

9.3

Tests About a Population Mean

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 .

Apr 12-10:57 AM

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.

 \bar{x} s_x

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.

Apr 12-10:59 AM

Carrying Out a Significance Test for μ

In Chapter 8, we introduced conditions that should be met before we construct a confidence interval for a population mean: Random, 10% when sampling without replacement, and Normal/Large Sample. These same three conditions must be verified before performing a significance test about a population mean.

Conditions For Performing A Significance Test About A Mean

- **Random:** The data come from a well-designed random sample or randomized experiment.
 - **10%:** When sampling without replacement, check that $n \leq (1/10)N$.
- **Normal/Large Sample:** The population has a Normal distribution or the sample size is large ($n \geq 30$). If the population distribution has unknown shape and $n < 30$, use a graph of the sample data to assess the Normality of the population. Do not use t procedures if the graph shows strong skewness or outliers.

Apr 12-10:59 AM

A classic rock radio station claims to play an average of 50 minutes of music every hour. However, it seems that every time you turn to this station, 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 50 minutes of music per hour.

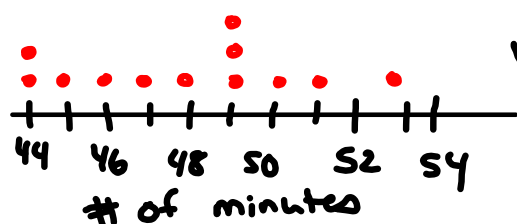
random ✓

10% ✓
> 120 hrs
per wk

Large Counts (Normal)

$n = 12$ ($n < 30$)

check graph



no strong skewness
no apparent outliers
reasonable to use t

Apr 13-10:23 AM

Carrying Out a Significance Test for μ

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.

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

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

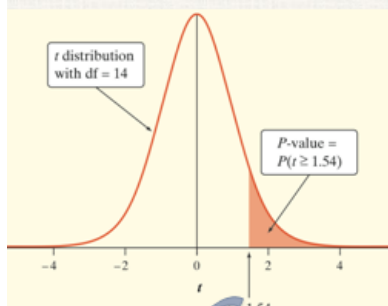
$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

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

Apr 12-11:01 AM

Carrying Out a Significance Test for μ

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.



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

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

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

$n = 15$
 $df = 14$

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

✓ 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.

$.05 < p\text{-val} < .10$

Apr 12-11:02 AM

In the previous example, we wanted to test $H_0: \mu = 50$ versus $H_a: \mu < 50$ using a sample of 12 hours of music.

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

$$\bar{X} = 47.9 \text{ min} \quad S_x = 2.81 \text{ min}$$

$$t = \frac{\bar{X} - \mu_0}{S_x / \sqrt{n}} = \frac{47.9 - 50}{2.81 / \sqrt{12}} = -2.59$$

$df = 11$

$$.01 < pval < .02$$

$$tcdf \left(\begin{matrix} \text{lower} \\ -10000 \end{matrix}, \begin{matrix} \text{upper} \\ -2.59 \end{matrix}, \begin{matrix} df \\ 11 \end{matrix} \right) = \text{pval} = .013$$

Apr 13-10:23 AM

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 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 $df = \infty$, 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.*

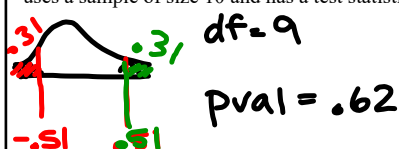
•Table B shows probabilities only for positive values of t . To find a P -value for a negative value of t , we use the symmetry of the t distributions.

Problem:

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

$$df = 74 \Rightarrow \text{use } df = 60$$

(b) Find the P -value for a test of: $\mu = 300$ versus $\mu \neq 300$ that uses a sample of size 10 and has a test statistic of $t = -0.51$.



$$tcdf$$

$$(-1000, -0.51, 9)$$

$$(0.51, 10000, 9)$$

$$pval > .25 \quad (2)$$

$$> .50$$

$$tcdf$$

$$\left(\begin{matrix} \text{lower} \\ 2.33 \end{matrix}, \begin{matrix} \text{upper} \\ 1000 \end{matrix}, \begin{matrix} df \\ 74 \end{matrix} \right)$$

Apr 12-11:02 AM

One point from the example deserves repeating: if the df you need isn't provided in **Table B**, use the next lower df that is available. It's no fair "rounding up" to a larger df . This is like pretending that your sample size is larger than it really is. Doing so would give you a smaller P -value than is true and would make you more likely to incorrectly reject H_0 when it's true (make a Type I error).

