CHAPTER 24

# Comparing Means

| | |
|---|---|
| **WHO** | AA alkaline batteries |
| **WHAT** | Length of battery life while playing a CD continuously |
| **UNITS** | Minutes |
| **WHY** | Class project |
| **WHEN** | 1998 |

**A S** *Video:* **Can Diet Prolong Life?** Watch a video that tells the story of an experiment. We'll analyze the data later in this chapter.

Should you buy generic rather than brand-name batteries? A Statistics student designed a study to test battery life. He wanted to know whether there was any real difference between brand-name batteries and a generic brand. To estimate the difference in mean lifetimes, he kept a battery-powered CD player[1] continuously playing the same CD, with the volume control fixed at 5, and measured the time until no more music was heard through the headphones. (He ran an initial trial to find out approximately how long that would take so that he didn't have to spend the first 3 hours of each run listening to the same CD.) For his trials he used six sets of AA alkaline batteries from two major battery manufacturers: a well-known brand name and a generic brand. He measured the time in minutes until the sound stopped. To account for changes in the CD player's performance over time, he randomized the run order by choosing sets of batteries at random. The table shows his data (times in minutes):

| Brand Name | Generic |
|---|---|
| 194.0 | 190.7 |
| 205.5 | 203.5 |
| 199.2 | 203.5 |
| 172.4 | 206.5 |
| 184.0 | 222.5 |
| 169.5 | 209.4 |

Studies that compare two groups are common throughout both science and industry. We might want to compare the effects of a new drug with the traditional therapy, the fuel efficiency of two car engine designs, or the sales of new products in two different test cities. In fact, battery manufacturers do research like this on their products and competitors' products themselves.

## Plot the Data

The natural display for comparing two groups is boxplots of the data for the two groups, placed side by side. Although we can't make a confidence interval

[1] Once upon a time, not so very long ago, there were no iPods. At the turn of the century, people actually carried CDs around—and devices to play them. We bet you can find one in your parents' closet.

or test a hypothesis from the boxplots themselves, you should always start with boxplots when comparing groups. Let's look at the boxplots of the battery test data.

It sure looks like the generic batteries lasted longer. And we can see that they were also more consistent. But is the difference large enough to change our battery-buying behavior? Can we be confident that the difference is more than just random fluctuation? That's why we need statistical inference.

The boxplot for the generic data identifies two possible outliers. That's interesting, but with only six measurements in each group, the outlier nomination rule is not very reliable. Both of the extreme values are plausible results, and the range of the generic values is smaller than the range of the brand-name values, even with the outliers. So we're probably better off just leaving these values in the data.

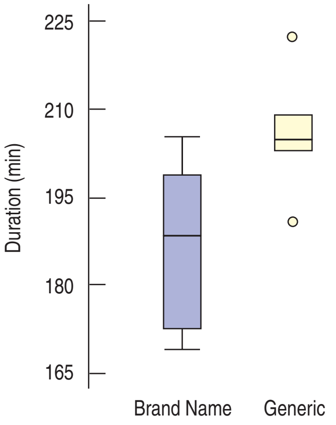


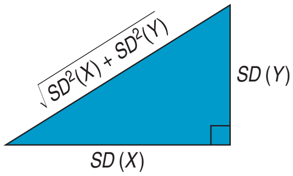
**FIGURE 24.1**

*Boxplots comparing the brand-name and generic batteries suggest a difference in duration.*

## Comparing Two Means

Comparing two means is not very different from comparing two proportions. In fact, it's not different in concept from any of the methods we've seen. Now, the population model parameter of interest is the difference between the *mean* battery lifetimes of the two brands, $\mu_1 - \mu_2$.

The rest is the same as before. The statistic of interest is the difference in the two observed means, $\bar{y}_1 - \bar{y}_2$. We'll start with this statistic to build our confidence interval, but we'll need to know its standard deviation and its sampling model. Then we can build confidence intervals and find P-values for hypothesis tests.



The Pythagorean Theorem of Statistics

We know that, for independent random variables, the variance of their *difference* is the *sum* of their individual variances, $Var(Y - X) = Var(Y) + Var(X)$. To find the standard deviation of the difference between the two independent sample means, we add their variances and then take a square root:

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{Var(\bar{y}_1) + Var(\bar{y}_2)}$$

$$= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}$$

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

Of course, we still don't know the true standard deviations of the two groups, $\sigma_1$ and $\sigma_2$, so as usual, we'll use the estimates, $s_1$ and $s_2$. Using the estimates gives us the *standard error*:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

We'll use the standard error to see how big the difference really is. Because we are working with means and estimating the standard error of their difference using the data, we shouldn't be surprised that the sampling model is a Student's *t*.

**FOR EXAMPLE** Finding the standard error of the difference in independent sample means

Can you tell how much you are eating from how full you are? Or do you need visual cues? Researchers[2] constructed a table with two ordinary 18 oz soup bowls and two identical-looking bowls that had been modified to slowly, imperceptibly, refill as they were emptied. They assigned experiment participants to the bowls randomly and served them tomato soup. Those eating from the ordinary bowls had their bowls refilled by ladle whenever they were one-quarter full. If people judge their portions by internal cues, they should eat about the same amount. How big a difference was there in the amount of soup consumed? The table summarizes their results.

| | Ordinary bowl | Refilling bowl |
|---|---|---|
| $n$ | 27 | 27 |
| $\bar{y}$ | 8.5 oz | 14.7 oz |
| s | 6.1 oz | 8.4 oz |

**Question:** How much variability do we expect in the difference between the two means? Find the standard error.

Participants were randomly assigned to bowls, so the two groups should be independent. It's okay to add variances.

$$SE(\bar{y}_{refill} - \bar{y}_{ordinary}) = \sqrt{\frac{s_r^2}{n_r} + \frac{s_o^2}{n_o}} = \sqrt{\frac{8.4^2}{27} + \frac{6.1^2}{27}} = 2.0 \text{ oz.}$$

The confidence interval we build is called a **two-sample *t*-interval** (for the difference in means). The corresponding hypothesis test is called a **two-sample *t*-test.** The interval looks just like all the others we've seen—the statistic plus or minus an estimated margin of error:

$$(\bar{y}_1 - \bar{y}_2) \pm ME$$
$$\text{where } ME = t^* \times SE(\bar{y}_1 - \bar{y}_2).$$

**z or t?**

If you know $\sigma$, use $z$.
(That's rare!)
Whenever you use $s$ to estimate $\sigma$, use $t$.

Compare this formula with the one for the confidence interval for the difference of two proportions we saw in Chapter 22 (page 505). The formulas are almost the same. It's just that here we use a Student's *t*-model instead of a Normal model to find the appropriate critical $t^*$-value corresponding to our chosen confidence level.

What are we missing? Only the degrees of freedom for the Student's *t*-model. Unfortunately, *that* formula is strange.

The deep, dark secret is that the sampling model isn't *really* Student's *t*, but only something close. The trick is that by using a special, adjusted degrees-of-freedom value, we can make it so close to a Student's *t*-model that nobody can tell the difference. The adjustment formula is straightforward but doesn't help our understanding much, so we leave it to the computer or calculator. (If you are curious and really want to see the formula, look in the footnote.[3])

[2] Brian Wansink, James E. Painter, and Jill North, "Bottomless Bowls: Why Visual Cues of Portion Size May Influence Intake," *Obesity Research*, Vol. 13, No. 1, January 2005.

