10.4 Single Sample: Tests Concerning a Single Mean

In this section, we formally consider tests of hypotheses on a single population mean. Many of the illustrations from previous sections involved tests on the mean, so the reader should already have insight into some of the details that are outlined here.

Tests on a Single Mean (Variance Known)

We should first describe the assumptions on which the experiment is based. The model for the underlying situation centers around an experiment with $X_1, X_2, \ldots, X_n$ representing a random sample from a distribution with mean μ and variance $\sigma^2 > 0$. Consider first the hypothesis

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0.$$

The appropriate test statistic should be based on the random variable $\bar{X}$. In Chapter 8, the Central Limit Theorem was introduced, which essentially states that despite the distribution of X, the random variable $\bar{X}$ has approximately a normal distribution with mean μ and variance σ^2/n for reasonably large sample sizes. So, $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$. We can then determine a critical region based on the computed sample average, $\bar{x}$. It should be clear to the reader by now that there will be a two-tailed critical region for the test.

Standardization of $\bar{X}$

It is convenient to standardize $\bar{X}$ and formally involve the **standard normal** random variable Z, where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We know that *under* H_0, that is, if $\mu = \mu_0$, $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ follows an $n(x; 0, 1)$ distribution, and hence the expression

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

can be used to write an appropriate nonrejection region. The reader should keep in mind that, formally, the critical region is designed to control α, the probability of type I error. It should be obvious that a *two-tailed signal* of evidence is needed to support H_1. Thus, given a computed value $\bar{x}$, the formal test involves rejecting H_0 if the computed *test statistic* z falls in the critical region described next.

Test Procedure for a Single Mean (Variance Known)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{or} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

If $-z_{\alpha/2} < z < z_{\alpha/2}$, do not reject H_0. Rejection of H_0, of course, implies acceptance of the alternative hypothesis $\mu \neq \mu_0$. With this definition of the critical region, it should be clear that there will be probability α of rejecting H_0 (falling into the critical region) when, indeed, $\mu = \mu_0$.

Although it is easier to understand the critical region written in terms of z, we can write the same critical region in terms of the computed average $\bar{x}$. The following can be written as an identical decision procedure:

$$\textit{reject } H_0 \text{ if } \bar{x} < a \text{ or } \bar{x} > b,$$

where

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, for a significance level α, the critical values of the random variable z and $\bar{x}$ are both depicted in Figure 10.9.

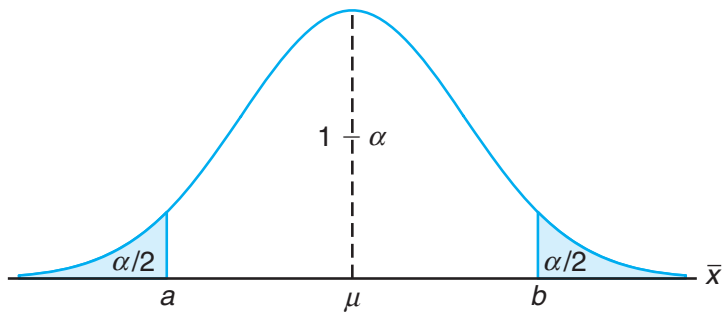

Figure 10.9: Critical region for the alternative hypothesis $\mu \neq \mu_0$.

Tests of one-sided hypotheses on the mean involve the same statistic described in the two-sided case. The difference, of course, is that the critical region is only in one tail of the standard normal distribution. For example, suppose that we seek to test

$$H_0\colon\ \mu = \mu_0,$$
$$H_1\colon\ \mu > \mu_0.$$

The signal that favors H_1 comes from *large values* of z. Thus, rejection of H_0 results when the computed $z > z_\alpha$. Obviously, if the alternative is $H_1\colon \mu < \mu_0$, the critical region is entirely in the lower tail and thus rejection results from $z < -z_\alpha$. Although in a one-sided testing case the null hypothesis can be written as $H_0\colon \mu \leq \mu_0$ or $H_0\colon \mu \geq \mu_0$, it is usually written as $H_0\colon \mu = \mu_0$.

The following two examples illustrate tests on means for the case in which σ is known.

Example 10.3: A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution:

1. H_0: $\mu = 70$ years.
2. H_1: $\mu > 70$ years.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 71.8$ years, $\sigma = 8.9$ years, and hence $z = \frac{71.8-70}{8.9/\sqrt{100}} = 2.02$.
6. Decision: Reject H_0 and conclude that the mean life span today is greater than 70 years.

The P-value corresponding to $z = 2.02$ is given by the area of the shaded region in Figure 10.10.

