Statistics

Marco Caserta
marco.caserta@ie.edu

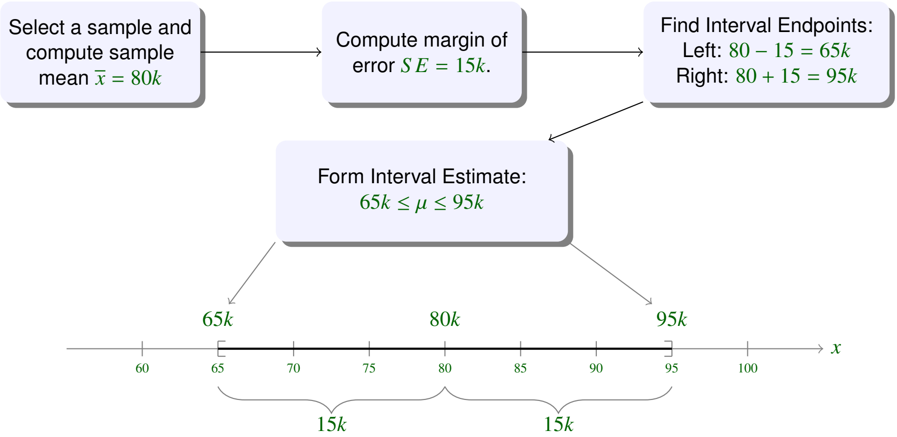
IE University

1 Confidence intervals

1 Confidence intervals

Where Are We Going?

- Sampling and Inferential Statistics: Imagine you want to know what is the expected salary of a graduate at IEU. How would you approach the problem?



Select a sample and compute sample mean $\overline{x} = 80k$

Compute margin of error $SE = 15k$.

Find Interval Endpoints:
Left: $80 - 15 = 65k$
Right: $80 + 15 = 95k$

Form Interval Estimate:
$65k \leq \mu \leq 95k$

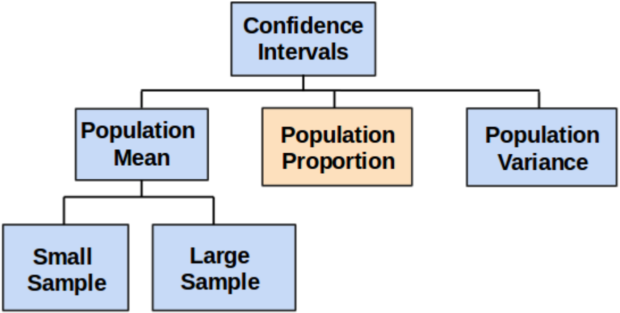
Confidence intervals

- A plausible range of values for the population parameter is called a CONFIDENCE INTERVAL.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.



Confidence Intervals

Population Mean

Population Proportion

Population Variance

Small Sample

Large Sample

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\bar{x} \pm 2 \times SE \quad = \quad 3.2 \pm 2 \times 0.25$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \end{aligned}$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (2.7, 3.7) \end{aligned}$$

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(A) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(B) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(C) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(D) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(A) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(B) *college students on average have been in between 2.7 and 3.7 exclusive relationships.*

(C) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(D) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

A more accurate interval

Confidence interval, a general formula

$$point\ estimate \pm z^{\star} \times SE$$

A more accurate interval

Confidence interval, a general formula

$$point\ estimate \pm z^{\star} \times SE$$

Conditions when the point estimate = $\bar{x}$:

1. Independence: Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, $n <$ 10% of population
2. Sample size / skew: $n \geq 30$ and population distribution should not be extremely skewed

A more accurate interval

Confidence interval, a general formula

$$point\ estimate \pm z^{\star} \times SE$$

Conditions when the point estimate = $\bar{x}$:

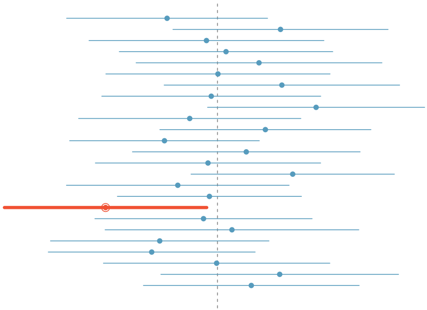
1. Independence: Observations in the sample must be independent
    - random sample/assignment
    - if sampling without replacement, $n <$ 10% of population
2. Sample size / skew: $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ later on.

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $\textit{point estimate} \pm 2 \times SE$.
- Then about 95% of those intervals would contain the true population mean (μ).
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.

