Chapter 9: Testing a Claim

Section 9.3

Tests About a Population Mean

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Chapter 9
Testing a Claim

- 9.1 Significance Tests: The Basics
- 9.2 Tests about a Population Proportion
- **9.3 Tests about a Population Mean**

+

Section 9.3
Tests About a Population Mean

Learning Objectives

After this section, you should be able to…

- ✓ CHECK conditions for carrying out a test about a population mean.
- ✓ CONDUCT a one-sample *t* test about a population mean.
- ✓ CONSTRUCT a confidence interval to draw a conclusion for a two-sided test about a population mean.
- ✓ PERFORM significance tests for paired data.

■ Introduction

Confidence intervals and significance tests for a population proportion *p* are based on *z*-values from the standard Normal distribution.

Inference about a population mean μ uses a *t* distribution with *n* - 1 degrees of freedom, except in the rare case when the population standard deviation σ is known.

We learned how to construct confidence intervals for a population mean in Section 8.3. Now we'll examine the details of testing a claim about an unknown parameter μ.

Carrying Out a Significance Test for *μ*

In an earlier example, a company claimed to have developed a new AAA battery that lasts longer than its regular AAA batteries. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. An SRS of 15 new batteries lasted an average of 33.9 hours with a standard deviation of 9.8 hours. Do these data give *convincing evidence* that the new batteries last longer on average?

To find out, we must perform a significance test of

$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

where μ = the true mean lifetime of the new deluxe AAA batteries.

Check Conditions:

Three conditions should be met before we perform inference for an unknown population mean: Random, Normal, and Independent. The Normal condition for means is

Population distribution is Normal or sample size is large ($n \geq 30$)

We often don't know whether the population distribution is Normal. But if the sample size is large ($n \geq 30$), we can safely carry out a significance test (due to the central limit theorem). If the sample size is small, we should examine the sample data for any obvious departures from Normality, such as skewness and outliers.

Alternate Example – Less music?

A classic rock radio station claims to play an average of 50 minutes of music every hour. However, every time you turn to this station it seems like there is a commercial playing. To investigate their claim, you randomly select 12 different hours during the next week and record what the radio station plays in each of the 12 hours. Here are the number of minutes of music in each of these hours:

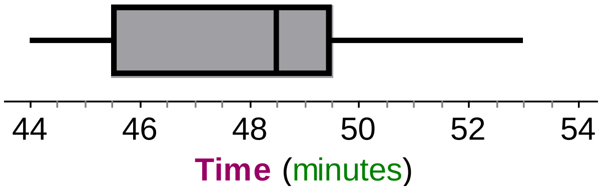
44	49	45	51	49	53	49	44	47	50	46	48

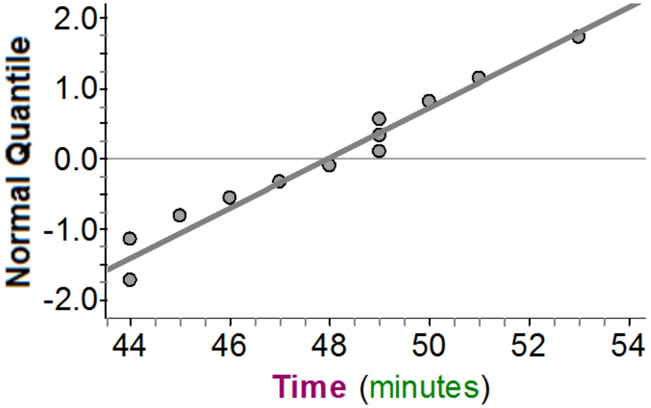
Problem: Check the conditions for carrying out a significance test of the company's claim that it plays an average of at least 50 minutes of music per hour.

Solution:

✓Random: A random sample of hours was selected.

✓Normal: We don't know if the population distribution of music times is approximately Normal and we don't have a large sample size, so we will graph the data and look for any departures from Normality.





The dotplot and boxplot are roughly symmetric with no outliers, and the Normal probability plot is close to linear, so it is reasonable to use *t* procedures for these data.

✓Independent: There are more than 10(12) = 120 hours of music played during the week.

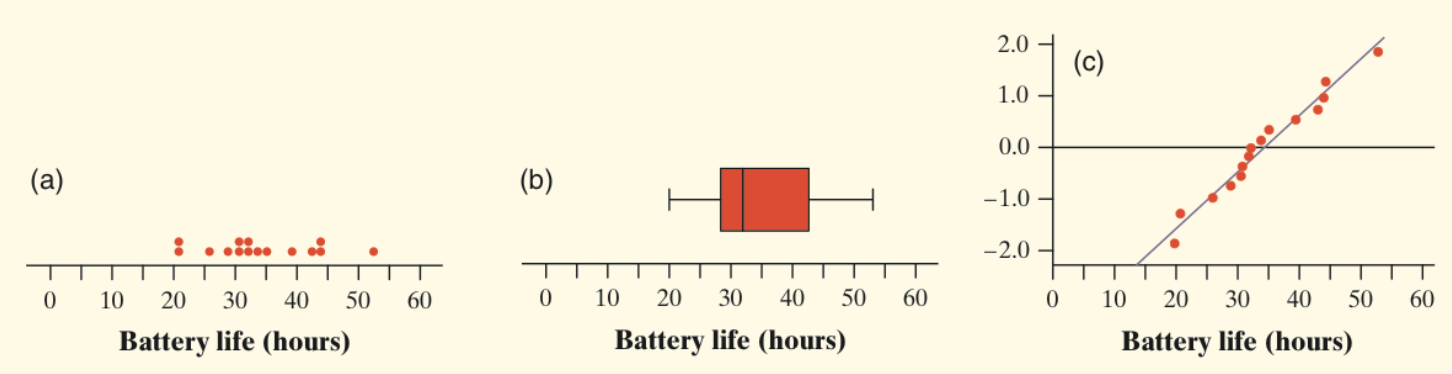
Carrying Out a Significance Test for *μ*

Check Conditions:

Three conditions should be met before we perform inference for an unknown population mean: Random, Normal, and Independent.

✓ ***Random*** The company tests an SRS of 15 new AAA batteries.

✓***Normal*** We don’t know if the population distribution of battery lifetimes for the company’s new AAA batteries is Normal. With such a small sample size (*n* = 15), we need to inspect the data for any departures from Normality.



The dotplot and boxplot show slight right-skewness but no outliers. The Normal probability plot is close to linear. We should be safe performing a test about the population mean lifetime *μ*.

✓***Independent*** Since the batteries are being sampled without replacement, we need to check the *10% condition*: there must be at least 10(15) = 150 new AAA batteries. This seems reasonable to believe.