Using Table A.3, we have

$$P = P(Z > 2.02) = 0.0217.$$

As a result, the evidence in favor of H_1 is even stronger than that suggested by a 0.05 level of significance. ∎

Example 10.4: A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kilograms against the alternative that $\mu \neq 8$ kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

Solution:

1. H_0: $\mu = 8$ kilograms.
2. H_1: $\mu \neq 8$ kilograms.
3. $\alpha = 0.01$.
4. Critical region: $z < -2.575$ and $z > 2.575$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 7.8$ kilograms, $n = 50$, and hence $z = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$.
6. Decision: Reject H_0 and conclude that the average breaking strength is not equal to 8 but is, in fact, less than 8 kilograms.

Since the test in this example is two tailed, the desired P-value is twice the area of the shaded region in Figure 10.11 to the left of $z = -2.83$. Therefore, using Table A.3, we have

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046,$$

which allows us to reject the null hypothesis that $\mu = 8$ kilograms at a level of significance smaller than 0.01. ∎

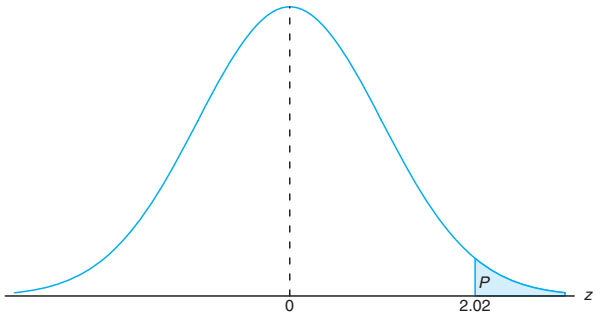

Figure 10.10: P-value for Example 10.3.

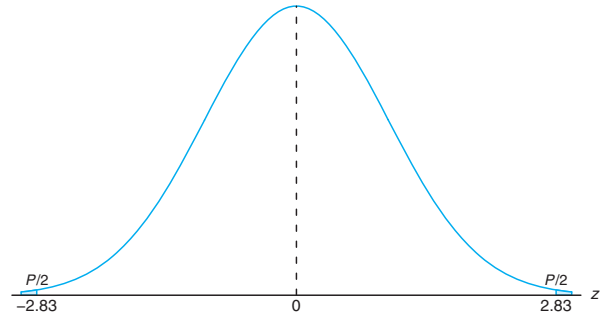

Figure 10.11: P-value for Example 10.4.

Relationship to Confidence Interval Estimation

The reader should realize by now that the hypothesis-testing approach to statistical inference in this chapter is very closely related to the confidence interval approach in Chapter 9. Confidence interval estimation involves computation of bounds within which it is "reasonable" for the parameter in question to lie. For the case of a single population mean μ with σ^2 known, the structure of both hypothesis testing and confidence interval estimation is based on the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

It turns out that the testing of H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$ at a significance level α *is equivalent to computing a* $100(1-\alpha)\%$ *confidence interval on* μ *and rejecting* H_0 *if* μ_0 *is outside the confidence interval.* If μ_0 is inside the confidence interval, the hypothesis is not rejected. The equivalence is very intuitive and quite simple to illustrate. Recall that with an observed value $\bar{x}$, failure to reject H_0 at significance level α implies that

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

which is equivalent to

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The equivalence of confidence interval estimation to hypothesis testing extends to differences between two means, variances, ratios of variances, and so on. As a result, the student of statistics should not consider confidence interval estimation and hypothesis testing as separate forms of statistical inference. For example, consider Example 9.2 on page 271. The 95% confidence interval on the mean is given by the bounds (2.50, 2.70). Thus, with the same sample information, a two-sided hypothesis on μ involving any hypothesized value between 2.50 and 2.70 will not be rejected. As we turn to different areas of hypothesis testing, the equivalence to the confidence interval estimation will continue to be exploited.

Tests on a Single Sample (Variance Unknown)

One would certainly suspect that tests on a population mean μ with σ^2 unknown, like confidence interval estimation, should involve the use of Student t-distribution. Strictly speaking, the application of Student t for both confidence intervals and hypothesis testing is developed under the following assumptions. The random variables $X_1, X_2, \ldots, X_n$ represent a random sample from a normal distribution with unknown μ and σ^2. Then the random variable $\sqrt{n}(\bar{X} - \mu)/S$ has a Student t-distribution with $n-1$ degrees of freedom. The structure of the test is identical to that for the case of σ known, with the exception that the value σ in the test statistic is replaced by the computed estimate S and the standard normal distribution is replaced by a t-distribution.

The t-Statistic for a Test on a Single Mean (Variance Unknown)

For the two-sided hypothesis

$$H_0: \ \mu = \mu_0,$$
$$H_1: \ \mu \neq \mu_0,$$

we reject H_0 at significance level α when the computed t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

exceeds $t_{\alpha/2,n-1}$ or is less than $-t_{\alpha/2,n-1}$.