Given the limitations of **Table B**, our advice is to use technology to find P -values when carrying out a significance test about a population mean.

19. TECHNOLOGY CORNER COMPUTING P-VALUES FROM t DISTRIBUTIONS ON THE CALCULATOR

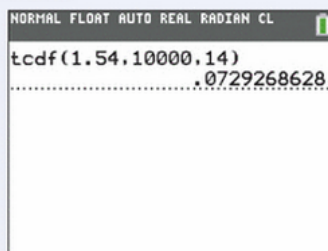
You can use the `tcdf` command on the TI-83/84 and TI-89 to calculate areas under a t distribution curve. The syntax is `tcdf(lower bound, upper bound, df)`.

Let's use the `tcdf` command to compute the P -values from the last two examples.

Better batteries: To find $P(t \geq 1.54)$,

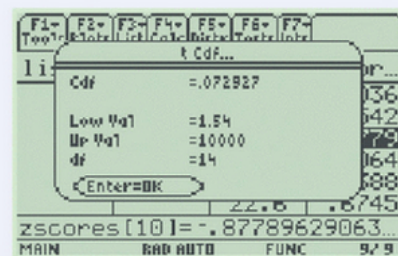
TI-83/84

- Press **2nd** **VAR** (DISTR) and choose `tcdf` (. OS 2.55 or later: In the dialog box, enter these values: lower: 1.54, upper: 10000, df : 14, choose Paste, and then press **ENTER**. Older OS: Complete the command `tcdf(1.54, 10000, 14)` and press **ENTER**.



TI-89

- In the Stats/List Editor, press **F5** (Distr) and choose `t Cdf...`
- In the dialog box, enter these values: Lower value: 1.54, Upper value: 10000, Deg of Freedom, df : 14, and then choose **ENTER**.



Apr 7-10:24 AM

CHECK YOUR UNDERSTANDING

The makers of Aspro brand aspirin want to be sure that their tablets contain the right amount of active ingredient (acetylsalicylic acid). So they inspect a random sample of 36 tablets from a batch in production. When the production process is working properly, Aspro tablets have an average of $\mu = 320$ milligrams (mg) of active ingredient. The amount of active ingredient in the 36 selected tablets has mean 319 mg and standard deviation 3 mg.

1 sample t-test for μ = true mean active ingred in batch of aspro tablets

- State appropriate hypotheses for a significance test in this setting.

$$H_0: \mu = 320 \quad H_a: \mu \neq 320$$

- Check that the conditions are met for carrying out the test.

random ✓

large counts ✓
 $n = 36$

10% ✓ > 360 in batch

- Calculate the test statistic. Show your work.

$$t = \frac{319 - 320}{3/\sqrt{36}} = -2$$

- Use **Table B** to find the P -value. Then use technology to get a more accurate result. What conclusion would you draw?

$$df = 35 \rightarrow \text{use } 30$$

$$tcdf(2, 10000, 35) = .0267$$

$$2(.0267) < pval < 2(.05)$$

$$.05 < pval < .1$$

$$pval = 2(.0267) = .0534$$

at $\alpha = .05$ Fail to Reject H_0

Apr 7-10:25 AM

Mar 13-1:37 PM

The One-Sample t -Test

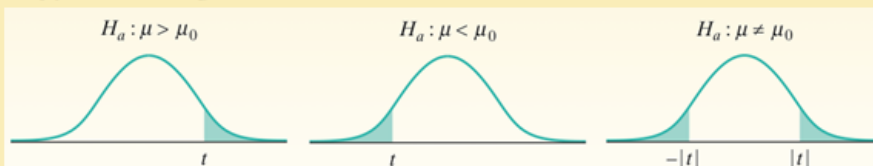
When the conditions are met, we can test a claim about a population mean μ using a **one-sample t test**.

