the 5 percent level, that the new medication is more effective than the old one?

Solution

Let μ_x denote the mean cholesterol reduction of a volunteer who is given the new medication, and let μ_y be the equivalent value for one given the control. If we want to see if the data were sufficient to prove that $\mu_x > \mu_y$, then we should use them to test

$$H_0\colon \mu_x \leq \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

The value of the test statistic is

$$\text{TS} = \frac{8.8 - 8.2}{\sqrt{4.5/50 + 5.4/50}} = 1.3484$$

Since this is a one-sided test where the null hypothesis will be rejected when TS is large, the p value equals the probability that a standard normal (which would be the approximate distribution of TS if $\mu_x = \mu_y$) is as large as 1.3484. That is, the p value of these data is

$$p \text{ value} = P\{Z \geq 1.3484\} = 0.089$$

Since the p value is greater than 0.05, the evidence is not strong enough to establish, at the 5 percent level of significance, that the new medication is more effective than the old. ■

In Example 10.3, note that we compared the new drug to a standard medication rather than to a placebo. Now, when a new drug is tested in situations where there is no accepted treatment, the drug should always be tested against a placebo. However, if there is a viable treatment already in place, then the new drug should be tested against it. This is obvious in very serious diseases, where there may be ethical questions related to prescribing a placebo. Also, in general, one always hopes to conclude that a new drug is better than the previous state-of-the-art drug as opposed to concluding that it is "better than nothing."

■ Example 10.4

A phenomenon quite similar to the placebo effect is often observed in industrial human-factor experiments. It has been noted that a worker's productivity usually increases when that worker becomes aware that she or he is being monitored. Because this phenomenon was documented and widely publicized after some studies on increasing productivity carried out at the Hawthorne plant of the Western Electric company, it is sometimes referred to as the *Hawthorne effect*. To counter this effect, industrial experiments often make use of a control group.

An industrial consultant has suggested a modification of the existing method for producing semiconductors. She claims that this modification will increase the number of semiconductors a worker can produce in a day. To test the effectiveness of her ideas, management has set up a small study. A group of 50 workers have been randomly divided into two groups. One of the groups, consisting of 30 workers, has been trained in the modification proposed by the consultant. The other group, acting as a control, has been trained in a different modification. These two modifications are considered by management to be roughly equal in complexity of learning and in time of implementation. In addition, management is quite certain that the alternative (to the one proposed by the consultant) modification would not have any real effect on productivity. Neither group was told whether it was learning the consultant's proposal or not.

The workers were then monitored for a period of time with the following results.

For those trained in the technique of the consultant:
  The average number of semiconductors produced per worker was 242.
  The sample variance was 62.2.
For those workers in the control group:
  The average number of semiconductors produced per worker was 234.
  The sample variance was 58.4.

Are these data sufficient to prove that the consultant's modification will increase productivity?

Solution

Let μ_x denote the mean number of semiconductors that would be produced over the period of the study by workers trained in the method of the consultant. Also let μ_y denote the mean number produced by workers given the alternative technique. To prove the consultant's claim that $\mu_x > \mu_y$, we need to test

$$H_0\colon \mu_x \leq \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

The data are

$$\begin{aligned} n &= 30 & m &= 20 \\ \overline{X} &= 242 & \overline{Y} &= 234 \\ S_x^2 &= 62.2 & S_y^2 &= 58.4 \end{aligned}$$

Thus the value of the test statistic is

$$\text{TS} = \frac{242 - 234}{\sqrt{62.2/30 + 58.4/20}} = 3.58$$

Hence, the p value of these data is

$$p \text{ value} = P\{Z \geq 3.58\} = 0.0002$$

Thus, the data are significant enough to prove that the consultant's modification was more effective than the one used by the control group. ■

Historical Perspective

The idea of using part of a sample as a control goes back a long way. In the 11th century, the Arabic doctor Avicenna laid down rules for medical experimentation on human subjects. Some of these touched on the use of controls. In 1626 Francis Bacon published an account of the effects of steeping wheat seeds in nine different mixtures, such as water mixed with cow dung, urine, and different types of wine, with unsteeped seed as a control. The greatest yield resulted when seed was steeped in urine.

The first general writing on experiments using controls was done by the British farmer Arthur Young. Young stressed that agricultural experiments must always compare a new treatment with a known one. He published his thoughts in 1771 in the book *A Course of Experimental Agriculture*.

When we are given raw data, rather than summary statistics, the sample means and sample variances can be calculated by a manual computation or by using a calculator or a computer program such as Program 3-1. These quantities should then be used to determine the value of the test statistic TS. Finally, the p value can then be obtained by using the normal probability table (Table D.1 in App. D).

Historical Perspective

The Hawthorne effect illustrates that the presence of an observer may affect the behavior of those being observed. As noted in Example 10.4, the recognition of this phenomenon grew out of research conducted during the 1920s at the Hawthorne plant of Western Electric. Investigators set out to determine how the productivity of workers at this plant could be improved. Their initial studies were designed to examine the effects of changes in lighting on the productivity of workers assembling telephone components. Gradual increases in lighting were made, and each change led to increased productivity. Productivity, in fact, continued to increase even when the lighting was made abnormally bright. Even more surprising was the fact that when the lighting was reduced, productivity still continued to rise.

■ Example 10.5

Test

$$H_0\colon \mu_x \le \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

for the following data:

X: 22, 21, 25, 29, 31, 18, 28, 33, 28, 26, 32,35, 27, 29, 26

Y: 14, 17, 22, 18, 19, 21, 24, 33, 28, 22, 27, 18, 21, 19, 33, 31

Solution

A simple calculation yields that

$$n = 15 \qquad m = 16$$

$$\overline{X} = 27.333 \qquad \overline{Y} = 22.938$$

$$S_x^2 = 21.238 \qquad S_y^2 = 34.329$$

Table 10.3 Tests of Means of Two Normal Populations Having Unknown Variances when Samples are Independent and Sample Sizes are Large

The sample mean and sample variance of a sample of size n from a normal population having mean μ_x and unknown variance σ_x^2 are, respectively, $\overline{X}$ and S_x^2. The sample mean and sample variance of a sample of size m from a second normal population having mean μ_y and unknown variance σ_y^2 are, respectively, $\overline{Y}$ and S_y^2. The two samples are independent, and both n and m are at least 20.

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_x^2/n+S_y^2/m}}$	Reject H_0 if $\lvert TS\rvert \geq z_{\alpha/2}$ Do not reject otherwise	$2P\{Z \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_x^2/n+S_y^2/m}}$	Reject H_0 if $TS \geq z_\alpha$ Do not reject otherwise	$P\{Z \geq v\}$

Hence the value of TS is

$$TS = \frac{4.395}{\sqrt{21.238/15 + 34.329/16}} = 2.33$$

Since this is a one-sided test that will call for rejection only at large values of TS, we have

$$p \text{ value} = P\{Z \geq 2.33\} = 0.01$$

Therefore, the hypothesis that the mean of the X population is no greater than that of the Y population would be rejected at all significance levels greater than or equal to 0.01. ■

Table 10.3 details both the two-sided and the one-sided tests presented in this section.

PROBLEMS

1. A high school is interested in determining whether two of its instructors are equally able to prepare students for a statewide examination in geometry. Seventy students taking geometry this semester were randomly divided into two groups of 35 each. Instructor 1 taught geometry to the first group, and instructor 2 to the second. At the end of

the semester, the students took the statewide examination, with the following results:

Class of instructor 1	Class of instructor 2
$\overline{X} = 72.6$	$\overline{Y} = 74.0$
$S_x^2 = 6.6$	$S_y^2 = 6.2$

Can we conclude from these results that the instructors are not equally able in preparing students for the examinations? Use the 5 percent level of significance. Give the null and alternative hypotheses and the resulting p value.

2. Sample weights (in pounds) of newborn babies born in two adjacent counties in western Pennsylvania yielded the following data:

$$n = 53 \qquad m = 44$$
$$\overline{X} = 6.8 \qquad \overline{Y} = 7.2$$
$$S^2 = 5.2 \qquad S^2 = 4.9$$

Consider a test of the hypothesis that the mean weight of newborns is the same in both counties. What is the resulting p value? How would you express your conclusions to an intelligent person who has not yet studied statistics?

3. An administrator of a large exercise spa is curious as to whether women members younger than 40 years old use the spa with the same frequency as do women members over age 40. Random samples of 30 women younger than 40 years of age and 30 women older than age 40 were chosen and the women tracked for the following month. The result was that the younger group had a sample mean of 3.6 visits with a sample standard deviation of 1.3 visits, while the older group had a sample mean of 3.8 visits with a sample standard deviation of 1.4 visits. Use these data to test the hypothesis that the mean number of visits of the population of older women is the same as that of younger women.

4. You are interested in testing the hypothesis that the mean travel time from your home to work in the morning is the same as the mean travel time from work back to home in the evening. To check this hypothesis, you recorded the times for 40 workdays. It turned out that the sample mean for the trip to work was 38 minutes with a sample standard deviation of 4 minutes, and the sample mean of the return trip home was 42 minutes with a sample standard deviation of 7 minutes.
   (a) What conclusion can you draw at the 5 percent level of significance?
   (b) What is the p value?

