Confidence Intervals

Gene Quinn

Confidence Intervals

In our prior work with confidence intervals, we assumed the population standard deviation σ was known.

Usually σ is known in one of two situations:

- We are dealing with a *standardized* measure such as IQ scores or SAT scores

Confidence Intervals

In our prior work with confidence intervals, we assumed the population standard deviation σ was known.

Usually σ is known in one of two situations:

- We are dealing with a *standardized* measure such as IQ scores or SAT scores
- A very reliable estimate of σ is available from some source such as a census of the population

Confidence Intervals

In our prior work with confidence intervals, we assumed the population standard deviation σ was known.

Usually σ is known in one of two situations:

- We are dealing with a *standardized* measure such as IQ scores or SAT scores
- A very reliable estimate of σ is available from some source such as a census of the population

Most of the time, neither of these is true. What should be done when σ is **not** known?

Confidence Intervals

In our prior work with confidence intervals, we assumed the population standard deviation σ was known.

Usually σ is known in one of two situations:

- We are dealing with a *standardized* measure such as IQ scores or SAT scores
- A very reliable estimate of σ is available from some source such as a census of the population

Most of the time, neither of these is true. What should be done when σ is **not** known?

When the assumption that σ is known is unreasonable, we can use the *sample* standard deviation s to estimate σ.

Confidence Intervals

Unfortunately, we cannot simply substitute s everywhere we find σ in our confidence interval formulas.

The statistic

$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

will have a normal or bell curve distribution if the underlying population has a normal distribution

Confidence Intervals

Unfortunately, we cannot simply substitute s everywhere we find σ in our confidence interval formulas.

The statistic

$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

will have a normal or bell curve distribution if the underlying population has a normal distribution

If the sample size is large, say 30 or more, the central limit theorem states that this statistic will have an *approximately* normal distribution.

Confidence Intervals

However, the statistic

$$\frac{\overline{x} - \mu}{s/\sqrt{n}}$$

does not have a normal distribution.

Confidence Intervals

However, the statistic

$$\frac{\overline{x} - \mu}{s/\sqrt{n}}$$

does not have a normal distribution.

The correct distribution was discovered in 1904 by William Gosset, an employee of the Guiness Brewery in Dublin.

Confidence Intervals

However, the statistic

$$\frac{\overline{x} - \mu}{s/\sqrt{n}}$$

does not have a normal distribution.

The correct distribution was discovered in 1904 by William Gosset, an employee of the Guiness Brewery in Dublin.

Forbidden to publish by company regulations, Gosset published his result under the pseudonym *student*

Confidence Intervals

However, the statistic

$$\frac{\overline{x} - \mu}{s/\sqrt{n}}$$

does not have a normal distribution.

The correct distribution was discovered in 1904 by William Gosset, an employee of the Guiness Brewery in Dublin.

Forbidden to publish by company regulations, Gosset published his result under the pseudonym *student*

As a result, the distribution became known as *Student's t-distribution*

The t Distribution

Actually the student's t-distribution (also known simply as the t distribution) is quite similiar to the normal distribution.

The t Distribution

Actually the student's t-distribution (also known simply as the t distribution) is quite similiar to the normal distribution.

In fact, they become indistinguishable as the sample size n becomes large.

The t Distribution

Actually the student's t-distribution (also known simply as the t distribution) is quite similiar to the normal distribution.

In fact, they become indistinguishable as the sample size n becomes large.

If you look at the last row of the t distribution table in the text, it is labeled "z" and matches the normal distribution.

The t Distribution

Actually the student's t-distribution (also known simply as the t distribution) is quite similiar to the normal distribution.

In fact, they become indistinguishable as the sample size n becomes large.

If you look at the last row of the t distribution table in the text, it is labeled "z" and matches the normal distribution.

In some versions of this table, the last row is labeled ∞ indicating that this is the limit as the sample size becomes large without bound.

The t Distribution

Actually the student's t-distribution (also known simply as the t distribution) is quite similiar to the normal distribution.

In fact, they become indistinguishable as the sample size n becomes large.

If you look at the last row of the t distribution table in the text, it is labeled "z" and matches the normal distribution.

In some versions of this table, the last row is labeled ∞ indicating that this is the limit as the sample size becomes large without bound.

When n is small, say < 30, the t distribution has slightly higher dispersion than the normal distribution.

Computing t Values

As with normal or z values, we can use a spreadsheet to compute values of the t distribution corresponding to a given α value.

Computing t Values

As with normal or z values, we can use a spreadsheet to compute values of the t distribution corresponding to a given α value.

The appropriate spreadsheet function is called $TINV$, and the value we are usually interested in is:

$$t_{\alpha/2} = TINV(\alpha, n-1)$$

where $1-\alpha$ is the level of confidence we want an n is the sample size.