One Sample t -Test for a Mean

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

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

Find the P -value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a in a t -distribution with $df = n - 1$



Apr 12-11:02 AM

Short Subs

Abby and Raquel like to eat sub sandwiches. However, they noticed that the lengths of the "6-inch sub" sandwiches they get at their favorite restaurant seemed shorter than the advertised length. To investigate, they randomly selected 24 different times during the next month and ordered a "6-inch" sub. Here are the actual lengths of each of the 24 sandwiches (in inches):

4.50	4.75	4.75	5.00	5.00	5.00	5.50	5.50
5.50	5.50	5.50	5.50	5.75	5.75	5.75	6.00
6.00	6.00	6.00	6.00	6.50	6.75	6.75	7.00

Problem:


(a) Do these data provide convincing evidence at the $\alpha = 0.10$ level that the sandwiches at this restaurant are shorter than advertised, on average?

(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 would mean in context.

1 sample t test for μ @ $\alpha = .10$

μ = true mean length of "6 inch" subs from this restaurant

Random \checkmark 10% > 240 sandwiches in a month Large Counts/Normal
 $n \geq 30$, graph
 no extreme skewness and no outliers so reasonable to use t


 length (in)

$H_0: \mu = 6 \text{ in}$ $H_a: \mu < 6 \text{ in}$

Apr 13-10:26 AM

$$t = -2.41 \quad p\text{-val} = .0123$$

$$\bar{x} = 5.68 \quad S_x = .657 \quad n = 24$$

$$df = 23$$

$$p\text{-val} .0123 < \alpha = .10 \quad \text{Reject } H_0$$

There is convincing evidence that the true mean length of "6" subs is less than 6 in. at this restaurant.

Type I α - Reject H_0 , when H_0 true
 Found the subs smaller,
 but really they were fine

Mar 9-8:51 AM

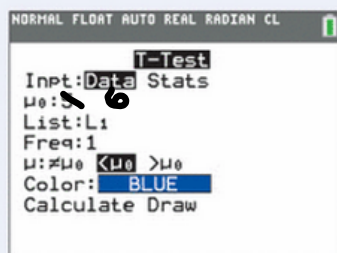
You can also use your calculator to carry out a one-sample t test. But be sure to read the AP[®] exam tip at the end of the Technology Corner.

20. TECHNOLOGY CORNER ONE-SAMPLE t TEST FOR A MEAN ON THE CALCULATOR

You can perform a one-sample t test using either raw data or summary statistics on the TI-83/84 or TI-89. Let's use the calculator to carry out the test of $H_0: \mu = 5$ versus $H_a: \mu < 5$ from the dissolved oxygen example. Start by entering the sample data in L1/list1. Then, to do the test:

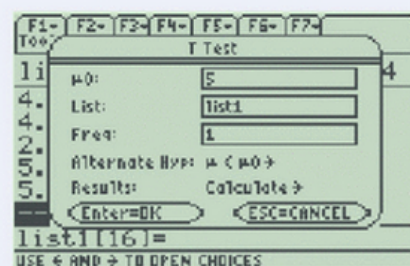
TI-83/84

- Press **[STAT]**, choose TESTS and T-Test.
- Adjust your settings as shown.

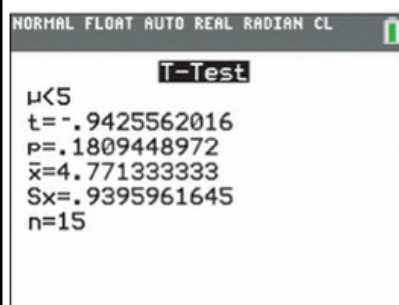


TI-89

- Press **[2nd]** **[F1]** ([F6]) and choose T-Test.
- Adjust your settings as shown.

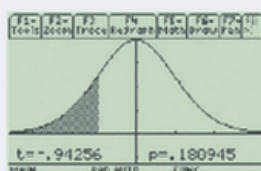
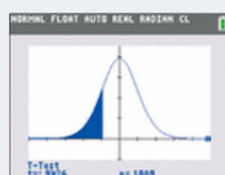


Apr 7-10:30 AM



The test statistic is $t = -0.94$ and the P -value is 0.1809.

