PART VI

Learning About the World

CHAPTER 23

Inferences About Means

WHO	Vehicles on Triphammer Road
WHAT	Speed
UNITS	Miles per hour
WHEN	April 11, 2000, 1 p.m.
WHERE	A small town in the northeastern United States
WHY	Concern over impact on residential neighborhood

Motor vehicle crashes are the leading cause of death for people between 4 and 33 years old. In the year 2006, motor vehicle accidents claimed the lives of 43,300 people in the United States. This means that, on average, motor vehicle crashes resulted in 119 deaths each day, or 1 death every 12 minutes. Speeding is a contributing factor in 31% of all fatal accidents, according to the National Highway Traffic Safety Administration.

Triphammer Road is a busy street that passes through a residential neighborhood. Residents there are concerned that vehicles traveling on Triphammer often exceed the posted speed limit of 30 miles per hour. The local police sometimes place a radar speed detector by the side of the road; as a vehicle approaches, this detector displays the vehicle's speed to its driver.

The local residents are not convinced that such a passive method is helping the problem. They wish to persuade the village to add extra police patrols to encourage drivers to observe the speed limit. To help their case, a resident stood where he could see the detector and recorded the speed of vehicles passing it during a 15-minute period one day. When clusters of vehicles went by, he noted only the speed of the front vehicle. Here are his data and the histogram.

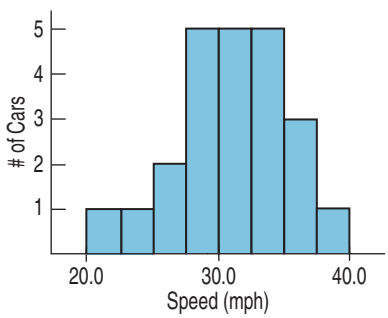


FIGURE 23.1
The speeds of cars on Triphammer Road seem to be unimodal and symmetric, at least at this scale.

Speed		
29	29	24
34	34	34
34	32	36
28	31	31
30	27	34
29	37	36
38	29	21
31	26	

We're interested both in estimating the true mean speed and in testing whether it exceeds the posted speed limit. Although the sample of vehicles is a convenience sample, not a truly random sample, there's no compelling reason to

believe that vehicles at one time of day are driving faster or slower than vehicles at another time of day,[1] so we can take the sample to be representative.

These data differ from data on proportions in one important way. Proportions are usually reported as summaries. After all, individual responses are just "success" and "failure" or "1" and "0." Quantitative data, though, usually report a value for each individual. When you have a value for each individual, you should remember the three rules of data analysis and plot the data, as we have done here.

We have quantitative data, so we summarize with means and standard deviations. Because we want to make inferences, we'll think about sampling distributions, too, and we already know most of the facts we need.

Getting Started

You've learned how to create confidence intervals and test hypotheses about proportions. We always center confidence intervals at our best guess of the unknown parameter. Then we add and subtract a margin of error. For proportions, that means $\hat{p} \pm ME$.

We found the margin of error as the product of the standard error, $SE(\hat{p})$, and a critical value, z^*, from the Normal table. So we had $\hat{p} \pm z^* SE(\hat{p})$.

We knew we could use z because the Central Limit Theorem told us (back in Chapter 18) that the sampling distribution model for proportions is Normal.

Now we want to do exactly the same thing for means, and fortunately, the Central Limit Theorem (still in Chapter 18) told us that the same Normal model works as the sampling distribution for means.

THE CENTRAL LIMIT THEOREM

When a random sample is drawn from any population with mean μ and standard deviation σ, its sample mean, $\bar{y}$, has a sampling distribution with the same *mean* μ but whose *standard deviation* is $\frac{\sigma}{\sqrt{n}}$ (and we write $\sigma(\bar{y}) = SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$).

No matter what population the random sample comes from, the *shape* of the sampling distribution is approximately Normal as long as the sample size is large enough. The larger the sample used, the more closely the Normal approximates the sampling distribution for the mean.

FOR EXAMPLE Using the CLT (as if we knew σ)

Based on weighing thousands of animals, the American Angus Association reports that mature Angus cows have a mean weight of 1309 pounds with a standard deviation of 157 pounds. This result was based on a very large sample of animals from many herds over a period of 15 years, so let's assume that these summaries are the population parameters and that the distribution of the weights was unimodal and reasonably symmetric.

Question: What does the CLT predict about the mean weight seen in random samples of 100 mature Angus cows?

(*continued*)

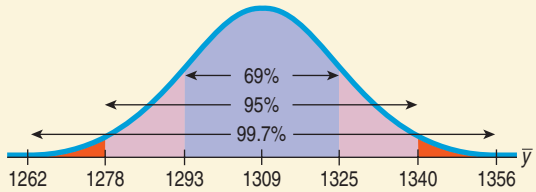
[1] Except, perhaps, at rush hour. But at that time, traffic is slowed. Our concern is with ordinary traffic during the day.

For Example (*continued*)

It's given that weights of all mature Angus cows have $\mu = 1309$ and $\sigma = 157$ pounds. Because $n = 100$ animals is a fairly large sample, I can apply the Central Limit Theorem. I expect the resulting sample means $\bar{y}$ will average 1309 pounds and have a standard deviation of $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{157}{\sqrt{100}} = 15.7$ pounds.

The CLT also says that the distribution of sample means follows a Normal model, so the 68–95–99.7 Rule applies. I'd expect that

- in 68% of random samples of 100 mature Angus cows, the mean weight will be between $1309 - 15.7 = 1293.3$ and $1309 + 15.7 = 1324.7$ pounds;
- in 95% of such samples, $1277.6 \le \bar{y} \le 1340.4$ pounds;
- in 99.7% of such samples, $1261.9 \le \bar{y} \le 1356.1$ pounds.



The CLT says that all we need to model the sampling distribution of $\bar{y}$ is a random sample of quantitative data.

And the true population standard deviation, σ.

Uh oh. That could be a problem. How are we supposed to know σ? With proportions, we had a link between the proportion value and the standard deviation of the sample proportion: $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$. And there was an obvious way to estimate the standard deviation from the data: $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$. But for means, $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$, so knowing $\bar{y}$ doesn't tell us anything about $SD(\bar{y})$. We know n, the sample size, but the population standard deviation, σ, could be *anything*. So what should we do? We do what any sensible person would do: We estimate the population parameter σ with s, the sample standard deviation based on the data. The resulting standard error is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.

Because we estimate the standard deviation of the sampling distribution model from the data, it's a *standard error*. So we use the $SE(\bar{y})$ notation. Remember, though, that it's just the estimated standard deviation of the sampling distribution model for means.

A century ago, people used this standard error with the Normal model, assuming it would work. And for large sample sizes it *did* work pretty well. But they began to notice problems with smaller samples. The sample standard deviation, s, like any other statistic, varies from sample to sample. And this extra variation in the standard error was messing up the P-values and margins of error.

William S. Gosset is the man who first investigated this fact. He realized that not only do we need to allow for the extra variation with larger margins of error and P-values, but we even need a new sampling distribution model. In fact, we need a whole *family* of models, depending on the sample size, n. These models are unimodal, symmetric, bell-shaped models, but the smaller our sample, the more we must stretch out the tails. Gosset's work transformed Statistics, but most people who use his work don't even know his name.

A S ***Activity:*** **Estimating the Standard Error.** What's the average age at which people have heart attacks? A confidence interval gives a good answer, but we must estimate the standard deviation from the data to construct the interval.

Gosset's *t*

Gosset had a job that made him the envy of many. He was the quality control engineer for the Guinness Brewery in Dublin, Ireland. His job was to make sure that the stout (a thick, dark beer) leaving the brewery was of high enough quality to meet the demands of the brewery's many discerning customers. It's easy to imagine why a large sample with many observations might be undesirable when testing stout, not to mention dangerous to one's health. So Gosset often used small

To find the sampling distribution of $\frac{\bar{y}}{s/\sqrt{n}}$, *Gosset simulated it* by hand. *He drew paper slips of small samples from a hat* hundreds of times *and computed the means and standard deviations with a mechanically cranked calculator. Today you could repeat in seconds on a computer the experiment that took him over a year. Gosset's work was so meticulous that not only did he get the shape of the new histogram approximately right, but he even figured out the exact* formula *for it from his sample. The formula was not confirmed mathematically until years later by Sir R. A. Fisher.*

samples of 3 or 4. But he noticed that with samples of this size, his tests for quality weren't quite right. He knew this because when the batches that he rejected were sent back to the laboratory for more extensive testing, too often they turned out to be OK.