Computing t Values

$$t_{\alpha/2} = TINV(\alpha, n-1)$$

Notice that the second argument to $TINV$ is $n-1$, one *less* than the sample size.

Computing t Values

$$t_{\alpha/2} = TINV(\alpha, n-1)$$

Notice that the second argument to $TINV$ is $n-1$, one *less* than the sample size.

Notice also that we do not divide α by two as we did with the $NORMSINV$ function.

Computing t Values

$$t_{\alpha/2} = TINV(\alpha, n-1)$$

Notice that the second argument to $TINV$ is $n-1$, one *less* than the sample size.

Notice also that we do not divide α by two as we did with the $NORMSINV$ function.

There is no theoretical reason for this, the spreadsheet designers just implemented the $TINV$ function differently from $NORMSINV$

Confidence Intervals

The lower bound of the confidence interval when σ is *unknown* is:

$$\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Confidence Intervals

The lower bound of the confidence interval when σ is *unknown* is:

$$\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The upper bound of the confidence interval is:

$$\overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Confidence Intervals

The lower bound of the confidence interval when σ is *unknown* is:

$$\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The upper bound of the confidence interval is:

$$\overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

As before the constant $t_{\alpha/2}$ is determined by the level of confidence $1 - \alpha$

Confidence Intervals

The lower bound in terms of spreadsheet functions is

$$\overline{x} - TINV(\alpha, n-1) \cdot \frac{s}{\sqrt{n}}$$

Confidence Intervals

The lower bound in terms of spreadsheet functions is

$$\overline{x} - TINV(\alpha, n-1) \cdot \frac{s}{\sqrt{n}}$$

The upper bound of the confidence interval is:

$$\overline{x} + TINV(\alpha, n-1) \cdot \frac{s}{\sqrt{n}}$$

Confidence Intervals for Proportions

One of the most common applications of confidence intervals involve estimates of population *proportions*

Confidence Intervals for Proportions

One of the most common applications of confidence intervals involve estimates of population *proportions*

Recall that we can often approximate a binomial random variable with a normal distribution.

Confidence Intervals for Proportions

One of the most common applications of confidence intervals involve estimates of population *proportions*

Recall that we can often approximate a binomial random variable with a normal distribution.

Using the normal approximation to the binomial, we can modify the results for confidence intervals for means to obtain confidence intervals for the population proportion.

These results assume $n\hat{p}(1-\hat{p}) \geq 10$ and the sample represents no more than 5% of the population.

Confidence Intervals for Proportions

The point estimate $\hat{p}$ of a population proportion p,

$$\hat{p} = \frac{x}{n}$$

will have a normal or bell curve distribution with

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

provided $np(1-p) \geq 10$.

Confidence Intervals for Proportions

The lower bound of the confidence interval for a *proportion* p is:

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Intervals for Proportions

The lower bound of the confidence interval for a *proportion* p is:

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The upper bound of the confidence interval is:

$$\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margin of Error

The **margin of error** associated with a $1-\alpha$ level confidence interval for a proportion p is

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = NORMSINV(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margin of Error

The **margin of error** associated with a $1-\alpha$ level confidence interval for a proportion p is

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = NORMSINV(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

As before the margin of error is $1/2$ the width of the confidence interval for p

Sample Size

To get a specified margin of error E, the required sample size is:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

Sample Size

To get a specified margin of error E, the required sample size is:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

In terms of a spreadsheet formula, the sample size is:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{NORMSINV(1 - \alpha/2)}{E} \right)^2$$

Sample Size

To get a specified margin of error E, the required sample size is:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

In terms of a spreadsheet formula, the sample size is:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{NORMSINV(1 - \alpha/2)}{E} \right)^2$$

In this formula, $\hat{p}$ represents an estimate of p from some other source, not from the sample which is still being planned.

Sample Size

Example: We want to estimate the proportion of students who graduate from high school in a certain state.

What size sample is required if the proportion is approximately $.75$ and the we want the estimate to be within 3% of p with 95% confidence?

Sample Size

Example: We want to estimate the proportion of students who graduate from high school in a certain state.

What size sample is required if the proportion is approximately $.75$ and the we want the estimate to be within 3% of p with 95% confidence?

Set $\hat{p} = .75$ and $\alpha = .05$. Then

$$n = (.75)(.25)\left(\frac{NORMSINV(0.975)}{0.03}\right)^2 = 800$$

Sample Size

It should be pointed out that the maximum required sample size occurs when $\hat{p} = 1/2$.

So, absent a reliable extimate of p, one can always assume the worst case of $p = 0.5$.

Sample Size

It should be pointed out that the maximum required sample size occurs when $\hat{p} = 1/2$.

So, absent a reliable extimate of p, one can always assume the worst case of $p = 0.5$.

The down side of this approach is that it may produce a much larger sample size than actually required.