If you specify "Draw," you see a t distribution curve ($df = 14$) with the lower tail shaded.



AP[®] EXAM TIP Remember: if you just give calculator results with no work, and one or more values are wrong, you probably won't get any credit for the "Do" step. If you opt for the calculator-only method, name the procedure (t test) and report the test statistic ($t = -0.94$), degrees of freedom ($df = 14$), and P -value (0.1809).

Note: If you are given summary statistics instead of the original data, you would select the option "Stats" instead of "Data" in the first screen.

Apr 7-10:31 AM



CHECK YOUR UNDERSTANDING

A college professor suspects that students at his school are getting less than 8 hours of sleep a night, on average. To test his belief, the professor asks a random sample of 28 students, "How much sleep did you get last night?" Here are the data (in hours):

9 6 8 6 8 8 6 6.5 6 7 9 4 3 4 5 6 11 6 3 6 6 10 7 8 4.5 9 7 7

1. Do these data provide convincing evidence at the $\alpha = 0.05$ significance level in support of the professor's suspicion?

Apr 7-10:31 AM

Two-Sided Tests and Confidence Intervals

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.

$$\text{Test statistic: } t = \frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}}$$

$$\text{Confidence interval: } \bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$$

- ✓ A two-sided test at significance level α (say, $\alpha = 0.05$) and a $100(1 - \alpha)\%$ confidence interval (a 95% confidence interval if $\alpha = 0.05$) give similar information about the population parameter.
- ✓ When the two-sided significance test at level α rejects $H_0: \mu = \mu_0$, the $100(1 - \alpha)\%$ 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 .

Apr 12-11:02 AM

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 cubes is designed to make cubes that are 29.5 millimeters (mm) wide, but the actual width varies a little. To ensure that 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.

Problem:

Collection 1	
	29.4846 mm
	50
Width	0.0900274 mm
	0.0127318 mm
S1 = mean ()	
S2 = count ()	
S3 = stdDev ()	
S4 = stdError ()	

\bar{x}
 n
 s
 s/\sqrt{n}

- (a) Interpret the standard deviation and the standard error provided by the computer output.
 (b) Do these data give convincing evidence that the mean width of cubes produced this hour is different than 29.5 mm?

a) **Standard deviation:** The widths of cubes are typically .09 mm from the mean width
Standard error: In a random sample of size 50, the sample mean will typically be .013 mm from the mean

Apr 13-10:30 AM

1 sample t test for μ

$$H_0: \mu = 29.5 \quad H_a: \mu \neq 29.5$$

random ✓

10% ✓
> 500 produced
in 1 hr

large counts ✓
 $n \geq 30$

$$t = \frac{29.4846 - 29.5}{.0900274/\sqrt{50}} = -1.21 \quad df = 49$$

pval = .2322 > α .05
Fail to Reject H_0

There is not convincing evidence the true width of cubes produced in this hour is different than 29.5 mm.

Apr 14-10:43 AM

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:	29.4846 mm
Std dev:	0.0900274 mm
Std error:	0.0127318 mm
Confidence level:	95.0 %
Estimate:	29.4846 mm +/- 0.0255855 mm
Range:	29.459 mm to 29.5102 mm

Problem:

- Interpret the confidence interval. Would you make the same conclusion with the confidence interval as you did with the significance test in the previous Alternate Example?
- Interpret the confidence level.

a) $H_0: \mu = 29.5$ is w/in C.I., so plausible value.

b) In many samples of size 50, about 95% of the resulting CI will capture the true mean width

Apr 13-10:31 AM



CHECK YOUR UNDERSTANDING

The health director of a large company is concerned about the effects of stress on the company's middle-aged male employees. According to the National Center for Health Statistics, the mean systolic blood pressure for males 35 to 44 years of age is 128. The health director examines the medical records of a random sample of 72 male employees in this age group. The Minitab output displays the results of a significance test and a confidence interval.

1. Do the results of the significance test give convincing evidence that the mean blood pressure for all the company's middle-aged male employees differs from the national average? Justify your answer.

2. Interpret the 95% confidence interval in context. Explain how the confidence interval leads to the same conclusion as in Question 1.

