7.1 Inference for the Mean of a Population

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$\mathrm{SE}_{\overline{x}} = \frac{s}{\sqrt{n}}$$

In the previous chapter, the standardized sample mean, or one-sample *z* statistic,

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

However, when we substitute the standard error $s/\sqrt{n}$ for the standard deviation of $\overline{x}$, the statistic does *not* have a Normal distribution. It has a distribution that is new to us, called a *t distribution*.

THE *t* DISTRIBUTIONS

Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample *t* statistic**

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

has the ***t* distribution** with $n - 1$ **degrees of freedom**.

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

Suppose that an SRS of size n is drawn from a population having unknown mean μ. A level C **confidence interval** for μ is

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**. The confidence level is exactly C when the population distribution is Normal and is approximately correct for large n in other cases.

Watching traditional television. The Nielsen Company is a global information and media company and one of the leading suppliers of media information. In their annual Total Audience Report, the Nielsen Company states that adults age 18 to 24 years old average 18.5 hours per week watching traditional television.[1] Does this average seem reasonable for college students? They tend to watch a lot of television, but given their unusual schedules, they may be more likely to binge-watch or stream episodes after they air. To investigate, let's construct a 95% confidence interval for the average time (hours per week) spent watching traditional television among full-time U.S. college students. We draw the following SRS of size 8 from this population:

$$3.016.510.540.55.533.50.06.5$$

The sample mean is

$$\overline{x} = \frac{3.0+16.5+\ldots+6.5}{8} = 14.5$$

and the standard deviation is

$$s = \sqrt{\frac{(3.0-14.5)^2+(16.5-14.5)^2+\ldots+(6.5-14.5)^2}{8-1}} = 14.854$$

with degrees of freedom $n - 1 = 7$. The standard error is

$$\text{SE}_{\overline{x}} = s/\sqrt{n} = 14.854/\sqrt{8} = 5.252$$

From Table D, we find $t^* = 2.365$. The 95% confidence interval is

$$\begin{aligned} \overline{x} \pm t^* \frac{s}{\sqrt{n}} &= 14.5 \pm 2.365 \frac{14.854}{\sqrt{8}} \\ &= 14.5 \pm (2.365)(5.252) \\ &= 14.5 \pm 12.421 \\ &= (2.08,\ 26.92) \end{aligned}$$

We are 95% confident that among U.S. college students the average time spent watching traditional television is between 2.1 and 26.9 hours per week.

df = 7

t^*	1.895	2.365	2.517
C	0.90	0.95	0.96

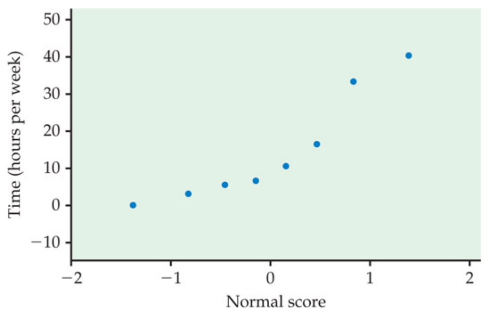


FIGURE 7.2 Normal quantile plot of data, Example 7.1.

In this example, we gave the interval $(2.1, 26.9)$ hours per week as our answer. Sometimes, we prefer to report the mean and margin of error: the mean time is 14.5 hours per week with a margin of error of 12.4 hours per week. This is

THE ONE-SAMPLE *t* TEST

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the **one-sample *t* statistic**

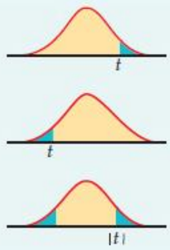
$$t = \frac{x - \mu_0}{s/\sqrt{n}}$$

In terms of a random variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

H_a: $\mu > \mu_0$ is $P(T \geq t)$

H_a: $\mu < \mu_0$ is $P(T \leq t)$

H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Significance test for watching traditional television. We want to test whether the average time that U.S. college students spend watching traditional television differs from the reported overall U.S. average of 18- to 24-year-olds at the 0.05 significance level. Specifically, we want to test

$$H_0: \mu = 18.5$$

$$H_a: \mu \neq 18.5$$

Recall that $n = 8$, $\overline{x} = 14.5$, and $s = 14.854$. The t test statistic is

$$t = \frac{x - \mu_0}{s/\sqrt{n}} = \frac{14.5 - 18.5}{14.854/\sqrt{8}} = -0.762$$

df = 7

p	0.25	0.20
t^*	0.711	0.896

This means that the sample mean $\overline{x} = 14.5$ is slightly more than 0.75 standard deviations below the null hypothesized value $\mu = 18.5$. Because the degrees of freedom are $n - 1 = 7$, this t statistic has the $t(7)$ distribution. Figure 7.3 shows that the P-value is $2P(T \geq 0.762)$, where T has the $t(7)$ distribution. From Table D, we see that $P(T \geq 0.711) = 0.25$ and $P(T \geq 0.896) = 0.20$.