5. The following experiment was conducted to compare the yields of two varieties of tomato plants. Thirty-six plants of each variety were randomly selected and planted in a field. The first variety produced an average yield of 12.4 kilograms per plant with a sample standard deviation of 1.6 kilograms. The second variety produced an average yield of 14.2 kilograms per plant with a sample standard deviation of 1.8 kilograms. Does this provide sufficient evidence to conclude that there is a difference in the mean yield for the two varieties? At what level of significance?
6. Data were collected to determine if there is a difference between the mean IQ scores of urban and rural students in upper Michigan. A random sample of 100 urban students yielded a sample mean score of 102.2 and a sample standard deviation of 11.8. A random sample of 60 rural students yielded a sample mean score of 105.3 with a sample standard deviation of 10.6. Are the data significant enough, at the 5 percent level, for us to reject the hypothesis that the mean scores of urban and rural students are the same?
7. In Prob. 6, are the data significant enough, at the 1 percent level, to conclude that the mean score of rural students in upper Michigan is greater than that of urban students? What are the null and the alternative hypotheses?
8. Suppose in Prob. 5 that the experimenter wanted to prove that the average yield of the second variety was greater than that of the first. What conclusion would have been drawn? Use a 5 percent level of significance.
9. A firm must decide between two different suppliers of lightbulbs. Management has decided to order from supplier A unless it can be "proved" that the mean lifetime of lightbulbs from supplier B is superior. A test of 28 lightbulbs from A and 32 lightbulbs from B yielded the following data as to the number of hours of use given by each lightbulb:

   A: 121, 76, 88, 103, 96, 89, 100, 112, 105, 101, 92, 98, 87, 75, 111, 118, 121, 96, 93, 82, 105, 78, 84, 96, 103, 119, 85, 84

   B: 127, 133, 87, 91, 81, 122, 115, 107, 109, 89, 82, 90, 81, 104, 109, 110, 106, 85, 93, 90, 100, 122, 117, 109, 98, 94, 103, 107, 101, 99, 112, 90

   At the 5 percent level of significance, which supplier should be used? Give the hypothesis to be tested and the resulting p value.
10. An administrator of a business school claims that the average salary of its graduates is, after 10 years, at least \$5000 higher than that of comparable graduates of a rival institution. To study this claim, a

random sample of 50 students who had graduated 10 years ago was selected, and the salaries of the graduates were determined. A similar sample of students from the rival institute was also chosen. Suppose the following data resulted:

College	Rival institution
$n = 50$	$m = 50$
$\overline{X} = 85.2$	$\overline{Y} = 74.8$
$S_x^2 = 26.4$	$S_y^2 = 24.5$

(a) To determine whether these data prove the administrator's claim, what should be the null and the alternative hypotheses?
(b) What is the resulting p value?
(c) What conclusions can you draw?

11. An attempt was recently made to verify whether women are being discriminated against, as far as wages are concerned, in a certain industry. To study this claim, a court-appointed researcher obtained a random sample of employees with 8 or more years' experience and with a history of regular employment during that time. With the unit of wages being $1, the following data on hourly pay resulted:

Female workers	Male workers
Sample size: 55	Sample size: 72
Sample mean: 10.80	Sample mean: 12.20
Sample variance: 0.90	Sample variance: 1.1

(a) What hypothesis should be tested? Give the null and the alternative hypotheses.
(b) What is the resulting p value?
(c) What does this prove?

12. The following data summary was obtained from a comparison of the lead content of human hair removed from adult individuals who had died between 1880 and 1920 with the lead content of present-day adults. The data were in units of micrograms, equal to one-millionth of a gram.

	1880–1920	Today
Sample size	30	100
Sample mean	48.5	26.6
Sample standard deviation	14.5	12.3

(a) Do these data establish, at the 1 percent level of significance, that the mean lead content of human hair is less today than it was in the years between 1880 and 1920? Clearly state what the null and alternative hypotheses are.

(b) What is the p value for the hypothesis tested in part (a)?

13. Forty workers were randomly divided into two sets of 20 each. Each set spent 2 weeks in a self-training program that was designed to teach a new production technique. The first set of workers was accompanied by a supervisor whose only job was to check that the workers were all paying attention. The second group was left on its own. After the program ended, the workers were tested. The results were as follows:

	Sample mean	Sample standard deviation
Supervised group	70.6	8.4
Unsupervised group	77.4	7.4

(a) Test the null hypothesis that supervision had no effect on the performance of the workers. Use the 1 percent level of significance.

(b) What is the p value?

(c) What would you conclude was the result of the supervision?

10.4 TESTING EQUALITY OF MEANS: SMALL-SAMPLE TESTS WHEN THE UNKNOWN POPULATION VARIANCES ARE EQUAL

Suppose again that we have independent samples from two normal populations:

$$X_1, \ldots, X_n \quad \text{and} \quad Y_1, \ldots, Y_m$$

and we are interested in testing hypotheses concerning the respective population means μ_x and μ_y. Unlike in the previous sections, we will suppose neither that the population variances are known nor that the sample sizes n and m are necessarily large.

In many situations, even though they are unknown, it is reasonable to suppose that the population variances σ_x^2 and σ_y^2 are approximately equal. So let us assume they are equal and denote their common value by σ^2. That is, suppose that

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

To obtain a test of the null hypothesis

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

when the population variances are equal, we start with the fact, shown in Sec. 10.2, that

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

has a standard normal distribution.

Thus, since $\sigma_x^2 = \sigma_y^2 = \sigma^2$, we see that when H_0 is true (and so $\mu_x - \mu_y = 0$), then $(\overline{X} - \overline{Y}) / \sqrt{\sigma^2/n + \sigma^2/m}$ has a standard normal distribution. That is,

When H_0 is true,

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma^2/n + \sigma^2/m}} \tag{10.2}$$

has a standard normal distribution.

The preceding result cannot be directly employed to test the null hypothesis of equal means since it involves the unknown parameter σ^2. As a result, we will first obtain an estimator of σ^2 and then determine the effect on the distribution of the quantity (10.2) when σ^2 is replaced by its estimator.

To obtain an estimator for σ^2, we make use of the fact that the sample variances S_x^2 and S_y^2 are both estimators of the common population variance σ^2. It is thus natural to combine, or *pool*, these two estimators. In other words, it is natural to consider a weighted average of the two sample variances. To determine the appropriate weights to attach to each one, recall that the sample variance from a sample of size, say, k has $k - 1$ degrees of freedom associated with it. From this we see that S_x^2 has $n - 1$ degrees of freedom associated with it, and S_x^2 has $m - 1$ degrees of freedom. Thus, we will use a pooled estimator that weights S_x^2 by the factor $(n-1)/(n-1+m-1)$ and weights S_y^2 by the factor $(m-1)/(n-1+m-1)$.

Definition *The estimator* S_p^2 *defined by*

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$$

is called the pooled estimator *of* σ^2.

Note that the larger the sample size, the greater the weight given to its sample variance in estimating σ^2. Also note that the pooled estimator will have $n - 1 + m - 1 = n + m - 2$ degrees of freedom attached to it.

If, in expression (10.2), we replace σ^2 by its pooled estimator S_p^2, then the resultant statistic can be shown, when H_0 is true, to have a t distribution with $n+m-2$ degrees of freedom. (This is directly analogous to what happens to the distribution of $\sqrt{n}\,(\overline{X}-\mu)/\sigma$ when the population variance σ^2 is replaced by the sample variance S^2—namely, this replacement changes the standard normal random variable $\sqrt{n}\,(\overline{X}-\mu)/\sigma$ to $\sqrt{n}\,(\overline{X}-\mu)/S$, which is a t random variable with $n-1$ degrees of freedom.)

From the preceding we see that to test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

one should first compute the value of the test statistic

$$\text{TS} = \frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2\,(1/n+1/m)}}$$

The significance-level-α test is then to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\text{TS}| \geq t_{n+m-2,\,\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

Alternatively the test can be run by determining the p value. If TS is observed to equal v, then the resulting p value of the test of H_0 against H_1 is given by

$$\begin{aligned} p \text{ value} &= P\{|T_{n+m-2}| \geq |v|\} \\ &= 2P\{T_{n+m-2} \geq |v|\} \end{aligned}$$

where T_{n+m-2} is a t random variable having $n+m-2$ degrees of freedom.

If we are interested in testing the one-sided hypothesis

$$H_0\colon \mu_x \leq \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

then H_0 will be rejected at large values of TS. Thus the significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS} \geq t_{n+m-2,\alpha} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If the value of the test statistics TS is v, then the p value is given by

$$p \text{ value} = P\{T_{n+m-2} \geq v\}$$

Program 10-1 will compute both the value of the test statistic and the corresponding p value for either a one-sided or a two-sided test.

■ Example 10.6

Twenty-two volunteers at a cold-research institute caught a cold after having been exposed to various cold viruses. A random selection of 10 volunteers were given tablets containing 1 gram of vitamin C. These tablets were taken 4 times a day. The control group, consisting of the other 12 volunteers, was given placebo tablets that looked and tasted exactly like the vitamin C ones. This was continued for each volunteer until a doctor, who did not know whether the volunteer was receiving vitamin C or the placebo, decided that the volunteer was no longer suffering from the cold. The length of time the cold lasted was then recorded.

At the end of this experiment, the following data resulted:

Treated with vitamin C	Treated with placebo	Treated with vitamin C	Treated with placebo
5.5	6.5	7.5	7.5
6.0	6.0	5.5	6.5
7.0	8.5	7.0	7.5
6.0	7.0	6.5	6.0
7.5	6.5		8.5
6.0	8.0		7.0

Do these data prove that taking 4 grams of vitamin C daily reduces the time that a cold lasts? At what level of significance?

Solution

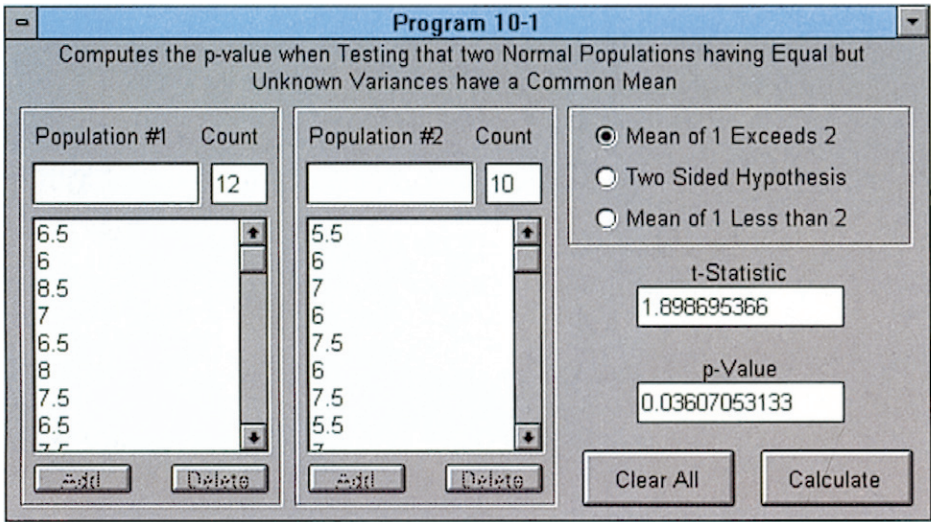
To prove the foregoing hypothesis, we need to reject the null hypothesis in a test of

$$H_0: \mu_p \leq \mu_c \quad \text{against} \quad H_1: \mu_p > \mu_c$$

where μ_c is the mean time a cold lasts when the vitamin C tablets are taken and μ_p is the mean time when the placebo is taken. Assuming that the variance of the length of the cold is the same for the vitamin C patients and the placebo patients, we test the hypothesis by running Program 10-1. This program computes the p value when testing that two normal populations having equal but unknown variances have a common mean.