The reader should recall from Chapters 8 and 9 that the t-distribution is symmetric around the value zero. Thus, this two-tailed critical region applies in a fashion similar to that for the case of known σ. For the two-sided hypothesis at significance level α, the two-tailed critical regions apply. For H_1: $\mu > \mu_0$, rejection results when $t > t_{\alpha,n-1}$. For H_1: $\mu < \mu_0$, the critical region is given by $t < -t_{\alpha,n-1}$.

Example 10.5: The Edison Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

Solution:

1. H_0: $\mu = 46$ kilowatt hours.
2. H_1: $\mu < 46$ kilowatt hours.
3. $\alpha = 0.05$.
4. Critical region: $t < -1.796$, where $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with 11 degrees of freedom.
5. Computations: $\bar{x} = 42$ kilowatt hours, $s = 11.9$ kilowatt hours, and $n = 12$. Hence,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decision: Do not reject H_0 and conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

Comment on the Single-Sample *t*-Test

The reader has probably noticed that the equivalence of the two-tailed t-test for a single mean and the computation of a confidence interval on μ with σ replaced by s is maintained. For example, consider Example 9.5 on page 275. Essentially, we can view that computation as one in which we have found all values of μ_0, the hypothesized mean volume of containers of sulfuric acid, for which the hypothesis H_0: $\mu = \mu_0$ *will not be rejected* at $\alpha = 0.05$. Again, this is consistent with the statement "Based on the sample information, values of the population mean volume between 9.74 and 10.26 liters are not unreasonable."

Comments regarding the normality assumption are worth emphasizing at this point. We have indicated that when σ is known, the Central Limit Theorem allows for the use of a test statistic or a confidence interval which is based on Z, the standard normal random variable. Strictly speaking, of course, the Central Limit Theorem, and thus the use of the standard normal distribution, does not apply unless σ is known. In Chapter 8, the development of the t-distribution was given. There we pointed out that normality on $X_1, X_2, \ldots, X_n$ was an underlying assumption. Thus, *strictly speaking*, the Student's t-tables of percentage points for tests or confidence intervals should not be used unless it is known that the sample comes from a normal population. In practice, σ can rarely be assumed to be known. However, a very good estimate may be available from previous experiments. Many statistics textbooks suggest that one can safely replace σ by s in the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

when $n \geq 30$ with a bell-shaped population and still use the Z-tables for the appropriate critical region. The implication here is that the Central Limit Theorem is indeed being invoked and one is relying on the fact that $s \approx \sigma$. Obviously, when this is done, the results must be viewed as approximate. Thus, a computed P-value (from the Z-distribution) of 0.15 may be 0.12 or perhaps 0.17, or a computed confidence interval may be a 93% confidence interval rather than a 95% interval as desired. Now what about situations where $n \leq 30$? The user cannot rely on s being close to σ, and in order to take into account the inaccuracy of the estimate, the confidence interval should be wider or the critical value larger in magnitude. The t-distribution percentage points accomplish this but are correct only when the sample is from a normal distribution. Of course, normal probability plots can be used to ascertain some sense of the deviation of normality in a data set.

For small samples, it is often difficult to detect deviations from a normal distribution. (Goodness-of-fit tests are discussed in a later section of this chapter.) For bell-shaped distributions of the random variables $X_1, X_2, \ldots, X_n$, the use of the t-distribution for tests or confidence intervals is likely to produce quite good results. When in doubt, the user should resort to nonparametric procedures, which are presented in Chapter 16.

Annotated Computer Printout for Single-Sample *t*-Test

It should be of interest for the reader to see an annotated computer printout showing the result of a single-sample t-test. Suppose that an engineer is interested in testing the bias in a pH meter. Data are collected on a neutral substance (pH = 7.0). A sample of the measurements were taken with the data as follows:

$$7.07 \quad 7.00 \quad 7.10 \quad 6.97 \quad 7.00 \quad 7.03 \quad 7.01 \quad 7.01 \quad 6.98 \quad 7.08$$

It is, then, of interest to test

$$H_0\text{: } \mu = 7.0,$$
$$H_1\text{: } \mu \neq 7.0.$$

In this illustration, we use the computer package *MINITAB* to illustrate the analysis of the data set above. Notice the key components of the printout shown in Figure 10.12. Of course, the mean $\bar{y}$ is 7.0250, StDev is simply the sample standard deviation $s = 0.044$, and SE Mean is the estimated standard error of the mean and is computed as $s/\sqrt{n} = 0.0139$. The t-value is the ratio

$$(7.0250 - 7)/0.0139 = 1.80.$$

```
pH-meter
  7.07   7.00   7.10   6.97   7.00   7.03   7.01   7.01   6.98   7.08
MTB > Onet 'pH-meter'; SUBC>   Test 7.

One-Sample T: pH-meter Test of mu = 7 vs not = 7
Variable  N    Mean   StDev  SE Mean      95% CI          T      P
pH-meter 10 7.02500 0.04403  0.01392 (6.99350, 7.05650) 1.80 0.106
```

Figure 10.12: *MINITAB* printout for one sample t-test for pH meter.