[3] $$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Are you sorry you looked? This formula usually doesn't even give a whole number. If you are using a table, you'll need a whole number, so round down to be safe. If you are using technology, it's even easier. The approximation formulas that computers and calculators use for the Student's *t*-distribution deal with degrees of freedom automatically.

A SAMPLING DISTRIBUTION FOR THE DIFFERENCE BETWEEN TWO MEANS

When the conditions are met, the sampling distribution of the standardized sample difference between the means of two independent groups,

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)},$$

can be modeled by a Student's *t*-model with a number of degrees of freedom found with a special formula. We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

# Assumptions and Conditions

Now we've got everything we need. Before we can make a two-sample *t*-interval or perform a two-sample *t*-test, though, we have to check the assumptions and conditions.

## Independence Assumption

**Independence Assumption:** The data in each group must be drawn independently and at random from a homogeneous population, or generated by a randomized comparative experiment. We can't expect that the data, taken as one big group, come from a homogeneous population, because that's what we're trying to test. But without randomization of some sort, there are no sampling distribution models and no inference. We can check two conditions:

**Randomization Condition:** Were the data collected with suitable randomization? For surveys, are they a representative random sample? For experiments, was the experiment randomized?

**10% Condition:** We usually don't check this condition for differences of means. We'll check it only if we have a very small population or an extremely large sample. We needn't worry about it at all for randomized experiments.

## Normal Population Assumption

As we did before with Student's *t*-models, we should check the assumption that the underlying populations are *each* Normally distributed. We check the . . .

**Nearly Normal Condition:** We must check this for *both* groups; a violation by either one violates the condition. As we saw for single sample means, the Normality Assumption matters most when sample sizes are small. For samples of $n < 15$ in either group, you should not use these methods if the histogram or Normal probability plot shows severe skewness. For $n$'s closer to 40, a mildly skewed histogram is OK, but you should remark on any outliers you find and not work with severely skewed data. When both groups are bigger than 40, the Central Limit Theorem starts to kick in no matter how the data are distributed, so the Nearly Normal Condition for the data matters less. Even in large samples, however, you should still be on the lookout for outliers, extreme skewness, and multiple modes.

## Independent Groups Assumption

**Independent Groups Assumption:** To use the two-sample *t* methods, the two groups we are comparing must be independent of each other. In fact, this test is

***Activity:* Does Restricting Diet Prolong Life?** This activity lets you construct a confidence interval to compare life spans of rats fed two different diets.

sometimes called the two *independent samples t*-test. No statistical test can verify this assumption. You have to think about how the data were collected. The assumption would be violated, for example, if one group consisted of husbands and the other group their wives. Whatever we measure on couples might naturally be related. Similarly, if we compared subjects' performances before some treatment with their performances afterward, we'd expect a relationship of each "before" measurement with its corresponding "after" measurement. In cases such as these, where the observational units in the two groups are related or matched, *the two-sample methods of this chapter can't be applied*. When this happens, we need a different procedure that we'll see in the next chapter.

## FOR EXAMPLE Checking assumptions and conditions

**Recap:** Researchers randomly assigned people to eat soup from one of two bowls: 27 got ordinary bowls that were refilled by ladle, and 27 others bowls that secretly refilled slowly as the people ate.

**Question:** Can the researchers use their data to make inferences about the role of visual cues in determining how much people eat?

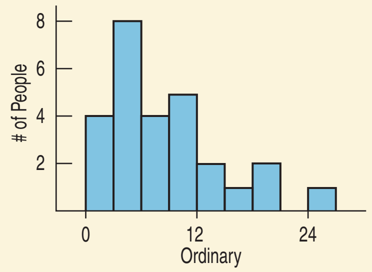
✓ **Independence Assumption:** The amount consumed by one person should be independent of the amount consumed by others.

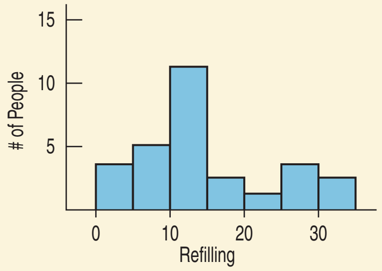
✓ **Randomization Condition:** Subjects were randomly assigned to the treatments.

✓ **Nearly Normal Condition:** The histograms for both groups look unimodal but somewhat skewed to the right. I believe both groups are large enough (27) to allow use of $t$-methods.

✓ **Independent Groups Assumption:** Randomization to treatment groups guarantees this.

It's okay to construct a two-sample $t$-interval for the difference in means.





Note: When you check the Nearly Normal Condition it's important that you include the graphs you looked at (histograms or Normal probability plots).

### An Easier Rule?

The formula for the degrees of freedom of the sampling distribution of the difference between two means is long, but the number of degrees of freedom is always at *least* the smaller of the two $n$'s, minus 1. Wouldn't it be easier to just use that value? You could, but *that* approximation can be a poor choice because it can give fewer than *half* the degrees of freedom you're entitled to from the correct formula.

### TWO-SAMPLE $t$-INTERVAL FOR THE DIFFERENCE BETWEEN MEANS

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups, $\mu_1 - \mu_2$. The confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2),$$

where the standard error of the difference of the means

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The critical value $t^*_{df}$ depends on the particular confidence level, $C$, that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

## FOR EXAMPLE Finding a confidence interval for the difference in sample means

**Recap:** Researchers studying the role of internal and visual cues in determining how much people eat conducted an experiment in which some people ate soup from bowls that secretly refilled. The results are summarized in the table.

We've already checked the assumptions and conditions, and have found the standard error for the difference in means to be $SE(\bar{y}_{refill} - \bar{y}_{ordinary}) = 2.0$ oz.

| | Ordinary bowl | Refilling bowl |
|---|---|---|
| $n$ | 27 | 27 |
| $\bar{y}$ | 8.5 oz | 14.7 oz |
| $s$ | 6.1 oz | 8.4 oz |

**Question:** What does a 95% confidence interval say about the difference in mean amounts eaten?

The observed difference in means is $\bar{y}_{refill} - \bar{y}_{ordinary} = (14.7 - 8.5) = 6.2$ oz

$$df = 47.46 \quad t^*_{47.46} = 2.011 \text{ (Table gives } t^*_{45} = 2.014.)$$

$$ME = t^* \times SE(\bar{y}_{refill} - \bar{y}_{ordinary}) = 2.011(2.0) = 4.02 \text{ oz}$$

The 95% confidence interval for $\mu_{refill} - \mu_{ordinary}$ is $6.2 \pm 4.02$, or (2.18, 10.22) oz.

I am 95% confident that people eating from a subtly refilling bowl will eat an average of between 2.18 and 10.22 more ounces of soup than those eating from an ordinary bowl.