Therefore, we conclude that the P-value is between $2 \times 0.20 = 0.40$ and $2 \times 0.25 = 0.50$. Software gives the exact value as $P = 0.4711$. These data are compatible with a mean of 18.5 hours per week. Under H_0, a difference this large or larger would occur about half the time simply due to chance. There is not enough evidence to reject the null hypothesis at the 0.05 level.

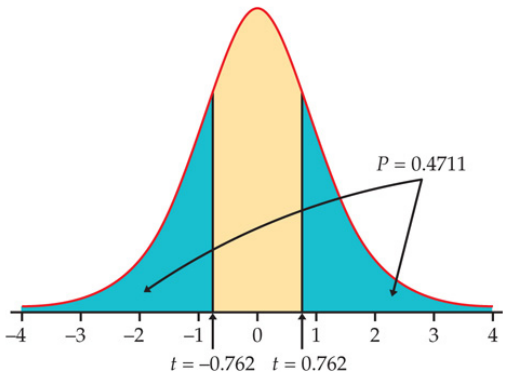


FIGURE 7.3 Sketch of the P-value calculation, Example 7.2.

In this example, we tested the null hypothesis $\mu = 18.5$ hours per week against the two-sided alternative $\mu \neq 18.5$ hours per week because we had no prior suspicion that the average among college students would be larger or smaller. If we had suspected that the average would be smaller (for example, expected more streaming of shows), we would have used a one-sided test.

One-sided test for watching traditional television. For the problem described in the previous example, we want to test whether the U.S. college student average is smaller than the overall U.S. population average. Here we test

$$H_0: \mu = 18.5$$

versus

$$H_a: \mu < 18.5$$

The t test statistic does not change: $t = -0.762$. As Figure 7.4 illustrates, however, the P-value is now $P(T \leq -0.762)$, half of the value in the previous example. From Table D, we can determine that $0.20 < P < 0.25$; software gives the exact value as $P = 0.2356$. Again, there

is not enough evidence to reject the null hypothesis in favor of the alternative at the 0.05 significance level.

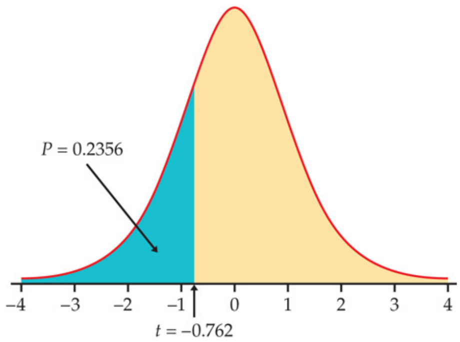


FIGURE 7.4 Sketch of the *P*-value calculation, Example 7.3.

Are the two means equivalent? Suppose the GE Healthcare researchers state that a mean difference less than 0.20 micron is not important. To see if the data support a mean difference within 0.00 ± 0.20 micron, we construct a 90% confidence interval for the mean difference.

The 51 differences have

$$\bar{x} = 0.0504 \quad \text{and} \quad s = 0.6943$$

To assess whether there is a difference between the measurements with and without this option, we test

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

Here, μ is the mean difference for the entire population of parts. The null hypothesis says that there is no difference, and H_a says that there is a difference, but does not specify a direction.

The one-sample t statistic is

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{0.0504}{0.6943/\sqrt{51}} = 0.52$$

The P-value is found from the $t(50)$ distribution. Remember that the degrees of freedom are 1 less than the sample size.

Table D shows that 0.52 lies to the left of the first column entry. This means the P-value is greater than 2(0.25) = 0.50. Software gives the exact value P = 0.6054. There is little evidence to suggest this option has an impact on the measurements. When reporting results, it is usual to omit the details of routine statistical procedures; our test would be reported in the form: "The difference in measurements was not statistically significant (t = 0.52, df = 50, P = 0.61)."

