STA258H5

Al Nosedal
and Alison Weir

Winter 2017

CONFIDENCE INTERVALS

- For one and two means
- For one and two proportions

Large-Sample Confidence interval for μ

Parameter : μ.

Confidence interval :

$$\bar{Y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Valid if

1. Random sample
2. Independent and identically distributed observations
3. n is large enough for CLT to apply

Interval Estimate of p

Draw a simple random sample of size n from a population with unknown proportion p of successes. An (approximate) confidence interval for p is:

$$\hat{p} \pm z_* \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

where z_* is a number coming from the Standard Normal that depends on the confidence level required.
Use this interval only when:

1. random sample
2. independent and identically distributed Bernoulli trials
3. n is "large" and
4. $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

Large-sample confidence interval for comparing two proportions

Draw an SRS of size n_1 from a population having proportion p_1 of successes and draw an independent SRS of size n_2 from another population having proportion p_2 of successes. When n_1 and n_2 are large, an approximate level C confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE$$

In this formula the standard error SE of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and z^* is the critical value for the standard Normal density curve with area C between $-z^*$ and z^*.

Small-Sample Confidence interval for μ

Parameter : μ.

Confidence interval ($\nu =$ df) :

$$\bar{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right), \qquad \nu = n - 1.$$

Valid if:

1. Random sample
2. independent and identically distributed observations
3. population **must** have Normal distribution (CLT does not apply)

Standard Error

When the standard deviation of a statistic is estimated from data, the result is called the *standard error* of the statistic. The standard error of the sample mean $\bar{x}$ is $\frac{s}{\sqrt{n}}$.

Conditions for inference comparing two means

- We have two SRSs, from two distinct populations. The samples are independent. That is, one sample has no influence on the other. Matching violates independence, for example. We measure the same response variable for both samples.
- Both populations are Normally distributed. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

Small-Sample Confidence interval for $\mu_1 - \mu_2$ (with equal variances)

Parameter : $\mu_1 - \mu_2$.

Confidence interval ($\nu =$ df) :

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
(requires that Normal samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$).

Small-Sample Confidence interval for $\mu_1 - \mu_2$ (unequal variances)

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1, and draw and independent SRS of size n_2 from another Normal population with unknown mean μ_2. A confidence interval for $\mu_1 - \mu_2$ is given by

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

Here t^* is the critical value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The degrees of freedom k are equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Degrees of freedom (Option 1)

Option 1. With software, use the statistic t with accurate critical values from the approximating t distribution.
The distribution of the two-sample t statistic is very close to the t distribution with degrees of freedom df given by

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

This approximation is accurate when both sample sizes n_1 and n_2 are 5 or larger.

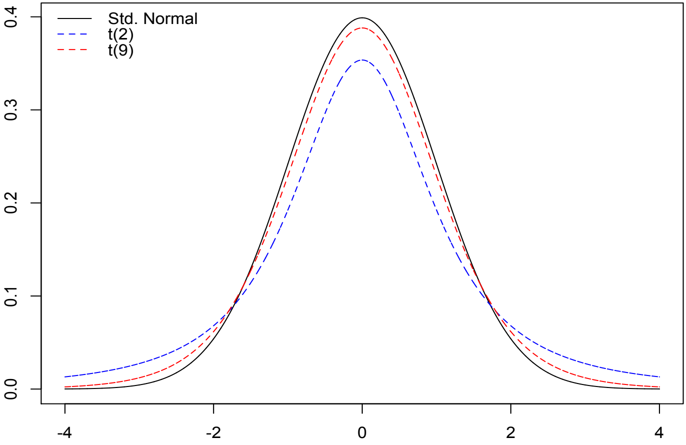
Degrees of freedom (Option2)

Option 2. Without software, use the statistic t with critical values from the t distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. These procedures are always conservative for any two Normal populations.