Carrying Out a Significance Test

Calculations: Test statistic and P-value

When performing a significance test, we do calculations assuming that the null hypothesis H_0 is true. The test statistic measures how far the sample result diverges from the parameter value specified by H_0, in standardized units. As before,

$$\text{test statistic} = \frac{\text{statistic - parameter}}{\text{standard deviation of statistic}}$$

For a test of H_0: $\mu = \mu_0$, our statistic is the sample mean. Its standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Because the population standard deviation σ is usually unknown, we use the sample standard deviation s_x in its place. The resulting test statistic has the standard error of the sample mean in the denominator

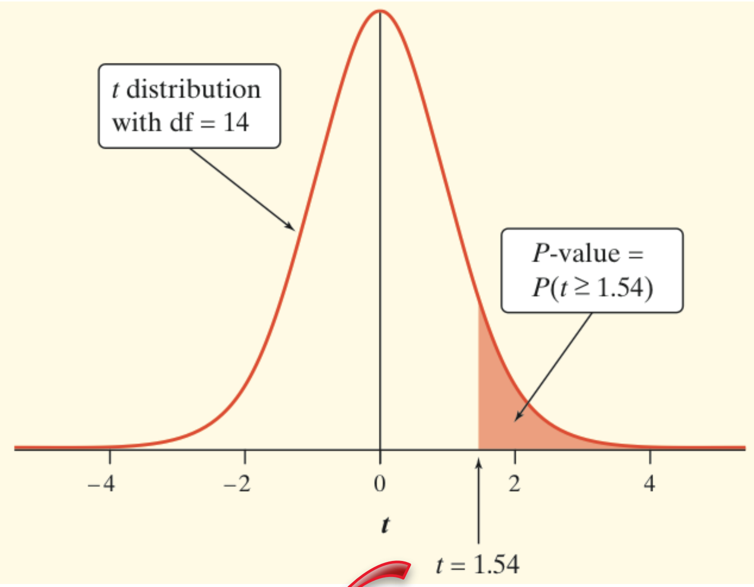
$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

When the Normal condition is met, this statistic has a *t* distribution with *n* - 1 degrees of freedom.

Carrying Out a Hypothesis Test

Tests About a Population Mean

The battery company wants to test H_0: $\mu = 30$ versus H_a: $\mu > 30$ based on an SRS of 15 new AAA batteries with mean lifetime and standard deviation $\bar{x} = 33.9$ hours and $s_x = 9.8$ hours.



$$\text{test statistic} = \frac{\text{statistic - parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \frac{33.9 - 30}{9.8 / \sqrt{15}} = 1.54$$

The *P*-value is the probability of getting a result this large or larger in the direction indicated by H_a, that is, $P(t \geq 1.54)$.

	Upper-tail probability *p*		
df	**.10**	**.05**	**.025**
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	3.131
	80%	90%	95%
	Confidence level *C*		

✓ Go to the *df* = 14 row.

✓ Since the *t* statistic falls between the values 1.345 and 1.761, the "Upper-tail probability *p*" is between 0.10 and 0.05.

✓ The *P*-value for this test is between 0.05 and 0.10.

Because the *P*-value exceeds our default α = 0.05 significance level, we can't conclude that the company's new AAA batteries last longer than 30 hours, on average.

Alternate Example – Less music?

- In the “Less music?” Alternate Example, you wanted to test H_0:µ = 50 versus
  H_a:µ < 50.

- **Problem:** Compute the test statistic and *P*-value for these data.

- **Solution:** The sample mean for these data is $\overline{x}$ = 47.9 with a standard deviation of S_X = 2.81. Thus, the test statistic is

$$t = \frac{47.9 - 50}{2.81 / \sqrt{12}} = -2.59$$

- Using Table B and a *t* distribution with 12 – 1 = 11 degrees of freedom, the *P*-value is *P*(*t* < –2.59). This probability is the same as *P*(*t* > 2.59) so the *P*-value is between 0.01 and 0.02.

Using Table B Wisely

- Table B gives a range of possible *P*-values for a significance. We can still draw a conclusion from the test in much the same way as if we had a single probability by comparing the range of possible *P*-values to our desired significance level.

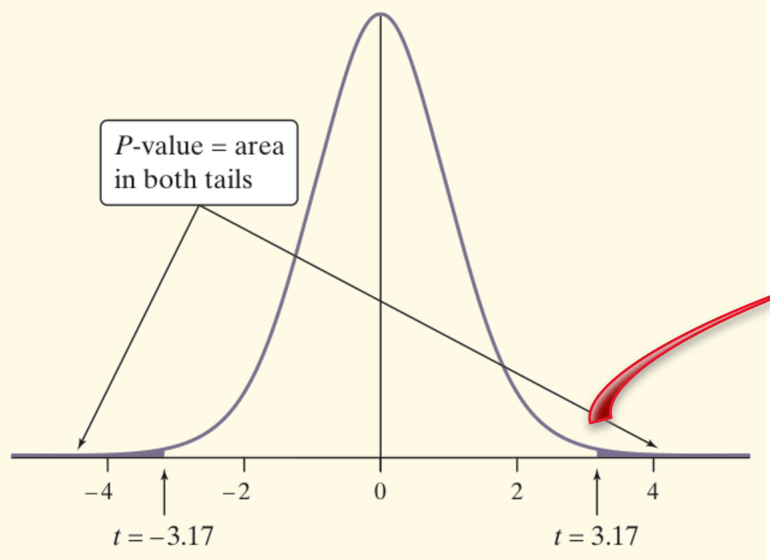
- Table B has other limitations for finding P-values. It includes probabilities only for *t* distributions with degrees of freedom from 1 to 30 and then skips to *df* = 40, 50, 60, 80, 100, and 1000. (The bottom row gives probabilities for *df* = ∞, which corresponds to the standard Normal curve.) *Note: If the df you need isn't provided in Table B, use the next lower df that is available.*

- Table B shows probabilities only for positive values of *t*. To find a *P*-value for a negative value of *t*, we use the symmetry of the *t* distributions.

Using Table B Wisely

Suppose you were performing a test of H_0: $\mu = 5$ versus H_a: $\mu \neq 5$ based on a sample size of n = 37 and obtained t = -3.17. Since this is a two-sided test, you are interested in the probability of getting a value of t less than -3.17 or greater than 3.17.

Due to the symmetric shape of the density curve, $P(t \leq -3.17) = P(t \geq 3.17)$. Since Table B shows only positive t-values, we must focus on t = 3.17.



Upper-tail probability *p*

df	**.005**	**.0025**	**.001**
29	2.756	3.038	3.396
30	2.750	3.030	3.385
40	2.704	2.971	3.307
	99%	99.5%	99.8%

Confidence level *C*

Since df = 37 – 1 = 36 is not available on the table, move across the df = 30 row and notice that t = 3.17 falls between 3.030 and 3.385.
The corresponding "Upper-tail probability p" is between 0.0025 and 0.001. For this two-sided test, the corresponding P-value would be between 2(0.001) = 0.002 and 2(0.0025) = 0.005.