Session						
One-Sample T						
Test of $\mu = 128$ vs not = 128						
N	Mean	StDev	SE Mean	95% CI	T	P
72	129.93	14.90	1.76	(126.43, 133.43)	1.10	0.275

Apr 7-10:35 AM

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**.

Apr 12-11:02 AM

Example: Paired data and one-sample t procedures

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 if being deprived of caffeine would lead to an increase in depression.

Apr 12-11:03 AM

Example: Paired data and one-sample t procedures

The table below contains data on the subjects' scores on a depression test. Higher scores show more symptoms of depression.

$\mu_d > 0$

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

Apr 12-11:03 AM

Example: Paired data and one-sample t procedures

Problem:

- (a) Why did researchers randomly assign the order in which subjects received placebo and caffeine?

Researchers want to be able to conclude that any statistically significant change in depression score is due to the treatments themselves and not to some other variable.

One obvious concern is the order of the treatments. Suppose that caffeine were given to all the subjects during the first 2-day period. What if the weather were nicer on these 2 days than during the second 2-day period when all subjects were given a placebo? Researchers wouldn't be able to tell if a large increase in the mean depression score is due to the difference in weather or due to the treatments.

Random assignment of the caffeine and placebo to the two time periods in the experiment should help ensure that no other variable (like the weather) is systematically affecting subjects' responses.

Apr 12-11:03 AM

Example: Paired data and one-sample t procedures

Problem:

(b) Carry out a test to investigate the researchers' question.

State:

If caffeine deprivation has no effect on depression, then we would expect the actual mean difference in depression scores to be 0.

We therefore want to test the hypotheses

$$H_0: \mu_d = 0$$

$$H_a: \mu_d > 0$$

where μ_d is the true mean difference (placebo - caffeine) in depression score for subjects like these. (We chose this order of subtraction to get mostly positive values.)

Because no significance level is given, we'll use $\alpha = 0.05$.

Apr 12-11:03 AM

Example: Paired data and one-sample t procedures

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

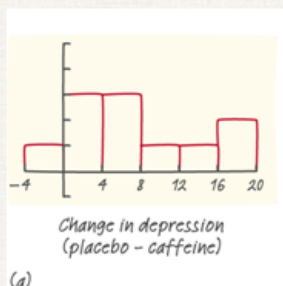
- Random: Researchers randomly assigned the treatments —placebo then caffeine, caffeine then placebo— to the subjects.

- 10% Condition: We aren't sampling, so it isn't necessary to check the 10% condition.

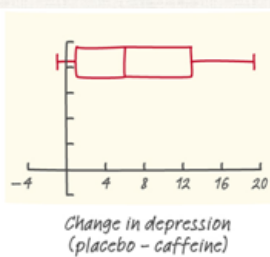
Apr 12-11:03 AM

Example: Paired data and one-sample t procedures

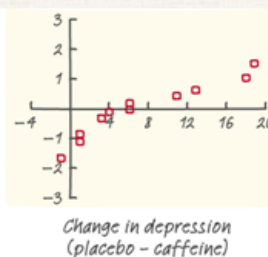
- Normal/Large Sample: We don't know whether the actual distribution of difference in depression scores (placebo - caffeine) is Normal. With such a small sample size ($n = 11$), we need to examine the data to see if it's safe to use t procedures.



(a)



(b)



(c)

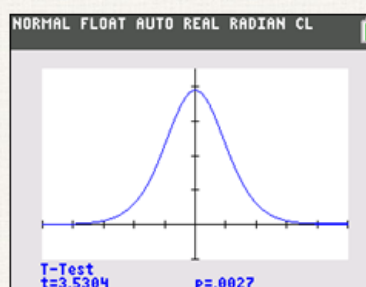
The histogram has an irregular shape with so few values; the boxplot shows some right-skewness but not outliers; and the Normal probability plot is slightly curved, indicating some slight skewness. With no outliers or strong skewness, the t procedures should be pretty accurate.

Apr 12-11:04 AM

Example: Paired data and one-sample t procedures

Do: We entered the differences in list1 and then used the calculator's t -test with "Draw" command,

- Test statistic $t = 3.53$
- P -value 0.0027, which is the area to the right of $t = 3.53$ on the t distribution curve with $df = 11 - 1 = 10$.