The P-value of 0.106 suggests results that are inconclusive. There is no evidence suggesting a strong rejection of H_0 (based on an α of 0.05 or 0.10), **yet one certainly cannot truly conclude that the pH meter is unbiased**. Notice that the sample size of 10 is rather small. An increase in sample size (perhaps another experiment) may sort things out. A discussion regarding appropriate sample size appears in Section 10.6.

10.5 Two Samples: Tests on Two Means

The reader should now understand the relationship between tests and confidence intervals, and can only heavily rely on details supplied by the confidence interval material in Chapter 9. Tests concerning two means represent a set of very important analytical tools for the scientist or engineer. The experimental setting is very much like that described in Section 9.8. Two independent random samples of sizes

n_1 and n_2, respectively, are drawn from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We know that the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has a standard normal distribution. Here we are assuming that n_1 and n_2 are sufficiently large that the Central Limit Theorem applies. Of course, if the two populations are normal, the statistic above has a standard normal distribution even for small n_1 and n_2. Obviously, if we can assume that $\sigma_1 = \sigma_2 = \sigma$, the statistic above reduces to

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

The two statistics above serve as a basis for the development of the test procedures involving two means. The equivalence between tests and confidence intervals, along with the technical detail involving tests on one mean, allow a simple transition to tests on two means.

The two-sided hypothesis on two means can be written generally as

$$H_0\colon \mu_1 - \mu_2 = d_0.$$

Obviously, the alternative can be two sided or one sided. Again, the distribution used is the distribution of the test statistic under H_0. Values $\bar{x}_1$ and $\bar{x}_2$ are computed and, for σ_1 and σ_2 known, the test statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

with a two-tailed critical region in the case of a two-sided alternative. That is, reject H_0 in favor of H_1: $\mu_1 - \mu_2 \neq d_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$. One-tailed critical regions are used in the case of the one-sided alternatives. The reader should, as before, study the test statistic and be satisfied that for, say, H_1: $\mu_1 - \mu_2 > d_0$, the signal favoring H_1 comes from large values of z. Thus, the upper-tailed critical region applies.

Unknown But Equal Variances

The more prevalent situations involving tests on two means are those in which variances are unknown. If the scientist involved is willing to assume that both distributions are normal and that $\sigma_1 = \sigma_2 = \sigma$, the *pooled t-test* (often called the two-sample t-test) may be used. The test statistic (see Section 9.8) is given by the following test procedure.

Two-Sample Pooled t-Test

For the two-sided hypothesis

$$H_0: \ \mu_1 = \mu_2,$$
$$H_1: \ \mu_1 \neq \mu_2,$$

we reject H_0 at significance level α when the computed t-statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

where

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

exceeds $t_{\alpha/2,n_1+n_2-2}$ or is less than $-t_{\alpha/2,n_1+n_2-2}$.

Recall from Chapter 9 that the degrees of freedom for the t-distribution are a result of pooling of information from the two samples to estimate σ^2. One-sided alternatives suggest one-sided critical regions, as one might expect. For example, for H_1: $\mu_1 - \mu_2 > d_0$, reject H_1: $\mu_1 - \mu_2 = d_0$ when $t > t_{\alpha,n_1+n_2-2}$.

Example 10.6: An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

***Solution*:** Let μ_1 and μ_2 represent the population means of the abrasive wear for material 1 and material 2, respectively.

1. H_0: $\mu_1 - \mu_2 = 2$.
2. H_1: $\mu_1 - \mu_2 > 2$.
3. $\alpha = 0.05$.
4. Critical region: $t > 1.725$, where $t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{1/n_1+1/n_2}}$ with $v = 20$ degrees of freedom.
5. Computations:

$$\bar{x}_1 = 85, \quad s_1 = 4, \quad n_1 = 12,$$
$$\bar{x}_2 = 81, \quad s_2 = 5, \quad n_2 = 10.$$

Hence

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$
$$P = P(T > 1.04) \approx 0.16. \qquad \text{(See Table A.4.)}$$

6. Decision: Do not reject H_0. We are unable to conclude that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units. ∎

Unknown But Unequal Variances

There are situations where the analyst is **not** able to assume that $\sigma_1 = \sigma_2$. Recall from Section 9.8 that, if the populations are normal, the statistic

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

has an approximate t-distribution with approximate degrees of freedom

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

As a result, the test procedure is to *not reject* H_0 when

$$-t_{\alpha/2,v} < t' < t_{\alpha/2,v},$$

with v given as above. Again, as in the case of the pooled t-test, one-sided alternatives suggest one-sided critical regions.