The sample 1 values are as follows: 6.5, 6, 8.5, 7, 6.5, 8, 7.5, 6.5, 7.5, 6, 8.5, and 7.

The sample 2 values are as follows: 5.5, 6, 7, 6, 7.5, 6, 7.5, 5.5, 7, and 6.5.



Program 10-1 computes the value of the t statistic as 1.898695366.

When we enter the values into Program 10-1, we make sure to note that the alternative hypothesis is not two-sided but rather is that the mean of sample 1 exceeds that of sample 2.

Consequently, the program computes the p value as 0.03607053133.

Thus H_0 would be rejected at the 5 percent level of significance.

Of course, if it was not convenient to run Program 10-1, then we could perform the test by first computing the values of the statistics $\overline{X}$, $\overline{Y}$, S_x^2, S_y^2, and S_p^2, where the X sample corresponds to those receiving a placebo and the Y sample to those receiving vitamin C. These computations give the values

$$\overline{X} = 7.125 \qquad \overline{Y} = 6.450$$

$$S_x^2 = 0.778 \qquad S_y^2 = 0.581$$

Therefore,

$$S_p^2 = \frac{11}{20}S_x^2 + \frac{9}{20}S_y^2 = 0.689$$

and the value of the test statistic is

$$\text{TS} = \frac{0.675}{\sqrt{0.689(1/12 + 1/10)}} = 1.90$$

Table 10.4 Tests of Means of Two Normal Populations Having Unknown Though Equal Variances when Samples are Independent

The sample mean and sample variance, respectively, of a sample size n from a normal population having mean μ_x and variance σ^2 are $\overline{X}$ and S_x^2. And the sample mean and sample variance of a sample of size m from a second normal population having mean μ_y and variance σ^2 are $\overline{Y}$ and S_y^2. The two samples are independent.

$$S_p^2 = \frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$$

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2(1/n+1/m)}}$	Reject H_0 if $\lvert TS\rvert \geq t_{n+m-2,\alpha/2}$ Do not reject otherwise	$2P\{T_{n+m-2} \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2(1/n+1/m)}}$	Reject H_0 if $TS \geq t_{n+m-2,\alpha}$ Do not reject otherwise	$P\{T_{n+m-2} \geq v\}$

Since, from Table D.2, $t_{20,0.05} = 1.725$, the null hypothesis is rejected at the 5 percent level of significance. That is, the evidence is significant, at the 5 percent level, in establishing that vitamin C reduces the mean time that a cold persists. ■

Table 10.4 details both the two-sided test and the one-sided test presented in this section.

PROBLEMS

In the following problems, assume that the population distributions are normal and have equal variances.

1. Twenty-five males between the ages of 25 and 30 who were participating in a well-known heart study carried out in Framingham, Massachusetts were randomly selected. Of these, 11 were smokers and 14 were not. The following data refer to readings of their systolic blood pressure:

Smokers	Nonsmokers	Smokers	Nonsmokers
124	130	131	127
134	122	133	135
136	128	125	120
125	129	118	122
133	118		120
127	122		115
135	116		123

Do the data indicate, at the 1 percent level of significance, a difference in mean systolic blood pressure levels for the populations represented by the two groups? If not, what about at the 5 percent level?

2. A study was instituted to learn how the diets of women changed during the winter and the summer. A random group of 12 women were observed during the month of July, and the percentage of each woman's calories that came from fat was determined. Similar observations were made on a different randomly selected group of size 12 during the month of January. Suppose the results were as follows:

July: 32.2, 27.4, 28.6, 32.4, 40.5, 26.2, 29.4, 25.8, 36.6, 30.3, 28.5, 32.0

January: 30.5, 28.4, 40.2, 37.6, 36.5, 38.8, 34.7, 29.5, 29.7, 37.2, 41.5, 37.0

Test the hypothesis that the mean fat intake is the same for both months. Use the

(a) 5 percent
(b) 1 percent

level of significance.

3. A consumer organization has compared the time it takes a generic pain reliever tablet to dissolve with the time it takes a name-brand tablet. Nine tablets of each were checked. The following data resulted:

Generic: 14.2, 14.7, 13.9, 15.3, 14.8, 13.6, 14.6, 14.9, 14.2

Name: 14.3, 14.9, 14.4, 13.8, 15.0, 15.1, 14.4, 14.7, 14.9

(a) Do the given data establish, at the 5 percent level of significance, that the name-brand tablet is quicker to dissolve?
(b) What about at the 10 percent level of significance?

4. To learn about the feeding habits of bats, a collection of 22 bats were tagged and tracked by radio. Of these 22 bats, 12 were female and 10 were male. The distances flown (in meters) between feedings were noted for each of the 22 bats, and the following summary statistics were obtained:

Female bats	**Male bats**
$n = 12$	$m = 10$
$\overline{X} = 180$	$\overline{Y} = 136$
$S_x = 92$	$S_y = 86$

Test the hypothesis that the mean distance flown between feedings is the same for the populations of male and female bats. Use the 5 percent level of significance.

5. To determine the effectiveness of a new method of teaching reading to young children, a group of 20 nonreading children were randomly divided into two groups of 10 each. The first group was taught by a standard method and the second group by an experimental method. At the end of the school term, a reading examination was given to each of the students, with the following summary statistics resulting:

Students using standard	Students using experimental
Average score = 65.6	Average score = 70.4
Standard deviation = 5.4	Standard deviation = 4.8

Are these data strong enough to prove, at the 5 percent level of significance, that the experimental method results in a higher mean test score?

6. Redo Prob. 2 of Sec. 10.3, assuming that the population variances are equal.
   (a) Would you reject the null hypothesis at the 5 percent level of significance?
   (b) How does the p value compare with the one previously obtained?

7. To learn about how diet affects the chances of getting diverticular disease, 20 vegetarians, 6 of whom had the disease, were studied. The total daily dietary fiber consumed by each of these individuals was determined, with the following results:

With disease	Without disease
$n = 6$	$m = 14$
$\overline{X} = 26.8$ grams	$\overline{Y} = 42.5$ grams
$S_x = 9.2$ grams	$S_y = 9.5$ grams

Test the hypothesis that the mean dietary fiber consumed daily is the same for the population of vegetarians having diverticular disease and the population of vegetarians who do not have this disease. Use the 5 percent level of significance.

8. It is "well known" that the average automobile commuter in the Los Angeles area drives more miles daily than does a commuter in the San Francisco Bay area. To see whether this "fact" is indeed true, a random sample of 20 Los Angeles area commuters and 20 San Francisco Bay area commuters were randomly chosen and their driving habits monitored. The following data relating to the average number and standard deviation of miles driven resulted.

Los Angeles commuter	San Francisco commuter
$\overline{X} = 57.4$	$\overline{X} = 52.8$
$S_x = 12.4$	$S_y = 13.8$

Do these data prove the hypothesis that the mean distance driven by Los Angeles commuters exceeds that of San Francisco commuters? Use the

(a) 10
(b) 5
(c) 1

percent level of significance.

9. The following are the results of independent samples of two different populations.

X: 10.3, 10.4, 11.3, 13.5, 12.7, 11.1, 10.9, 9.7, 14.5, 13.3

Y: 12.4, 11.7, 13.5, 12.9, 13.4, 15.5, 16.3, 13.7, 14.3

Test the null hypothesis that the two population means are equal against the alternative that they are unequal, at the

(a) 10 percent
(b) 5 percent
(c) 1 percent

level of significance.

10. A manager is considering instituting an additional 15-minute coffee break if it can be shown to decrease the number of errors that employees commit. The manager divided a sample of 20 employees into two groups of 10 each. Members of one group followed the same work schedule as before, but the members of the other group were given a 15-minute coffee break in the middle of the day. The following data give the total number of errors committed by each of the 20 workers over the next 20 working days.

Coffee break group: 8, 7, 5, 8, 10, 9, 7, 8, 4, 5

No-break group: 7, 6, 14, 12, 13, 8, 9, 6, 10, 9

Test the hypothesis, at the 5 percent level of significance, that instituting a coffee break does not reduce the mean number of errors. What is your conclusion?

10.5 PAIRED-SAMPLE t TEST

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are samples of the same size from different normal populations having respective means μ_x and μ_y. In certain situations there will be a relationship between the data values X_i and Y_i. Because of this relationship, the pairs of data values $X_i, Y_i, i = 1, \ldots, n$, will not be independent; so we will not be able to use the methods of previous sections to test hypotheses concerning μ_x and μ_y.

■ Example 10.7

Suppose we are interested in learning about the effect of a newly developed gasoline detergent additive on automobile mileage. To gather information, seven cars have been assembled, and their gasoline mileages (in units of miles per gallon) have been determined. For each car this determination is made both when gasoline without the additive is used and when gasoline with the additive is used. The data can be represented as follows:

Car	Mileage without additive	Mileage with additive
1	24.2	23.5
2	30.4	29.6
3	32.7	32.3
4	19.8	17.6
5	25.0	25.3
6	24.9	25.4
7	22.2	20.6

For instance, car 1 got 24.2 miles per gallon by using gasoline without the additive and only 23.5 miles per gallon by using gasoline with the additive, whereas car 4 obtained 19.8 miles per gallon by using gasoline without the additive and 17.6 miles per gallon by using gasoline with the additive.