Conclude: (a) Because the P -value is 0.0027, we reject $H_0: \mu_d = 0$.

We have convincing evidence that the true mean difference (placebo - caffeine) in depression score is positive for subjects like these.

Apr 12-11:04 AM

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 a regular lane. To collect their data, they randomly selected 15 times during a week, went to the same store, and bought the same item. One of the students used the express lane and the other used the closest regular lane. To decide which lane each of them would use, they flipped a coin. They entered the lanes at the same time, paid with the same method, and recorded the time in seconds it took them to complete the transaction.

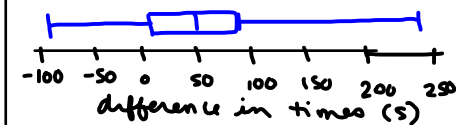
Problem: Carry out a test to see if there is convincing evidence that the express lane is faster, on average.

Paired t -test for μ_d

μ_d = true mean difference (regular - express) in time required to purchase item at this supermarket

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

random ✓ 10% ✓ large counts
 > 150 times to visit store $n < 30$ check graph



$$\bar{x} = 42.7 \quad s_x = 84 \quad t = \frac{42.7 - 0}{84/\sqrt{15}} = 1.966$$

$$df = 14 \quad p\text{-val} = .035 < \alpha = .05$$

Reject H_0

There is convincing evidence the express lane is faster, on avg at this store

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

Apr 13-10:33 AM



CHECK YOUR UNDERSTANDING

Refer to the Data Exploration from Chapter 4 on page 257. Do the data give convincing evidence at the $\alpha = 0.05$ significance level that filling tires with nitrogen instead of air decreases pressure loss?

DATA EXPLORATION Nitrogen in tires—a lot of hot air?

Most automobile tires are inflated with compressed air, which consists of about 78% nitrogen. Aircraft tires are filled with pure nitrogen, which is safer than air in case of fire. Could filling automobile tires with nitrogen improve safety, performance, or both?

Consumers Union designed a study to test whether nitrogen-filled tires would maintain pressure better than air-filled tires. They obtained two tires from each of several brands and then filled one tire in each pair with air and one with nitrogen. All tires were inflated to a pressure of 30 pounds per square inch and then placed outside for a year. At the end of the year, Consumers Union measured the pressure in each tire. The amount of pressure lost (in pounds per square inch) during the year for the air-filled and nitrogen-filled tires of each brand is shown in the table below.³⁰

Brand	Air	Nitrogen	Brand	Air	Nitrogen
BF Goodrich Traction T/A HR	7.6	7.2	Pirelli P6 Four Seasons	4.4	4.2
Bridgestone HP50 (Sears)	3.8	2.5	Sumitomo HTR H4	1.4	2.1
Bridgestone Potenza G009	3.7	1.6	Yokohama Avid H4S	4.3	3.0
Bridgestone Potenza RE950	4.7	1.5	BF Goodrich Traction T/A V	5.5	3.4
Bridgestone Potenza EL400	2.1	1.0	Bridgestone Potenza RE950	4.1	2.8
Continental Premier Contact H	4.9	3.1	Continental ContiExtreme Contact	5.0	3.4
Cooper Lifeline Touring SLE	5.2	3.5	Continental ContiProContact	4.8	3.3
Dayton Daytona HR	3.4	3.2	Cooper Lifeline Touring SLE	3.2	2.5
Falken Ziex ZE-512	4.1	3.3	General Exclaim UHP	6.8	2.7
Fuzion Hri	2.7	2.2	Hankook Ventus V4 H105	3.1	1.4
General Exclaim	3.1	3.4	Michelin Energy MXV4 Plus	2.5	1.5
Goodyear Assurance TripleTred	3.8	3.2	Michelin Pilot Exalto A/S	6.6	2.2
Hankook Optimo H418	3.0	0.9	Michelin Pilot HX MXM4	2.2	2.0
Kumho Solus KH16	6.2	3.4	Pirelli P6 Four Seasons	2.5	2.7
Michelin Energy MXV4 Plus	2.0	1.8	Sumitomo HTR ⁺	4.4	3.7
Michelin Pilot XGT H4	1.1	0.7			

Does filling tires with nitrogen instead of compressed air reduce pressure loss? Give appropriate graphical and numerical evidence to support your answer.