Gosset checked the stout's quality by performing hypothesis tests. He knew that the test would make some Type I errors and reject about 5% of the *good* batches of stout. However, the lab told him that he was in fact rejecting about 15% of the good batches. Gosset knew something was wrong, and it bugged him.

Gosset took time off to study the problem (and earn a graduate degree in the emerging field of Statistics). He figured out that when he used the standard error, $\frac{s}{\sqrt{n}}$, as an estimate of the standard deviation, the shape of the sampling model changed. He even figured out what the new model should be and called it a t-distribution.

The Guinness Company didn't give Gosset a lot of support for his work. In fact, it had a policy against publishing results. Gosset had to convince the company that he was not publishing an industrial secret, and (as part of getting permission to publish) he had to use a pseudonym. The pseudonym he chose was "Student," and ever since, the model he found has been known as **Student's *t*.**

Gosset's model is always bell-shaped, but the details change with different sample sizes. So the Student's t-models form a whole *family* of related distributions that depend on a parameter known as **degrees of freedom.** We often denote degrees of freedom as df and the model as t_{df}, with the degrees of freedom as a subscript.

A Confidence Interval for Means

To make confidence intervals or test hypotheses for means, we need to use Gosset's model. Which one? Well, for means, it turns out the right value for degrees of freedom is $df = n - 1$.

NOTATION ALERT:

Ever since Gosset, *t* has been reserved in Statistics for his distribution.

A PRACTICAL SAMPLING DISTRIBUTION MODEL FOR MEANS

When certain assumptions and conditions[2] are met, the standardized sample mean,

$$t = \frac{\bar{y} - \mu}{SE(\bar{y})},$$

follows a Student's t-model with $n - 1$ degrees of freedom. We estimate the standard deviation with

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

When Gosset corrected the model for the extra uncertainty, the margin of error got bigger, as you might have guessed. When you use Gosset's model instead of the Normal model, your confidence intervals will be just a bit wider and your P-values just a bit larger. That's the correction you need. By using the t-model, you've compensated for the extra variability in precisely the right way.

[2] You can probably guess what they are. We'll see them in the next section.

NOTATION ALERT:

When we found critical values from a Normal model, we called them z^*. When we use a Student's t-model, we'll denote the critical values t^*.

A S ***Activity:*** **Student's *t* in Practice.** Use a statistics package to find a t-based confidence interval; that's how it's almost always done.

ONE-SAMPLE t-INTERVAL FOR THE MEAN

When the assumptions and conditions[3] are met, we are ready to find the confidence interval for the population mean, μ. The confidence interval is

$$\bar{y} \pm t^*_{n-1} \times SE(\bar{y}),$$

where the standard error of the mean is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.

The critical value t^*_{n-1} depends on the particular confidence level, C, that you specify and on the number of degrees of freedom, $n - 1$, which we get from the sample size.

FOR EXAMPLE A one-sample t-interval for the mean

In 2004, a team of researchers published a study of contaminants in farmed salmon.[4] Fish from many sources were analyzed for 14 organic contaminants. The study expressed concerns about the level of contaminants found. One of those was the insecticide mirex, which has been shown to be carcinogenic and is suspected to be toxic to the liver, kidneys, and endocrine system. One farm in particular produced salmon with very high levels of mirex. After those outliers are removed, summaries for the mirex concentrations (in parts per million) in the rest of the farmed salmon are:

$$n = 150 \qquad \bar{y} = 0.0913 \text{ ppm} \qquad s = 0.0495 \text{ ppm.}$$

Question: What does a 95% confidence interval say about mirex?

$df = 150 - 1 = 149$

$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{0.0495}{\sqrt{150}} = 0.0040$

$t^*_{149} \approx 1.977$ (from table T, using 140 df)

(actually, $t^*_{149} \approx 1.976$ from technology)

So the confidence interval for μ is $\bar{y} \pm t^*_{149} \times SE(\bar{y}) = 0.0913 \pm 1.977(0.0040)$

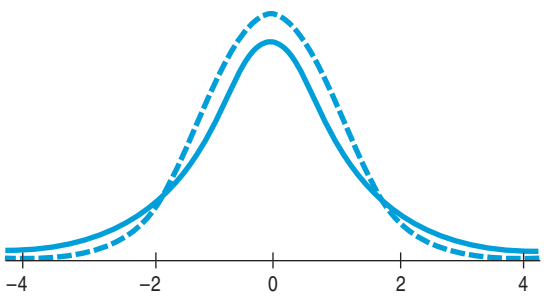
$= 0.0913 \pm 0.0079$

$= (0.0834, 0.0992)$

I'm 95% confident that the mean level of mirex concentration in farm-raised salmon is between 0.0834 and 0.0992 parts per million.

FIGURE 23.2

The t-model (solid curve) on 2 degrees of freedom has fatter tails than the Normal model (dashed curve). So the 68–95–99.7 Rule doesn't work for t-models with only a few degrees of freedom.



A S ***Activity:*** **Student's Distributions.** Interact with Gosset's family of t-models. Watch the shape of the model change as you slide the degrees of freedom up and down.

[3] Yes, the same ones, and they're still coming in the next section.

[4] Ronald A. Hites, Jeffery A. Foran, David O. Carpenter, M. Coreen Hamilton, Barbara A. Knuth, and Steven J. Schwager, "Global Assessment of Organic Contaminants in Farmed Salmon," *Science* 9 January 2004: Vol. 303., no. 5655, pp. 226–229.

TI-*n*spire

The *t*-models. See how *t*-models change as you change the degrees of freedom.

Student's *t*-models are unimodal, symmetric, and bell-shaped, just like the Normal. But *t*-models with only a few degrees of freedom have much fatter tails than the Normal. (That's what makes the margin of error bigger.) As the degrees of freedom increase, the *t*-models look more and more like the Normal. In fact, the *t*-model with infinite degrees of freedom is exactly Normal.[5] This is great news if you happen to have an infinite number of data values. Unfortunately, that's not practical. Fortunately, above a few hundred degrees of freedom it's very hard to tell the difference. Of course, in the rare situation that we *know* σ, it would be foolish not to use that information. And if we don't have to estimate σ, we can use the Normal model.

***z* or *t*?**

If you know σ, use z. (That's rare!)

Whenever you use s to estimate σ, use t.

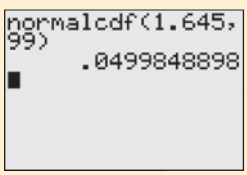
When σ is known Administrators of a hospital were concerned about the prenatal care given to mothers in their part of the city. To study this, they examined the gestation times of babies born there. They drew a sample of 25 babies born in their hospital in the previous 6 months. Human gestation times for healthy pregnancies are thought to be well-modeled by a Normal with a mean of 280 days and a standard deviation of 14 days. The hospital administrators wanted to test the mean gestation time of their sample of babies against the known standard. For this test, they should use the established value for the standard deviation, 14 days, rather than estimating the standard deviation from their sample. Because they use the model parameter value for σ, they should base their test on the Normal model rather than Student's t.

TI Tips

Finding *t*-model probabilities and critical values

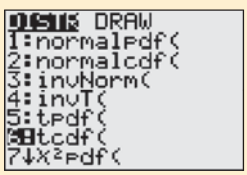
Finding Probabilities

You already know how to use your TI to find probabilities for Normal models using z-scores and `normalcdf`. What about t-models? Yes, the calculator can work with them, too.

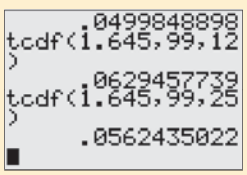


You know from your experience with confidence intervals that $z = 1.645$ cuts off the upper 5% in a Normal model. Use the TI to check that. From the `DISTR` menu, enter `normalcdf(1.645,99)`. Only 0.04998? Close enough for statisticians!

We might wonder about the probability of observing a t-value greater than 1.645, but we can't find that. There's only one Normal model, but there are many t-models, depending on the number of degrees of freedom. We need to be more specific.

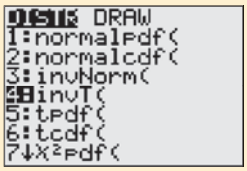


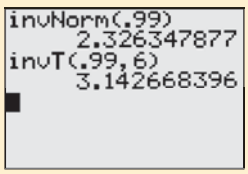
Let's find the probability of observing a t-value greater than 1.645 when there are 12 degrees of freedom. That we can do. Look in the `DISTR` menu again. See it? Yes, `tcdf`. That function works essentially like `normalcdf`, but after you enter the left and right cutoffs you must also specify the number of degrees of freedom. Try `tcdf(1.645,99,12)`.