Tests About a Population Mean

Alternate Exercise – More practice with Table B

Problem:

(a) Find the *P*-value for a test of H_0: $\mu = 10$ versus H_a: $\mu > 10$ that uses a sample of size 75 and has a test statistic of $t = 2.33$.

Since Table B does not include a line for df = 75, we will be conservative and use df = 60. Since $t = 2.33$ is between 2.099 and 2.390, the *P*-value is between 0.01 and 0.02.

(b) Find the *P*-value for a test of H_0: $\mu = 10$ versus H_a: $\mu \neq 10$ that uses a sample of size 10 and has a test statistic of *t* = -–0.51.

Using the line for df = 9, we find that the value 0.51 is smaller than any *t*-value provided in that line. This means that the area under the *t*-curve to the right of 0.51 is greater than 0.25. Since this is a two-sided test, the *P*-value must be at least 0.50.

The One-Sample *t* Test

When the conditions are met, we can test a claim about a population mean μ using a **one-sample t test**.

One-Sample *t* Test

Choose an SRS of size *n* from a large population that contains an unknown mean μ. To test the hypothesis $H_0 : \mu = \mu_0$, compute the one-sample *t* statistic

Find

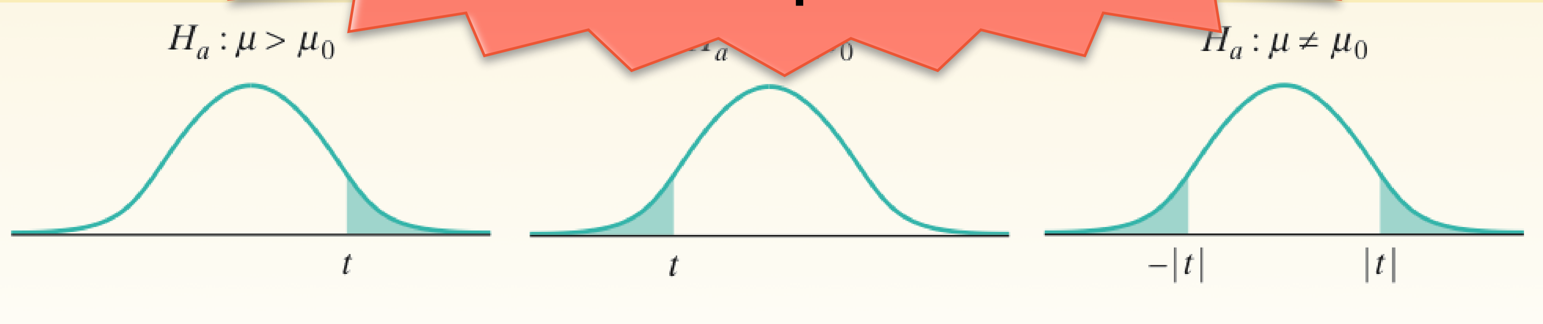
or la

distribu

rge

a *t*-

Use this test only when (1) the population distribution is Normal or the sample is large ($n \geq 30$), and (2) the population is at least 10 times as large as the sample.



Example: Healthy Streams

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water’s ability to support aquatic life. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter:

4.53	5.04	3.29	5.23	4.13	5.50	4.83	4.40
5.42	6.38	4.01	4.66	2.87	5.73	5.55	

A dissolved oxygen level below 5 mg/l puts aquatic life at risk.

State: We want to perform a test at the $\alpha = 0.05$ significance level of

$$H_0: \mu = 5$$
$$H_a: \mu < 5$$

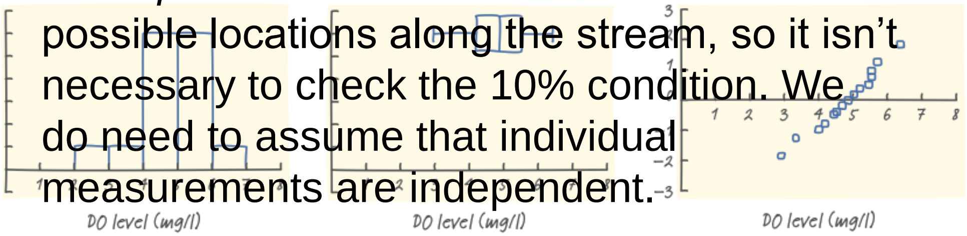
where μ is the actual mean dissolved oxygen level in this stream.

Plan: If conditions are met, we should do a one-sample t test for μ.

✓*Random* The researcher measured the DO level at 15 randomly chosen locations.

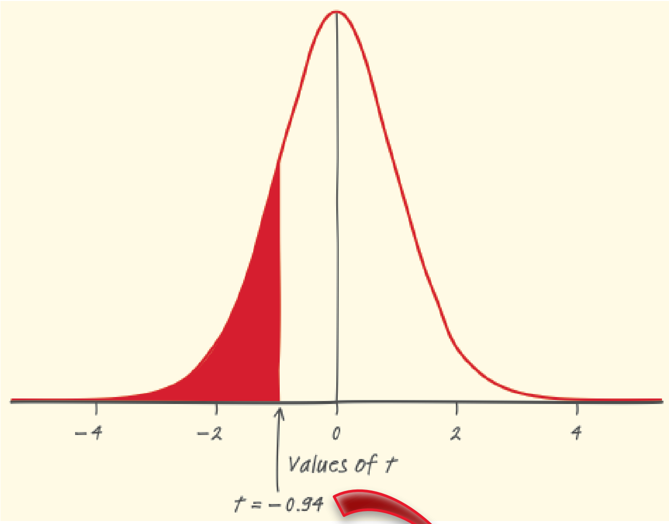
✓*Normal* We don’t know whether the population distribution of DO levels at all points along the stream is Normal. With such a small sample size (n = 15), we need to look at the data to see if it’s safe to use t procedures. The histogram looks roughly symmetric; the boxplot shows no outliers; and the Normal probability plot is fairly linear. With no outliers or strong skewness, the t procedures should be pretty accurate even if the population distribution isn’t Normal.

✓*Independent* There is an infinite number of possible locations along the stream, so it isn’t necessary to check the 10% condition. We do need to assume that individual measurements are independent.



Example: Healthy Streams

Do: The sample mean and standard deviation are $\bar{x} = 4.771$ and $s_x = 0.9396$



Test statistic $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \dfrac{4.771 - 5}{0.9396 / \sqrt{15}} = -0.94$

***P*-value** The *P*-value is the area to the left of *t* = -0.94 under the *t* distribution curve with df = 15 – 1 = 14.