The standard error is

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.6943}{\sqrt{51}} = 0.0972$$

so the margin of error is

$$m = t^* \times \text{SE}_{\bar{x}} = (1.676)\ (0.0972) = 0.1629$$

where the critical value t^* = 1.676 comes from Table D using 50 degrees of freedom. The confidence interval is

df = 50

t^*	1.676	2.009
C	90%	95%

$$\begin{aligned} \bar{x} \pm m &= 0.0504 \pm 0.1629 \\ &= (-0.112,\ 0.2133) \end{aligned}$$

This interval is *not* entirely within the 0.00 ± 0.20 micron region that the researchers state is not important. Thus, we *cannot* conclude at the 5% signficance level that the two means are equivalent. Because the observed mean difference is close to zero and well within the

"equivalent region," the company may want to consider a larger study to improve precision.

ONE SAMPLE TEST OF EQUIVALENCE

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test, at significance level α, if μ is within a range of equivalency to μ_0, specified by the interval $\mu_0 \pm \delta$:

1. Compute the confidence interval with $C = 1 - 2\alpha$.
2. Compare this interval with the range of equivalency.

If the confidence interval falls entirely within $\mu_0 \pm \delta$, conclude that μ is equivalent to μ_0. If the confidence interval is outside the equivalency range or contains values both within and outside the range, conclude the μ is not equivalent to μ_0.

Fortunately, the t procedures are quite robust against non-Normality of the population except in the case of outliers or strong skewness. Larger samples improve the accuracy of P-values and critical values from the t distributions when the population is not Normal. This is true for two reasons:

central limit theorem, p. 298
law of large numbers, p. 250

1. The sampling distribution of the sample mean $\bar{x}$ from a large sample is close to Normal (that's the central limit theorem). Normality of the individual observations is of little concern when the sample is large.
2. As the sample size n grows, the sample standard deviation s will be an accurate estimate of σ whether or not the population has a Normal distribution. This fact is closely related to the law of large numbers.

- *Sample size less than 15:* Use t procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use t.

- *Sample size at least 15 and less than 40:* The t procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The t procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

SECTION 7.1 SUMMARY

- Significance tests and confidence intervals for the mean μ of a Normal population are based on the sample mean $\bar{x}$ of an SRS. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.
- The **standard error** of the sample mean is

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- The standardized sample mean, or **one-sample z statistic**,

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

has the $N(0, 1)$ distribution. If the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$ is replaced by the **standard error** $s/\sqrt{n}$, the **one-sample *t* statistic**

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

has the ***t* distribution** with $n - 1$ degrees of freedom.

- There is a t distribution for every positive **degrees of freedom *k***. All are symmetric distributions similar in shape to Normal distributions. The $t(k)$ distribution approaches the $N(0, 1)$ distribution as k increases.
- A level C **confidence interval for the mean** μ of a Normal population is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n - 1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**.

- Significance tests for H_0: $\mu = \mu_0$ are based on the t statistic. P-values or fixed significance levels are computed from the $t(n - 1)$ distribution.
- A matched pairs analysis is needed when subjects or experimental units are matched in pairs or when there are two measurements on each individual or

experimental unit and the question of interest concerns the difference between the two measurements.

- The one-sample procedures are used to analyze **matched pairs** data by first taking the differences within the matched pairs to produce a single sample.
- One-sample **equivalence testing** assesses whether a population mean μ is practically different from a hypothesized mean μ_0. This test requires a threshold δ, which represents the largest difference between μ and μ_0 such that the means are considered equivalent.
- The t procedures are relatively **robust** against non-Normal populations. The t procedures are useful for non-Normal data when $15 \leq n < 40$ unless the data show outliers or strong skewness. When $n \geq 40$, the t procedures can be used even for clearly skewed distributions.

7.2 Comparing Two Means

Population	Variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

We want to compare the two population means, either by giving a confidence interval for $\mu_1 - \mu_2$ or by testing the hypothesis of no difference, H_0: $\mu_1 = \mu_2$.

Inference is based on two independent SRSs, one from each population. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	n_1	$\overline{x}_1$	s_1
2	n_2	$\overline{x}_2$	s_2

Throughout this section, the subscripts 1 and 2 show the population to which a parameter or a sample statistic refers.