Now, it is easy to see that two factors will determine a car's mileage per gallon. One factor is whether the gasoline includes the additive, and the second factor is the car itself. For this reason we should not treat the two samples as being independent; rather, we should consider paired data. ■

Suppose we want to test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

where the two samples consist of the paired data $X_i, Y_i, = 1, \ldots, n$. We can test this null hypothesis that the population means are equal by looking at the differences between the data values in a pairing. That is, let

$$D_i = X_i - Y_i \quad i = 1, \ldots, n$$

Now,

$$E\left[D_i\right] = E\left[X_i\right] - E\left[Y_i\right]$$

or, with $\mu_d = E[D_i]$,

$$\mu_d = \mu_x - \mu_y$$

The hypothesis that $\mu_x = \mu_y$ is therefore equivalent to the hypothesis that $\mu_d = 0$. Thus we can test the hypothesis that the population means are equal by testing

$$\mathrm{H}_0\colon \mu_d = 0 \quad \text{against} \quad \mathrm{H}_1\colon \mu_d \neq 0$$

Assuming that the random variables $D_1, \ldots, D_n$ constitute a sample from a normal population, we can test this null hypothesis by using the t test described in Sec. 9.4. That is, if we let $\overline{D}$ and S_d denote, respectively, the sample mean and sample standard deviation of the data $D_1, \ldots, D_n$, then the test statistic TS is given by

$$\mathrm{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

The significance-level-α test will be to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } |\mathrm{TS}| \geq t_{n-1,\alpha/2} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

where the value of $t_{n-1,\alpha/2}$ can be obtained from Table D.2.

Equivalently, the test can be performed by computing the value of the test statistic TS, say it is equal to v, and then computing the resulting p value, given by

$$p \text{ value} = P\{|T_{n-1}| \geq |v|\} = 2P\{T_{n-1} \geq |v|\}$$

where T_{n-1} is a t random variable with $n-1$ degrees of freedom. If a personal computer is available, then Program 9-1 can be used to determine the value of the test statistic and the resulting p value. The successive data values entered in this program should be $D_1, D_2, \ldots, D_n$ and the value of μ_0 (the null hypothesis value for the mean of D) entered should be 0.

■ Example 10.8

Using the data of Example 10.7, test, at the 5 percent level of significance, the null hypothesis that the additive does not change the mean number of miles obtained per gallon of gasoline.

Solution

If it is not convenient to run Program 9-1, we can use the data to compute first the differences D_i and then the summary statistics $\overline{D}$ and S_d. Using the data differences

$$0.7, 0.8, 0.4, 2.2, -0.3, -0.5, 1.6$$

results in the values

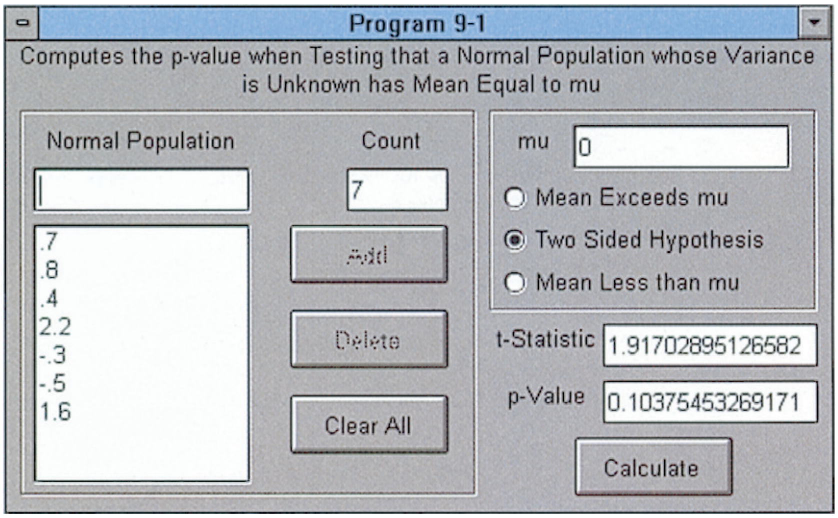
$$\overline{D} = 0.7 \quad S_d = 0.966$$

Therefore, the value of the test statistic is

$$\text{TS} = \frac{\sqrt{7}\,(0.7)}{0.966} = 1.917$$

Since, from Table D.2, $t_{6,0.025} = 2.447$, the null hypothesis that the mean mileage is the same whether or not the gasoline used contains the additive is not rejected at the 5 percent level of significance.

If a personal computer is available, then we can solve the problem by running Program 9-1. This yields the following:



Thus the null hypothesis will not even be rejected at the 10 percent level of significance. ■

One-sided tests concerning the two population means are similarly obtained. For instance, to test

$$H_0\colon \mu_x \le \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

we use the data $D_1, \ldots, D_n$ and test

$$H_0: \mu_d \leq 0 \quad \text{against} \quad H_1: \mu_d > 0$$

Again with the test statistic

$$\text{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

the significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS} > t_{n-1,\alpha} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

Equivalently, if the value of TS is v, then the p value is

$$p \text{ value} = P\{T_{n-1} \geq v\}$$

Program 9-1 can be used again to determine the value of the test statistic and the resulting p value. (If summary statistics $\overline{D}$ and S_d are given, then the p value can be obtained by calculating v, the value of the test statistic, and then running Program 8-1 to determine $P\{T_{n-1} \geq v\}$.)

■ Example 10.9

The management of a chain of stores wanted to determine whether advertising tended to increase its sales of women's shoes. To do so, management determined the number of shoe sales at six stores during a two-week period. While there were no advertisements in the first week, advertising was begun at the beginning of the second week. Assuming that any change in sales is due solely to the advertising, do the resulting data prove that advertising increases the mean number of sales? Use the 1 percent level of significance.

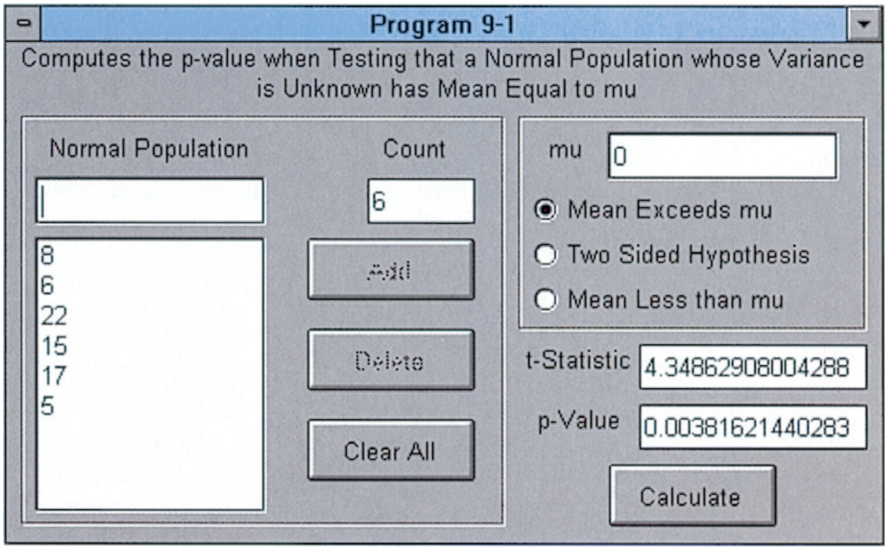
Store	First-week sales	Second-week sales
1	46	54
2	54	60
3	74	96
4	60	75
5	63	80
6	45	50

Solution

Letting D_i denote the increase in sales at store i, we need to check if the data are significant enough to establish that $\mu_d > 0$. Hence, we should test

$$H_0: \mu_d \leq 0 \quad \text{against} \quad H_1: \mu_d > 0$$

Using the data values 8, 6, 22, 15, 17, 5, we run Program 9-1 to obtain the following:



Thus the hypothesis that advertising does not result in increased sales is rejected at any significance level greater than or equal to 0.0038. Therefore, it is rejected at the 1 percent level of significance. ■

PROBLEMS

1. The following data relate to the heart rates (in beats per minute) of 12 individuals both before and after using chewing tobacco. The subjects were regular users of this substance.

Subject	Heart rate before use	Heart rate after use
1	73	77
2	67	69
3	68	73
4	60	70
5	76	74
6	80	88
7	73	76
8	77	82
9	66	69
10	58	61
11	82	84
12	78	80

(a) Test the hypothesis, at the 5 percent level of significance, that chewing smokeless tobacco does not result in a change in the mean heart rate of the population of regular users of chewing tobacco.

(b) What is the resulting p value?

2. A shoe salesman claims that using his company's running shoes will result, on average, in faster times. To check this claim, a track coach assembled a team of 10 sprinters. The coach randomly divided the runners into two groups of size 5. The members of the first group then ran 100 yards, using their usual running shoes; the members of the second group ran 100 yards, using the company's shoes. After time was given for rest, the group who ran with their usual shoes changed into the company's shoes and members of the other group changed to their usual shoes. Then they all ran another dash of 100 yards. The following data resulted:

	Racer									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Time (old shoes)	10.5	10.3	11.0	10.9	11.3	9.9	10.1	10.7	12.2	11.1
Time (new shoes)	10.3	10.0	10.6	11.1	11.0	9.8	10.2	10.5	11.8	10.5

Do these data prove the claim of the salesman that the company's new shoes result, on average, in lower times? Use the 10 percent level of significance. What about at the 5 percent level?

3. Use the t test on the following paired data to test

$$H_0: \mu_x = \mu_y \quad \text{against} \quad H_1: \mu_x \neq \mu_y$$

at the 5 percent level of significance.

	i										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
X_i	122	132	141	127	141	119	124	131	145	140	135
Y_i	134	126	133	122	155	116	118	137	140	133	142

4. A question of medical interest is whether jogging leads to a reduction in systolic blood pressure. To learn about this question, eight nonjogging volunteers have agreed to begin a 1-month jogging program. At the end of the month their blood pressures were determined and compared with earlier values, with the following data resulting:

	Subject							
	1	2	3	4	5	6	7	8
Blood pressure before	134	122	118	130	144	125	127	133
Blood pressure after	130	120	123	127	138	121	132	135

(a) Suppose you want to see if these data are significant enough to prove that jogging for 1 month will tend to reduce the systolic blood pressure. Give the null and alternative hypotheses.
(b) Do the data prove the hypothesis in (a) at the 5 percent level of significance?
(c) Do the data prove that the hypothesis is false?
(d) How would you present the results of this experiment to a medical person who is not trained in statistics?