Upper-tail probability *p*

df	.25	.20	.15
13	.694	.870	1.079
14	.692	.868	1.076
15	.691	.866	1.074
	50%	60%	70%

Confidence level *C*

Conclude: The *P*-value, is between 0.15 and 0.20. Since this is greater than our α = 0.05 significance level, we fail to reject H_0. We don't have enough evidence to conclude that the mean DO level in the stream is less than 5 mg/l.

Since we decided not to reject H_0, we could have made a Type II error (failing to reject H_0 when H_0 is false). If we did, then the mean dissolved oxygen level μ in the stream is actually less than 5 mg/l, but we didn't detect that with our significance test.

Alternate Example: Construction zones

- Every road has one at some point—construction zones that have much lower speed limits. To see if drivers obey these lower speed limits, a police officer used a radar gun to measure the speed (in miles per hour, or mph) of a random sample of 10 drivers in a 25 mph construction zone. Here are the results:

27 33 32 21 30 30 29 25 27 34

Problem:

- (a) Can we conclude that the average speed of drivers in this construction zone is greater than the posted 25 mph speed limit?
- **Solution:**

State: We want to perform a test at the $\alpha = 0.05$ significance level of H_0: $\mu = 25$ and H_a: $\mu > 25$ where μ is the actual mean dissolved oxygen level in this stream.

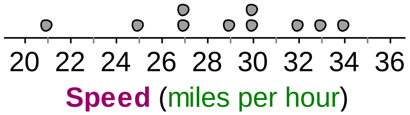
Plan: If conditions are met, we should do a one-sample *t* test for μ.

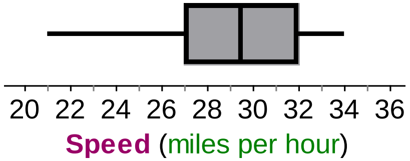
✓*Random* A random sample of drivers was selected.

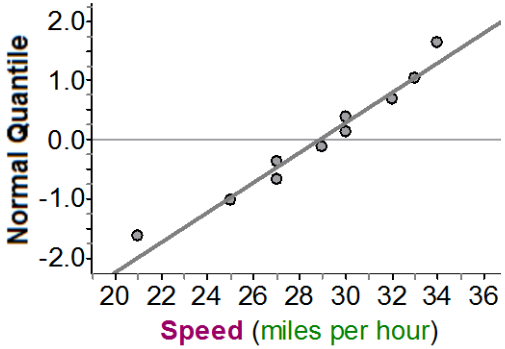
✓*Normal* We don't know if the population distribution of speeds is approximately Normal and we don't have a large sample size, so we will graph the data and look for any departures from Normality.

The dotplot and boxplot are only slightly skewed to the left with no outliers, and the Normal probability plot looks roughly linear, so it is reasonable to use *t* procedures for these data.

✓Independent: There are more than 10(10) = 100 drivers that go through this construction zone.







Alternate Example: Construction zones

Tests About a Population Mean

Do: The sample mean and standard deviation are $\bar{x} = 28.8 mph$ and $s_x = 3.9 mph$

Test statistic $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \dfrac{28.8 - 25}{3.94 / \sqrt{10}} = 3.05$

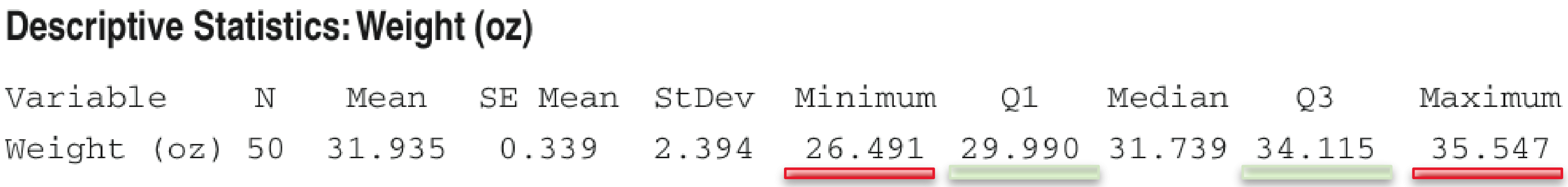
***P*-value** *P*(*t* > 3.05) using the *t* distribution with 10 – 1 = 9 degrees of freedom. Using technology, the *P*-value = tcdf(3.05,100,9) = 0.0069.

Conclude: Since the *P*-value is less than (0.0069 < 0.05), we reject the null hypothesis. There is convincing evidence that the true average speed of drivers in this construction zone is greater than 25 mph.

b) Given your conclusion in part (a), which kind of mistake—a Type I or a Type II error—could you have made? Explain what this mistake means in this context.
Since we rejected the null hypothesis, it is possible that we made a Type I error. In other words, it is possible that we concluded that the average speed of drivers in this construction zone is greater than 25 mph when in reality it isn't.

Two-Sided Tests

At the Hawaii Pineapple Company, managers are interested in the sizes of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A new irrigation system was installed in this field after the growing season. Managers wonder whether this change will affect the mean weight of future pineapples grown in the field. To find out, they select and weigh a random sample of 50 pineapples from this year's crop. The Minitab output below summarizes the data. Determine whether there are any outliers.



Descriptive Statistics: Weight (oz)

```
Variable      N    Mean   SE Mean  StDev  Minimum    Q1    Median    Q3    Maximum
Weight (oz) 50  31.935   0.339   2.394   26.491  29.990  31.739  34.115   35.547
```

✓ $IQR = Q_3 - Q_1 = 34.115 - 29.990 = 4.125$

✓Any data value greater than $Q_3 + 1.5(IQR)$ or less than $Q_1 - 1.5(IQR)$ is considered an outlier.

$$Q_3 + 1.5(IQR) = 34.115 + 1.5(4.125) = 40.3025$$
$$Q_1 - 1.5(IQR) = 29.990 - 1.5(4.125) = 23.0825$$

✓Since the maximum value 35.547 is less than 40.3025 and the minimum value 26.491 is greater than 23.0825, there are no outliers.

Two-Sided Tests

State: We want to test the hypotheses

$$H_0: \mu = 31$$
$$H_a: \mu \neq 31$$

where μ = the mean weight (in ounces) of all pineapples grown in the field this year. Since no significance level is given, we'll use $\alpha = 0.05$.

Plan: If conditions are met, we should do a one-sample *t* test for μ.