The upper tail probability for t_{12} is 0.063, higher than the Normal model's 0.05. That should make sense to you—remember, t-models are a bit fatter in the tails, so more of the distribution lies beyond the 1.645 cutoff. (That means we'll have to go a little wider to make a 90% confidence interval.)

[5] Formally, in the limit as n goes to infinity.





Check out what happens when there are more degrees of freedom, say, 25. The command `tcdf(1.645,99,25)` yields a probability of 0.056. That's closer to 0.05, for a good reason: *t*-models look more and more like the Normal model as the number of degrees of freedom increases.

Finding Critical Values

Your calculator can also determine the critical value of *t* that cuts off a specified percentage of the distribution, using `invT`. It works just like `invNorm`, but for *t* we also have to specify the number of degrees of freedom (of course).

Suppose we have 6 degrees of freedom and want to create a 98% confidence interval. A confidence level of 98% leaves 1% in each tail of our model, so we need to find the value of *t* corresponding to the 99th percentile. If a Normal model were appropriate, we'd use $z = 2.33$. (Try it: `invNorm(.99)`). Now think. How should the critical value for *t* compare?

If you thought, "It'll be larger, because *t*-models are more spread out," you're right. Check with your TI, remembering to specify our 6 degrees of freedom: `invT(.99,6)`. Were you surprised, though, that the critical value of *t* is so much larger?

So think once more. How would the critical value of *t* differ if there were 60 degrees of freedom instead of only 6? When you think you know, check it out on your TI.

Understanding *t*

Use your calculator to play around with `tcdf` and `invT` a bit. Try to develop a clear understanding of how *t*-models compare to the more familiar Normal model. That will help you as you learn to use *t*-models to make inferences about means.

Assumptions and Conditions

Gosset found the *t*-model by simulation. Years later, when Sir Ronald A. Fisher[6] showed mathematically that Gosset was right, he needed to make some assumptions to make it work. These are the assumptions we need to use the Student's *t*-models.

Independence Assumption

Independence Assumption: The data values should be independent. There's really no way to check independence of the data by looking at the sample, but we should think about whether the assumption is reasonable.

Randomization Condition: The data arise from a random sample or suitably randomized experiment. Randomly sampled data—and especially data from a Simple Random Sample—are ideal.

When a sample is drawn without replacement, technically we ought to confirm that we haven't sampled a large fraction of the population, which would threaten the independence of our selections. We check the

10% Condition: The sample is no more than 10% of the population.

In practice, though, we often don't mention the 10% Condition for means. Why not? When we made inferences about proportions, this condition was crucial

[6] We met Fisher back in Chapter 21. You can see his picture on page 486.

We Don't *Want* to Stop

We check conditions hoping that we can make a meaningful analysis of our data. The conditions serve as *disqualifiers*—we keep going unless there's a serious problem. If we find minor issues, we note them and express caution about our results. If the sample is not an SRS, but we believe it's representative of some populations, we limit our conclusions accordingly. If there are outliers, rather than stop, we perform the analysis both with and without them. If the sample looks bimodal, we try to analyze subgroups separately. Only when there's major trouble—like a strongly skewed small sample or an obviously nonrepresentative sample—are we unable to proceed at all.

because we usually had large samples. But for means our samples are generally smaller, so the independence problem arises only if we're sampling from a small population (and then there's a correction formula we could use—but let's not get into that here). And sometimes we're dealing with a randomized experiment; then there's no sampling at all.

Normal Population Assumption

Student's t-models won't work for data that are badly skewed. How skewed is too skewed? Well, formally, we assume that the data are from a population that follows a Normal model. Practically speaking, there's no way to be certain this is true.

And it's almost certainly *not* true. Models are idealized; real data are, well, real—*never* Normal. The good news, however, is that even for small samples, it's sufficient to check the . . .

Nearly Normal Condition: The data come from a distribution that is unimodal and symmetric.

Check this condition by making a histogram or Normal probability plot. The importance of Normality for Student's t depends on the sample size. Just our luck: It matters most when it's hardest to check.[7]

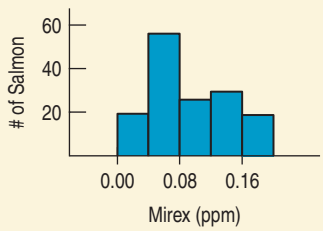
For very small samples ($n < 15$ or so), the data should follow a Normal model pretty closely. Of course, with so little data, it's rather hard to tell. But if you do find outliers or strong skewness, don't use these methods.

For moderate sample sizes (n between 15 and 40 or so), the t methods will work well as long as the data are unimodal and reasonably symmetric. Make a histogram.

When the sample size is larger than 40 or 50, the t methods are safe to use unless the data are extremely skewed. Be sure to make a histogram. If you find outliers in the data, it's always a good idea to perform the analysis twice, once with and once without the outliers, even for large samples. They may well hold additional information about the data that deserves special attention. If you find multiple modes, you may well have different groups that should be analyzed and understood separately.

FOR EXAMPLE — Checking assumptions and conditions for Student's t

Recap: Researchers purchased whole farmed salmon from 51 farms in eight regions in six countries. The histogram shows the concentrations of the insecticide mirex in 150 farmed salmon.



Question: Are the assumptions and conditions for inference satisfied?

✓ **Independence Assumption:** The fish were raised in many different places, and samples were purchased independently from several sources.

✓ **Randomization Condition:** The fish were selected randomly from those available for sale.

(*continued*)

[7] There are formal tests of Normality, but they don't really help. When we have a small sample—just when we really care about checking Normality—these tests have very little power. So it doesn't make much sense to use them in deciding whether to perform a t-test. We don't recommend that you use them.

For Example (*continued*)

- ✔ **10% Conditions:** There's lots of fish in the sea (and at the fish farms); 150 is certainly far fewer than 10% of the population.
- ✔ **Nearly Normal Condition:** The histogram of the data is unimodal. Although it may be somewhat skewed to the right, this is not a concern with a sample size of 150.

It's okay to use these data for inference about farm-raised salmon.

JUST CHECKING

Every 10 years, the United States takes a census. The census tries to count every resident. There are two forms, known as the "short form," answered by most people, and the "long form," slogged through by about one in six or seven households chosen at random. According to the Census Bureau (www.census.gov), ". . . each estimate based on the long form responses has an associated confidence interval."

1. Why does the Census Bureau need a confidence interval for long-form information but not for the questions that appear on both the long and short forms?
2. Why must the Census Bureau base these confidence intervals on *t*-models?

The Census Bureau goes on to say, "These confidence intervals are wider . . . for geographic areas with smaller populations and for characteristics that occur less frequently in the area being examined (such as the proportion of people in poverty in a middle-income neighborhood)."

3. Why is this so? For example, why should a confidence interval for the mean amount families spend monthly on housing be wider for a sparsely populated area of farms in the Midwest than for a densely populated area of an urban center? How does the formula show this will happen?

To deal with this problem, the Census Bureau reports long-form data only for ". . . geographic areas from which about two hundred or more long forms were completed—which are large enough to produce good quality estimates. If smaller weighting areas had been used, the confidence intervals around the estimates would have been significantly wider, rendering many estimates less useful . . ."

4. Suppose the Census Bureau decided to report on areas from which only 50 long forms were completed. What effect would that have on a 95% confidence interval for, say, the mean cost of housing? Specifically, which values used in the formula for the margin of error would change? Which would change a lot and which would change only slightly?
5. Approximately how much wider would that confidence interval based on 50 forms be than the one based on 200 forms?

STEP-BY-STEP EXAMPLE — A One-Sample *t*-Interval for the Mean

Let's build a 90% confidence interval for the mean speed of all vehicles traveling on Triphammer Road. The interval that we'll make is called the **one-sample *t*-interval.**

Question: What can we say about the mean speed of all cars on Triphammer Road?

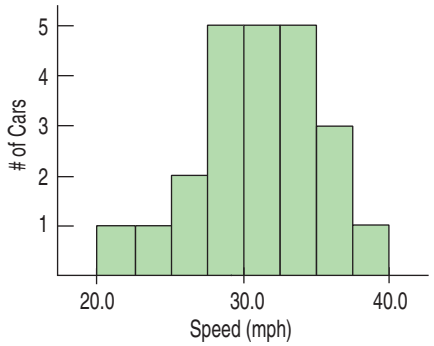
Plan State what we want to know. Identify the parameter of interest.