5. The following table gives the scores on a test of intelligence for 14 pairs of monozygotic (commonly called *identical*) twins who were separated at birth. One member of each pair was raised by a biological parent, while the other was raised in a home that did not contain either of their biological parents. The IQ test used is known in the psychological literature as the "dominoes" IQ test.

Twin raised by mother or father	Twin raised by neither parent	Twin raised by mother or father	Twin raised by neither parent
23	18	22	15
30	25	31	23
25	28	29	27
18	22	24	26
19	14	28	19
25	34	31	30
28	36	27	28

(a) Test the hypothesis that the mean IQ test score of a twin is not affected by whether he or she is raised by a biological parent. Use the 5 percent level of significance.
(b) What conclusions, if any, can be drawn from your hypothesis test?

6. Consider Prob. 2 of Sec. 10.4. Suppose that the same women were used for both months and that the data in each of the columns referred to the same woman's fat intake during the summer and winter.

(a) Test the hypothesis that there is no difference in fat intake during summer and winter. Use the 5 percent level of significance.
(b) Repeat (a), this time using the 1 percent level.

7. The following are scores on two IQ tests of 12 university students. One of the tests was taken before the student had a course in statistics, and the other was taken after.

Student	IQ score before course	IQ score after course
1	104	111
2	125	120
3	127	138
4	102	113
5	140	142
6	122	130
7	118	114
8	110	121
9	126	135
10	138	145
11	116	118
12	125	125

Use these data to test the hypothesis that a student's score on an IQ test will not tend to be any different after the student takes a statistics course. Use the 5 percent level of significance.

8. To see whether there are any differences in starting salaries for women and men law school graduates, a set of eight law firms was selected. For each of these firms a recently hired woman and a recently hired man were randomly chosen. The following starting salary information resulted from interviewing those chosen.

	Company							
	1	2	3	4	5	6	7	8
Woman's salary	52	53.2	78	75	62.5	72	39	49
Man's salary	54	55.5	78	81	64.5	70	42	51

Use the given data to test the hypothesis, at the 10 percent level of significance, that the starting salary is the same for both sexes.

9. To study the effectiveness of a certain commercial liquid protein diet, the Food and Drug Administration sampled nine individuals who were entering the program. Their weights both immediately before they

entered and six months after they completed the two-week program were recorded. The following data resulted:

Person	Weight before	Weight after
1	197	185
2	212	220
3	188	180
4	226	217
5	170	185
6	194	197
7	233	219
8	166	170
9	205	202

Suppose we want to determine if these data prove that the diet is effective, in the sense that the expected weight loss after six months is positive.

(a) What is the null hypothesis to be tested, and what is the alternative?

(b) Do the data prove that the diet works? Use the 5 percent level.

10. The following are the motor vehicle death rates per 100 million vehicle miles for a random selection of states in 1985, 1989, and 2001.

State	1985 Rate	1989 Rate	2001 Rate
Arkansas	3.4	3.3	2.1
Colorado	2.4	1.9	1.7
Indiana	2.6	1.9	1.3
Kentucky	2.6	2.4	1.8
Massachusetts	1.9	1.7	0.9
Ohio	2.1	2.1	1.3
Tennessee	3.4	2.3	1.8
Wyoming	2.7	2.3	2.3

Source: *Accident Facts,* National Safety Council, Chicago.

(a) Do the data establish, at the 5 percent level of significance, that the motor vehicle death rate was lower in 1989 than in 1985?

(b) Do the data establish, at the 5 percent level of significance, that the death rate was lower in 2001?

(c) What is the p value for the tests in parts (a) and (b)?

11. The following data give the marriage rates per 1000 population in a random sample of countries.

Crude Marriage Rates for Selected Countries
(per 1000 population)

Country	1999	1998	1997	1990
Australia	6.0	–	5.8	6.9
Austria	4.8	4.8	5.1	5.8
Belgium	4.3	4.4	4.7	6.6
Bulgaria	4.2	4.3	4.1	6.7
Czech Republic	5.2	5.4	5.6	8.4
Denmark	6.6	6.5	6.4	6.1
Finland	4.7	4.5	4.6	4.8
Germany	5.2	5.1	5.2	6.5
Greece	6.4	5.5	5.7	5.8
Hungary	4.5	4.5	4.6	6.4
Ireland	4.9	–	4.3	5.0
Israel	5.9	–	5.6	7.0
Japan	6.3	6.3	6.2	5.8
Luxembourg	4.9	–	4.8	6.2
Netherlands	5.6	–	5.5	6.4
New Zealand	5.3	–	5.3	7.0
Norway	5.3	–	–	5.2
Poland	5.7	5.4	5.3	6.7
Portugal	6.8	6.7	6.5	7.3
Romania	6.5	6.4	6.5	8.3
Russia	5.8	5.8	6.3	8.9
Sweden	4.0	3.5	3.7	4.7
Switzerland	4.9	–	5.3	6.9

Test the hypothesis that the worldwide marriage rates in 1999 are greater than those in 1990.

10.6 TESTING EQUALITY OF POPULATION PROPORTIONS

Consider two large populations, and let p_1 and p_2 denote, respectively, the proportions of the members of these two populations that have a certain characteristic of interest. Suppose that we are interested in testing the hypothesis that these proportions are equal against the alternative that they are unequal. That is, we are interested in testing

$$H_0: p_1 = p_2 \quad \text{against} \quad H_1: p_1 \neq p_2$$

To test this null hypothesis, suppose that independent random samples, of respective sizes n_1 and n_2, are drawn from the populations. Let X_1 and X_2 represent the number of elements in the two samples that have the characteristic.

Let $\hat{p}_1$ and $\hat{p}_2$ denote, respectively, the proportions of the members of the two samples that have the characteristic. That is, $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$. Since $\hat{p}_1$ and $\hat{p}_2$ are the respective estimators of p_1 and p_2, it is evident that we want to reject H_0 when $\hat{p}_1$ and $\hat{p}_2$ are far apart, that is, when $|\hat{p}_1 - \hat{p}_2|$ is sufficiently large. To see how far apart they need be to justify rejection of H_0, first we need to determine the probability distribution of $\hat{p}_1 - \hat{p}_2$.

Recall from Sec. 7.5 that the mean and variance of the proportion of the first sample that has the characteristic is given by

$$E[\hat{p}_1] = p_1 \qquad \mathrm{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}$$

and, similarly, for the second sample,

$$E[\hat{p}_2] = p_2 \qquad \mathrm{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

Thus we see that

$$\begin{aligned} E[\hat{p}_1 - \hat{p}_2] &= E[\hat{p}_1] - E[\hat{p}_2] \\ &= p_1 - p_2 \\ \mathrm{Var}(\hat{p}_1 - \hat{p}_2) &= \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

In addition, if we suppose that n_1 and n_2 are reasonably large, then $\hat{p}_1$ and $\hat{p}_2$ will have an approximately normal distribution, and thus so will their difference $\hat{p}_1 - \hat{p}_2$. As a result, the standardized variable

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

will have a distribution that is approximately that of a standard normal random variable.

Now suppose that H_0 is true, and so the proportions are equal. Let p denote their common value; that is, $p_1 = p_2 = p$. In this case $p_1 = p_2 = 0$, and so the quantity

$$W = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)/n_1 + p(1-p)/n_2}} \tag{10.3}$$

will have an approximately standard normal distribution. We cannot, however, base our test directly on W, for it depends on the unknown quantity p. However, we can estimate p by noting that of the combined sample of size $n_1 + n_2$ there are a total of $X_1 + X_2 = n_1\hat{p}_1 + n_2\hat{p}_2$ elements that have the characteristic of interest. Therefore, when H_0 is true and each population has the same proportion of its members with the characteristic, the natural estimator of that common proportion p is as follows.

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

The estimator $\hat{p}$ is called the *pooled* estimator of p.

We will now substitute the estimator $\hat{p}$ for the unknown parameter p in Eq. (10.3) for W and base our test on the resulting expression. That is, we will use the test statistic

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

It can be shown that for reasonably large values of n_1 and n_2 (both being at least 30 should suffice), TS will, when H_0 is true, have a distribution that is approximately equal to the standard normal distribution. Thus, the significance-level-α test of

$$H_0\colon p_1 = p_2 \quad \text{against} \quad H_1\colon p_1 \neq p_2$$

is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\text{TS}| \geq z_{\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

The test can also be performed by determining the value of the test statistic, say it is equal to v, and then determining the p value given by

$$p \text{ value} = P\{|Z| \geq |v|\} = 2P\{Z \geq |v|\}$$

where Z is (as always) a standard normal random variable.

■ Example 10.10

In criminal proceedings a convicted defendant is sometimes sent to prison by the presiding judge and is sometimes not. A question has arisen in legal circles as to whether a judge's decision is affected by (1) whether the defendant pleaded guilty or (2) whether he or she pleaded innocent but was subsequently found guilty. The following data refer to individuals, all having previous prison records, convicted of second-degree robbery.

74, out of 142 who pleaded guilty, went to prison

61, out of 72 who pleaded not guilty, went to prison

Do these data indicate that a convicted individual's chance of being sent to prison depends on whether she or he had pleaded guilty?

Solution

Let p_1 denote the probability that a convicted individual who pleaded guilty will be sent to prison, and let p_2 denote the corresponding probability for one who pleaded innocent but was adjudged guilty. To see if the data are significant enough to prove that $p_1 \neq p_2$, we need to test

$$\mathrm{H}_0\colon p_1 = p_2 \quad \text{against} \quad \mathrm{H}_1\colon p_1 \neq p_2$$

The data yield

$$n_1 = 142 \qquad \hat{p}_1 = \frac{74}{142} = 0.5211$$

$$n_2 = 72 \qquad \hat{p}_2 = \frac{61}{72} = 0.8472$$

The value of the pooled estimator $\hat{p}$ is

$$\hat{p} = \frac{74 + 61}{142 + 72} = 0.6308$$

and the value of the test statistic is

$$\mathrm{TS} = \frac{0.5211 - 0.8472}{\sqrt{(1/142 + 1/72)\,(0.6308)\,(1 - 0.6308)}} = -4.67$$

The p value is given by

$$p \text{ value} = 2P\{Z \geq 4.67\} \approx 0$$

For such a small p value the null hypothesis will be rejected. That is, we can conclude that the decision of a judge, with regard to whether a convicted defendant should be sent to prison, is indeed affected by whether that defendant pleaded guilty or innocent. (We cannot, however, conclude that pleading guilty is a good strategy for a defendant as far as avoiding prison is concerned. The reason we cannot is that a defendant who pleads innocent has a chance of being acquitted.) ■

Our next example illustrates the difficulties that abound in modeling real phenomena.