Apr 13-8:24 AM

Using Tests Wisely

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.

How Large a Sample Do I Need?

- A smaller significance level requires stronger evidence to reject the null hypothesis.
- Higher power gives a better chance of detecting a difference when it really exists.
- At any significance level and desired power, detecting a small difference between the null and alternative parameter values requires a larger sample than detecting a large difference.

Computer-Based Learning

The superintendent of a large school district wants to buy a computer-based math program for his district. Because the program is very expensive, he will randomly select some classes to pilot test the program. The response variable will be the average increase in grade-level equivalent for the students in the class. For example, if a student went from 3.1 (third grade, 1 month) to 4.3 (fourth grade, 3 months) by the end of the year, the student increased 1.2 grade levels during the year. The hypotheses that the superintendent will test are: $H_0: \mu = 1$ vs. $H_a: \mu > 1$, where μ = the mean increase in grade-level equivalent for classes in the district. To determine the number of classes he should use, the superintendent considers the following:

- Significance level: The superintendent wants to avoid a Type I error (and spending money on a program that doesn't help), so he chooses $\alpha = 0.01$.
- Effect size: A mean increase of one month (1.1) would be important to detect.
- Power: The superintendent wants a probability of at least 0.80 that a test will detect an increase of one month in grade-level equivalent.

Apr 12-11:04 AM

The following Activity gives you a chance to investigate the sample size needed to achieve a power of 0.9 in the bone mineral content study.

ACTIVITY: Investigating Power

In this Activity, you will use the *Statistical Power* applet at the book's Web site to determine the sample size needed for the exercise study of the previous example. Based on the results of a previous study, researchers are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the 6-month period. We'll start by seeing whether or not 25 subjects are enough.



1. Go to www.whfreeman.com/tps5e and launch the *Statistical Power* applet. Enter the values: $H_0: \mu = 0$, $H_a: \mu > 0$, $\sigma = 2$, $n = 25$, $\alpha = 0.05$, and alternate $\mu = 1$. Then click "Update." What is the power? As a class, discuss what this number means in simple terms.
2. Change the significance level to 0.01. What effect does this have on the power of the test to detect $\mu = 1$? Why?
3. The researchers decide that they are willing to risk a 5% chance of making a Type I error. Change the significance level back to $\alpha = 0.05$. Now increase the sample size to 30. What happens to the power? Why?
4. Keep increasing the sample size until the power is at least 0.90. What minimum sample size should the researchers use for their study?
5. Would the researchers need a smaller or a larger sample size to detect a mean increase of 1.5% in TBBMC? A 0.85% increase? Use the applet to investigate.
6. Summarize what you have learned about how significance level, effect size, and power influence the sample size needed for a significance test.

Apr 13-8:26 AM

Using Tests Wisely

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.

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.

Apr 12-11:04 AM

Improving SAT scores

A national chain of SAT-preparation schools wants to know if using a smartphone app in addition to their regular program will help increase student scores more than using just the regular program. On average, the students in their regular program increase their scores by 128 points during the three-month class. To investigate using the smartphone app, they have 5000 students use the app along with the regular program and measure their improvement. Then they will test the following hypotheses: $H_0: \mu = 128$ vs. $H_a: \mu > 128$, where μ is the true mean improvement in SAT score for students at this school.

After three months, the average improvement was $\bar{x} = 130$ with a standard deviation of $s_x = 65$. The test statistic is $t = 2.18$ with a P -value of 0.0148. The increase was statistically significant, but not practically important. An increase of only 2 points on the SAT isn't a very big deal.

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 that we will make at least 1 Type I error in these 20 tests. So finding 1 significant result in 20 is definitely not a surprise.