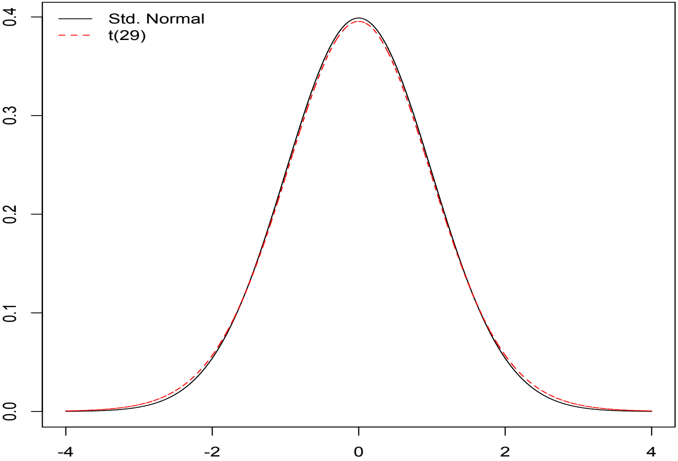
The t distributions

- The density curves of the t distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the t distributions is a bit greater than of the Standard Normal distribution. The t distributions have more probability in the tails and less in the center than does the Standard Normal. This is true because substituting the estimate s for the fixed parameter σ introduces more variation into the statistic.
- As the degrees of freedom increase, the t density curve approaches the $N(0,1)$ curve ever more closely. This happens because s estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.

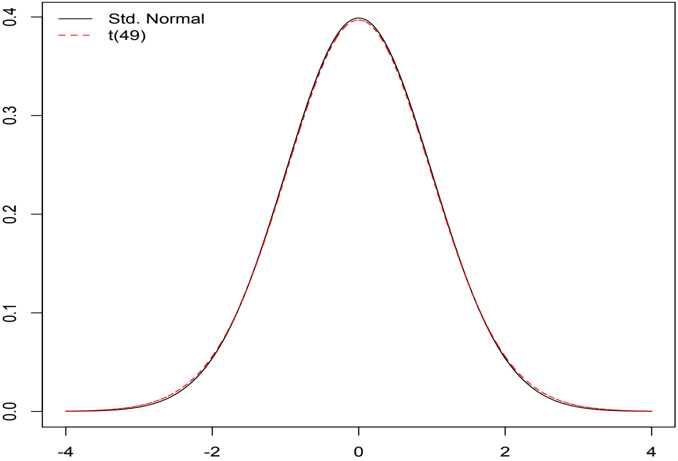
Density curves



Density curves



Density curves



Example: Ancient air

The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:
63.4 65 64.4 63.3 54.8 64.5 60.8 49.1 51.0
Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere. Use a 90% confidence interval to estimate the mean percent of nitrogen in ancient air (Our present-day atmosphere is about 78.1% nitrogen).

Solution

μ = mean percent of nitrogen in ancient air. We will estimate μ with a 90% confidence interval.
With $\bar{x} = 59.5888$, $s = 6.2552$, and $t^* = 1.860$ ($df = 9 - 1 = 8$), the 90% confidence interval for μ is

$$59.5888 \pm 1.860 \left(\frac{6.2552}{\sqrt{9}} \right)$$

$$59.5888 \pm 3.8782$$

$$55.7106 \text{ to } 63.4670$$

```
# Step 1. Entering data;

nitrogen=c(63.4 ,65,64.4,63.3,54.8,
64.5,60.8,49.1,51.0);

# Step 2. Constructing CI;

t.test(nitrogen,conf.level=0.90);
```

```
##
##  One Sample t-test
##
## data:  nitrogen
## t = 28.5785, df = 8, p-value = 2.43e-09
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  55.71155 63.46622
## sample estimates:
## mean of x
##  59.58889
```

Logging in the rain forest

"Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning". These words begin a report on a statistical study of the effects of logging in Borneo. Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:
Unlogged: 22 18 22 20 15 21 13 13 19 13 19 15
Logged : 17 4 18 14 18 15 15 10 12
Use the data to give a 99% confidence interval for the difference in mean number of species between unlogged and logged plots.

Solution

1. Find $\bar{x}_1 - \bar{x}_2$; $\bar{x}_1 - \bar{x}_2 = 17.5 - 13.6666 = 3.8334$
2. Find SE = Standard Error. We have that: $s_1 = 3.5290$, $s_2 = 4.5$, $n_1 = 12$ and $n_2 = 9$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3.5290^2}{12} + \frac{4.5^2}{9}} = 1.8132$$

3. Find $m = t^* SE$. From Table, we have $df = 8$ and 99% confidence level, then $t^* = 3.355$. Hence, $m = (3.355)(1.8132) = 6.0832$.
4. Find Confidence Interval.