■ Example 10.11

Predicting a child's gender Suppose we are interested in determining a model for predicting the gender of future children in families. The simplest model would be to suppose that each new birth, no matter what the present makeup of the family, will be a boy with some probability p_0. (Interestingly enough, existing data indicate that p_0 would be closer to 0.51 than to 0.50.)

Somewhat surprisingly, this simple model does not hold up when real data are considered. For instance, data on the gender of members of French families were given by Malinvaud in 1955. Consider families having four or more children. Malinvaud reported 36,694 such families whose first three children were girls (that is, the three eldest children were girls), and he reported 42,212 such families whose first three children were all boys. Malinvaud's data indicated that in those families whose first three children were all girls, their next child was a boy 49.6 percent of the time, whereas in the families whose first three children were all boys, the next one was a boy 52.3 percent of the time.

Let p_1 denote the probability that the next child of a family presently composed of three girls is a boy, and let p_2 denote the corresponding probability for a family presently composed of three boys. If we use the given data to test

$$H_0\colon p_1 = p_2 \quad \text{against} \quad H_1\colon p_1 \neq p_2$$

then we have

$$n_1 = 36{,}694 \qquad n_2 = 42,212$$
$$\hat{p}_1 = 0.496 \qquad \hat{p}_2 = 0.523$$

and so

$$\hat{p} = \frac{36{,}694\,(0.496) + 42,212\,(0.523)}{36{,}694 + 42{,}212} = 0.51044$$

Therefore, the value of the test statistic is

$$\mathrm{TS} = \frac{0.496 - 0.523}{\sqrt{(1/36{,}694 + 1/42{,}212)\,(0.5104)\,(1 - 0.5104)}} = -7.567$$

Since $|\mathrm{TS}| \geq z_{0.005} = 2.58$, the null hypothesis that the probability that the next child is a boy is the same regardless of whether the present family is made up of three girls or of three boys is rejected at the 1 percent level of significance. Indeed, the p value of these data is

$$p \text{ value} = P\{|Z| \geq 7.567\} = 2P\{Z \geq 7.567\} \approx 0$$

This shows that any model that assumes that the probability of the gender of an unborn does not depend on the present makeup of the family is not consistent with existing data. (One model that is consistent with the given data is to suppose that each family has its own probability that a newborn will be a boy, with this probability remaining the same no matter what the present makeup of the family. This probability, however, differs from family to family.) ■

The ideal way to test the hypothesis that the results of two different treatments are identical is to randomly divide a group of people into a set that will receive the first treatment and one that will receive the second. However, such randomization is not always possible. For instance, if we want to study whether drinking alcohol increases the risk of prostate cancer, we cannot instruct a randomly chosen sample to drink alcohol. An alternative way to study the hypothesis is to use an *observational* study that begins by randomly choosing a set of drinkers and one of nondrinkers. These sets are followed for a period of time and the resulting data are then used to test the hypothesis that members of the two groups have the same risk for prostate cancer.

Our next example illustrates another way of performing an observational study.

■ Example 10.12

In 1970, the researchers Herbst, Ulfelder, and Poskanzer (H-U-P) suspected that vaginal cancer in young women, a rather rare disease, might be caused by one's mother having taken the drug diethylstilbestrol (usually referred to as DES) while pregnant. To study this possibility, the researchers could have performed an observational study by searching for a (treatment) group of women whose mothers took DES when pregnant and a (control) group of women whose mothers did not. They could then observe these groups for a period of time and use the resulting data to test the hypothesis that the probabilities of contracting vaginal cancer are the same for both groups. However, because vaginal cancer is so rare (in both groups), such a study would require a large number of individuals in both groups and would probably have to continue for many years to obtain significant results. Consequently, H-U-P decided on a different type of observational study. They uncovered 8 women between the ages of 15 and 22 who had vaginal cancer. Each of these women (called cases) was then matched with 4 others, called *referents* or *controls*. Each of the referents of a case was free of the cancer and was born within 5 days in the same hospital and in the same type of room (either private or public) as the case. Arguing that if DES had no effect on vaginal cancer then the probability, call it p_c, that the mother of a case took DES would be the same as the probability, call it p_r, that the mother of a referent took DES, the researchers H-U-P decided to test

$$H_0\colon p_c = p_r \quad \text{against} \quad H_1\colon p_c \neq p_r$$

Discovering that 7 of the 8 cases had mothers who took DES while pregnant whereas none of the 32 referents had mothers who took the drug, the researchers concluded that there was a strong association between

DES and vaginal cancer (see Herbst, A., Ulfelder, H., and Poskanzer, D., "Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women," *New England Journal of Medicine*, **284**, 878–881, 1971). (The p value for these data is approximately 0.) ■

If we are interested in verifying the one-sided hypothesis that p_1 is larger than p_2, then we should take that to be the alternative hypothesis and so test

$$\mathrm{H}_0\colon p_1 \le p_2 \quad \text{against} \quad \mathrm{H}_1\colon p_1 > p_2$$

The same test statistic TS as used before is still employed, but now we reject H_0 only when TS is large (since this occurs when $\hat{p}_1 - \hat{p}_2$ is large). Thus, the one-sided significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if TS} \ge z_\alpha \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

Alternatively, if the value of the test statistic TS is v, then the resulting p value is

$$p \text{ value} = P\{Z \ge v\}$$

where Z is a standard normal.

Remark *The test of*

$$\mathrm{H}_0\colon p_1 \le p_2 \quad \textit{against} \quad \mathrm{H}_1\colon p_1 > p_2$$

is the same as

$$\mathrm{H}_0\colon p_1 = p_2 \quad \textit{against} \quad \mathrm{H}_1\colon p_1 > p_2$$

This is so because in both cases we want to reject H_0 *when* $\hat{p}_1 - \hat{p}_2$ *is so large that such a large value would have been highly unlikely if* p_1 *were not greater than* p_2.

■ Example 10.13

A manufacturer has devised a new method for producing computer chips. He feels that this new method will reduce the proportion of chips that turn out to have defects. To verify this, 320 chips were produced by the new method and 360 by the old. The result was that 76 of the former and 94 of the latter were defective. Is this significant enough evidence for the manufacturer to conclude that the new method will produce a smaller proportion of defective chips? Use the 5 percent level of significance.

Solution

Let p_1 denote the probability that a chip produced by the old method will be defective, and let p_2 denote the corresponding probability for a chip produced by the new method. To conclude that $p_1 > p_2$, we need to reject H_0 when testing

$$H_0: p_1 \le p_2 \quad \text{against} \quad H_1: p_1 > p_2$$

The data are

$$n_1 = 360 \qquad n_2 = 320$$
$$\hat{p}_1 = \frac{94}{360} = 0.2611 \qquad \hat{p}_2 = \frac{76}{320} = 0.2375$$

The value of the pooled estimator is thus

$$\hat{p} = \frac{94 + 76}{360 + 320} = 0.25$$

Hence, the value of the test statistic is

$$\text{TS} = \frac{0.2611 - 0.2375}{\sqrt{(1/360 + 1/320)\,(0.25)\,(0.75)}} = 0.7094$$

Since $z_{0.05} = 1.645$, we cannot reject the null hypothesis at the 5 percent level of significance. That is, the evidence is not significant enough for us to conclude that the new method will produce a smaller percentage of defective chips than the old method.

The p value for the data is

$$p \text{ value} = P\{Z \ge 0.7094\} = 0.239$$

indicating that a value of TS at least as large as the one observed will occur 24 percent of the time when the two probabilities are equal. ■

Table 10.5 details the tests considered in this section.

Table 10.5 Tests Concerning Two Binomial Probabilities

The proportions of members of two populations that have a certain characteristic are p_1 and p_2. A random sample of size n_1 is chosen from the first population, and an independent random sample of size n_2 is chosen from the second population. The numbers of members of the two samples with the characteristic are X_1 and X_2, respectively.

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

$$\hat{p} = \frac{X_1+X_2}{n_1+n_2}$$

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$p_1 = p_2$	$p_1 \ne p_2$	$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if $\lvert\text{TS}\rvert \ge z_{\alpha/2}$ Do not reject otherwise	$2P\{Z \ge \lvert v\rvert\}$
$p_1 \le p_2$	$p_1 > p_2$	$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if TS $\ge z_\alpha$ Do not reject otherwise	$P\{Z \ge v\}$

Statistics In Perspective

Do Not Misinterpret a Rejection

We must be careful when deciding what a rejection of the null hypothesis really means, for often interpretations are given that are not warranted by the available data. For instance, suppose a hypothesis test was performed to study whether the probabilities that a patient does not survive an operation are the same at hospitals A and B. Suppose that a random sample of the operations performed at hospital A yielded that 72 out of 480 patients operated on did not survive, whereas a sample at hospital B yielded that 30 of 360 did not survive. While we can certainly conclude from these data that the survival probabilities are unequal, we cannot conclude that hospital A is not doing as good a job as hospital B, for without additional data we cannot rule out such possibilities as that hospital A is performing more high-risk operations than is B and that is the reason it has a lower survival rate.

For another example that indicates how careful we must be when interpreting the meaning of a rejected hypothesis, consider a hypothetical study of the salaries of male and female salespeople at a large corporation. Suppose that a random sample of 50 male and 50 female employees indicated that the average salary of the men was $40,000 per year whereas that of the women was $36,000. Assuming that the sample variances were small, a test of the hypothesis that the mean salary was the same for both populations would be rejected. But what could we conclude from this? For instance, would we be justified in concluding that the women are being discriminated against? The answer is that we cannot come to such a conclusion with the information presented, for there are many possible explanations for the apparent differences in mean salary.

One possibility might be that the mix of experienced and inexperienced workers is different for the two sexes. For instance, taking into account whether an employee had worked for more or less than 5 years might have produced the following data.