## STEP-BY-STEP EXAMPLE A Two-Sample *t*-Interval

Judging from the boxplot, the generic batteries seem to have lasted about 20 minutes longer than the brand-name batteries. Before we change our buying habits, what should we expect to happen with the next batteries we buy?

**Question:** How much longer might the generic batteries last?

THINK

**Plan** State what we want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the difference in the means, not the individual group means.

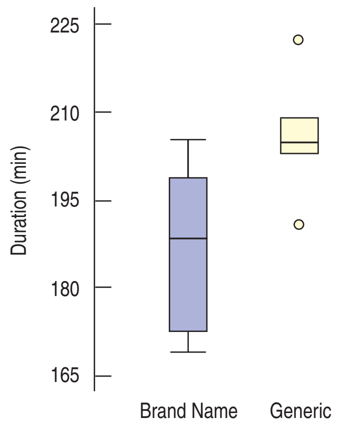
Identify the *population(s)* about which you wish to make statements. We hope to make decisions about purchasing batteries, so we're interested in all the AA batteries of these two brands.

Identify the variables and review the W's.

REALITY CHECK

From the boxplots, it appears our confidence interval should be centered near a difference of 20 minutes. We don't have a lot of intuition about how far the interval should extend on either side of 20.

I have measurements of the lifetimes (in minutes) of 6 sets of generic and 6 sets of brand-name AA batteries from a randomized experiment. I want to find an interval that is likely, with 95% confidence, to contain the true difference $\mu_G - \mu_B$ between the mean lifetime of the generic AA batteries and the mean lifetime of the brand-name batteries.



**Model** Think about the appropriate assumptions and check the conditions to be sure that a Student's *t*-model for the sampling distribution is appropriate.

For very small samples like these, we often don't worry about the 10% Condition.

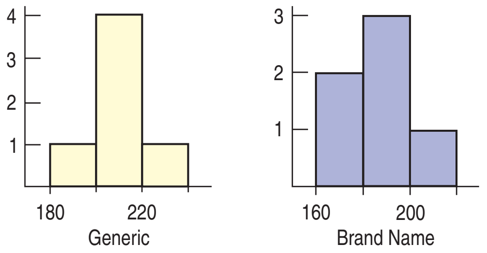
✓ **Randomization Condition:** The batteries were selected at random from those available for sale. Not exactly an SRS, but a reasonably representative random sample.

✓ **Independence Assumption:** The batteries were packaged together, so they may not be independent. For example, a storage problem might affect all the batteries in the same pack. Repeating the study for several different packs of batteries would make the conclusions stronger.

✓ **Independent Groups Assumption:** Batteries manufactured by two different companies and purchased in separate packages should be independent.

✓ **Nearly Normal Condition:** The samples are small, but the histograms look unimodal and symmetric:

Make a picture. Boxplots are the display of choice for comparing groups, but now we want to check the *shape* of distribution of each group. Histograms or Normal probability plots do a better job there.



State the sampling distribution model for the statistic. Here the degrees of freedom will come from that messy approximation formula.

Under these conditions, it's okay to use a Student's t-model.

Specify your method.

I'll use a **two-sample *t*-interval.**

SHOW

**Mechanics** Construct the confidence interval.

Be sure to include the units along with the statistics. Use meaningful subscripts to identify the groups.

I know $n_G = 6$ $\quad n_B = 6$

$$\bar{y}_G = 206.0 \text{ min} \qquad \bar{y}_B = 187.4 \text{ min}$$
$$s_G = 10.3 \text{ min} \qquad s_B = 14.6 \text{ min}$$

Use the sample standard deviations to find the standard error of the sampling distribution.

The groups are independent, so

$$SE(\bar{y}_G - \bar{y}_B) = \sqrt{SE^2(\bar{y}_G) + SE^2(\bar{y}_B)}$$
$$= \sqrt{\frac{s_G^2}{n_G} + \frac{s_B^2}{n_B}}$$
$$= \sqrt{\frac{10.3^2}{6} + \frac{14.6^2}{6}}$$

We have three choices for degrees of freedom. The best alternative is to let the computer or calculator use the approximation formula for df. This gives a fractional degree of freedom (here df = 8.98), and technology can find a corresponding critical value. In this case, it is $t^* = 2.263$.

$$= \sqrt{\frac{106.09}{6} + \frac{213.16}{6}}$$

$$= \sqrt{53.208}$$

$$= 7.29 \text{ min.}$$

Or we could round the approximation formula's df value down to an integer so we can use a $t$ table. That gives 8 df and a critical value $t^* = 2.306$.

The easy rule says to use only $6 - 1 = 5$ df. That gives a critical value $t^* = 2.571$. The corresponding confidence interval is about 14% wider—a high price to pay for a small savings in effort.

df (from technology[4]) = 8.98

The corresponding critical value for a 95% confidence level is $t^* = 2.263$.

So the margin of error is

$$\begin{aligned} ME &= t^* \times SE(\bar{y}_G - \bar{y}_B) \\ &= 2.263(7.29) \\ &= 16.50 \text{ min.} \end{aligned}$$

The 95% confidence interval is

$$(206.0 - 187.4) \pm 16.5 \text{ min.}$$

$$\text{or } 18.6 \pm 16.5 \text{ min.}$$

$$= (2.1, 35.1) \text{ min.}$$

TELL

**Conclusion** Interpret the confidence interval in the proper context.

Less formally, you could say, "I'm 95% confident that generic batteries last an average of 2.1 to 35.1 minutes longer than brand-name batteries."

I am 95% confident that the interval from 2.1 minutes to 35.1 minutes captures the mean amount of time by which generic batteries outlast brand-name batteries for this task. If generic batteries are cheaper, there seems little reason not to use them. If it is more trouble or costs more to buy them, then I'd consider whether the additional performance is worth it.

## Another One Just Like the Other Ones?

A S ***Activity:* Find Two-Sample *t*-Intervals.** Who wants to deal with that ugly df formula? We usually find these intervals with a statistics package. Learn how here.

Yes. That's been our point all along. Once again we see a statistic plus or minus the margin of error. And the ME is just a critical value times the standard error. Just look out for that crazy degrees of freedom formula.

**TI Tips**

### Creating the confidence interval

If you have been successful using your TI to make confidence intervals for proportions and 1-sample means, then you can probably already use the 2-sample function just fine. But humor us while we do one. Please?

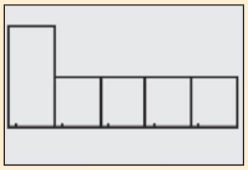
[4] If you try to find the degrees of freedom with that messy approximation formula (We dare you! It's in the footnote on page 562) using the values above, you'll get 8.99. The minor discrepancy is because we rounded the standard deviations to the nearest 10th.

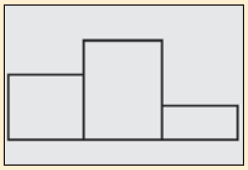
### Find a confidence interval for the difference in means, given data from two independent samples.

- Let's do the batteries. Always think about whether the samples are independent. If not, stop right here. These procedures are appropriate only for independent groups.
- Enter the data into two lists.