- The mean of the difference $\overline{x}_1 - \overline{x}_2$ is the difference between the means $\mu_1 - \mu_2$. This follows from the addition rule for means and the fact that the mean of any $\overline{x}$ is the same as the mean μ of the population.
- The variance of the difference $\overline{x}_1 - \overline{x}_2$ is the sum of their variances, which is

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

This follows from the addition rule for variances. Because the samples are independent, their sample means $\overline{x}_1$ and $\overline{x}_2$ are independent random variables.

- If the two population distributions are both Normal, then the distribution of $\overline{x}_1 - \overline{x}_2$ is also Normal. This is true because each sample mean alone is Normally distributed and because a difference between independent Normal random variables is also Normal.

Based on information from the National Health and Nutrition Examination Survey, we assume that the heights (in inches) of 10-year-old girls are $N(56.9, 2.8)$ and the heights of 10-year-old boys are $N(56.0, 3.5)$.[22] The heights

(12 girls and 8 boys)

of the students in our class are assumed to be random samples from these populations. The two distributions are shown in Figure 7.11(a).

The difference $\bar{x}_1 - \bar{x}_2$ between the female and male mean heights varies in different random samples. The sampling distribution has mean

$$\mu_1 - \mu_2 = 56.9 - 56.0 = 0.9 \text{ inches}$$

and variance

$$\begin{aligned} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} &= \frac{2.8^2}{12} + \frac{3.5^2}{8} \\ &= 2.18 \end{aligned}$$

The standard deviation of the difference in sample means is, therefore, $\sqrt{2.18} = 1.48$ inches.

If the heights vary Normally, the difference in sample means is also Normally distributed. The distribution of the difference in heights is shown in Figure 7.11(b). We standardize $\bar{x}_1 - \bar{x}_2$ by subtracting its mean (0.9) and dividing by its standard deviation (1.48). Therefore, the probability that the girls, on average, are taller than the boys is

$$\begin{aligned} P(\bar{x}_1 - \bar{x}_2 > 0) &= P\left(\frac{(\bar{x}_1 - \bar{x}_2) - 0.9}{1.48} > \frac{0 - 0.9}{1.48}\right) \\ &= P(Z > -0.61) = 0.7291 \end{aligned}$$

TWO-SAMPLE *z* STATISTIC

Suppose that $\bar{x}_1$ is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the **two-sample *z* statistic**

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

has the standard Normal $N(0, 1)$ sampling distribution.

The two-sample *t* procedures

Suppose now that the population standard deviations σ_1 and σ_2 are not known. We estimate them by the sample standard deviations s_1 and s_2 from our two samples. Following the pattern of the one-sample case, we substitute the standard errors for the standard deviations used in the two-sample *z* statistic. The result is the *two-sample t statistic:*

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately, this statistic does *not* have a *t* distribution. A *t* distribution replaces the $N(0, 1)$ distribution only when a single standard deviation (σ) in a *z* statistic is replaced by its sample standard deviation (s). In this case, we replace two standard deviations (σ_1 and σ_2) by their estimates (s_1 and s_2), which does not produce a statistic having a *t* distribution.

THE TWO-SAMPLE *t* CONFIDENCE INTERVAL

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. The **confidence interval for $\mu_1 - \mu_2$** given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least *C* no matter what the population standard deviations may be. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is the **margin of error**. Here, t^* is the value for the $t(k)$ density curve with area *C* between $-t^*$ and t^*. The value of the degrees of freedom *k* is approximated by software, or we use the smaller of $n_1 - 1$ and $n_2 - 1$. Similarly, we can use either software or the conservative approach with Table D to approximate the value of t^*.

Group	n	$\overline{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

To describe the size of the treatment effect, let's construct a confidence interval for the difference between the treatment group and the control group means. The interval is

$$(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (51.48 - 41.52)$$

$$\pm t^* \sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}$$

$$= 9.96 \pm 4.31t^*$$

The second degrees of freedom approximation uses the t (20) distribution.

21−1

df = 20

t^*	1.725	2.086	2.197
C	0.90	0.95	0.96

Table D gives $t^* = 2.086$. With this approximation, we have

$$9.96 \pm (4.31 \times 2.086) = 9.96 \pm 8.99 = (1.0, 18.9)$$

We estimate the mean improvement to be about 10 points, with a margin of error of almost 9 points. Unfortunately, the data do not allow a very precise estimate of the size of the average improvement.