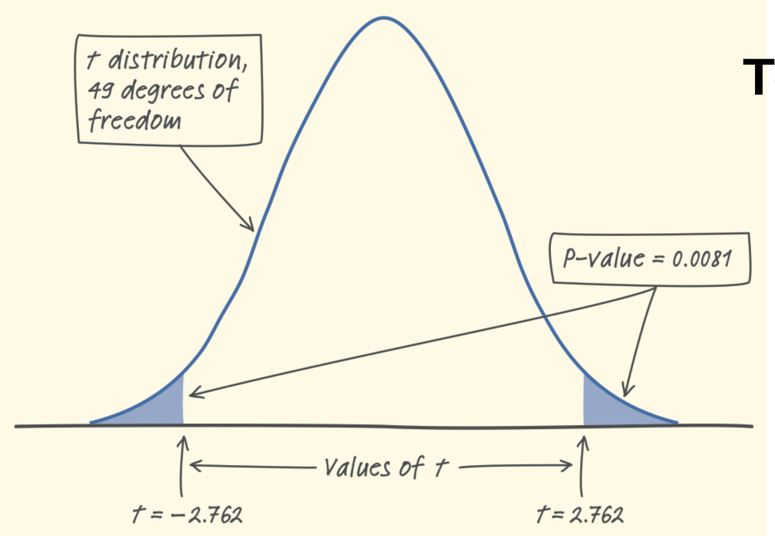
✓*Random* The data came from a random sample of 50 pineapples from this year's crop.

✓*Normal* We don't know whether the population distribution of pineapple weights this year is Normally distributed. But $n = 50 \geq 30$, so the large sample size (and the fact that there are no outliers) makes it OK to use *t* procedures.

✓*Independent* There need to be at least 10(50) = 500 pineapples in the field because managers are sampling without replacement (*10% condition*). We would expect many more than 500 pineapples in a "large field."

Two-Sided Tests

Do: The sample mean and standard deviation are $\bar{x} = 31.935$ and $s_x = 2.394$



Test statistic $t = \dfrac{\bar{x} - \mu_0}{s_x/\sqrt{n}} = \dfrac{31.935 - 31}{2.394/\sqrt{50}} = 2.762$

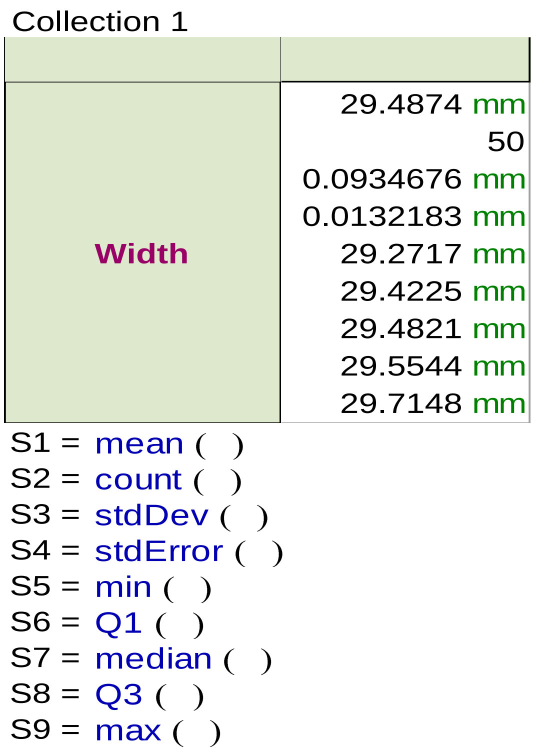
***P*-value** The *P*-value for this two-sided test is the area under the *t* distribution curve with 50 - 1 = 49 degrees of freedom. Since Table B does not have an entry for df = 49, we use the more conservative df = 40. The upper tail probability is between 0.005 and 0.0025 so the desired *P*-value is between 0.01 and 0.005.

	Upper-tail probability *p*		
df	**.005**	**.0025**	**.001**
30	2.750	3.030	3.385
40	2.704	2.971	3.307
50	2.678	2.937	3.261
	99%	99.5%	99.8%
	Confidence level *C*		

Conclude: Since the *P*-value is between 0.005 and 0.01, it is less than our *α* = 0.05 significance level, so we have enough evidence to reject H_0 and conclude that the mean weight of the pineapples in this year’s crop is not 31 ounces.

Alternate Example – Don’t break the ice

In the children’s game Don’t Break the Ice, small plastic ice cubes are squeezed into a square frame. Each child takes turns tapping out a cube of “ice” with a plastic hammer hoping that the remaining cubes don’t collapse. For the game to work correctly, the cubes must be big enough so that they hold each other in place in the plastic frame but not so big that they are too difficult to tap out. The machine that produces the plastic ice cubes is designed to make cubes that are 29.5 millimeters (mm) wide, but the actual width varies a little. To make sure the machine is working well, a supervisor inspects a random sample of 50 cubes every hour and measures their width. The Fathom output below summarizes the data from a sample taken during one hour.



```
Collection 1

          29.4874 mm
               50
        0.0934676 mm
        0.0132183 mm
Width     29.2717 mm
          29.4225 mm
          29.4821 mm
          29.5544 mm
          29.7148 mm

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = min ( )
S6 = Q1 ( )
S7 = median ( )
S8 = Q3 ( )
S9 = max ( )
```

Problem:

(a) Interpret the standard deviation and the standard error provided by the computer output.

Solution:

(a) Standard deviation: The widths of the cubes are about 0.093 mm from the mean width, on average. Standard error: In random samples of size 50, the sample mean will be about 0.013 mm from the true mean, on average.

Tests About a Population Mean

Alternate Example – Don’t break the ice

(b) Do these data give convincing evidence that the mean width of cubes produced this hour is not 29.5 mm?

State: We want to test the following hypotheses at the = 0.05 significance level:

$H_0: \mu = 29.5$

$H_a: \mu \neq 29.5$

where *μ* = the true mean width of plastic ice cubes

Plan: If conditions are met, we should do a one-sample *t* test for *μ*.

✓*Random* A random sample of plastic ice cubes was selected.

✓*Normal* We have a large sample size ($n = 50 \geq 30$), so it is OK to use *t* procedures.

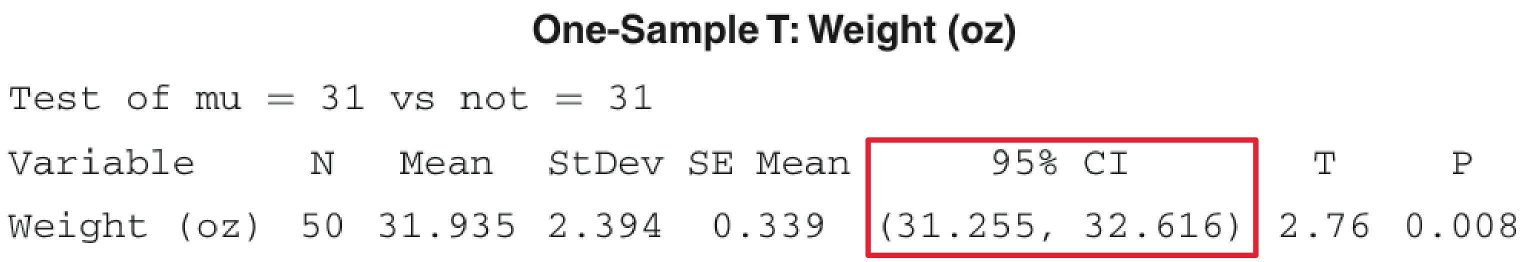
✓Independent: It is reasonable to assume that there are more than 10(50) = 500 cubes produced by this machine each hour.