Paired Observations

A study of the two-sample t-test or confidence interval on the difference between means should suggest the need for experimental design. Recall the discussion of experimental units in Chapter 9, where it was suggested that the conditions of the two populations (often referred to as the two treatments) should be assigned randomly to the experimental units. This is done to avoid biased results due to systematic differences between experimental units. In other words, in hypothesis-testing jargon, it is important that any significant difference found between means be due to the different conditions of the populations and not due to the experimental units in the study. For example, consider Exercise 9.40 in Section 9.9. The 20 seedlings play the role of the experimental units. Ten of them are to be treated with nitrogen and 10 with no nitrogen. It may be very important that this assignment to the "nitrogen" and "no-nitrogen" treatments be random to ensure that systematic differences between the seedlings do not interfere with a valid comparison between the means.

In Example 10.6, time of measurement is the most likely choice for the experimental unit. The 22 pieces of material should be measured in random order. We

need to guard against the possibility that wear measurements made close together in time might tend to give similar results. **Systematic** (nonrandom) **differences in experimental units** *are not expected*. However, random assignments guard against the problem.

References to planning of experiments, randomization, choice of sample size, and so on, will continue to influence much of the development in Chapters 13, 14, and 15. Any scientist or engineer whose interest lies in analysis of real data should study this material. The pooled t-test is extended in Chapter 13 to cover more than two means.

Testing of two means can be accomplished when data are in the form of paired observations, as discussed in Chapter 9. In this pairing structure, the conditions of the two populations (treatments) are assigned randomly within homogeneous units. Computation of the confidence interval for $\mu_1 - \mu_2$ in the situation with paired observations is based on the random variable

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}},$$

where $\bar{D}$ and S_d are random variables representing the sample mean and standard deviation of the differences of the observations in the experimental units. As in the case of the *pooled t-test*, the assumption is that the observations from each population are normal. This two-sample problem is essentially reduced to a one-sample problem by using the computed differences $d_1, d_2, \ldots, d_n$. Thus, the hypothesis reduces to

$$H_0\colon\ \mu_D = d_0.$$

The computed test statistic is then given by

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}.$$

Critical regions are constructed using the t-distribution with $n - 1$ degrees of freedom.

Problem of Interaction in a Paired *t*-Test

Not only will the case study that follows illustrate the use of the paired t-test but the discussion will shed considerable light on the difficulties that arise when there is an interaction between the treatments and the experimental units in the paired t structure. Recall that interaction between factors was introduced in Section 1.7 in a discussion of general types of statistical studies. The concept of interaction will be an important issue from Chapter 13 through Chapter 15.

There are some types of statistical tests in which the existence of interaction results in difficulty. The paired t-test is one such example. In Section 9.9, the paired structure was used in the computation of a confidence interval on the difference between two means, and the advantage in pairing was revealed for situations in which the experimental units are homogeneous. The pairing results in a reduction in σ_D, the standard deviation of a difference $D_i = X_{1i} - X_{2i}$, as discussed in

Section 9.9. If interaction exists between treatments and experimental units, the advantage gained in pairing may be substantially reduced. Thus, in Example 9.13 on page 293, the no interaction assumption allowed the difference in mean TCDD levels (plasma vs. fat tissue) to be the same across veterans. A quick glance at the data would suggest that there is no significant violation of the assumption of no interaction.

In order to demonstrate how interaction influences $\text{Var}(D)$ and hence the quality of the paired t-test, it is instructive to revisit the ith difference given by $D_i = X_{1i} - X_{2i} = (\mu_1 - \mu_2) + (\epsilon_1 - \epsilon_2)$, where X_{1i} and X_{2i} are taken on the ith experimental unit. If the pairing unit is homogeneous, the errors in X_{1i} and in X_{2i} should be similar and not independent. We noted in Chapter 9 that the positive covariance between the errors results in a reduced $\text{Var}(D)$. Thus, the size of the difference in the treatments and the relationship between the errors in X_{1i} and X_{2i} contributed by the experimental unit will tend to allow a significant difference to be detected.

What Conditions Result in Interaction?