THE TWO-SAMPLE t SIGNIFICANCE TEST

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. To test the hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$, compute the **two-sample t statistic**

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values for the $t(k)$ distribution, where the degrees of freedom k either are approximated by software or are the smaller of $n_1 - 1$ and $n_2 - 1$.

Is there an improvement? For the DRP study described in Example 7.11 (page 437), we hope to show that the treatment (Group 1) performs better than the control (Group 2). For a formal significance test, the hypotheses are

$$H_0: \mu_1 = \mu_2$$

Group	n	$\overline{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

$$H_a: \mu_1 > \mu_2$$

The two-sample t test statistic is

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{51.48 - 41.52}{\sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}} = 2.31$$

The P-value for the one-sided test is $P(T \geq 2.31)$. For the second approximation, the degrees of freedom k are equal to the smaller of

$$n_1 - 1 = 21 - 1 = 20 \quad \text{and} \quad n_2 - 1 = 23 - 1 = 22$$

df = 20

p	0.02	0.01
t^*	2.197	2.528

Comparing 2.31 with the entries in Table D for 20 degrees of freedom, we see that P lies between 0.01 and 0.02.

The data strongly suggest that directed reading activity improves the DRP score ($t = 2.31$, df = 20, $0.01 < P < 0.02$).

Timing of food intake and weight loss. There is emerging evidence of a relationship between timing of feeding and weight regulation. In one study, researchers followed 402 obese or overweight individuals through a 20-week weight-loss treatment.[26] To investigate the timing of food intake, participants were grouped into early eaters and late eaters, based on the timing of their main meal. Here are the summary statistics of their weight loss over the 20 weeks, in kilograms (kg):

Group	n	$\overline{x}$	s
Early eater	202	9.9	5.8
Late eater	200	7.7	6.1

The early eaters lost more weight on average. Can we conclude that these two groups are not the same? Or is this observed difference merely what we could expect to see given the variation among participants?

While other evidence suggests that early eaters should lose more weight, the researchers did not specify a direction for the difference. Thus, the hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Because the samples are large, we can confidently use the t procedures even though we lack the detailed data and so cannot verify the Normality condition.

The two-sample t statistic is

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{9.9 - 7.7}{\sqrt{\frac{5.8^2}{202} + \frac{6.1^2}{200}}} = 3.71$$

The conservative approach finds the P-value by comparing 3.71 to critical values for the $t(199)$ distribution because the smaller sample has 200 observations. Because Table D does not contain a row for 199 degrees of freedom, we will be even more conservative and use the first row in the table with degrees of freedom less than 199. This means we'll use the $t(100)$ distribution to compute the P-value.

Our calculated value of t is larger than the $p = 0.0005$ entry in the table. We must double the table tail area p because the alternative is two-sided, so we conclude that the P-value is less than 0.001. The data give conclusive evidence that early eaters lost more weight, on average, than late eaters ($t = 3.71$, df = 100, $P < 0.001$).

df = 100

p	0.0005
t^*	3.390

In this example the exact *P*-value is very small because $t = 3.71$ says that the observed difference in means is over 3.5 standard errors above the hypothesized difference of zero ($\mu_1 = \mu_2$). In this study, the researchers also compared energy intake and energy expenditure between late and early eaters. Despite the observed weight loss difference of 2.2 kg, no significant differences in these variables were found.

The pooled two-sample *t* procedures

There is one situation in which a *t* statistic for comparing two means has exactly a *t* distribution. This is when the two Normal population distributions have the *same* standard deviation. As we've done with other *t* statistics, we will first develop the *z* statistic and then, from it, the *t* statistic. In this case, notice that we need to substitute only a single standard error when we go from the *z* to the *t* statistic. This is why the resulting *t* statistic has a *t* distribution.

Call the common—and still unknown—standard deviation of both populations σ. Both sample variances s_1^2 and s_2^2 estimate σ^2. The best way to combine these two estimates is to average them with weights equal to their degrees of freedom. This gives more weight to the sample variance from the larger sample, which is reasonable. The resulting estimator of σ^2 is

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

pooled estimator of σ^2

This is called the **pooled estimator of σ^2** because it combines the information in both samples.

When both populations have variance σ^2, the addition rule for variances says that $\bar{x}_1 - \bar{x}_2$ has variance equal to the *sum* of the individual variances, which is

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

The standardized difference between means in this equal-variance case is, therefore,

$$z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

This is a special two-sample *z* statistic for the case in which the populations have the same σ. Replacing the unknown σ by the estimate s_p gives a *t* statistic. The degrees of freedom are $n_1 + n_2 - 2$, the sum of the degrees of freedom of the two sample variances. This t statistic is the basis of the pooled two-sample t inference procedures.