Identify the variables and review the W's.

I want to find a 90% confidence interval for the mean speed, μ, of vehicles driving on Triphammer Road. I have data on the speeds of 23 cars there, sampled on April 11, 2000.

Make a picture. Check the distribution shape and look for skewness, multiple modes, and outliers.

Here's a histogram of the 23 observed speeds.



REALITY CHECK

The histogram centers around 30 mph, and the data lie between 20 and 40 mph. We'd expect a confidence interval to place the population mean within a few mph of 30.

Model Think about the assumptions and check the conditions.

Note that with this small sample we probably didn't need to check the 10% Condition.

On the other hand, doing so gives us a chance to think about what the population is.

- ✓ **Independence Assumption:** This is a convenience sample, but care was taken to select cars that were not driving near each other, so their speeds are plausibly independent.
- ✓ **Randomization Condition:** Not really met. This is a convenience sample, but I have reason to believe that it is representative.
- ✓ **10% Condition:** The cars I observed were fewer than 10% of all cars that travel Triphammer Road.
- ✓ **Nearly Normal Condition:** The histogram of the speeds is unimodal and symmetric.

State the sampling distribution model for the statistic.

The conditions are satisfied, so I will use a Student's t-model with

$$(n - 1) = 22 \text{ degrees of freedom}$$

Choose your method.

and find a **one-sample t-interval for the mean.**

SHOW

Mechanics Construct the confidence interval.

Be sure to include the units along with the statistics.

Calculating from the data (see page 530):

$$n = 23 \text{ cars}$$
$$\bar{y} = 31.0 \text{ mph}$$
$$s = 4.25 \text{ mph}.$$

The standard error of $\bar{y}$ is

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{4.25}{\sqrt{23}} = 0.886 \text{ mph}.$$

The critical value we need to make a 90% interval comes from a Student's t table, a computer program, or a calculator. We have $23 - 1 = 22$ degrees of freedom. The selected confidence level says that we want 90% of the probability to be caught in the middle, so we exclude 5% in *each* tail, for a total of 10%. The degrees

The 90% critical value is $t^*_{22} = 1.717$, so the margin of error is

$$ME = t^*_{22} \times SE(\bar{y})$$
$$= 1.717(0.886)$$
$$= 1.521 \text{ mph}.$$

The 90% confidence interval for the mean speed is 31.0 ± 1.5 mph.

of freedom and 5% tail probability are all we need to know to find the critical value.

REALITY CHECK The result looks plausible and in line with what we thought.

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

When we construct confidence intervals in this way, we expect 90% of them to cover the true mean and 10% to miss the true value. That's what "90% confident" means.

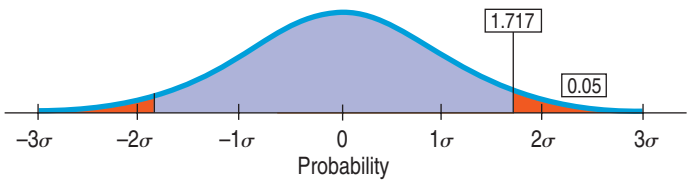
I am 90% confident that the interval from 29.5 mph to 32.5 mph contains the true mean speed of all vehicles on Triphammer Road.

Caveat: This was not a random sample of vehicles. It was a convenience sample taken at one time on one day. And the participants were not blinded. Drivers could see the police device, and some may have slowed down. I'm reluctant to extend this inference to other situations.

TI-*nspire*

Intervals for Means. Generate confidence intervals from many samples to see how often they successfully capture the true mean.

Here's the part of the Student's *t* table that gives the critical value we needed for the Step-by-Step confidence interval. (See Table T in the back of the book.) To find a critical value, locate the row of the table corresponding to the degrees of freedom and the column corresponding to the probability you want. Our 90% confidence interval leaves 5% of the values on either side, so look for 0.05 at the top of the column or 90% at the bottom. The value in the table at that intersection is the critical value we need: 1.717.



	0.25	*0.2*	*0.15*	*0.1*	*0.05*	*0.025*	*0.02*
19	.6876	.8610	1.066	1.328	1.729	2.093	2.205
20	.6870	.8600	1.064	1.325	1.725	2.086	2.197
21	.6864	.8591	1.063	1.323	1.721	2.080	2.189
22	.6858	.8583	1.061	1.321	1.717	2.074	2.183
23	.6853	.8575	1.060	1.319	1.714	2.069	2.177
24	.6848	.8569	1.059	1.318	1.711	2.064	2.172
25	.6844	.8562	1.058	1.316	1.708	2.060	2.167
26	.6840	.8557	1.058	1.315	1.706	2.056	2.162
27	.6837	.8551	1.057	1.314	1.703	2.052	2.158
C				80%	90%	95%	

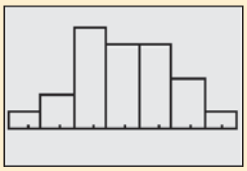
As degrees of freedom increase, the shape of Student's *t*-models changes more gradually. Table T at the back of the book includes degrees of freedom between 100 and 1000 selected so that you can pin down the P-value for just about any df. If your df's aren't listed, take the cautious approach by using the next lower value or use technology.

A S ***Activity:* Building *t*-Intervals with the *t*-Table.** Interact with an animated version of Table T.

Of course, you can also create the confidence interval with computer software or a calculator.

TI Tips

Finding a confidence interval for a mean

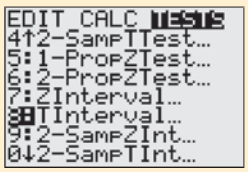


Yes, your calculator can create a confidence interval for a mean. And it's so easy we'll do two!

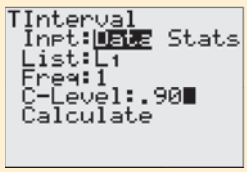
Find a confidence interval given a set of data

- Type the speeds of the 23 Triphammer cars into L1. Go ahead; we'll wait.

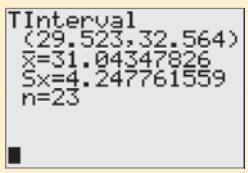
29	34	34	28	30	29	38	31	29	34	32	31
27	37	29	26	24	34	36	31	34	36	21	



- Set up a STATPLOT to create a histogram of the data so you can check the nearly Normal condition. Looks okay—unimodal and roughly symmetric.
- Under STAT TESTS choose 8:TInterval.
- Choose Inpt:Data, then specify that your data is List:L1.
- For these data the frequency is 1. (If your data have a frequency distribution stored in another list, you would specify that.)
- Choose the confidence level you want.
- Calculate the interval.



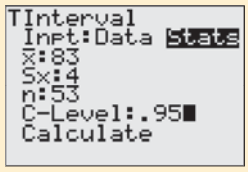
There's the 90% confidence interval. That was easy—but remember, the calculator only does the *Show*. Now you have to *Tell* what it means.

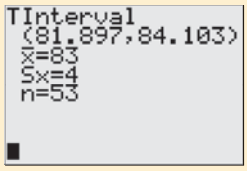


No data? Find a confidence interval given the sample's mean and standard deviation

Sometimes instead of the original data you just have the summary statistics. For instance, suppose a random sample of 53 lengths of fishing line had a mean strength of 83 pounds and standard deviation of 4 pounds. Let's make a 95% confidence interval for the mean strength of this kind of fishing line.

- Without the data you can't check the Nearly Normal Condition. But 53 is a moderately large sample, so assuming there were no outliers, it's okay to proceed. You need to say that.
- Go back to STAT TESTS and choose 8:TInterval again. This time indicate that you wish to enter the summary statistics. To do that, select Stats, then hit ENTER.
- Specify the sample mean, standard deviation, and sample size.
- Choose a confidence level and Calculate the interval.
- If (repeat, IF . . .) strengths of fishing lines follow a Normal model, we are 95% confident that this kind of line has a mean strength between 81.9 and 84.1 pounds.





More Cautions About Interpreting Confidence Intervals

Confidence intervals for means offer new tempting wrong interpretations. Here are some things you *shouldn't* say:

A S ***Activity:* Intuition for *t*-based Intervals.** A narrated review of Student's *t*.

- ***Don't say,*** "*90% of all the vehicles* on Triphammer Road drive at a speed between 29.5 and 32.5 mph." The confidence interval is about the *mean* speed, not about the speeds of *individual* vehicles.

So What *Should* We Say?