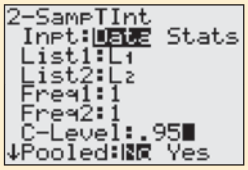
| | | | | | | |
|---|---|---|---|---|---|---|
| *NameBrand* in L1: | 194.0 | 205.5 | 199.2 | 172.4 | 184.0 | 169.5 |
| *Generic* in L2: | 190.7 | 203.5 | 203.5 | 206.5 | 222.5 | 209.4 |

- Make histograms of the data to check the Nearly Normal Condition. We see that L1's histogram doesn't look so good. But remember—this is a very small data set. The bars represent only one or two values each. It's not unusual for the histogram to look a little ragged. Try resetting the WINDOW to a range of 160 to 220 with XScl=20, and Ymax=4. Redraw the GRAPH. Looks better.

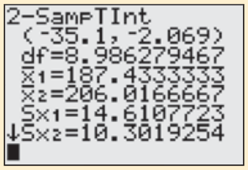




- It's your turn to try this. Check L2. Go on, do it.
- Under STAT TESTS choose 0:2-SampTInt.
- Specify that you are using the Data in L1 and L2, specify 1 for both frequencies, and choose the confidence level you want.
- Pooled? We'll discuss this issue later in the chapter, but the easy advice is: Just Say No.
- To Calculate the interval, you need to scroll down one more line.



```
2-SampTInt
 Inpt:Data Stats
 List1:L1
 List2:L2
 Freq1:1
 Freq2:1
 C-Level:.95
↓Pooled:No Yes
```



```
2-SampTInt
 (-35.1,-2.069)
 df=8.986279467
 x̄1=187.4333333
 x̄2=206.0166667
 Sx1=14.6107723
↓Sx2=10.3019254
```

Now you have the 95% confidence interval. See df? The calculator did that messy degrees of freedom calculation for you. You have to love that!

Notice that the interval bounds are negative. That's because the TI is doing $\mu_1 - \mu_2$, and the generic batteries (L2) lasted longer. No harm done—you just need to be careful to interpret that result correctly when you *Tell* what the confidence interval means.

### No data? Find a confidence interval using the sample statistics.

In many situations we don't have the original data, but must work with the summary statistics from the two groups. As we saw in the last chapter, you can still have your TI create the confidence interval with 0:2-SampTInt by choosing the Inpt:Stats option. Enter both means, standard deviations, and sample sizes, then Calculate. We show you the details in the next TI Tips.

## JUST CHECKING

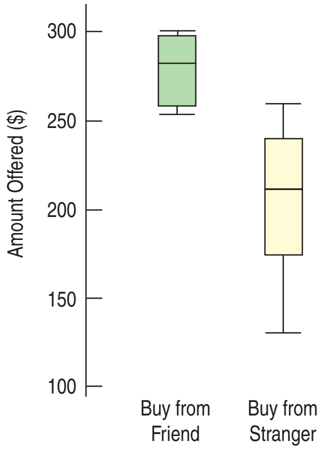
Carpal tunnel syndrome (CTS) causes pain and tingling in the hand, sometimes bad enough to keep sufferers awake at night and restrict their daily activities. Researchers studied the effectiveness of two alternative surgical treatments for CTS (Mackenzie, Hainer, and Wheatley, *Annals of Plastic Surgery*, 2000). Patients were randomly assigned to have endoscopic or open-incision surgery. Four weeks later the endoscopic surgery patients demonstrated a mean pinch strength of 9.1 kg compared to 7.6 kg for the open-incision patients.

1. Why is the randomization of the patients into the two treatments important?
2. A 95% confidence interval for the difference in mean strength is about (0.04 kg, 2.96 kg). Explain what this interval means.
3. Why might we want to examine such a confidence interval in deciding between these two surgical procedures?
4. Why might you want to see the data before trusting the confidence interval?

## Testing the Difference Between Two Means

If you bought a used camera in good condition from a friend, would you pay the same as you would if you bought the same item from a stranger? A researcher at Cornell University (J. J. Halpern, "The Transaction Index: A Method for Standardizing Comparisons of Transaction Characteristics Across Different Contexts," *Group Decision and Negotiation*, 6: 557–572) wanted to know how friendship might affect simple sales such as this. She randomly divided subjects into two groups and gave each group descriptions of items they might want to buy. One group was told to imagine buying from a friend whom they expected to see again. The other group was told to imagine buying from a stranger.

Here are the prices they offered for a used camera in good condition:



| | |
|---|---|
| **WHO** | University students |
| **WHAT** | Prices offered for a used camera |
| **UNITS** | $ |
| **WHY** | Study of the effects of friendship on transactions |
| **WHEN** | 1990s |
| **WHERE** | U.C. Berkeley |

**Price Offered for a Used Camera ($)**

| Buying from a Friend | Buying from a Stranger |
|---|---|
| 275 | 260 |
| 300 | 250 |
| 260 | 175 |
| 300 | 130 |
| 255 | 200 |
| 275 | 225 |
| 290 | 240 |
| 300 | |

The researcher who designed this study had a specific concern. Previous theories had doubted that friendship had a measurable effect on pricing. She hoped to find an effect on friendship. This calls for a hypothesis test—in this case a **two-sample *t*-test for the difference between means.**[5]

## A Test for the Difference Between Two Means

A S ***Activity:* The Two-Sample *t*-Test.** How different are beef hot dogs and chicken hot dogs? Test whether measured differences are statistically significant.

You already know enough to construct this test. The test statistic looks just like the others we've seen. It finds the difference between the observed group means and compares this with a hypothesized value for that difference. We'll call that hypothesized difference $\Delta_0$ ("delta naught"). It's so common for that hypothesized difference to be zero that we often just assume $\Delta_0 = 0$. We then compare the difference in the means with the standard error of that difference. We already know that for a difference between independent means, we can find P-values from a Student's *t*-model on that same special number of degrees of freedom.

### TWO-SAMPLE *t*-TEST FOR THE DIFFERENCE BETWEEN MEANS

The conditions for the two-sample *t*-test for the difference between the means of two independent groups are the same as for the two-sample *t*-interval. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

[5] Because it is performed so often, this test is usually just called a "two-sample *t*-test."

**NOTATION ALERT:**

$\Delta_0$—delta naught—isn't so standard that you can assume everyone will understand it. We use it because it's the Greek letter (good for a parameter) "D" for "difference." You should say "delta naught" rather than "delta zero"—that's standard for parameters associated with null hypotheses.

where the hypothesized difference is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

When the conditions are met and the null hypothesis is true, this statistic can be closely modeled by a Student's *t*-model with a number of degrees of freedom given by a special formula. We use that model to obtain a P-value.

## STEP-BY-STEP EXAMPLE A Two-Sample *t*-Test for the Difference Between Two Means

The usual null hypothesis is that there's no difference in means. That's just the right null hypothesis for the camera purchase prices.

**Question:** Is there a difference in the price people would offer a friend rather than a stranger?

THINK

**Plan** State what we want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the difference in the means, not the individual group means.

Identify the variables and check the W's.

I want to know whether people are likely to offer a different amount for a used camera when buying from a friend than when buying from a stranger. I wonder whether the difference between mean amounts is zero. I have bid prices from 8 subjects buying from a friend and 7 buying from a stranger, found in a randomized experiment.