Test statistic $t = \dfrac{29.4874 - 29.5}{0.0934 \big/ \sqrt{50}} = -0.95$

Conclude: Since the *P*-value is greater than (0.3468 > 0.05), we fail to reject the null hypothesis. There is not convincing evidence that the true width of the plastic ice cubes produced this hour is different from 29.5 mm.

Confidence Intervals Give More Information

Minitab output for a significance test and confidence interval based on the pineapple data is shown below. The test statistic and *P*-value match what we got earlier (up to rounding).



One-Sample T: Weight (oz)

```
Test of mu = 31 vs not = 31

Variable      N    Mean   StDev SE Mean        95% CI           T      P
Weight (oz)  50  31.935  2.394   0.339  (31.255, 32.616)  2.76  0.008
```

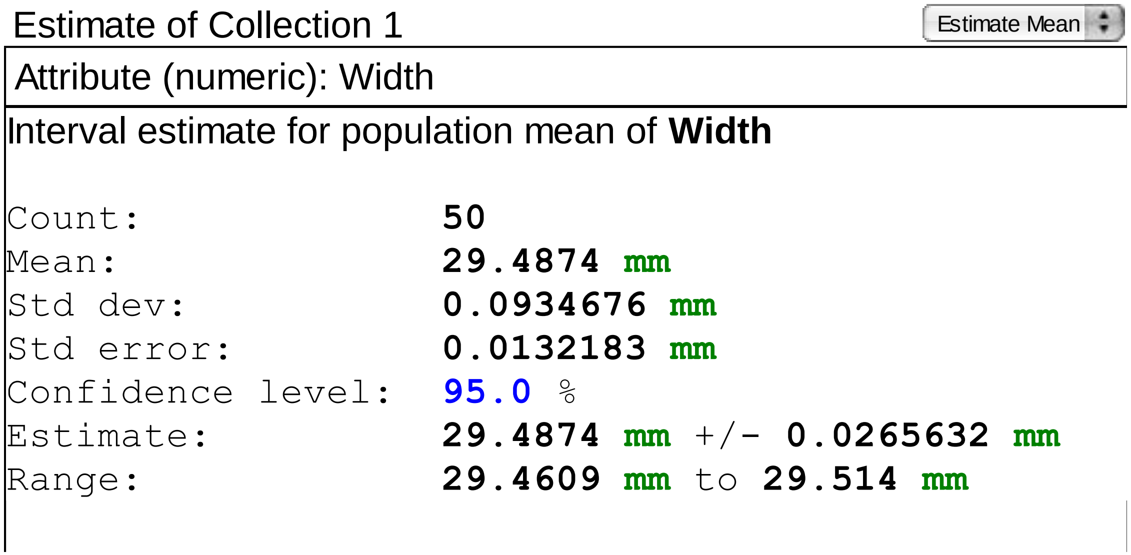
The 95% confidence interval for the mean weight of all the pineapples grown in the field this year is 31.255 to 32.616 ounces. We are 95% confident that this interval captures the true mean weight µ of this year's pineapple crop.

As with proportions, there is a link between a two-sided test at significance level α and a 100(1 – *α*)% confidence interval for a population mean *µ*.

For the pineapples, the two-sided test at α =0.05 rejects H_0: μ = 31 in favor of H_a: $\mu \neq 31$. The corresponding 95% confidence interval does not include 31 as a plausible value of the parameter *µ*. In other words, the test and interval lead to the same conclusion about H_0. But the confidence interval provides much more information: *a set of plausible values for the population mean*.

Alternate Example – Don’t break the ice

Here is Fathom output for a 95% confidence interval for the true mean width of plastic ice cubes produced this hour.



Estimate of Collection 1

Attribute (numeric): Width

Interval estimate for population mean of **Width**

```
Count:              50
Mean:               29.4874 mm
Std dev:            0.0934676 mm
Std error:          0.0132183 mm
Confidence level:   95.0 %
Estimate:           29.4874 mm +/- 0.0265632 mm
Range:              29.4609 mm to 29.514 mm
```

Problem:

(a) Interpret the confidence interval. Would you make the same conclusion with the confidence interval as you did with the significance test in the previous example?

(b) Interpret the confidence level.

Solution:

(a) We are 95% confident that the interval from 29.4609 mm to 29.514 mm captures the true mean width of plastic ice cubes produced this hour. Since the interval includes 29.5 as a plausible value for the true mean width, we do not have convincing evidence that the true mean is not 29.5 mm. This is the same conclusion we made in the significance test earlier.

(b) If we were to take many random samples of 50 plastic ice cubes and make 95% confidence intervals for the true mean width of the cubes, then about 95% of the intervals we construct will include the true mean width.

Confidence Intervals and Two-Sided Tests

The connection between two-sided tests and confidence intervals is even stronger for means than it was for proportions. That's because both inference methods for means use the standard error of the sample mean in the calculations.

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$ Confidence interval: $\bar{x} \pm t^* \dfrac{s_x}{\sqrt{n}}$

✓ A two-sided test at significance level *α* (say, *α* = 0.05) and a 100(1 – *α*)% confidence interval (a 95% confidence interval if *α* = 0.05) give similar information about the population parameter.

✓ When the two-sided significance test at level α rejects H_0: $\mu = \mu_0$, the 100(1 – *α*)% confidence interval for *μ* will not contain the hypothesized value μ_0 .

✓ When the two-sided significance test at level *α* fails to reject the null hypothesis, the confidence interval for *μ* will contain μ_0 .

Inference for Means: Paired Data

Comparative studies are more convincing than single-sample investigations. For that reason, one-sample inference is less common than comparative inference. Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, result in **paired data**.

When paired data result from measuring the same quantitative variable twice, as in the job satisfaction study, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample *t* procedures to perform inference about the mean difference μ_d.

These methods are sometimes called **paired *t* procedures**.

Paired *t* Test

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from coffee, colas, and other substances with caffeine for the duration of the experiment. During one two-day period, subjects took capsules containing their normal caffeine intake. During another two-day period, they took placebo capsules. The order in which subjects took caffeine and the placebo was randomized. At the end of each two-day period, a test for depression was given to all 11 subjects. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression.