Since 90% of random samples yield an interval that captures the true mean, we *should* say, "I am 90% confident that the interval from 29.5 to 32.5 mph contains the mean speed of all the vehicles on Triphammer Road." It's also okay to say something less formal: "I am 90% confident that the average speed of all vehicles on Triphammer Road is between 29.5 and 32.5 mph." Remember: *Our uncertainty is about the interval, not the true mean.* The interval varies randomly. The true mean speed is neither variable nor random—just unknown.

- ***Don't say,*** "We are 90% confident that *a randomly selected vehicle* will have a speed between 29.5 and 32.5 mph." This false interpretation is also about individual vehicles rather than about the *mean* of the speeds. We are 90% confident that the *mean* speed of all vehicles on Triphammer Road is between 29.5 and 32.5 mph.
- ***Don't say,*** "The mean speed of the vehicles is 31.0 mph 90% *of the time*." That's about means, but still wrong. It implies that the true mean varies, when in fact it is the confidence interval that would have been different had we gotten a different sample.
- Finally, ***don't say,*** "90% *of all samples* will have mean speeds between 29.5 and 32.5 mph." That statement suggests that *this* interval somehow sets a standard for every other interval. In fact, this interval is no more (or less) likely to be correct than any other. You could say that 90% of all possible samples will produce intervals that actually do contain the true mean speed. (The problem is that, because we'll never know where the true mean speed really is, we can't know if our sample was one of those 90%.)
- ***Do say,*** "90% of intervals that could be found in this way would cover the true value." Or make it more personal and say, "I am 90% confident that the true mean speed is between 29.5 and 32.5 mph."

Make a Picture, Make a Picture, Make a Picture

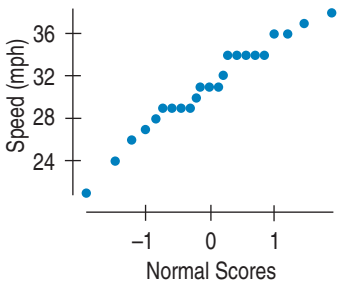


FIGURE 23.3

A Normal probability plot of speeds looks reasonably straight.

The only reasonable way to check the Nearly Normal Condition is with graphs of the data. Make a histogram of the data and verify that its distribution is unimodal and symmetric and that it has no outliers. You may also want to make a Normal probability plot to see that it's reasonably straight. You'll be able to spot deviations from the Normal model more easily with a Normal probability plot, but it's easier to understand the particular nature of the deviations from a histogram.

If you have a computer or graphing calculator doing the work, there's no excuse not to look at *both* displays as part of checking the Nearly Normal Condition.

A Test for the Mean

The residents along Triphammer Road have a more specific concern. It appears that the mean speed along the road is higher than it ought to be. To get the police to patrol more frequently, though, they'll need to show that the true mean speed is *in fact greater* than the 30 mph speed limit. This calls for a hypothesis test called the **one-sample *t*-test for the mean.**

You already know enough to construct this test. The test statistic looks just like the others we've seen. It compares the difference between the observed statistic and a hypothesized value to the standard error of the observed statistic. We already know that, for means, the appropriate probability model to use for P-values is Student's t with $n - 1$ degrees of freedom.

We're ready to go:

A S ***Activity:* A *t*-Test for Wind Speed.** Watch the video in the preceding activity, and then use the interactive tool to test whether there's enough wind for electricity generation at a site under investigation.

ONE-SAMPLE *t*-TEST FOR THE MEAN

The assumptions and conditions for the one-sample *t*-test for the mean are the same as for the one-sample *t*-interval. We test the hypothesis $H_0: \mu = \mu_0$ using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}.$$

The standard error of $\bar{y}$ is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.

When the conditions are met and the null hypothesis is true, this statistic follows a Student's *t*-model with $n - 1$ degrees of freedom. We use that model to obtain a P-value.

FOR EXAMPLE A one-sample *t*-test for the mean

Recap: Researchers tested 150 farm-raised salmon for organic contaminants. They found the mean concentration of the carcinogenic insecticide mirex to be 0.0913 parts per million, with standard deviation 0.0495 ppm. As a safety recommendation to recreational fishers, the Environmental Protection Agency's (EPA) recommended "screening value" for mirex is 0.08 ppm.

Question: Are farmed salmon contaminated beyond the level permitted by the EPA? (We've already checked the conditions; see pages 537–8.)

$$H_0: \mu = 0.08$$
$$H_A: \mu > 0.08$$

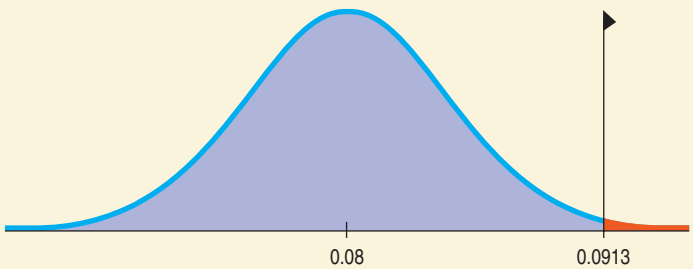
These data satisfy the conditions for inference; I'll do a one-sample *t*-test for the mean:

$n = 150, df = 149$

$\bar{y} = 0.0913, s = 0.0495$

$$SE(\bar{y}) = \frac{0.0495}{\sqrt{150}} = 0.0040$$

$$t_{149} = \frac{0.0913 - 0.08}{0.0040} = 2.825$$



$P(t_{149} > 2.825) = 0.0027$ (from technology).

With a P-value that low, I reject the null hypothesis and conclude that, in farm-raised salmon, the mirex contamination level does exceed the EPA screening value.

STEP-BY-STEP EXAMPLE A One-Sample *t*-Test for the Mean

Let's apply the one-sample *t*-test to the Triphammer Road car speeds. The speed limit is 30 mph, so we'll use that as the null hypothesis value.

Question: Does the mean speed of all cars exceed the posted speed limit?

THINK

Plan State what we want to know. Make clear what the population and parameter are.

I want to know whether the mean speed of vehicles on Triphammer Road exceeds the posted speed limit of 30 mph. I have a sample of 23 car speeds on April 11, 2000.

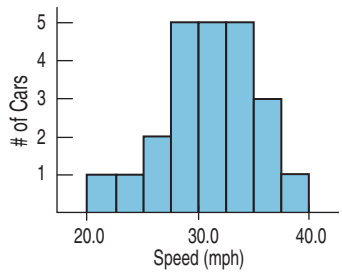
Identify the variables and review the W's.

Hypotheses The null hypothesis is that the true mean speed is equal to the limit. Because we're interested in whether the vehicles are speeding, the alternative is one-sided.

H_0: Mean speed, $\mu = 30$ mph

H_A: Mean speed, $\mu > 30$ mph

Make a picture. Check the distribution for skewness, multiple modes, and outliers.



REALITY CHECK

The histogram of the observed speeds is clustered around 30, so we'd be surprised to find that the mean was much higher than that. (The fact that 30 is within the confidence interval that we've just found confirms this suspicion.)

Model Think about the assumptions and check the conditions.

- ✔ **Independence Assumption:** These cars are a convenience sample, but they were selected so no two cars were driving near each other, so I am justified in believing that their speeds are independent.
- ✔ **Randomization Condition:** Although I have a convenience sample, I have reason to believe that it is a representative sample.
- ✔ **Nearly Normal Condition:** The histogram of the speeds is unimodal and reasonably symmetric.

(We won't worry about the 10% Condition—it's a small sample.)

State the sampling distribution model. (Be sure to include the degrees of freedom.)

The conditions are satisfied, so I'll use a Student's t-model with $(n - 1) = 22$ degrees of freedom to do a **one-sample t-test for the mean.**

Choose your method.

SHOW

Mechanics Be sure to include the units when you write down what you know from the data.

From the data,

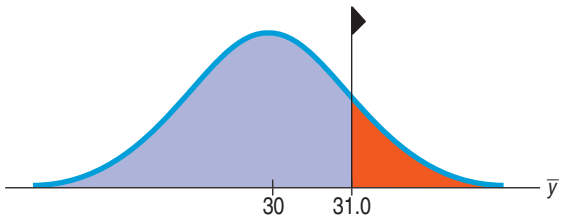
$$n = 23 \text{ cars}$$
$$\bar{y} = 31.0 \text{ mph}$$
$$s = 4.25 \text{ mph}$$

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{4.25}{\sqrt{23}} = 0.886 \text{ mph}.$$

We use the null model to find the P-value. Make a picture of the *t*-model centered at $\mu = 30$. Since this is an upper-tail test, shade the region to the right of the observed mean speed.