**Hypotheses** State the null and alternative hypotheses. The research claim is that friendship changes what people are willing to pay.[6] The natural null hypothesis is that friendship makes no difference.

We didn't start with any knowledge of whether friendship might increase or decrease the price, so we choose a two-sided alternative.

$H_0$: The difference in mean price offered to friends and the mean price offered to strangers is zero:

$$\mu_F - \mu_S = 0.$$

$H_A$: The difference in mean prices is not zero:

$$\mu_F - \mu_S \neq 0.$$

**Model** Think about the assumptions and check the conditions. (Note that, because this is a randomized experiment, we haven't sampled at all, so the 10% Condition does not apply.)

✓ **Randomization Condition:** The experiment was randomized. Subjects were assigned to treatment groups at random.

✓ **Independence Assumption:** This is an experiment, so there is no need for the subjects to be randomly selected from any

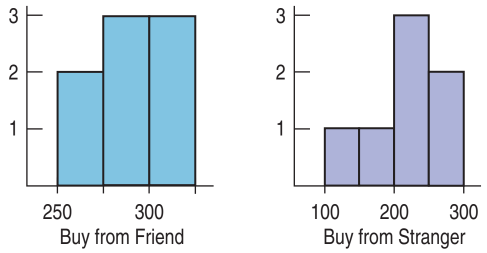
[6] This claim is a good example of what is called a "research hypothesis" in many social sciences. The only way to check it is to deny that it's true and see where the resulting null hypothesis leads us.

particular population. All we need to check is whether they were assigned randomly to treatment groups.

✓ **Independent Groups Assumption:** Randomizing the experiment gives independent groups.

✓ **Nearly Normal Condition:** Histograms of the two sets of prices are roughly unimodal and symmetric:

Make a picture. Boxplots are the display of choice for comparing groups, as seen on page 561. We also want to check the shapes of the distribution. Histograms or Normal probability plots do a better job for that.



State the sampling distribution model.

Specify your method.

The assumptions are reasonable and the conditions are okay, so I'll use a Student's *t*-model to perform a **two-sample *t*-test.**

**Mechanics** List the summary statistics. Be sure to use proper notation.

From the data:

$$n_F = 8 \qquad n_S = 7$$
$$\bar{y}_F = \$281.88 \qquad \bar{y}_S = \$211.43$$
$$s_F = \$18.31 \qquad s_S = \$46.43$$

Use the null model to find the P-value. First determine the standard error of the difference between sample means.

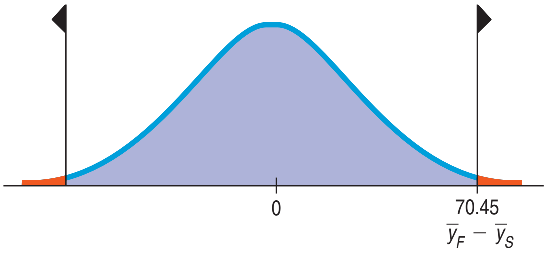
For independent groups,

$$SE(\bar{y}_F - \bar{y}_S) = \sqrt{SE^2(\bar{y}_F) + SE^2(\bar{y}_S)}$$
$$= \sqrt{\frac{s_F^2}{n_F} + \frac{s_S^2}{n_S}}$$
$$= \sqrt{\frac{18.31^2}{8} + \frac{46.43^2}{7}}$$
$$= 18.70$$

The observed difference is

$$(\bar{y}_F - \bar{y}_S) = 281.88 - 211.43 = \$70.45$$

Make a picture. Sketch the *t*-model centered at the hypothesized difference of zero. Because this is a two-tailed test, shade the region to the right of the observed difference and the corresponding region in the other tail.



Find the *t*-value.

$$t = \frac{(\bar{y}_F - \bar{y}_S) - (0)}{SE(\bar{y}_F - \bar{y}_S)} = \frac{70.45}{18.70} = 3.77$$

A statistics program or graphing calculator finds the P-value using the fractional degrees of freedom from the approximation formula.

$df = 7.62$ (from technology)

$$\text{P-value} = 2P(t_{7.62} > 3.77) = 0.006$$

TELL

**Conclusion** Link the P-value to your decision about the null hypothesis, and state the conclusion in context.

Be cautious about generalizing to items whose prices are outside the range of those in this study.

If there were no difference in the mean prices, a difference this large would occur only 6 times in 1000. That's too rare to believe, so I reject the null hypothesis and conclude that people are likely to offer a friend more than they'd offer a stranger for a used camera (and possibly for other, similar items).

TI Tips

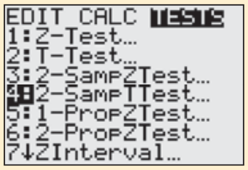
## Testing a hypothesis about a difference in means

Now let's use the TI to do a hypothesis test for the difference of two means—independent, of course! (Have we said that enough times yet?)

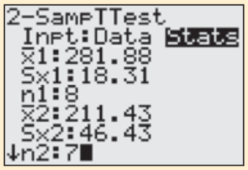
### Test a hypothesis when you know the sample statistics.

We'll demonstrate by using the statistics from the camera-pricing example. A sample of 8 people suggested they'd sell the camera to a friend for an average price of $281.88 with standard deviation $18.31. An independent sample of 7 other people would charge a stranger an average of $211.43 with standard deviation $46.43. Does this represent a significant difference in prices?

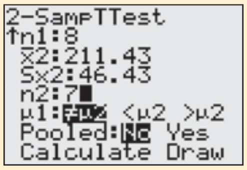
- From the `STAT TESTS` menu select `4:2-SampTTest`.
- Specify `Inpt:Stats`, and enter the appropriate sample statistics.
- You have to scroll down to complete the specifications. This is a two-tailed test, so choose alternative `≠μ2`.
- `Pooled`? Just say `No`. (We did promise to explain that and we will, coming up next.)
- Ready . . . set . . . `Calculate`!



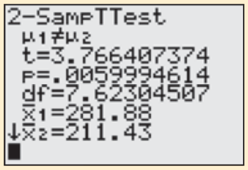
```
EDIT CALC TESTS
1:Z-Test…
2:T-Test…
3:2-SampZTest…
4:2-SampTTest…
5:1-PropZTest…
6:2-PropZTest…
7↓ZInterval…
```



```
2-SampTTest
 Inpt:Data Stats
 x̄1:281.88
 Sx1:18.31
 n1:8
 x̄2:211.43
 Sx2:46.43
↓n2:7
```



```
2-SampTTest
↑n1:8
 x̄2:211.43
 Sx2:46.43
 n2:7
 μ1:≠μ2 <μ2 >μ2
 Pooled:No Yes
 Calculate Draw
```



```
2-SampTTest
 μ1≠μ2
 t=3.766407374
 p=.0059994614
 df=7.62304507
 x̄1=281.88
↓x̄2=211.43
```

The TI reports a calculated value of $t = 3.77$ and a P-value of 0.006. It's hard to tell who your real friends are.

### By now we probably don't have to tell you how to do a `2-SampTTest` starting with data in lists.

So we won't.