Let us consider a situation in which the experimental units are not homogeneous. Rather, consider the ith experimental unit with random variables X_{1i} and X_{2i} that are not similar. Let ϵ_{1i} and ϵ_{2i} be random variables representing the errors in the values X_{1i} and X_{2i}, respectively, at the ith unit. Thus, we may write

$$X_{1i} = \mu_1 + \epsilon_{1i} \text{ and } X_{2i} = \mu_2 + \epsilon_{2i}.$$

The errors with expectation zero may tend to cause the response values X_{1i} and X_{2i} to move in opposite directions, resulting in a negative value for $\text{Cov}(\epsilon_{1i}, \epsilon_{2i})$ and hence negative $\text{Cov}(X_{1i}, X_{2i})$. In fact, the model may be complicated even more by the fact that $\sigma_1^2 = \text{Var}(\epsilon_{1i}) \neq \sigma_2^2 = \text{Var}(\epsilon_{2i})$. The variance and covariance parameters may vary among the n experimental units. Thus, unlike in the homogeneous case, D_i will tend to be quite different across experimental units due to the heterogeneous nature of the difference in $\epsilon_1 - \epsilon_2$ among the units. This produces the interaction between treatments and units. In addition, for a specific experimental unit (see Theorem 4.9),

$$\sigma_D^2 = \text{Var}(D) = \text{Var}(\epsilon_1) + \text{Var}(\epsilon_2) - 2\ \text{Cov}(\epsilon_1, \epsilon_2)$$

is inflated by the negative covariance term, and thus the advantage gained in pairing in the homogeneous unit case is lost in the case described here. While the inflation in $\text{Var}(D)$ will vary from case to case, there is a danger in some cases that the increase in variance may neutralize any difference that exists between μ_1 and μ_2. Of course, a large value of $\bar{d}$ in the t-statistic may reflect a treatment difference that overcomes the inflated variance estimate, s_d^2.

Case Study 10.1: **Blood Sample Data**: In a study conducted in the Forestry and Wildlife Department at Virginia Tech, J. A. Wesson examined the influence of the drug succinylcholine on the circulation levels of androgens in the blood. Blood samples were taken from wild, free-ranging deer immediately after they had received an intramuscular injection of succinylcholine administered using darts and a capture gun. A second blood sample was obtained from each deer 30 minutes after the

first sample, after which the deer was released. The levels of androgens at time of capture and 30 minutes later, measured in nanograms per milliliter (ng/mL), for 15 deer are given in Table 10.2.

Assuming that the populations of androgen levels at time of injection and 30 minutes later are normally distributed, test at the 0.05 level of significance whether the androgen concentrations are altered after 30 minutes.

Table 10.2: Data for Case Study 10.1

	Androgen (ng/mL)		
Deer	**At Time of Injection**	**30 Minutes after Injection**	d_i
1	2.76	7.02	4.26
2	5.18	3.10	−2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	−2.45
11	11.78	11.10	−0.68
12	3.90	3.74	−0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

***Solution*:** Let μ_1 and μ_2 be the average androgen concentration at the time of injection and 30 minutes later, respectively. We proceed as follows:

1. H_0: $\mu_1 = \mu_2$ or $\mu_D = \mu_1 - \mu_2 = 0$.
2. H_1: $\mu_1 \neq \mu_2$ or $\mu_D = \mu_1 - \mu_2 \neq 0$.
3. $\alpha = 0.05$.
4. Critical region: $t < -2.145$ and $t > 2.145$, where $t = \frac{\overline{d} - d_0}{s_D/\sqrt{n}}$ with $v = 14$ degrees of freedom.
5. Computations: The sample mean and standard deviation for the d_i are

$$\overline{d} = 9.848 \qquad \text{and} \qquad s_d = 18.474.$$

Therefore,

$$t = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06.$$

6. Though the t-statistic is not significant at the 0.05 level, from Table A.4,

$$P = P(|T| > 2.06) \approx 0.06.$$

As a result, there is some evidence that there is a difference in mean circulating levels of androgen.

The assumption of no interaction would imply that the effect on androgen levels of the deer is roughly the same in the data for both treatments, i.e., at the time of injection of succinylcholine and 30 minutes following injection. This can be expressed with the two factors switching roles; for example, the difference in treatments is roughly the same across the units (i.e., the deer). There certainly are some deer/treatment combinations for which the no interaction assumption seems to hold, but there is hardly any strong evidence that the experimental units are homogeneous. However, the nature of the interaction and the resulting increase in $\text{Var}(\bar{D})$ appear to be dominated by a substantial difference in the treatments. This is further demonstrated by the fact that 11 of the 15 deer exhibited positive signs for the computed d_i and the negative d_i (for deer 2, 10, 11, and 12) are small in magnitude compared to the 12 positive ones. Thus, it appears that the mean level of androgen is significantly higher 30 minutes following injection than at injection, and the conclusions may be stronger than $p = 0.06$ would suggest.