$\bar{x}_1 - \bar{x}_2 \pm t^* SE = 3.8334 \pm 6.0832 = -2.2498$ to 9.9166.

R gives a 99% confidence interval of -1.520400 to 9.187067 species, using $df = 14.793$.

```
# Step 1. Entering data;

unlogged=c(22,18,22,20,15,21,13,13,19,13,19,15);

logged=c(17,4,18,14,18,15,15,10,12);

# Step 2. Confidence Interval;

t.test(unlogged,logged,conf.level=0.99);
```

```
##
##  Welch Two Sample t-test
##
## data:  unlogged and logged
## t = 2.1141, df = 14.793, p-value = 0.05192
## alternative hypothesis: true difference in means is not equ
## 99 percent confidence interval:
##  -1.520400  9.187067
## sample estimates:
## mean of x mean of y
##  17.50000  13.66667
```

Example: Direct and Broker-Purchased Mutual Funds

Millions of investors buy mutual funds, choosing from thousands of possibilities. Some funds can be purchased directly from banks or other financial institutions whereas others must be purchased through brokers, who charge a fee for this service. This raises the question, Can investors do better by buying mutual funds directly than by purchasing mutual funds through brokers? To help answer this question, a group of researchers randomly sampled the annual returns from mutual funds that can be acquired directly and mutual funds that are bought through brokers and recorded the net annual returns, which are the returns on investment after deducting all relevant fees.

Example: Direct and Broker-Purchased Mutual Funds (cont.)

From the data, the following statistics were calculated:

$n_1 = 50$
$n_2 = 50$
$\bar{x}_1 = 6.63$
$\bar{x}_2 = 3.72$
$s_1^2 = 37.49$
$s_2^2 = 43.34$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The pooled variance estimator is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(49)37.49 + (49)43.34}{50 + 50 - 2} = 40.42$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The number of degrees of freedom of the test statistic is

$$\nu = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The confidence interval estimator of the difference between two means with equal population variance is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

or

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The 95% confidence interval estimate of the difference between the return for directly purchased mutual funds and the mean return for broker-purchased mutual funds is

$$(6.63 - 3.72) \pm 1.984\sqrt{40.42\left(\frac{1}{50} + \frac{1}{50}\right)}.$$

$$2.91 \pm 2.52.$$

The lower and upper limits are 0.39 and 5.43.

Example: Direct and Broker-Purchased Mutual Funds (cont.)

We estimate that the return on directly purchased mutual funds is on average between 0.38 and 5.43 percentage points larger than broker-purchased mutual funds.

Problem

A survey of 611 office workers investigated telephone answering practices, including how often each office worker was able to answer incoming telephone calls and how often incoming telephone calls went directly to voice mail. A total of 281 office workers indicated that they never need voice mail and are able to take every telephone call.

a. What is the point estimate of the proportion of the population of office workers who are able to take every telephone call?

b. At 90% confidence, what is the margin of error?

c. What is the 90% confidence interval for the proportion of the population of office workers who are able to take every telephone call?

Solution

a. $\hat{p} = \frac{281}{611} = 0.46$

b. Margin of error =

$z_* \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1.65 \sqrt{\frac{(0.46)(0.54)}{611}} = 1.65(0.0201) = 0.0332$

c. $\hat{p} \pm z_* \left(\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \right)$

$0.46 \pm .0332$

$(0.4268, 0.4932)$

R Code

```
prop.test(281,611,conf.level=0.90);

##
##  1-sample proportions test with continuity correction
##
## data:  281 out of 611, null probability 0.5
## X-squared = 3.7709, df = 1, p-value = 0.05215
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.4261763 0.4939896
## sample estimates:
##         p
## 0.4599018
```

Example

From Journal for the Scientific Study of Religion
Researchers classified 434 LA residents by (a) ethnicity and (b) whether the person has felt the presence of someone who has died. Counts are in the table below.

Ethnicity	Felt	Count
Black	Yes	62
Black	No	47
Japanese	Yes	32
Japanese	No	78
Mexican	Yes	61
Mexican	No	53
White	Yes	38
White	No	63

Construct a 90% confidence interval for the proportion of Japanese who have felt the presence of someone who has died.