The *t*-statistic calculation is just a standardized value, like *z*. We subtract the hypothesized mean and divide by the standard error.

The P-value is the probability of observing a sample mean as large as 31.0 (or larger) *if* the true mean were 30.0, as the null hypothesis states. We can find this P-value from a table, calculator, or computer program.



$$t = \frac{\bar{y} - \mu_0}{SE(\bar{y})} = \frac{31.0 - 30.0}{0.886} = 1.13$$

REALITY CHECK We're not surprised that the difference isn't statistically significant.

(The observed mean is 1.13 standard errors above the hypothesized value.)

$$P\text{-value} = P(t_{22} > 1.13) = 0.136$$

TELL

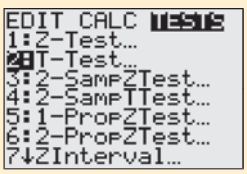
Conclusion Link the P-value to your decision about H_0, and state your conclusion in context.

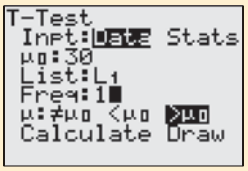
Unfortunately for the residents, there is no course of action associated with failing to reject this particular null hypothesis.

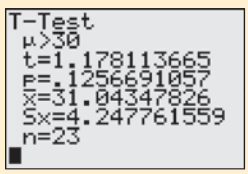
The P-value of 0.136 says that if the true mean speed of vehicles on Triphammer Road were 30 mph, samples of 23 vehicles can be expected to have an observed mean of at least 31.0 mph 13.6% of the time. That P-value is not small enough for me to reject the hypothesis that the true mean is 30 mph at any reasonable alpha level. I conclude that there is not enough evidence to say the average speed is too high.

TI Tips

Testing a hypothesis about a mean







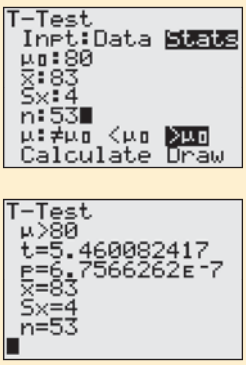
Testing a Hypothesis Given a Set of Data

Still have the Triphammer Road auto speeds in L1? Good. Let's use the TI to see if the mean is significantly higher than 30 mph (you've already checked the histogram to verify the nearly Normal condition, of course).

- Go to the STAT TESTS menu, and choose 2:T-Test.
- Tell it you want to use the stored Data.
- Enter the mean of the null model, and indicate where the data are.
- Since this is an upper tail test, choose the $>\mu_0$ option.
- Calculate.

There's everything you need to know: the summary statistics, the calculated value of *t*, and the P-value of 0.126. (*t* and P differ slightly from the values in our worked example because when we did it by hand we rounded off the mean and standard deviation. No harm done.)

As always, the *Tell* is up to you.



Testing a Hypothesis Given the Sample's Mean and Standard Deviation

Don't have the actual data? Just summary statistics? No problem, assuming you can verify the necessary conditions. In the last TI Tips we created a confidence interval for the strength of fishing line. We had test results for a random sample of 53 lengths of line showing a mean strength of 83 pounds and a standard deviation of 4 pounds. Is there evidence that this kind of fishing line exceeds the "80-lb test" as labeled on the package?

We bet you know what to do even without our help. Try it before you read on.

- Go back to `2:T-Test`.
- You're entering `Stats` this time.
- Specify the hypothesized mean and the sample statistics.
- Choose the alternative being tested (upper tail here).
- `Calculate`.

The results of the calculator's mechanics show a large t and a really small P-value (0.0000007). We have very strong evidence that the mean breaking strength of this kind of fishing line is over the 80 pounds claimed by the manufacturer.

Significance and Importance

Recall that "statistically significant" does not mean "actually important" or "meaningful," even though it sort of sounds that way. In this example, it does seem that speeds may be a bit above 30 miles per hour. If so, it's possible that a larger sample would show statistical significance.

But would that be the right decision? The difference between 31 miles per hour and 30 miles per hour doesn't seem meaningful, and rejecting the null hypothesis wouldn't change that. Even with a statistically significant result, it would be hard to convince the police that vehicles on Triphammer Road were driving at dangerously fast speeds. It would probably also be difficult to persuade the town that spending more money to lower the average speed on Triphammer Road would be a good use of the town's resources. Looking at the confidence interval, we can say with 90% confidence that the mean speed is somewhere between 29.5 and 32.5 mph. Even in the worst case, if the mean speed is 32.5 mph, would this be a bad enough situation to convince the town to spend more money? Probably not. It's always a good idea when we test a hypothesis to also check the confidence interval and think about the likely values for the mean.

JUST CHECKING

In discussing estimates based on the long-form samples, the Census Bureau notes, "The disadvantage . . . is that . . . estimates of characteristics that are also reported on the short form will not match the [long-form estimates]."

The short-form estimates are values from a complete census, so they are the "true" values—something we don't usually have when we do inference.

6. Suppose we use long-form data to make 95% confidence intervals for the mean age of residents for each of 100 of the Census-defined areas. How many of these 100 intervals should we expect will fail to include the true mean age (as determined from the complete short-form Census data)?

7. Based only on the long-form sample, we might test the null hypothesis about the mean household income in a region. Would the power of the test increase or decrease if we used an area with more long forms?

Intervals and Tests

The 90% confidence interval for the mean speed was 31.0 mph ± 1.5, or (29.5 mph, 32.5 mph). If someone hypothesized that the mean speed was really 30 mph, how would you feel about it? How about 35 mph?

Because the confidence interval included the speed limit of 30 mph, it certainly looked like 30 mph might be a plausible value for the true mean speed of the vehicles on Triphammer Road. In fact, 30 mph gave a P-value of 0.136—too large to reject the null hypothesis. We should have seen this coming. The hypothesized mean of 30 mph lies *within the confidence interval*. It's one of the reasonable values for the mean.

Confidence intervals and significance tests are built from the same calculations. In fact, they are really complementary ways of looking at the same question. Here's the connection: The confidence interval contains all the null hypothesis values we can't reject with these data.

More precisely, a level C confidence interval contains *all* of the plausible null hypothesis values that would *not* be rejected by a two-sided hypothesis test at alpha level $1 - C$. So a 95% confidence interval matches a $1 - 0.95 = 0.05$ level two-sided test for these data.

Confidence intervals are naturally two-sided, so they match exactly with two-sided hypothesis tests. When, as in our example, the hypothesis is one-sided, the corresponding alpha level is $(1 - C)/2$.

Fail to reject Our 90% confidence interval was 29.5 to 32.5 mph. If any of these values had been the null hypothesis for the mean, then the corresponding hypothesis test at $\alpha = 0.05$ (because $\frac{1 - 0.90}{2} = 0.05$) would not have been able to reject the null. That is, the corresponding one-sided P-value for our observed mean of 31 mph would be greater than 0.05. So, we would not reject any hypothesized value between 29.5 and 32.5 mph.

Sample Size

A S ***Activity:* The Real Effect of Small Sample Size.** We know that smaller sample sizes lead to wider confidence intervals, but is that just because they have fewer degrees of freedom?

How large a sample do we need? The simple answer is "more." But more data cost money, effort, and time, so how much is enough? Suppose your computer just took an hour to download a movie you wanted to watch. You're not happy. You hear about a program that claims to download movies in under a half hour. You're interested enough to spend $29.95 for it, but only if it really delivers. So you get the free evaluation copy and test it by downloading that movie 5 different times. Of course, the mean download time is not exactly 30 minutes as claimed. Observations vary. If the margin of error were 8 minutes, though, you'd probably be able to decide whether the software is worth the money. Doubling the sample size would require another 5 hours of testing and would reduce your margin of error to a bit under 6 minutes. You'll need to decide whether that's worth the effort.

As we make plans to collect data, we should have some idea of how small a margin of error we need to be able to draw a conclusion or detect a difference we want to see. If the size of the effect we're studying is large, then we may be able to tolerate a larger *ME*. If we need great precision, however, we'll want a smaller *ME*, and, of course, that means a larger sample size.

Armed with the *ME* and confidence level, we can find the sample size we'll need. Almost.