## JUST CHECKING

Recall the experiment comparing patients 4 weeks after surgery for carpal tunnel syndrome. The patients who had endoscopic surgery demonstrated a mean pinch strength of 9.1 kg compared to 7.6 kg for the open-incision patients.

**5.** What hypotheses would you test?

**6.** The P-value of the test was less than 0.05. State a brief conclusion.

**7.** The study reports work on 36 "hands," but there were only 26 patients. In fact, 7 of the endoscopic surgery patients had both hands operated on, as did 3 of the open-incision group. Does this alter your thinking about any of the assumptions? Explain.

## FOR EXAMPLE A two-sample *t*-test

Many office "coffee stations" collect voluntary payments for the food consumed. Researchers at the University of Newcastle upon Tyne performed an experiment to see whether the image of eyes watching would change employee behavior.[7] They alternated pictures (seen here) of eyes looking at the viewer with pictures of flowers each week on the cupboard behind the "honesty box." They measured the consumption of milk to approximate the amount of food consumed and recorded the contributions (in £) each week per liter of milk. The table summarizes their results.

**Question:** Do these results provide evidence that there really is a difference in honesty even when it's only photographs of eyes that are "watching"?

| | Eyes | Flowers |
|---|---|---|
| *n* (# weeks) | 5 | 5 |
| $\bar{y}$ | 0.417 £/l | 0.151 £/l |
| *s* | 0.1811 £/l | 0.067 £/l |

$$H_0: \mu_{eyes} - \mu_{flowers} = 0$$
$$H_A: \mu_{eyes} - \mu_{flowers} \neq 0$$

✓ **Independence Assumption:** The amount paid by one person should be independent of the amount paid by others.

✓ **Randomization Condition:** This study was observational. Treatments alternated a week at a time and were applied to the same group of office workers.

✓ **Nearly Normal Condition:** I don't have the data to check, but it seems unlikely there would be outliers in either group. I could be more certain if I could see histograms for both groups.

✓ **Independent Groups Assumption:** The same workers were recorded each week, but week-to-week independence is plausible.

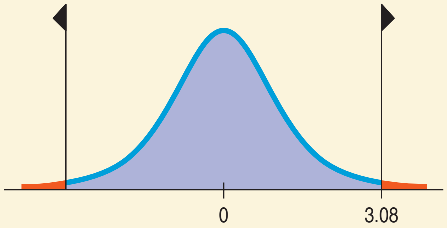
It's okay to do a two-sample *t*-test for the difference in means:

$$SE(\bar{y}_{eyes} - \bar{y}_{flowers}) = \sqrt{\frac{s^2_{eyes}}{n_{eyes}} + \frac{s^2_{flowers}}{n_{flowers}}} = \sqrt{\frac{0.1811^2}{5} + \frac{0.067^2}{5}} = 0.0864$$

$$df = 5.07$$

$$t_5 = \frac{(\bar{y}_{eyes} - \bar{y}_{flowers}) - 0}{SE(\bar{y}_{eyes} - \bar{y}_{flowers})} = \frac{0.417 - 0.151}{0.0864} = 3.08$$

$$P(|t_5| > 3.08) = 0.027$$



Assuming the data were free of outliers, the very low P-value leads me to reject the null hypothesis. This study provides evidence that people will leave higher average voluntary payments for food if pictures of eyes are "watching."

(Note: In Table T we can see that at 5 df, $t = 3.08$ lies between the critical values for $P = 0.02$ and $P = 0.05$, so we could report $P < 0.05$.)

[7] Melissa Bateson, Daniel Nettle, and Gilbert Roberts, "Cues of Being Watched Enhance Cooperation in a Real-World Setting," *Biol. Lett. doi*:10.1098/rsbl.2006.0509.

## Back into the Pool

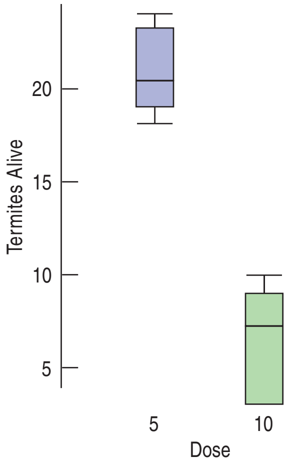
Remember that when we know a proportion, we know its standard deviation. When we tested the null hypothesis that two proportions were equal, that link meant we could assume their variances were equal as well. This led us to pool our data to estimate a standard error for the hypothesis test.

For means, there is also a pooled $t$-test. Like the two-proportions $z$-test, this test assumes that the variances in the two groups are equal. But be careful: Knowing the mean of some data doesn't tell you anything about their variance. And knowing that two means are equal doesn't say anything about whether their variances are equal. If we were willing to *assume* that their variances are equal, we could pool the data from two groups to estimate the common variance. We'd estimate this pooled variance from the data, so we'd still use a Student's $t$-model. This test is called a **pooled *t*-test (for the difference between means).**

Pooled $t$-tests have a couple of advantages. They often have a few more degrees of freedom than the corresponding two-sample test and a much simpler degrees of freedom formula. But these advantages come at a price: You have to pool the variances and think about another assumption. The assumption of equal variances is a strong one, is often not true, and is difficult to check. For these reasons, we recommend that you use a two-sample $t$-test instead.

The pooled $t$-test is the theoretically correct method only when we have a good reason to believe that the variances are equal. And (as we will see shortly) there are times when this makes sense. Keep in mind, however, that it's never wrong *not* to pool.

## *The Pooled *t*-Test



Termites cause billions of dollars of damage each year, to homes and other buildings, but some tropical trees seem to be able to resist termite attack. A researcher extracted a compound from the sap of one such tree and tested it by feeding it at two different concentrations to randomly assigned groups of 25 termites.[8] After 5 days, 8 groups fed the lower dose had an average of 20.875 termites alive, with a standard deviation of 2.23. But 6 groups fed the higher dose had an average of only 6.667 termites alive, with a standard deviation of 3.14. Is this a large enough difference to declare the sap compound effective in killing termites? In order to use the pooled $t$-test, we must make the **Equal Variance Assumption** that the variances of the two populations from which the samples have been drawn are equal. That is, $\sigma_1^2 = \sigma_2^2$. (Of course, we could think about the standard deviations being equal instead.) The corresponding **Similar Spreads Condition** really just consists of looking at the boxplots to check that the spreads are not wildly different. We were going to make boxplots anyway, so there's really nothing new here.

Once we decide to pool, we estimate the common variance by combining numbers we already have:

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

$$s_{pooled}^2 = \frac{(8 - 1)2.23^2 + (6 - 1)3.14^2}{(8 - 1) + (6 - 1)} = 7.01$$

(If the two sample sizes are equal, this is just the average of the two variances.)

Now we just substitute this pooled variance in place of each of the variances in the standard error formula.

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} = s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{7.01}{8} + \frac{7.01}{6}} = 1.43$$

[8] Adam Messer, Kevin McCormick, Sunjaya, H. H. Hagedorm, Ferny Tumbel, and J. Meinwald, "Defensive role of tropical tree resins: antitermitic sesquiterpenes from Southeast Asian Dipterocarpaceae," *J Chem Ecology*, 16:122, pp. 3333–3352.