Results of a caffeine deprivation study

Subject	Depression (caffeine)	Depression (placebo)	Difference (placebo – caffeine)
1	5	16	**11**
2	5	23	**18**
3	4	5	**1**
4	3	7	**4**
5	8	14	**6**
6	5	24	**19**
7	0	6	**6**
8	0	3	**3**
9	2	15	**13**
10	11	12	**1**
11	1	0	**- 1**

State: If caffeine deprivation has no effect on depression, then we would expect the actual mean difference in depression scores to be 0. We want to test the hypotheses

$$H_0: \mu_d = 0$$
$$H_a: \mu_d > 0$$

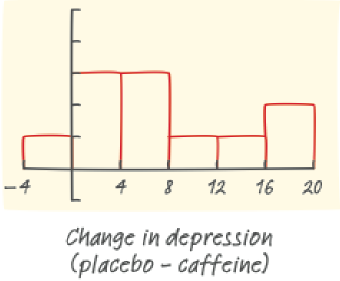
where μ_d = the true mean difference (placebo – caffeine) in depression score. Since no significance level is given, we'll use $\alpha = 0.05$.

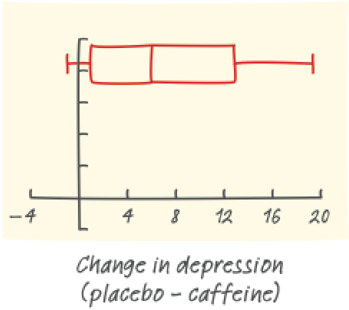
Paired *t* Test

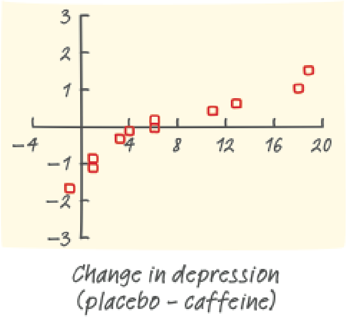
Plan: If conditions are met, we should do a paired *t* test for μ_d.

✓*Random* researchers randomly assigned the treatment order—placebo then caffeine, caffeine then placebo—to the subjects.

✓*Normal* We don't know whether the actual distribution of difference in depression scores (placebo - caffeine) is Normal. With such a small sample size (n = 11), we need to examine the data to see if it's safe to use *t* procedures.







The histogram has an irregular shape with so few values; the boxplot shows some right-skewness but not outliers; and the Normal probability plot looks fairly linear. With no outliers or strong skewness, the *t* procedures should be pretty accurate.

✓*Independent* We aren't sampling, so it isn't necessary to check the *10% condition*. We will assume that the changes in depression scores for individual subjects are independent. This is reasonable if the experiment is conducted properly.

Paired *t* Test

Do: The sample mean and standard deviation are $\bar{x}_d = 7.364$ and $s_d = 6.918$

Test statistic $t = \dfrac{\bar{x}_d - \mu_0}{s_d / \sqrt{n}} = \dfrac{7.364 - 0}{6.918 / \sqrt{11}} = 3.53$

***P*-value** According to technology, the area to the right of *t* = 3.53 on the *t* distribution curve with df = 11 – 1 = 10 is 0.0027.

Conclude: With a *P*-value of 0.0027, which is much less than our chosen *α* = 0.05, we have convincing evidence to reject H_0: $\mu_d = 0$. We can therefore conclude that depriving these caffeine-dependent subjects of caffeine caused an average increase in depression scores.

Alternate Example – Is the express lane faster?

- For their second semester project in AP Statistics, Libby and Kathryn decided to investigate which line was faster in the supermarket, the express lane or the regular lane. To collect their data, they randomly selected 15 times during a week, went to the same store, and bought the same item. However, one of them used the express lane and the other used a regular lane. To decide which lane each of them would use, they flipped a coin. If it was heads, Libby used the express lane and Kathryn used the regular lane. If it was tails, Libby used the regular lane and Kathryn used the express lane. They entered their randomly assigned lanes at the same time and each recorded the time in seconds it took them to complete the transaction.

Time in Express Lane (seconds)	Time in Regular Lane (seconds)
337	342
226	472
502	456
408	529
151	181
284	339
150	229
357	263
349	332
257	352
321	341
383	397
565	694
363	324
85	127

Problem: Carry out a test to see if there is convincing evidence that the express lane is faster.

Solution: Since these data are paired, we will consider the differences in time (regular – express). Here are the 15 differences. In this case, a positive difference means that the express lane was faster.

5, 246, –46, 121, 30, 55, 79, –94, –17, 95, 20, 14, 129, –39, 42

State: We want to test the following hypotheses at the = 0.05 significance level:

$$H_0: \mu_d = 0$$
$$H_a: \mu_d > 0$$

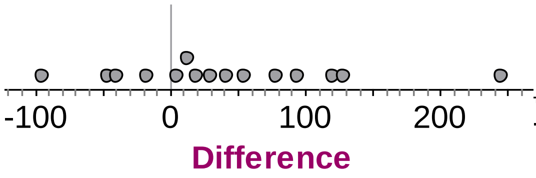
where μ_d = = the true mean difference (regular – express) in time required to purchase an item at the supermarket.

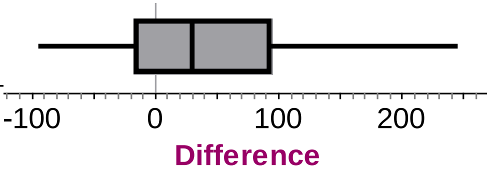
Alternate Example – Is the express lane faster?

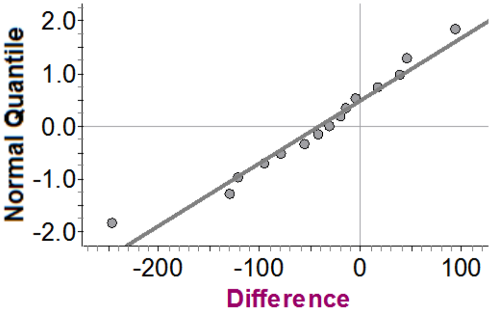
Plan: If conditions are met, we should do a paired *t* test for μ_d.

✓*Random* researchers randomly assigned the treatment order—placebo then caffeine, caffeine then placebo—to the subjects.

✓*Normal* We don't know if the population distribution of differences is approximately Normal and we don't have a large sample size, so we will graph the differences and look for any departures from Normality.







The dotplot and boxplot are slightly skewed to the right with no outliers, and the Normal probability plot looks roughly linear, so it is reasonable to use *t* procedures for these data.

✓*Independent* Since we randomly selected the times to conduct the study from an infinite number of possible times, the differences should be independent.

Alternate Example – Is the express lane faster?