R Code

```
prop.test(x=32, n= 32+78, conf.level=0.90, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  32 out of 32 + 78, null probability 0.5
## X-squared = 19.2364, df = 1, p-value = 1.155e-05
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.2253802 0.3664766
## sample estimates:
##         p
## 0.2909091
```

Example: How to quit smoking

Nicotine patches are often used to help smokers quit. Does giving medicine to fight depression help? A randomized double-blind experiment assigned 244 smokers who wanted to stop to receive nicotine patches and another 245 to receive both a patch and the antidepression drug bupropion. Results: After a year, 40 subjects in the nicotine patch group had abstained from smoking, as had 87 in the patch-plus-drug group. Give a 99% confidence interval for the difference (treatment minus control) in the proportion of smokers who quit.

Solution

$\hat{p}_1 = \frac{40}{244} \approx 0.1639$ and $\hat{p}_2 = \frac{87}{245} \approx 0.3551$.
The standard error is
$SE = \sqrt{\frac{(0.1639)(1-0.1639)}{244} + \frac{(0.3551)(1-0.3551)}{245}} \approx 0.0387$.
The 99% confidence interval is:

$$(0.3551 - 0.1639) \pm 2.576(0.0387)$$

Lower Confidence Limit $= 0.1912 - 0.0996 = 0.0915$
Upper Confidence Limit $= 0.1912 + 0.0996 = 0.2908$

R Code

```r
successes=c(87, 40);

totals=c(245, 244);

prop.test(successes,totals, conf.level=0.99,
correct=FALSE);
```

R Code

```
##
##  2-sample test for equality of proportions without continui
##  correction
##
## data:  successes out of totals
## X-squared = 23.2371, df = 1, p-value = 1.432e-06
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.09152484 0.29081039
## sample estimates:
##    prop 1    prop 2
## 0.3551020 0.1639344
```

R Code

```r
successes=c(40,87);

totals=c(244,245);

prop.test(successes,totals, conf.level=0.99,
correct=FALSE);
```

R Code

```
##
##  2-sample test for equality of proportions without continui
##  correction
##
## data:  successes out of totals
## X-squared = 23.2371, df = 1, p-value = 1.432e-06
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.29081039 -0.09152484
## sample estimates:
##    prop 1    prop 2
## 0.1639344 0.3551020
```

Homework?

The operations manager of a production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. After observing 120 workers assembling similar devices, she noticed that their average time was 16.2 minutes (with standard deviation 3.6 minutes). Construct a 92% confidence interval for the mean assembly time. State all necessary assumptions.

Tax Collected from Audited Returns

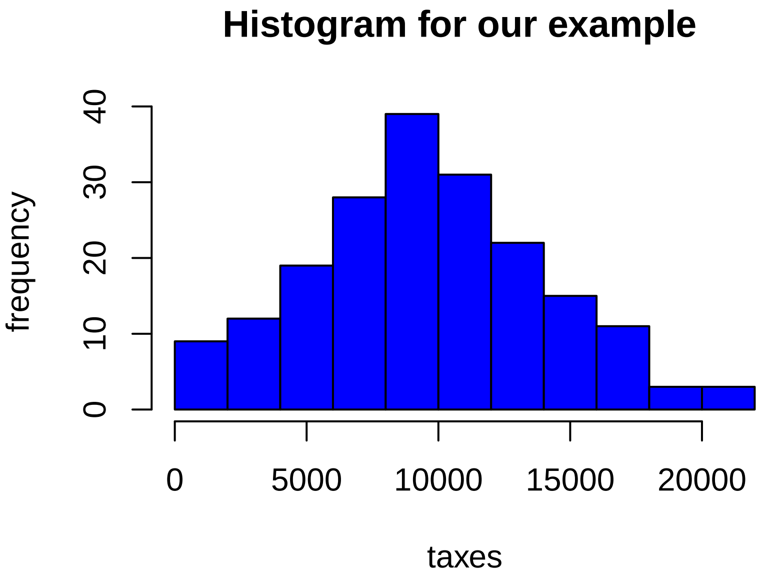
In 2010, 142,823,000 tax returns were filed in the United States. The Internal Revenue Service (IRS) examined 1.107%, or 1,581,000, of them to determine if they were correctly done. To determine how well the auditors are performing, a random sample of these returns was drawn and the additional tax was reported, see file taxes.csv. Estimate with 95% confidence the mean additional income tax collected from the 1,581,000 files audited.