THE POOLED TWO-SAMPLE t PROCEDURES

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. Suppose also that the two populations have the same standard deviation. A level C **confidence interval for $\mu_1 - \mu_2$ is**

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here, t^* is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^*. The quantity

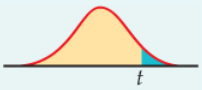
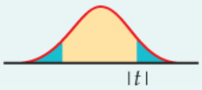
$$t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is the **margin of error.**

To test the hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$, compute the **pooled two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In terms of a random variable T having the $t(n_1 + n_2 - 2)$ distribution, the P-value for a test of H_0 against

H_a: $\mu_1 - \mu_2 > \Delta_0$	is	$P(T \geq t)$	
H_a: $\mu_1 - \mu_2 < \Delta_0$	is	$P(T \leq t)$	
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	is	$2P(T \geq \|t\|)$	

Does increased calcium reduce blood pressure? Take Group 1 to be the calcium group and Group 2 to be the placebo group. The evidence that calcium lowers blood pressure more than a placebo is assessed by testing

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Here are the summary statistics for the decrease in blood pressure:

Group	Treatment	n	$\overline{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	–0.273	5.901

The calcium group shows a drop in blood pressure, and the placebo group has a small increase. The sample standard deviations do not rule out equal population standard deviations. A difference this large will often arise by chance in samples this small. We are willing to assume equal population standard deviations. The pooled sample variance is

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

$$= \frac{(10-1)8.743^2+(11-1)5.901^2}{10+11-2} = 54.536$$

so that

$$s_p = \sqrt{54.536} = 7.385$$

The pooled two-sample t statistic is

$$t = \frac{(\overline{x}_1-\overline{x}_2)-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$= \frac{5.000-(-0.273)}{7.385\sqrt{\frac{1}{10}+\frac{1}{11}}}$$

$$= \frac{5.273}{3.227} = 1.634$$

df = 19

p	0.10	0.05
t^*	1.328	1.729

The P-value is $P(T \geq 1.634)$, where T has the $t(19)$ distribution.

From Table D, we can see that P falls between the $\alpha = 0.10$ and $\alpha = 0.05$ levels. Statistical software gives the exact value $P = 0.059$. The experiment found evidence that calcium reduces blood pressure, but the evidence falls a bit short of the traditional 5% and 1% levels.

How different are the calcium and placebo groups? We estimate that the effect of calcium supplementation is the difference between the sample means of the calcium and the placebo groups, $\overline{x}_1 - \overline{x}_2 = 5.273$ mm Hg. A 90% confidence interval for $\mu_1 - \mu_2$ uses the critical value $t^* = 1.729$ from the $t\,(19)$ distribution. The interval is

$$(\overline{x}_1 - \overline{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [5.000 - (-0.273)] \pm (1.729)$$

$$= 5.273 \pm 5.579$$

We are 90% confident that the difference in means is in the interval $(-0.306, 10.852)$. The calcium treatment reduced blood pressure by about 5.3 mm Hg more than a placebo on the average, but the margin of error for this estimate is 5.6 mm Hg.

SECTION 7.2 SUMMARY

- Significance tests and confidence intervals for the difference between the means μ_1 and μ_2 of two Normal populations are based on the difference $\overline{x}_1 - \overline{x}_2$ between the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.
- When independent SRSs of sizes n_1 and n_2 are drawn from two Normal populations with parameters μ_1, σ_1 and μ_2, σ_2 the **two-sample *z* statistic**

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the $N(0, 1)$ distribution.

- The **two-sample *t* statistic**

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does *not* have a *t* distribution. However, good approximations are available.

- **Conservative inference procedures** for comparing μ_1 and μ_2 are obtained from the two-sample *t* statistic by using the $t\,(k)$ distribution with degrees of freedom *k* equal to the smaller of $n_1 - 1$ and $n_2 - 1$.
- **More accurate probability values** can be obtained by estimating the degrees of freedom from the data. This is the usual procedure for statistical software.
- An approximate level *C* **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area *C* between $-t^*$ and t^*, where *k* is computed from the data by software or is the smaller of $n_1 - 1$ and $n_2 - 1$. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is the **margin of error**.