Year of employment	Number	Average salary ($)
Men:		
Less than 5	10	34,000
More than 5	40	41,500
Total	50	40,500
Women:		
Less than 5	40	34,500
More than 5	10	42,000
Total	50	36,500

For instance, a total of 10 of the 50 women have been employed more than 5 years, and their average salary is $42,000 per year. Thus, we see that even though the average salary of the men is higher than that of the women, when time of employment is taken into account, the female employees are actually receiving higher salaries than their male counterparts.

PROBLEMS

1. Two methods have been proposed for producing transistors. If method 1 resulted in 20 unacceptable transistors out of a total of 100 produced and method 2 resulted in 12 unacceptable transistors out of a total of 100 produced, can we conclude that the proportions of unacceptable transistors that will be produced by the two methods are different?
   (a) Use the 5 percent level of significance.
   (b) What about at the 10 percent level of significance?
2. A random sample of 220 female and 210 male coffee drinkers were questioned. The result was that 71 of the women and 58 of the men indicated a preference for decaffeinated coffee. Do these data establish, at the 5 percent level of significance, that the proportion of female coffee drinkers who prefer decaffeinated coffee differs from the corresponding proportion for men? What is the p value?
3. An automobile insurance company selected random samples of 300 single male policyholders and 300 married male policyholders, all between the ages of 25 and 30. It recorded the number who had reported accidents at some time within the past 3 years. The resulting data were that 19 percent of the single policyholders and 12 percent of the married ones had reported an accident.
   (a) Does this establish, at the 10 percent level of significance, that there is a difference in these two types of policyholders?
   (b) What is the p value for the test in part (a)?
4. A large swine flu vaccination program was instituted in 1976. Approximately 50 million of the roughly 220 million North Americans received the vaccine. Of the 383 persons who subsequently contracted swine flu, 202 had received the vaccine.
   (a) Test the hypothesis, at the 5 percent level, that the probability of contracting swine flu is the same for the vaccinated portion of the population as for the unvaccinated.
   (b) Do the results of part (a) indicate that the vaccine itself was causing the flu? Can you think of any other possible explanations?
5. Two insect sprays are to be compared. Two rooms of equal size are sprayed, one with spray 1 and the other with spray 2. Then 100 insects are released in each room, and after 2 hours the dead insects are counted. Suppose the result is 64 dead insects in the room sprayed with spray 1 and 52 dead insects in the other room.
   (a) Is the evidence significant enough for us to reject, at the 5 percent level, the hypothesis that the two sprays have equal ability to kill insects?
   (b) What is the p value of the test in part (a)?

6. Random samples of 100 residents from San Francisco and 100 from Los Angeles were chosen, and the residents were questioned about whether they favored raising the driving age. The result was that 56 of those from San Francisco and 45 of those from Los Angeles were in favor.
   (a) Are these data strong enough to establish, at the 10 percent level of significance, that the proportions of the population in the two cities that are in favor are different?
   (b) What about at the 5 percent level?
7. In 1983, a random sample of 1000 scientists included 212 female scientists. On the other hand, a random sample of 1000 scientists drawn in 1990 included 272 women. Use these data to test the hypothesis, at the 5 percent level of significance, that the proportion of scientists who are female was the same in 1983 as in 1990. Also find the p value.
8. Example 10.11 considered a model for predicting a child's gender. One generalization of that model would be to suppose that a child's gender depends only on the number of previous children in the family and on the number of these who are boys. If this were so, then the gender of the third child in families whose children presently consist of one boy and one girl would not depend on whether the order of the first two children was boy–girl or girl–boy. The following data give the gender of the third child in families whose first two children were a boy and a girl. It distinguishes whether the boy or the girl was older. (Boy–girl means, for instance, that the older child was a boy.)

Boy–girl families	**Girl–boy families**
412 boys	560 boys
418 girls	544 girls

   Use the given data to test the hypothesis that the sex of a third child in a family presently having one boy and one girl does not depend upon the gender birth order of the two older siblings. Use the 5 percent level of significance.
9. According to the National Center for Health Statistics, there were a total of 330,535 African American females and 341,441 African American males born in 1988. Also in that year, 1,483,487 white females and 1,562,675 white males were born. Use these data to test the hypothesis that the proportion of all African American babies who are female is equal to the proportion of all white babies who are female. Use the 5 percent level of significance. Also find the p value.
10. Suppose a random sample of 480 heart-bypass operations at hospital A showed that 72 patients did not survive, whereas a random sample of 360 operations at hospital B showed that 30 patients did not

survive. Find the p value of the test of the hypothesis that the survival probabilities are the same at the two hospitals.

11. A birthing class run by the University of California has recently added a lecture on the importance of the use of automobile car seats for children. This decision was made after a study of the results of an experiment in which the lecture was given in some of the birthing classes and not in others. A follow-up interview, carried out 1 year later, questioned 82 couples who had heard the lecture and 120 who had not. A total of 78 of the couples who had heard the lecture stated that they always used an infant car seat, whereas a total of 90 of those couples not attending the lecture made the same claim.
    (a) Assuming the accuracy of the given information, is the difference significant enough to conclude that instituting the lecture will result in increased use of car seats? Use the 5 percent level of significance.
    (b) What is the p value?
12. In a study of the effect of two chemotherapy treatments on the survival of patients with multiple myeloma, each of 156 patients were equally likely to be given either one of the two treatments. As reported by Lipsitz, Dear, Laird, and Molenberghs in a 1998 paper in *Biometrics*, the result of this was that 39 of the 72 patients given the first treatment and 44 of the 84 patients given the second treatment survived for over five years.
    (a) Use these data to test the null hypothesis that the two treatments are equally effective.
    (b) Is the fact that 72 of the patients received one of the treatments while 84 received the other consistent with the claim that the determination of the treatment to be given to each patient was made in a totally random fashion?
13. To see how effective a newly developed vaccine is against the common cold, 204 workers at a ski resort were randomly divided into two groups of size 102 each. Members of the first group were given the vaccine throughout the winter months, while members of the second group were given a placebo. By the end of the winter season, it turned out that 29 individuals who had been receiving the vaccine caught at least one cold, compared to 34 of those receiving the placebo. Does this prove, at the 5 percent level of significance, that the vaccine is effective in preventing colds?
14. The American Cancer Society recently sampled 2500 adults and determined that 738 of them were smokers. A similar poll of 2000 adults carried out in 1986 yielded a total of 640 smokers. Do these figures prove that the proportion of adults who smoke has decreased since 1986?

(a) Use the 5 percent level of significance.
(b) Use the 1 percent level of significance.

15. In a recent study of 22,000 male physicians, half were given a daily dose of aspirin while the other half were given a placebo. The study was continued for a period of 6 years. During this time 104 of those taking the aspirin and 189 of those taking the placebo suffered heart attacks. Does this result indicate that taking a daily dose of aspirin decreases the risk of suffering a heart attack? Give the null hypothesis and the resulting p value.

16. In the 1970s, the U.S. Veterans Administration conducted an experiment comparing coronary artery bypass surgery with medical drug therapy as treatments for coronary artery disease. The experiment involved 596 patients, of whom 286 were randomly assigned to receive surgery, with the remaining 310 assigned to drug therapy. A total of 252 of those receiving surgery and a total of 270 of those receiving drug therapy were still alive 3 years after treatment. Use these data to test the hypothesis that the survival probabilities are equal.

KEY TERMS

Two-sample tests: Tests concerning the relationships of parameters from two separate populations.

Paired-sample tests: Tests where the data consist of pairs of dependent variables.

SUMMARY

I. Testing Equality of Population Means: Independent Samples. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent samples from normal populations having respective parameters μ_x, σ_x^2 and μ_y, σ_y^2.

Case 1: *σ_x^2 and σ_y^2 are known.*

To test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \neq \mu_y$$

use the test statistic

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |TS| \ge z_{\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If the value of TS is v, then

$$p \text{ value} = P\{|Z| \ge |v|\} = 2P\{Z \ge |v|\}$$

where Z is a standard normal random variable.

The significance-level-α test of

$$H_0\colon \mu_x \le \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

uses the same test statistic. The test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } \; TS \ge z_{\alpha} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If TS $= v$, then the p value is

$$p \text{ value} = P\{Z \ge v\}$$

Case 2: *σ_x^2 and σ_y^2 are unknown and n and m are large.*

To test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \ne \mu_y$$

or

$$H_0\colon \mu_x \le \mu_y \quad \text{against} \quad H_1\colon \mu_x > \mu_y$$

use the test statistic

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$$

where S_x^2 and S_y^2 are the respective sample variances. The test statistic, the significance-level-α test, and the p value are then exactly the same as in case 1.