Reading our data

```
# Step 1. Entering data;

taxes_file<-read.csv(file="taxes.csv",header=TRUE)

names(taxes_file);

## [1] "Taxes"

taxes = taxes_file$Taxes;
```

Histogram

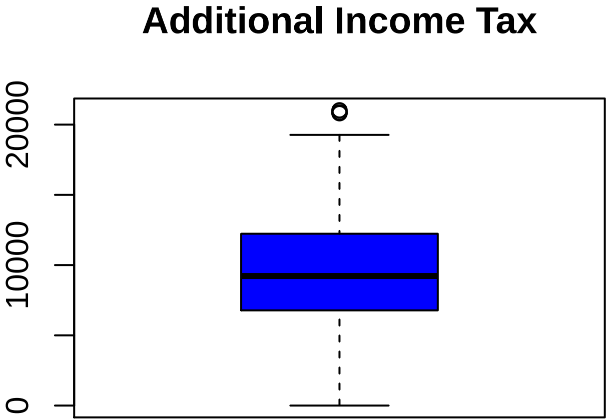
```r
hist(taxes,
main="Histogram for our example ",
xlab="taxes", ylab="frequency",
col="blue");
```


Histogram for our example
frequency
0
10
20
30
40
0
5000
10000
15000
20000
taxes

Boxplot

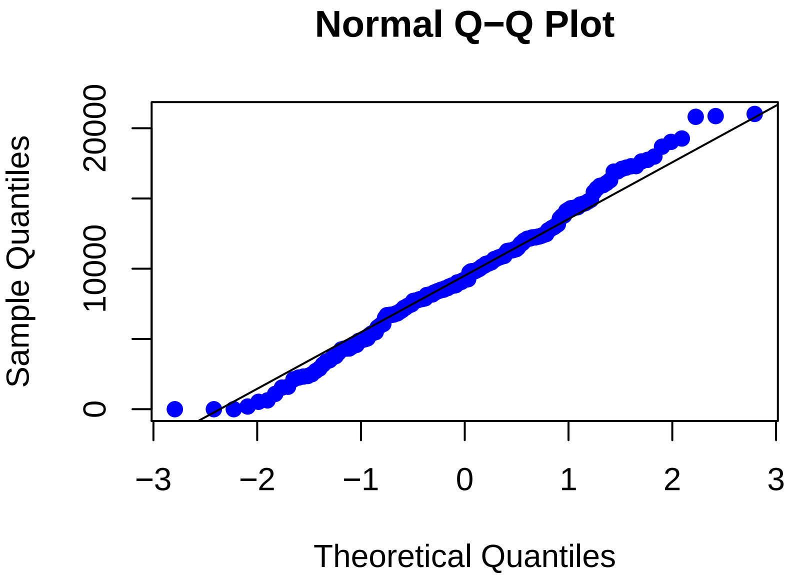
```
boxplot(taxes,
main="Additional Income Tax",
col="blue");
```

Boxplot



R Code

```r
## Q-Q plot (using R function);

qqnorm(taxes,col="blue",pch=19);
qqline(taxes);
```


Normal Q−Q Plot
Sample Quantiles
0
10000
20000
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Finding CI

```r
# Step 2. Constructing CI;

t.test(taxes,conf.level=0.95);
```

Finding CI

```
##
##  One Sample t-test
##
## data:  taxes
## t = 29.3449, df = 191, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   8886.932 10167.721
## sample estimates:
## mean of x
##  9527.326
```

We estimate that the mean additional tax collected lies between \$8,887 and \$10,168 (with 95% confidence).

A few final comments

When we introduced the Student t-distribution, we pointed out that the t-statistic is Student t-distributed if the population from which we've sampled is Normal. However, statisticians have shown that the mathematical process that derived the Student t-distribution is **robust**, which means that if the population is non-Normal, the results of the confidence interval estimate are still valid provided that the population is **not extremely non-Normal**. Our histogram, boxplot, and Q-Q plot suggest that our variable of interest is not extremely non-Normal, and in fact, may be Normal.