We know that for a mean, $ME = t^*_{n-1} \times SE(\bar{y})$ and that $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, so we can determine the sample size by solving this equation for n:

$$ME = t^*_{n-1}\frac{s}{\sqrt{n}}.$$

The good news is that we have an equation; the bad news is that we won't know most of the values we need to solve it. When we thought about sample size for proportions back in Chapter 19, we ran into a similar problem. There we had to guess a working value for p to compute a sample size. Here, we need to know s. We don't know s until we get some data, but we want to calculate the sample size *before* collecting the data. We might be able to make a good guess, and that is often good enough for this purpose. If we have no idea what the standard deviation might be, or if the sample size really matters (for example, because each additional individual is very expensive to sample or experiment on), it might be a good idea to run a small *pilot study* to get some feeling for the standard deviation.

That's not all. Without knowing n, we don't know the degrees of freedom and we can't find the critical value, t^*_{n-1}. One common approach is to use the corresponding z^* value from the Normal model. If you've chosen a 95% confidence level, then just use 2, following the 68–95–99.7 Rule. If your estimated sample size is, say, 60 or more, it's probably okay—z^* was a good guess. If it's smaller than that, you may want to add a step, using z^* at first, finding n, and then replacing z^* with the corresponding t^*_{n-1} and calculating the sample size once more.

Sample size calculations are *never* exact. The margin of error you find *after* collecting the data won't match exactly the one you used to find n. The sample size formula depends on quantities that you won't have until you collect the data, but using it is an important first step. Before you collect data, it's always a good idea to know whether the sample size is large enough to give you a good chance of being able to tell you what you want to know.

FOR EXAMPLE Finding sample size

A company claims its program will allow your computer to download movies quickly. We'll test the free evaluation copy by downloading a movie several times, hoping to estimate the mean download time with a margin of error of only 8 minutes. We think the standard deviation of download times is about 10 minutes.

Question: How many trial downloads must we run if we want 95% confidence in our estimate with a margin of error of only 8 minutes?

Using $z^* = 1.96$, solve

$$8 = 1.96\frac{10}{\sqrt{n}}$$

$$\sqrt{n} = \frac{1.96 \times 10}{8} = 2.45$$

$$n = (2.45)^2 = 6.0025$$

That's a small sample size, so I'll use $(6 - 1) = 5$ degrees of freedom[8] to substitute an appropriate t^* value. At 95%, $t^*_5 = 2.571$. Solving the equation one more time:

$$8 = 2.571\frac{10}{\sqrt{n}}$$

[8] Ordinarily we'd round the sample size *up*. But at this stage of the calculation, rounding *down* is the safer choice. Can you see why?

$$\sqrt{n} = \frac{2.571 \times 10}{8} \approx 3.214$$

$$n = (3.214)^2 \approx 10.33$$

To make sure the ME is no larger, I'll round up, which gives $n = 11$ runs. So, to get an ME of 8 minutes, I'll find the downloading times for 11 movies.

Degrees of Freedom

Some calculators offer an alternative button for standard deviation that divides by n instead of $n - 1$. Why don't you stick a wad of gum over the "n" button so you won't be tempted to use it? Use $n - 1$.

The number of degrees of freedom, $(n - 1)$, might have reminded you of the value we divide by to find the standard deviation of the data (since, in fact, it's the same number). When we introduced that formula, we promised to say a bit more about why we divide by $n - 1$ rather than by n. The reason is closely tied to the reasoning behind the t-distribution.

If only we knew the true population mean, μ, we would find the sample standard deviation as

$$s = \sqrt{\frac{\sum (y - \mu)^2}{n}} \qquad \text{(Equation 23.1)}^9$$

We use $\bar{y}$ instead of μ, though, and that causes a problem. For any sample, the data values will generally be closer to their own sample mean than to the true population mean, μ. Why is that? Imagine that we take a random sample of 10 high school seniors. The mean SAT verbal score is 500 in the United States. But the sample mean, $\bar{y}$, for *these* 10 seniors won't be exactly 500. Are the 10 seniors' scores closer to 500 or $\bar{y}$? They'll always be closer to their own average $\bar{y}$. If we used $\sum (y - \bar{y})^2$ instead of $\sum (y - \mu)^2$ in Equation 23.1 to calculate s, our standard deviation estimate would be too small. How can we fix it? The amazing mathematical fact is that we an compensate for the smaller sum exactly by dividing by $n - 1$ instead of by n. So that's all the $n - 1$ is doing in the denominator of s. And we call $n - 1$ the degrees of freedom.

WHAT CAN GO WRONG?

The most fundamental issue you face is knowing when to use Student's t methods.

- **Don't confuse proportions and means.** When you treat your data as categorical, counting successes and summarizing with a sample proportion, make inferences using the Normal model methods you learned about in Chapters 19 through 22. When you treat your data as quantitative, summarizing with a sample mean, make your inferences using Student's t methods.

Student's t methods work only when the Normality Assumption is true. Naturally, many of the ways things can go wrong turn out to be different ways that the Normality

(*continued*)

[9] Statistics textbooks usually have equation numbers so they can talk about equations by name. We haven't needed equation numbers yet, but we admit it's useful here, so this is our first.

Assumption can fail. It's always a good idea to look for the most common kinds of failure. It turns out that you can even fix some of them.

- **Beware of multimodality.** The Nearly Normal Condition clearly fails if a histogram of the data has two or more modes. When you see this, look for the possibility that your data come from two groups. If so, your best bet is to try to separate the data into different groups. (Use the variables to help distinguish the modes, if possible. For example, if the modes seem to be composed mostly of men in one and women in the other, split the data according to sex.) Then you could analyze each group separately.
- **Beware of skewed data.** Make a Normal probability plot and a histogram of the data. If the data are very skewed, you might try re-expressing the variable. Re-expressing may yield a distribution that is unimodal and symmetric, more appropriate for Student's *t* inference methods for means. Re-expression cannot help if the sample distribution is not unimodal. Some people may object to re-expressing the data, but unless your sample is very large, you just can't use the methods of this chapter on skewed data.
- **Set outliers aside.** Student's *t* methods are built on the mean and standard deviation, so we should beware of outliers when using them. When you make a histogram to check the Nearly Normal Condition, be sure to check for outliers as well. If you find some, consider doing the analysis twice, both with the outliers excluded and with them included in the data, to get a sense of how much they affect the results.

  The suggestion that you can perform an analysis with outliers removed may be controversial in some disciplines. Setting aside outliers is seen by some as "cheating." But an analysis of data with outliers left in place is *always* wrong. The outliers violate the Nearly Normal Condition and also the implicit assumption of a homogeneous population, so they invalidate inference procedures. An analysis of the non-outlying points, along with a separate discussion of the outliers, is often much more informative and can reveal important aspects of the data.

  How can you tell whether there are outliers in your data? The "outlier nomination rule" of boxplots can offer some guidance, but it's just a rule of thumb and not an absolute definition. The best practical definition is that a value is an outlier if removing it substantially changes your conclusions about the data. You won't want a single value to determine your understanding of the world unless you are very, very sure that it is absolutely correct. Of course, when the outliers affect your conclusion, this can lead to the uncomfortable state of not really knowing what to conclude. Such situations call for you to use your knowledge of the real world and your understanding of the data you are working with.[10]

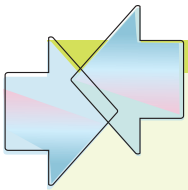
  Of course, Normality issues aren't the only risks you face when doing inferences about means. Remember to *Think* about the usual suspects.
- **Watch out for bias.** Measurements of all kinds can be biased. If your observations differ from the true mean in a systematic way, your confidence interval may not capture the true mean. And there is no sample size that will save you. A bathroom scale that's 5 pounds off will be 5 pounds off even if you weigh yourself 100 times and take the average. We've seen several sources of bias in surveys, and measurements can be biased, too. Be sure to think about possible sources of bias in your measurements.
- **Make sure cases are independent.** Student's *t* methods also require the sampled values to be mutually independent. We check for random sampling and the 10% Condition. You should also think hard about whether there are likely violations of independence in the data collection method. If there are, be very cautious about using these methods.
- **Make sure that data are from an appropriately randomized sample.** Ideally, all data that we analyze are drawn from a simple random sample or generated by a randomized experiment. When they're not, be careful about making inferences from them. You

As tempting as it is to get rid of annoying values, you can't just throw away outliers and not discuss them. It isn't appropriate to lop off the highest or lowest values just to improve your results.

[10] An important reason for you to know Statistics rather than let someone else analyze your data.

may still compute a confidence interval correctly, or get the mechanics of the P-value right, but this might not save you from making a serious mistake in inference.

- **Interpret your confidence interval correctly.** Many statements that sound tempting are, in fact, misinterpretations of a confidence interval for a mean. You might want to have another look at some of the common mistakes, explained on pages 541–2. Keep in mind that a confidence interval is about the mean of the population, not about the means of samples, individuals in samples, or individuals in the population.