Annotated Computer Printout for Paired *t*-Test

Figure 10.13 displays a *SAS* computer printout for a paired t-test using the data of Case Study 10.1. Notice that the printout looks like that for a single sample t-test and, of course, that is exactly what is accomplished, since the test seeks to determine if $\overline{d}$ is significantly different from zero.

Analysis Variable : Diff

N	Mean	Std Error	t Value	Pr > |t|
15	9.8480000	4.7698699	2.06	0.0580

Figure 10.13: *SAS* printout of paired t-test for data of Case Study 10.1.

Summary of Test Procedures

As we complete the formal development of tests on population means, we offer Table 10.3, which summarizes the test procedure for the cases of a single mean and two means. Notice the approximate procedure when distributions are normal and variances are unknown but not assumed to be equal. This statistic was introduced in Chapter 9.

10.6 Choice of Sample Size for Testing Means

In Section 10.2, we demonstrated how the analyst can exploit relationships among the sample size, the significance level α, and the power of the test to achieve a certain standard of quality. In most practical circumstances, the experiment should be planned, with a choice of sample size made prior to the data-taking process if possible. The sample size is usually determined to achieve good power for a fixed α and fixed specific alternative. This fixed alternative may be in the

Table 10.3: Tests Concerning Means

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n-1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n-1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

form of $\mu - \mu_0$ in the case of a hypothesis involving a single mean or $\mu_1 - \mu_2$ in the case of a problem involving two means. Specific cases will provide illustrations.

Suppose that we wish to test the hypothesis

$$H_0\colon\ \mu = \mu_0,$$
$$H_1\colon\ \mu > \mu_0,$$

with a significance level α, when the variance σ^2 is known. For a specific alternative, say $\mu = \mu_0 + \delta$, the power of our test is shown in Figure 10.14 to be

$$1 - \beta = P(\bar{X} > a \text{ when } \mu = \mu_0 + \delta).$$

Therefore,

$$\begin{aligned}\beta &= P(\bar{X} < a \text{ when } \mu = \mu_0 + \delta)\\ &= P\left[\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \frac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_0 + \delta\right].\end{aligned}$$

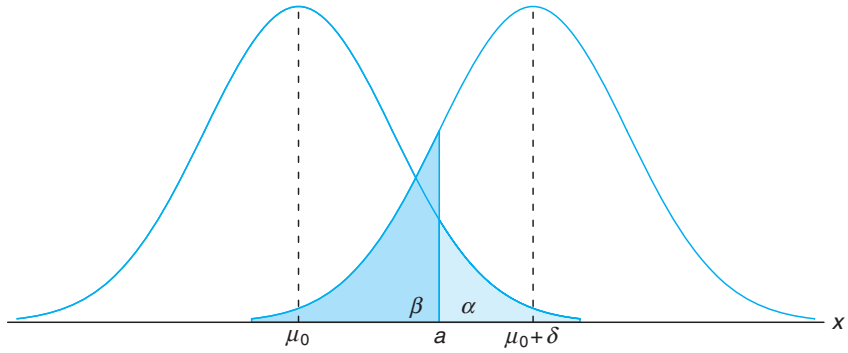

Figure 10.14: Testing $\mu = \mu_0$ versus $\mu = \mu_0 + \delta$.

Under the alternative hypothesis $\mu = \mu_0 + \delta$, the statistic

$$\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}}$$

is the standard normal variable Z. So

$$\beta = P\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}\right) = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right),$$

from which we conclude that

$$-z_\beta = z_\alpha - \frac{\delta\sqrt{n}}{\sigma},$$

and hence

$$\text{Choice of sample size:} \quad n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2},$$

a result that is also true when the alternative hypothesis is $\mu < \mu_0$.

In the case of a two-tailed test, we obtain the power $1 - \beta$ for a specified alternative when

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Example 10.7: Suppose that we wish to test the hypothesis

$$H_0\text{: } \mu = 68 \text{ kilograms},$$
$$H_1\text{: } \mu > 68 \text{ kilograms}$$

for the weights of male students at a certain college, using an $\alpha = 0.05$ level of significance, when it is known that $\sigma = 5$. Find the sample size required if the power of our test is to be 0.95 when the true mean is 69 kilograms.

***Solution*:** Since $\alpha = \beta = 0.05$, we have $z_\alpha = z_\beta = 1.645$. For the alternative $\beta = 69$, we take $\delta = 1$ and then

$$n = \frac{(1.645 + 1.645)^2(25)}{1} = 270.6.$$

Therefore, 271 observations are required if the test is to reject the null hypothesis 95% of the time when, in fact, μ is as large as 69 kilograms.