Do: The sample mean difference is $\bar{x} = 42.7$ with a standard deviation of $s_x = 84.0$ seconds.

Test statistic $$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \frac{42.7 - 0}{84.0 / \sqrt{15}} = 1.97$$

***P*-value** *P*(*t* > 1.97) using the t distribution with 15 – 1 = 14 degrees of freedom. Using technology, the *P*-value = tcdf(1.97,100,14) = 0.034.

Conclude: Since the *P*-value is less than (0.034 < 0.05), we reject the null hypothesis. There is convincing evidence that express lane is faster than the regular lane.

Using Tests Wisely

Significance tests are widely used in reporting the results of research in many fields. New drugs require significant evidence of effectiveness and safety. Courts ask about statistical significance in hearing discrimination cases. Marketers want to know whether a new ad campaign significantly outperforms the old one, and medical researchers want to know whether a new therapy performs significantly better. In all these uses, statistical significance is valued because it points to an effect that is unlikely to occur simply by chance.

Carrying out a significance test is often quite simple, especially if you use a calculator or computer. Using tests wisely is not so simple. Here are some points to keep in mind when using or interpreting significance tests.

Statistical Significance and Practical Importance
When a null hypothesis (“no effect” or “no difference”) can be rejected at the usual levels ($\alpha = 0.05$ or $\alpha = 0.01$), there is good evidence of a difference. But that difference may be very small. When large samples are available, even tiny deviations from the null hypothesis will be significant.

Test About a Population Mean

Using Tests Wisely

Don’t Ignore Lack of Significance

There is a tendency to infer that there is no difference whenever a *P*-value fails to attain the usual 5% standard. In some areas of research, small differences that are detectable only with large sample sizes can be of great practical significance. When planning a study, verify that the test you plan to use has a high probability (power) of detecting a difference of the size you hope to find.

Statistical Inference Is Not Valid for All Sets of Data

Badly designed surveys or experiments often produce invalid results. Formal statistical inference cannot correct basic flaws in the design. Each test is valid only in certain circumstances, with properly produced data being particularly important.

Beware of Multiple Analyses

Statistical significance ought to mean that you have found a difference that you were looking for. The reasoning behind statistical significance works well if you decide what difference you are seeking, design a study to search for it, and use a significance test to weigh the evidence you get. In other settings, significance may have little meaning.

Alternate Example – Curriculum Reform

- Suppose that a large school district implemented a new math curriculum for the current school year. To see if the new curriculum is effective, the district randomly selects 500 students and compares their scores on a standardized test after the curriculum change to their scores on the same test before the curriculum change. The mean improvement was $\overline{X}_d$= 0.9 with a standard deviation of S_d = 12.0.

Problem:

(a) Calculate the test statistic and *P*-value for a test of : $H_0 : \mu_d = 0$ versus $H_a : \mu_d > 0.$

(b) Are the results significant at the 5% level?

(c) Can we conclude that the new curriculum is the cause of the apparent increase in scores?

Solution: (a) $t = \frac{0.9 - 0}{12 / \sqrt{500}} = 1.68$ P-value = tcdf(1.68,100,499) = 0.047

(b) Since the *P*-value is less than 0.05, the results are significant at the 5% level. The improvement was larger than what we could expect to happen by chance. However, an increase of 0.9 points may or may not be practically significant.

(c) No. Even though the increase was statistically significant, we cannot conclude that the new curriculum was the cause of the increase since there was no control group for comparison and no random assignment of treatments.

Test About a Population Mean

Alternate Example – Caffeine and pulse rates

When an AP Statistics class did the caffeine and pulse rates experiment described in the Alternate Examples from section 4.2, the confidence interval for the difference in average pulse rate changes was (–2.52, 4.59). This means that drinking soda with caffeine can increase your pulse rate by as much as 4.59 beats per minute or decrease it by as much as 2.52 beats per minute compared to drinking soda with no caffeine. It seems likely that caffeine does have an effect on pulse rates, but we need more data to estimate this effect with more precision.

Alternate Example – GPA in AP Statistics

Suppose that you wanted to know the average GPA for students at your school who are enrolled in AP Statistics. Since this isn't a large population, you conduct a census and record the GPA for each student. Is it appropriate to construct a one-sample *t* interval for the population mean GPA? No. If we have GPAs for every student in the population, we can calculate exactly. We only use a confidence interval (or significance test) when we need to account for the variability caused by random sampling or random assignment to treatments.

Alternate Example – More cell phones and brain cancer

Suppose that 20 significance tests were conducted and in each case the null hypothesis was true. What is the probability that we avoid a Type I error in all 20 tests? If we are using a 5% significance level, each individual test has a 0.95 probability of avoiding a Type I error. Assuming that the results of the tests are independent, the probability of avoiding Type I errors in each of the tests is (0.95)(0.95) (0.95) = $(0.95)^{20}$ = 0.36. This means that there is a 64% chance we will make at least 1 Type I error in these 20 tests. So, finding one significant result in 20 is definitely not a surprise.

Section 9.3
Tests About a Population Mean

Summary

In this section, we learned that…

- ✓ Significance tests for the mean *μ* of a Normal population are based on the sampling distribution of the sample mean. Due to the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.

- ✓ If we somehow know σ, we can use a *z* test statistic and the standard Normal distribution to perform calculations. In practice, we typically do not know σ. Then, we use the **one-sample *t* statistic**

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

with *P*-values calculated from the *t* distribution with *n* - 1 degrees of freedom.

Section 9.3
Tests About a Population Mean

Summary

- ✓ The **one-sample *t* test** is approximately correct when

  Random The data were produced by random sampling or a randomized experiment.

  Normal The population distribution is Normal OR the sample size is large ($n \geq 30$).

  Independent Individual observations are independent. When sampling without replacement, check that the population is at least 10 times as large as the sample.

- ✓ Confidence intervals provide additional information that significance tests do not—namely, a range of plausible values for the parameter μ. A two-sided test of H_0: $\mu = \mu_0$ at significance level α gives the same conclusion as a 100(1 – α)% confidence interval for μ.
- ✓ Analyze **paired data** by first taking the difference within each pair to produce a single sample. Then use one-sample *t* procedures.

Section 9.3
Tests About a Population Mean

Summary

- ✓ Very small differences can be highly significant (small *P*-value) when a test is based on a large sample. A statistically significant difference need not be practically important.
- ✓ Lack of significance does not imply that H_0 is true. Even a large difference can fail to be significant when a test is based on a small sample.
- ✓ Significance tests are not always valid. Faulty data collection, outliers in the data, and other practical problems can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

Looking Ahead…

In the next Chapter…

We'll learn how to compare two populations or groups.

We'll learn about

- ✓ **Comparing Two Proportions**
- ✓ **Comparing Two Means**