CONNECTIONS

The steps for finding a confidence interval or hypothesis test for means are just like the corresponding steps for proportions. Even the form of the calculations is similar. As the z-statistic did for proportions, the t-statistic tells us how many standard errors our sample mean is from the hypothesized mean. For means, though, we have to estimate the standard error separately. This added uncertainty changes the model for the sampling distribution from z to t.

As with all of our inference methods, the randomization applied in drawing a random sample or in randomizing a comparative experiment is what generates the sampling distribution. Randomization is what makes inference in this way possible at all.

The new concept of degrees of freedom connects back to the denominator of the sample standard deviation calculation, as shown earlier.

There's just no escaping histograms and Normal probability plots. The Nearly Normal Condition required to use Student's t can be checked best by making appropriate displays of the data. Back when we first used histograms, we looked at their shape and, in particular, checked whether they were unimodal and symmetric, and whether they showed any outliers. Those are just the features we check for here. The Normal probability plot zeros in on the Normal model a little more precisely.

WHAT HAVE WE LEARNED?

We first learned to create confidence intervals and test hypotheses about proportions. Now we've turned our attention to means, and learned that statistical inference for means relies on the same concepts; only the mechanics and our model have changed.

- We've learned that what we can say about a population mean is inferred from data, using the mean of a representative random sample.
- We've learned to describe the sampling distribution of sample means using a new model we select from the Student's t family based on our degrees of freedom.
- We've learned that our ruler for measuring the variability in sample means is the standard error $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.
- We've learned to find the margin of error for a confidence interval using that ruler and critical values based on a Student's t-model.
- And we've also learned to use that ruler to test hypotheses about the population mean.

Above all, we've learned that the reasoning of inference, the need to verify that the appropriate assumptions are met, and the proper interpretation of confidence intervals and P-values all remain the same regardless of whether we are investigating means or proportions.

Terms

Student's t Degrees of freedom (df)

533. A family of distributions indexed by its degrees of freedom. The t-models are unimodal symmetric, and bell shaped, but generally have fatter tails and a narrower center than the Normal model. As the degrees of freedom increase, t-distributions approach the Normal.

One-sample t-interval for the mean

534. A one-sample t-interval for the population mean is

$$\bar{y} \pm t^*_{n-1} \times SE(\bar{y}), \text{ where } SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

The critical value t^*_{n-1} depends on the particular confidence level, C, that you specify and on the number of degrees of freedom, $n - 1$.

One-sample t-test for the mean

543. The one-sample t-test for the mean tests the hypothesis $H_0: \mu = \mu_0$ using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}.$$

The standard error of $\bar{y}$ is

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

Skills

- Know the assumptions required for t-tests and t-based confidence intervals.
- Know how to examine your data for violations of conditions that would make inference about the population mean unwise or invalid.
- Understand that a confidence interval and a hypothesis test are essentially equivalent. You can do a two-tailed hypothesis test at level of significance α with a $1 - \alpha$ confidence interval, or a one-tailed test with a $1 - 2\alpha$ confidence interval.

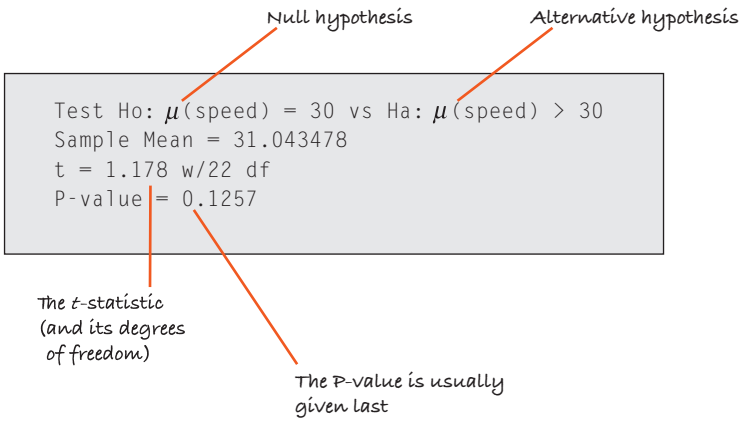
- Be able to compute and interpret a t-test for the population mean using a statistics package or working from summary statistics for a sample.
- Be able to compute and interpret a t-based confidence interval for the population mean using a statistics package or working from summary statistics for a sample.

- Be able to explain the meaning of a confidence interval for a population mean. Make clear that the randomness associated with the confidence level is a statement about the interval bounds and not about the population parameter value.
- Understand that a 95% confidence interval does not trap 95% of the sample values.
- Be able to interpret the result of a test of a hypothesis about a population mean.
- Know that we do not "accept" a null hypothesis if we cannot reject it. We say that we fail to reject it.
- Understand that the P-value of a test does not give the probability that the null hypothesis is correct.

INFERENCE FOR MEANS ON THE COMPUTER

Statistics packages offer convenient ways to make histograms of the data. Even better for assessing near-Normality is a Normal probability plot. When you work on a computer, there is simply no excuse for skipping the step of plotting the data to check that it is nearly Normal. *Beware:* Statistics packages don't agree on whether to place the Normal scores on the x-axis (as we have done) or the y-axis. Read the axis labels.

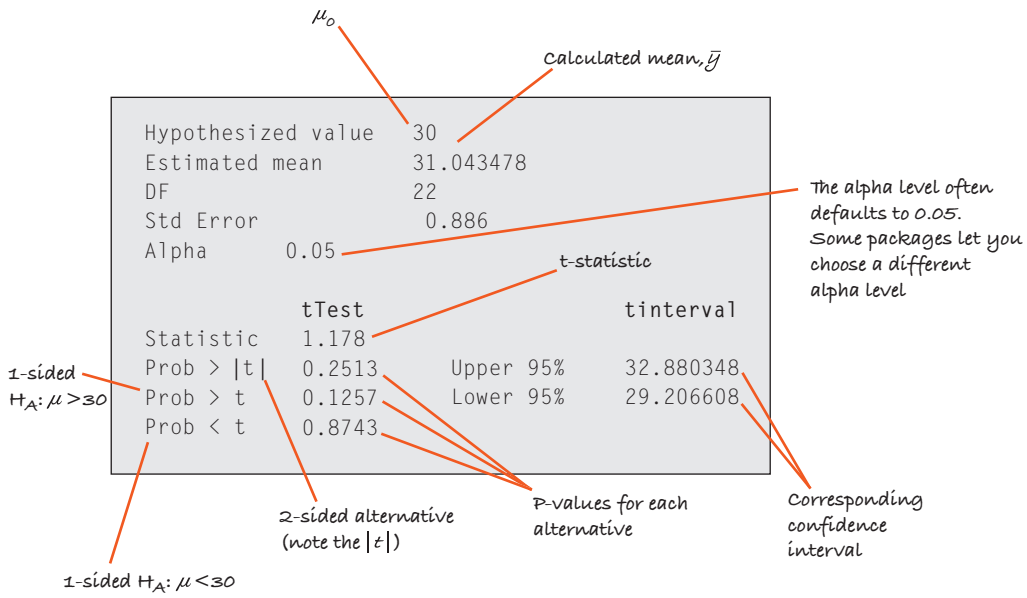
Any standard statistics package can compute a hypothesis test. Here's what the package output might look like in general (although no package we know gives the results in exactly this form):[11]



A S ***Activity:*** **Student's *t* in Practice.** We almost always use technology to do inference with Student's *t*. Here's a chance to do that as you investigate several questions.

The package computes the sample mean and sample standard deviation of the variable and finds the P-value from the *t*-distribution based on the appropriate number of degrees of freedom. All modern statistics packages report P-values. The package may also provide additional information such as the sample mean, sample standard deviation, *t*-statistic value, and degrees of freedom. These are useful for interpreting the resulting P-value and telling the difference between a meaningful result and one that is merely statistically significant. Statistics packages that report the estimated standard deviation of the sampling distribution usually label it "standard error" or "SE."

Inference results are also sometimes reported in a table. You may have to read carefully to find the values you need. Often, test results and the corresponding confidence interval bounds are given together. And often you must read carefully to find the alternative hypotheses. Here's an example of that kind of output:



The commands to do inference for means on common statistics programs and calculators are not always obvious. (By contrast, the resulting output is usually clearly labeled and easy to read.) The guides for each program can help you start navigating.

[11] Many statistics packages keep as many as 16 digits for all intermediate calculations. If we had kept as many, our results in the Step-By-Step section would have been closer to these.