Two-Sample Case

A similar procedure can be used to determine the sample size $n = n_1 = n_2$ required for a specific power of the test in which two population means are being compared. For example, suppose that we wish to test the hypothesis

$$H_0: \ \mu_1 - \mu_2 = d_0,$$
$$H_1: \mu_1 - \mu_2 \neq d_0,$$

when σ_1 and σ_2 are known. For a specific alternative, say $\mu_1 - \mu_2 = d_0 + \delta$, the power of our test is shown in Figure 10.15 to be

$$1 - \beta = P(|\bar{X}_1 - \bar{X}_2| > a \text{ when } \mu_1 - \mu_2 = d_0 + \delta).$$

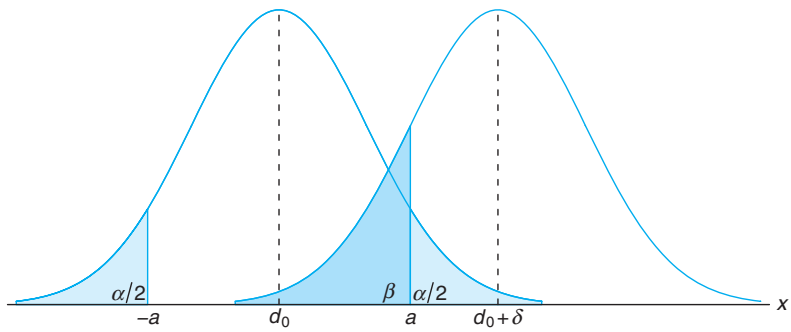

Figure 10.15: Testing $\mu_1 - \mu_2 = d_0$ versus $\mu_1 - \mu_2 = d_0 + \delta$.

Therefore,

$$\begin{aligned}\beta &= P(-a < \bar{X}_1 - \bar{X}_2 < a \text{ when } \mu_1 - \mu_2 = d_0 + \delta) \\ &= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right. \\ &\qquad \left. < \frac{a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \text{ when } \mu_1 - \mu_2 = d_0 + \delta\right].\end{aligned}$$

Under the alternative hypothesis $\mu_1 - \mu_2 = d_0 + \delta$, the statistic

$$\frac{\bar{X}_1 - \bar{X}_2 - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

is the standard normal variable Z. Now, writing

$$-z_{\alpha/2} = \frac{-a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \quad \text{and} \quad z_{\alpha/2} = \frac{a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

we have

$$\beta = P\left[-z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < Z < z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right],$$

from which we conclude that

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

and hence

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

For the one-tailed test, the expression for the required sample size when $n = n_1 = n_2$ is

$$\text{Choice of sample size:} \quad n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

When the population variance (or variances, in the two-sample situation) is unknown, the choice of sample size is not straightforward. In testing the hypothesis $\mu = \mu_0$ when the true value is $\mu = \mu_0 + \delta$, the statistic

$$\frac{\bar{X} - (\mu_0 + \delta)}{S/\sqrt{n}}$$

does not follow the t-distribution, as one might expect, but instead follows the **noncentral *t*-distribution**. However, tables or charts based on the noncentral t-distribution do exist for determining the appropriate sample size if some estimate of σ is available or if δ is a multiple of σ. Table A.8 gives the sample sizes needed to control the values of α and β for various values of

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu - \mu_0|}{\sigma}$$

for both one- and two-tailed tests. In the case of the two-sample t-test in which the variances are unknown but assumed equal, we obtain the sample sizes $n = n_1 = n_2$ needed to control the values of α and β for various values of

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu_1 - \mu_2 - d_0|}{\sigma}$$

from Table A.9.

Example 10.8: In comparing the performance of two catalysts on the effect of a reaction yield, a two-sample t-test is to be conducted with $\alpha = 0.05$. The variances in the yields

are considered to be the same for the two catalysts. How large a sample for each catalyst is needed to test the hypothesis

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

if it is essential to detect a difference of 0.8σ between the catalysts with probability 0.9?

***Solution*:** From Table A.9, with $\alpha = 0.05$ for a two-tailed test, $\beta = 0.1$, and

$$\Delta = \frac{|0.8\sigma|}{\sigma} = 0.8,$$

we find the required sample size to be $n = 34$.

In practical situations, it might be difficult to force a scientist or engineer to make a commitment on information from which a value of Δ can be found. The reader is reminded that the Δ-value quantifies the kind of difference between the means that the scientist considers important, that is, a difference considered *significant* from a scientific, not a statistical, point of view. Example 10.8 illustrates how this choice is often made, namely, by selecting a fraction of σ. Obviously, if the sample size is based on a choice of $|\delta|$ that is a small fraction of σ, the resulting sample size may be quite large compared to what the study allows.