The formula for degrees of freedom for the Student's $t$-model is simpler, too. It was so complicated for the two-sample $t$ that we stuck it in a footnote.[9] Now it's just df $= n_1 + n_2 - 2$.

$$t = \frac{20.875 - 6.667}{1.43} = 9.935$$

Substitute the pooled-$t$ estimate of the standard error and its degrees of freedom into the steps of the confidence interval or hypothesis test, and you'll be using the pooled-$t$ method. For the termites, $\bar{y}_1 - \bar{y}_2 = 14.208$, giving a $t$-value $= 9.935$ with 12 df and a P-value $\leq 0.0001$.

Of course, if you decide to use a pooled-$t$ method, you must defend your assumption that the variances of the two groups are equal.

### POOLED $t$-TEST AND CONFIDENCE INTERVAL FOR MEANS

The conditions for the pooled $t$-test for the difference between the means of two independent groups (commonly called a "pooled $t$-test") are the same as for the two-sample $t$-test with the additional assumption that the variances of the two groups are the same. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

where the hypothesized difference, $\Delta_0$, is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}} = s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where the pooled variance is

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

When the conditions are met and the null hypothesis is true, we can model this statistic's sampling distribution with a Student's $t$-model with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom. We use that model to obtain a P-value for a test or a margin of error for a confidence interval.

The corresponding confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\text{df}} \times SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2),$$

where the critical value $t^*$ depends on the confidence level and is found with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom.

***Activity:* The Pooled *t*-Test.** It's those hot dogs again. The same interactive tool can handle a pooled $t$-test, too. Take it for a spin here.

## Is the Pool All Wet?

We're testing whether the means are equal, so we admit that we don't *know* whether they are equal. Doesn't it seem a bit much to just *assume* that the variances are equal? Well, yes—but there are some special cases to consider. So when *should* you use pooled-$t$ methods rather than two-sample $t$ methods?

Never.

What, never?

Well, hardly ever.

[9] But not this one. See page 562.

You see, when the variances of the two groups are in fact equal, the two methods give pretty much the same result. (For the termites, the two-sample $t$ statistic is barely different—9.436 with 8 df—and the P-value is still $< 0.001$.) Pooled methods have a small advantage (slightly narrower confidence intervals, slightly more powerful tests) mostly because they usually have a few more degrees of freedom, but the advantage is slight.

When the variances are *not* equal, the pooled methods are just not valid and can give poor results. You have to use the two-sample methods instead.

Because the advantages of pooling are small, and you are allowed to pool only rarely (when the Equal Variances Assumption is met), ***don't.***

**It's never wrong *not* to pool.**

As the sample sizes get bigger, the advantages that come from a few more degrees of freedom make less and less difference. So the advantage (such as it is) of the pooled method is greatest when the samples are small—just when it's hardest to check the conditions. And the difference in the degrees of freedom is greatest when the variances are not equal—just when you can't use the pooled method anyway. Our advice is to use the two-sample $t$ methods to compare means.

Pooling may make sense in a randomized comparative experiment. We start by assigning our experimental units to treatments at random, as the experimenter did with the termites. We know that at the start of the experiment each treatment group is a random sample from the same population,[10] so each treatment group begins with the same population variance. In this case, assuming that the variances are equal after we apply the treatment is the same as assuming that the treatment doesn't change the variance. When we test whether the true means are equal, we may be willing to go a bit farther and say that the treatments made no difference *at all.* For example, we might suspect that the treatment is no different from the placebo offered as a control. Then it's not much of a stretch to assume that the variances have remained equal. It's still an assumption, and there are conditions that need to be checked (make the boxplots, make the boxplots, make the boxplots), but at least it's a plausible assumption.

This line of reasoning is important. The methods used to analyze comparative experiments *do* pool variances in exactly this way and defend the pooling with a version of this argument. The chapter on Analysis of Variance on the DVD introduces these methods.

## WHAT CAN GO WRONG?

- **Watch out for paired data.** The Independent Groups Assumption deserves special attention. If the samples are not independent, you can't use these two-sample methods. This is probably the main thing that can go wrong when using these two-sample methods. The methods of this chapter can be used *only* if the observations in the two groups are *independent.* Matched-pairs designs in which the observations are deliberately related arise often and are important. The next chapter deals with them.
- **Look at the plots.** The usual (by now) cautions about checking for outliers and non-Normal distributions apply, of course. The simple defense is to make and examine boxplots. You may be surprised how often this simple step saves you from the wrong or even absurd conclusions that can be generated by a single undetected outlier. You don't want to conclude that two methods have very different means just because one observation is atypical.

[10] That is, the population of experimental subjects. Remember that to be valid, experiments do not need a representative sample drawn from a population because we are not trying to estimate a population model parameter.

**Do what we say, not what we do . . .** Precision machines used in industry often have a bewildering number of parameters that have to be set, so experiments are performed in an attempt to try to find the best settings. Such was the case for a hole-punching machine used by a well-known computer manufacturer to make printed circuit boards. The data were analyzed by one of the authors, but because he was in a hurry, he didn't look at the boxplots first and just performed *t*-tests on the experimental factors. When he found extremely small P-values even for factors that made no sense, he plotted the data. Sure enough, there was one observation 1,000,000 times bigger than the others. It turns out that it had been recorded in microns (millionths of an inch), while all the rest were in inches.

## CONNECTIONS

The structure and reasoning of inference methods for comparing two means are very similar to what we used for comparing two proportions. Here we must estimate the standard errors independent of the means, so we use Student's *t*-models rather than the Normal.

We first learned about side-by-side boxplots in Chapter 5. There we made general statements about the shape, center, and spread of each group. When we compared groups, we asked whether their centers looked different compared to how spread out the distributions were. Here we've made that kind of thinking precise, with confidence intervals for the difference and tests of whether the means are the same.

We use Student's *t* as we did for single sample means, and for the same reasons: We are using standard errors from the data to estimate the standard deviation of the sample statistic. As before, to work with Student's *t*-models, we need to check the Nearly Normal Condition. Histograms and Normal probability plots are the best methods for such checks.

As always, we've decided whether a statistic is large by comparing it with its standard error. In this case, our statistic is the difference in means.

We pooled data to find a standard deviation when we tested the hypothesis of equal proportions. For that test, the assumption of equal variances was a consequence of the null hypothesis that the proportions were equal, so it didn't require an extra assumption. When two proportions are equal, so are their variances. But means don't have a linkage with their corresponding variances; so to use pooled-*t* methods, we must make the additional assumption of equal variances. When we can make this assumption, the pooled variance calculations are very similar to those for proportions, combining the squared deviations of each group from its own mean to find a common variance.

## WHAT HAVE WE LEARNED?

**Are the means of two groups the same? If not, how different are they? We've learned to use statistical inference to compare the means of two independent groups.**

- We've seen that confidence intervals and hypothesis tests about the difference between two means, like those for an individual mean, use *t*-models.
- Once again we've seen the importance of checking assumptions that tell us whether our method will work.
- We've seen that, as when comparing proportions, finding the standard error for the difference in sample means depends on believing that our data come from independent groups. Unlike proportions, however, pooling is usually not the best choice here.
- And we've seen once again that we can add variances of independent random variables to find the standard deviation of the difference in two independent means.
- Finally, we've learned that the reasoning of statistical inference remains the same; only the mechanics change.