Case 3: *σ_x^2 are σ_y^2 assumed to be unknown but equal.*

To test

$$H_0\colon \mu_x = \mu_y \quad \text{against} \quad H_1\colon \mu_x \ne \mu_y$$

use the test statistic

$$\mathrm{TS} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$$

where S_p^2, called the *pooled estimator* of the common variance, is given by

$$S_p^2 = \frac{n-1}{n+m-2}S_x^2 + \frac{m-1}{n+m-2}S_y^2$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } |\mathrm{TS}| \geq t_{n+m-2,\alpha/2} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

If $\mathrm{TS} = v$, then the p value is

$$p \text{ value} = 2P\{t_{n+m-2} \geq |v|\}$$

In the preceding, T_{n+m-2} is a t random variable having $n+m-2$ degrees of freedom, and $t_{n+m-2,\alpha}$ is such that

$$P\{T_{n+m-2} \geq t_{n+m-2,\alpha}\} = \alpha$$

To test

$$\mathrm{H}_0\colon \mu_x \leq \mu_y \quad \text{against} \quad \mathrm{H}_1\colon \mu_x > \mu_y$$

use the same test statistic. The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } \mathrm{TS} \geq t_{n+m-2,\alpha} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

If $\mathrm{TS} = v$, then

$$p \text{ value} = P\{T_{n+m-2} \geq v\}$$

II. Testing Equality of Population Means: Paired Samples. Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are samples from populations having respective means of μ_x and μ_y. Suppose also that these samples are not independent but that the n pairs of random variables X_i and Y_i are dependent, $i = 1, \ldots, n$. Let, for each i,

$$D_i = X_i - Y_i$$

and suppose that $D_1, \ldots, D_n$ constitute a sample from a normal population. Let

$$\mu_d = E\left[D_i\right] = \mu_x - \mu_y$$

To test

$$\mathrm{H}_0\colon \mu_x = \mu_y \quad \text{against} \quad \mathrm{H}_1\colon \mu_x \neq \mu_y$$

test the equivalent hypothesis

$$\mathrm{H}_0\colon \mu_d = 0 \quad \text{against} \quad \mathrm{H}_1\colon \mu_d \neq 0$$

Testing that the two samples have equal means is thus equivalent to testing that a normal population has mean 0. This latter hypothesis is tested by using the t test presented in Sec. 9.4. The test statistic is

$$\mathrm{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

and the significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } |\mathrm{TS}| \geq t_{n-1,\alpha/2} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

If $\mathrm{TS} = v$, then

$$p \text{ value} = 2P\{T_{n-1} \geq |v|\}$$

To test the one-sided hypothesis

$$\mathrm{H}_0\colon \mu_x \leq \mu_y \quad \text{against} \quad \mathrm{H}_1\colon \mu_x > \mu_y$$

use the test statistic

$$\mathrm{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } \mathrm{TS} \geq t_{n-1,\alpha} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

If $\mathrm{TS} = v$, then

$$p \text{ value} = P\{T_{n-1} \geq v\}$$

III. Testing Equality of Population Proportions. Consider two large populations and a certain characteristic possessed by some members of these populations. Let p_1 and p_2 denote, respectively, the proportions of the members of the first and second populations that possess this characteristic. Suppose that a random sample of size n_1 is chosen from population 1 and that one of size n_2 is chosen from population 2. Let X_1 and X_2 denote, respectively, the numbers of members of these samples that possess the characteristic.

Let

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

denote the proportions of the samples that have the characteristic, and let

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

denote the proportion of the combined samples with the characteristic.

To test

$$H_0\colon p_1 = p_2 \quad \text{against} \quad H_1\colon p_1 \neq p_2$$

use the test statistic

$$\text{TS} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$$

The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\text{TS}| \geq z_{\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If the value of TS is v, then

$$p \text{ value} = 2P\{Z \geq |v|\}$$

To test

$$H_0\colon p_1 \leq p_2 \quad \text{against} \quad H_1\colon p_1 > p_2$$

use the same test statistic. The significance-level-α test is to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } \text{TS} \geq z_{\alpha} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

If $\text{TS} = v$, then

$$p \text{ value} = P\{Z \geq v\}$$

Remark *In the foregoing, as in all the text, Z always refers to a standard normal random variable, and* z_α *is such that*

$$P\{Z \geq z_\alpha\} = \alpha$$

REVIEW PROBLEMS

1. The following data concerning the birth weight (in grams) of newborns resulted from a study that attempted to determine the effect of maternal smoking on unborn babies.

Nonsmokers	Smokers
$n = 1820$	$m = 1340$
$\overline{X} = 3480$ grams	$\overline{Y} = 3260$ grams
$S_x = 9.2$ grams	$S_y = 10.4$ grams

   (a) Test the hypothesis, at the 5 percent level of significance, that the mean weight of a newborn is the same whether or not the mother is a smoker.
   (b) What is the p value in part (a)?

2. A study was initiated to compare two treatments for reducing the possibility of rejection in heart transplants. The first treatment involves giving the patient sodium salicylate, and the second calls for this drug to be given in conjunction with a second drug, azathioprine. The study was conducted on male rats, with one type of rat being used as heart donor and a second type being used as recipient. (The use of different types of rats ensured that recipients would not survive too long.) The variable of interest is the survival time in days after receipt of the transplanted heart. The following summary statistics were obtained.

Sodium salicylate	Sodium salicylate with azathioprine
$n = 14$	$m = 12$
$\overline{X} = 15.2$ days	$\overline{Y} = 14$ days
$S_x = 9.2$ days	$S_y = 9.0$ days

   Test the hypothesis, at the 5 percent level, that both treatments are equally effective in the population of rats.

3. A recent study concerning knee injuries of football players compared two types of football shoes. Out of a randomly chosen group of 1440 players, 240 used multicleated shoes and 1200 used more conventional

football shoes. All played on natural grass. Of those using the multi-cleated shoes, 13 suffered knee injuries. Of those using conventional shoes, 78 suffered knee injuries.

(a) Test the hypothesis that the probability of a knee injury is the same for both groups of players. Use the 5 percent level of significance.
(b) What is the p value in part (a)?
(c) Are the given data strong enough to establish that the multi-cleated shoes are superior to the conventional ones in terms of reducing the probability of a knee injury?
(d) In part (c), at what levels of significance would the evidence be strong enough?

4. Use the first 60 data values in App. A. Test, at the 5 percent level of significance, the hypothesis that men's and women's mean
   (a) Cholesterol
   (b) Blood pressure
   are equal.

5. The following data come from an experiment performed by Charles Darwin and reported in his 1876 book *The Effects of Cross- and Self-Fertilization in the Vegetable Kingdom*. The data were first analyzed by Darwin's cousin Francis Galton. Galton's analysis was, however, in error. A correct analysis was eventually done by R. A. Fisher.

   Darwin's experiment dealt with 15 pairs of *Zea mays*, a type of corn plant. One plant in each pair had been cross-fertilized while the other plant had been self-fertilized. The pairs were grown in the same pot, and their heights were measured. The data were as follows:

Pair	Cross-fertilized plant	Self-fertilized plant
1	23.5	17.375
2	12	20.375
3	21	20
4	22	20
5	19.125	18.375
6	21.5	18.625
7	22.125	18.625
8	20.375	15.25
9	18.25	16.5
10	21.625	18
11	23.25	16.25
12	21	18
13	22.125	12.75
14	23	15.5
15	12	18

(a) Test, at the 5 percent level of significance, the hypothesis that the mean height of cross-fertilized *Zea mays* corn plants is equal to that of self-fertilized *Zea mays* plants.

(b) Determine the p value for the test of the hypothesis in part (a).

6. A continuing debate in public health circles concerns the dangers of being exposed to dioxin, an environmental contaminant. A German study published in the October 19, 1991, issue of *The Lancet*, a British medical journal, considered records of workers at a herbicide manufacturing plant that made use of dioxin. A control group consisted of workers at a nearby gas supply company who had similar medical profiles. The following data relating to the number of workers who had died from cancer were obtained.

	Control group	Dioxin-exposed group
Sample size	1583	1242
Number dying from cancer	113	123

(a) Test the hypothesis that the probability of dying from cancer is the same for the two groups. Use the 1 percent level of significance.

(b) Find the p value for the test of part (a).

7. Consider Prob. 6. Of the 1583 gas company workers whose records were studied, there were a total of 1184 men and 399 women. Of these individuals, 93 men and 20 women died of cancer. Test the hypothesis, at the 5 percent level, that the probability of dying from cancer is the same for workers of both sexes.

8. A random sample of 56 women revealed that 38 were in favor of gun control. A random sample of 64 men revealed that 32 were in favor. Use these data to test the hypothesis that the proportion of men and the proportion of women in favor of gun control are the same. Use the 5 percent level of significance. What is the p value?

9. Use the data presented in Review Prob. 19 of Chap. 8 to test the hypothesis that the chances of scoring a run are the same when there is one out and a player on second base and when there are no outs and a player on first base.

10. The following data concern 100 randomly chosen professional baseball games and 100 randomly chosen professional football games in the 1990–91 season. The data present, for the two sports, the number of games in which the team leading at the three-quarter mark (end of the seventh inning in baseball and end of the third quarter in football) ended up losing the game.

Sport	Number of games	Number of games lost by leader
Baseball	92	6
Football	93	21

Find the p value of the test of the hypothesis that the probability of the leading team's losing the game is the same in both sports. (*Note*: The number of games is not 100 because 8 of the baseball games and 7 of the football games were tied at the three-quarter point.)

11. The following relates to the same set of sample games reported in Prob. 10. It details the number of games in which the home team was the winner.

Sport	Number of games	Number of games that home team won
Baseball	100	53
Football	99	57

Test the hypothesis, at the 5 percent level of significance, that the proportion of games won by the home team is the same in both sports.

12. Suppose that a test of H_0: $\mu_x = \mu_y$ against H_1: $\mu_x \neq \mu_y$ results in rejecting H_0 at the 5 percent level of significance. Which of the following statements is (are) true?
    (a) The difference in sample means was statistically significant at the 1 percent level of significance.
    (b) The difference in sample means was statistically significant at the 10 percent level of significance.
    (c) The difference in sample means is equal to the difference in population means.

13. To verify the hypothesis that blood lead levels tend to be higher for children whose parents work in a factory that uses lead in the manufacturing process, researchers examined lead levels in the blood of 33 children whose parents worked in a battery manufacturing factory. (Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., and Saah, M., "Lead Absorption in Children of Employees in a Lead-Related Industry," *American Journal of Epidemiology*, **115**, 549–555, 1982.) Each of these children were then *matched* by another child who was of similar age, lived in a similar neighborhood, had a similar exposure to traffic, but whose parent did not work with lead. The blood levels of the 33 cases (sample 1) as well as those of the 33 controls (sample 2) were then used to test the hypothesis that the average blood levels of these groups are the same. If the resulting sample means and sample

standard deviations were

$$\overline{X}_1 = 0.015, \quad S_1 = 0.004, \quad \overline{X}_2 = 0.006, \quad S_2 = 0.006$$

find the resulting p value. Assume a common variance.

14. A scientist looking into the effect of smoking on heart disease has chosen a large random sample of smokers and another of nonsmokers. She plans to study these two groups for 5 years to see if the number of heart attacks among the members of the smokers group is significantly greater than the number among the nonsmokers. Such a result, the scientist believes, should be strong evidence of an association between smoking and heart attacks.
    Would the scientist be justified in her conclusion if the following were true?
    1. Older people are at greater risk of heart disease than are younger people.
    2. As a group, smokers tend to be somewhat older than nonsmokers.

    Explain how the experimental design can be improved so that meaningful conclusions can be drawn.