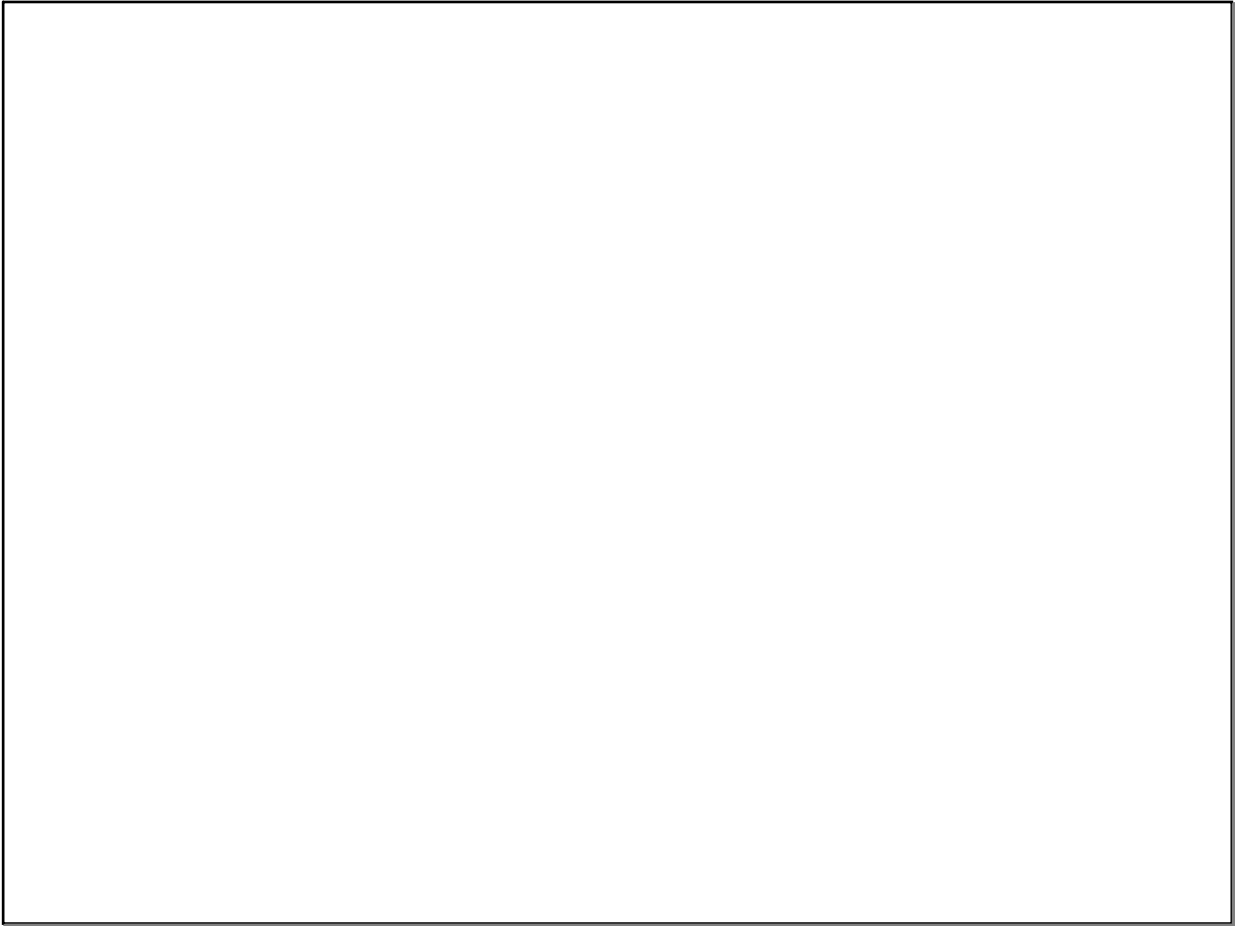
To avoid this problem, we can adjust the significance level for each individual test so that the cumulative probability of avoiding Type I error is 0.95. Using the same logic as the last calculation, we should solve the following for :

$$(1 - \alpha)(1 - \alpha)(1 - \alpha) = (1 - \alpha)^{20} = 0.95$$

$$\alpha = 1 - \sqrt[20]{0.95} = 0.00256$$

Thus, if we plan to do 20 tests, use a significance level of $\alpha = 0.00256$ to get the probability of at least one Type I error to be 0.05

Apr 13-10:37 AM



Mar 12-1:06 PM