- Significance tests for $H_0 : \mu_1 - \mu_2 = \Delta_0$ use the **two-sample *t* statistic**

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The *P*-value is approximated using the $t\,(k)$ distribution where *k* is estimated from the data using software or is the smaller of $n_1 - 1$ and $n_2 - 1$.

- The guidelines for practical use of two-sample *t* procedures are similar to those for one-sample *t* procedures. Equal sample sizes are recommended.
- If we can assume that the two populations have equal variances, **pooled two-sample *t* procedures** can be used. These are based on the **pooled estimator**

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

of the unknown common variance and the $t\,(n_1 + n_2 - 2)$ distribution. We do not recommend this procedure for regular use.

7.3 Additional Topics on Inference

Sample size for confidence intervals We can arrange to have both high confidence and a small margin of error by choosing an appropriate sample size. Let's first focus on the one-sample t confidence interval. Its margin of error is

$$m = t^* \mathrm{SE}_{\overline{x}} = t^* \frac{s}{\sqrt{n}}$$

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR FOR A MEAN μ

The level C confidence interval for a mean μ will have an expected margin of error less than or equal to a specified value m when the sample size is such that

$$m \geq t^* s^* / \sqrt{n}$$

Here t^* is the critical value for confidence level C with $n - 1$ degrees of freedom, and s^* is the guessed value for the population standard deviation

1. Get an initial sample size by replacing t^* with z^*. Compute $n = (z^* s^*/m)^2$ and round up to the nearest integer.
2. Use this sample size to obtain t^*, and check if $m \geq t^* s^* / \sqrt{n}$.
3. If the requirement is satisfied, then this n is the needed sample size. If the requirement is not satisfied, increase n by 1 and return to Step 2.

Planning a survey of college students. In Example 7.1 (page 411), we calculated a 95% confidence interval for the mean hours per week a college student watches traditional television. The margin of error based on an SRS of $n = 8$ students was 12.42 hours. Suppose that a new study is being planned and the goal is to have a margin of error of five hours. How many students need to be sampled?

The sample standard deviation in Example 7.1 is $s = 14.854$ hours. To be conservative, we'll guess that the population standard deviation is 17.5 hours.

1. To compute an initial n, we replace t^* with z^*. This results in

$$n = \left(\frac{z^* s^*}{m}\right)^2 = \left[\frac{1.96(17.5)}{5}\right]^2 = 47.06$$

   Round up to get $n = 48$.

2. We now check to see if this sample size satisfies the requirement when we switch back to t^*. For $n = 48$, we have $n - 1 = 47$ degrees of freedom and $t^* = 2.011$. Using this value, the expected margin of error is

$$2.011\,(17.5) / \sqrt{48} = 5.08$$

   This is larger than $m = 5$, so the requirement is not satisfied.

3. The following table summarizes these calculations for some larger values of n.

n	$t^*s^*/\sqrt{n}$
49	5.03
50	4.97
51	4.92

The requirement is first satisfied when $n = 50$. Thus, we need to sample at least $n = 50$ students for the expected margin of error to be no more than five hours.

SECTION 7.3 SUMMARY

- The **sample size** required to obtain a confidence interval with an expected margin of error no larger than m for a population mean satisfies the constraint

$$m \geq t^* s^* / \sqrt{n}$$

  where t^* is the critical value for the desired level of confidence with $n - 1$ degrees of freedom, and s^* is the guessed value for the population standard deviation.

- The sample sizes necessary for a two-sample confidence interval can be obtained using a similar constraint, but guesses of both standard deviations and an estimate for the degrees of freedom are required. We suggest using the smaller of $n_1 - 1$ and $n_2 - 1$ for degrees of freedom.
- The **power** of the one-sample t test can be calculated like that of the z test, using an approximate value for both σ and s.
- The **power** of the two-sample t test is found by first finding the critical value for the significance test, the degrees of freedom, and the **noncentrality parameter** for the alternative of interest. These are used to calculate the power from a **noncentral *t* distribution**. A Normal approximation works quite well.

Calculating margins of error for various study designs and conditions is an alternative procedure for evaluating designs.

- The **sign test** is a **distribution-free test** because it uses probability calculations that are correct for a wide range of population distributions.
- The sign test for "no treatment effect" in matched pairs counts the number of positive differences. The P-value is computed from the $B(n, 1/2)$ distribution, where n is the number of non-0 differences. The sign test is less powerful than the t test in cases where use of the t test is justified.