## Terms

**Two-sample *t* methods**

562. Two-sample *t* methods allow us to draw conclusions about the difference between the means of two independent groups. The two-sample methods make relatively few assumptions about the underlying populations, so they are usually the method of choice for comparing two sample means. However, the Student's *t*-models are only approximations for their true sampling distribution. To make that approximation work well, the two-sample *t* methods have a special rule for estimating degrees of freedom.

**Two-sample *t*-interval for the difference between means**

564. A confidence interval for the difference between the means of two independent groups found as

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\text{df}} \times SE(\bar{y}_1 - \bar{y}_2)$$

where

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the number of degrees of freedom is given by a special formula (see footnote 3 on page 562).

**Two-sample *t*-test for the difference between means**

569. A hypothesis test for the difference between the means of two independent groups. It tests the null hypothesis

$$\text{H}_0\text{: } \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference, $\Delta_0$, is almost always 0, using the statistic

$$t_{\text{df}} = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)},$$

with the number of degrees of freedom given by the special formula.

**Pooling**

574. Data from two or more populations may sometimes be combined, or *pooled,* to estimate a statistic (typically a pooled variance) when we are willing to assume that the estimated value is the same in both populations. The resulting larger sample size may lead to an estimate with lower sample variance. However, pooled estimates are appropriate only when the required assumptions are true.

**Pooled-*t* methods**

575. Pooled-*t* methods provide inferences about the difference between the means of two independent populations under the assumption that both populations have the same standard deviation. When the assumption is justified, pooled-*t* methods generally produce slightly narrower confidence intervals and more powerful significance tests than two-sample *t* methods. When the assumption is not justified, they generally produce worse results—sometimes substantially worse.

We recommend that you use two-sample *t* methods instead.

## Skills

THINK

- Be able to recognize situations in which we want to do inference on the difference between the means of two independent groups.
- Know how to examine your data for violations of conditions that would make inference about the difference between two population means unwise or invalid.
- Be able to recognize when a pooled-*t* procedure might be appropriate and be able to explain why you decided to use a two-sample method anyway.

SHOW

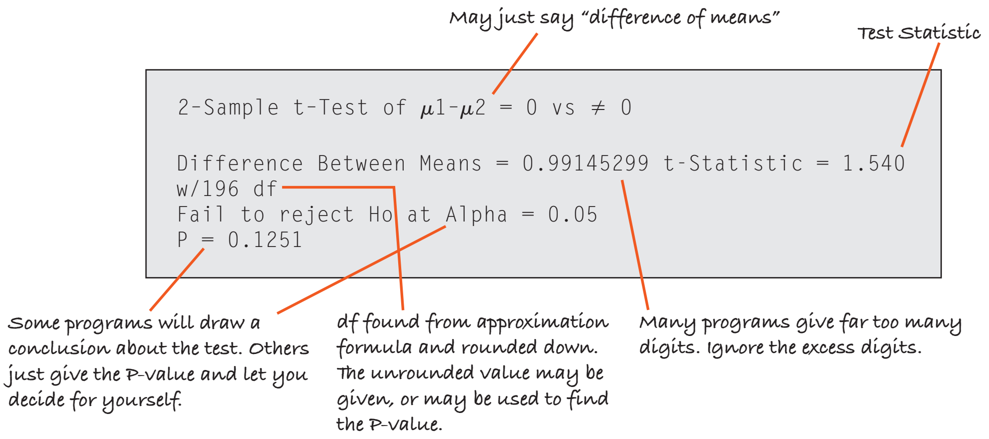
- Be able to perform a two-sample *t*-test using a statistics package or calculator (at least for finding the degrees of freedom).

TELL

- Be able to interpret a test of the null hypothesis that the means of two independent groups are equal. (If the test is a pooled *t*-test, your interpretation should include a defense of your assumption of equal variances.)

## TWO-SAMPLE METHODS ON THE COMPUTER

Here's some typical computer package output with comments:



```
2-Sample t-Test of μ1-μ2 = 0 vs ≠ 0

Difference Between Means = 0.99145299 t-Statistic = 1.540
w/196 df
Fail to reject Ho at Alpha = 0.05
P = 0.1251
```

May just say "difference of means"

Test Statistic

Some programs will draw a conclusion about the test. Others just give the P-value and let you decide for yourself.

df found from approximation formula and rounded down. The unrounded value may be given, or may be used to find the P-value.

Many programs give far too many digits. Ignore the excess digits.

Most statistics packages compute the test statistic for you and report a P-value corresponding to that statistic. And, of course, statistics packages make it easy to examine the boxplots and histograms of the two groups, so you have no excuse for skipping this important check.

Some statistics software automatically tries to test whether the variances of the two groups are equal. Some automatically offer both the two-sample-*t* and pooled-*t* results. Ignore the test for the variances; it has little power in any situation in which its results could matter. If the pooled and two-sample methods differ in any important way, you should stick with the two-sample method. Most likely, the Equal Variance Assumption needed for the pooled method has failed.

The degrees of freedom approximation usually gives a fractional value. Most packages seem to round the approximate value down to the next smallest integer (although they may actually compute the P-value with the fractional value, gaining a tiny amount of power).

## EXERCISES

**1. Dogs and calories.** In July 2007, *Consumer Reports* examined the calorie content of two kinds of hot dogs: meat (usually a mixture of pork, turkey, and chicken) and all beef. The researchers purchased samples of several different brands. The meat hot dogs averaged 111.7 calories, compared to 135.4 for the beef hot dogs. A test of the null hypothesis that there's no difference in mean calorie content yields a P-value of 0.124. Would a 95% confidence interval for $\mu_{Meat} - \mu_{Beef}$ include 0? Explain.

**2. Dogs and sodium.** The *Consumer Reports* article described in Exercise 1 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don't differ in the mean amounts of sodium yields a P-value of 0.11. Would a 95% confidence interval for $\mu_{Meat} - \mu_{Beef}$ include 0? Explain.

**3. Dogs and fat.** The *Consumer Reports* article described in Exercise 1 also listed the fat content (in grams) for samples of beef and meat hot dogs. The resulting 90% confidence interval for $\mu_{Meat} - \mu_{Beef}$ is $(-6.5, -1.4)$.

a) The endpoints of this confidence interval are negative numbers. What does that indicate?

b) What does the fact that the confidence interval does not contain 0 indicate?

c) If we use this confidence interval to test the hypothesis that $\mu_{Meat} - \mu_{Beef} = 0$, what's the corresponding alpha level?

**4. Washers.** In June 2007, *Consumer Reports* examined top-loading and front-loading washing machines, testing samples of several different brands of each type. One of the variables the article reported was "cycle time", the number of minutes it took each machine to wash a load of clothes. Among the machines rated good to excellent, the 98% confidence interval for the difference in mean cycle time $(\mu_{Top} - \mu_{Front})$ is $(-40, -22)$.

a) The endpoints of this confidence interval are negative numbers. What does that indicate?