- A point estimate is UNBIASED if the sampling distribution of the estimate is centered at the parameter it estimates (CLT)
- A MINIMUM VARIANCE UNBIASED ESTIMATOR is the unbiased estimator with the smallest variance
- $\bar{x}$ is a minimum variance unbiased estimator for μ



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

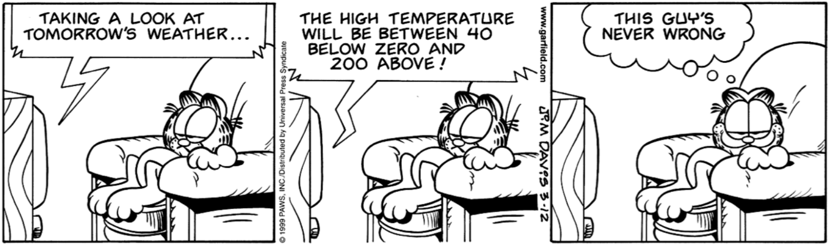
A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

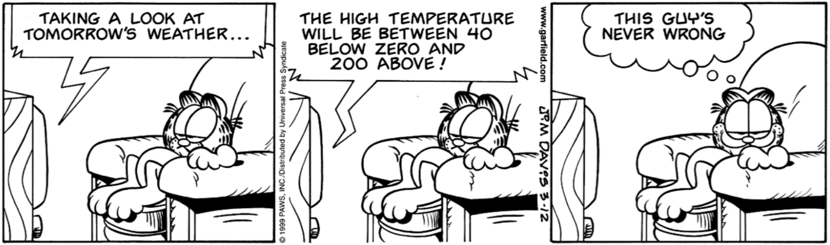

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

If the interval is too wide it may not be very informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

$$\textit{point estimate} \pm z^\star \times SE$$

- In a confidence interval, $z^\star \times SE$ is called the MARGIN OF ERROR, (ME) and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(A) $Z = 2.05$

(B) $Z = 1.96$

(C) $Z = 2.33$

(D) $Z = -2.33$

(E) $Z = -1.65$

Which of the below Z scores is the appropriate $z^{\star}$ when calculating a 98% confidence interval?

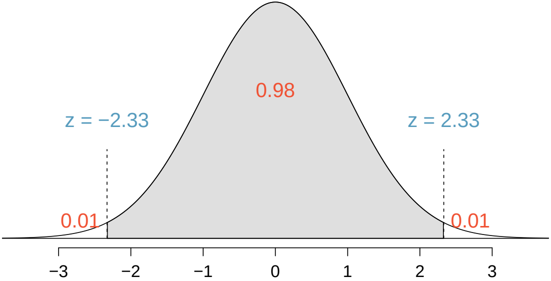
(A) $Z = 2.05$

(B) $Z = 1.96$

(C) $Z = 2.33$

(D) $Z = -2.33$

(E) $Z = -1.65$



Case Study: AirNoShows

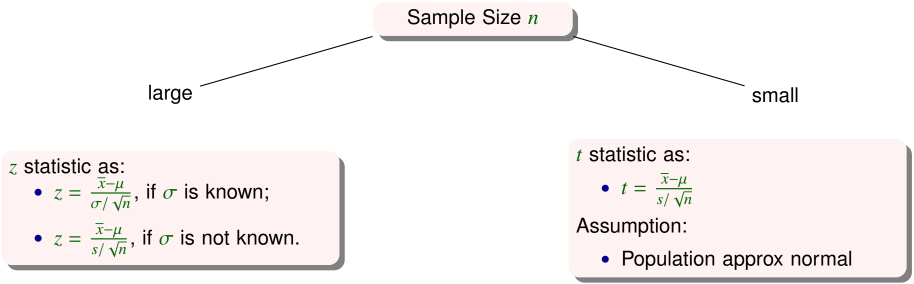
Data File: AIRNOSHOWS.xls

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its average number of unoccupied seats per flight over the past year. To accomplish this, the record of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sample flights.

1. Obtain the most important descriptive statistics using Excel.
2. Estimate the mean number of unoccupied seats per flight during the past year, using a 90% confidence interval.

- Use descriptive statistics to gain a quick overview of the data $\bar{x} = 11.6, s = 4.1, s/\sqrt{225} = 0.27$
- Are assumptions for large-sample CI satisfied?
- Manually compute the 90% CI (To obtain $z_{\alpha/2}$: Excel: `NORMSINV(0.95)=1.65`) [11.15, 12.04]
- Verify it with `Excel`: `CONFINDENCE.NORM(`α`, s, n)` $z_{\alpha/2} s/\sqrt{n} = 0.45$
- What does the 90% CI *really* mean?

Confidence Intervals for Population Mean: t-statistic



Sample Size n

large

small

z statistic as:

- $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, if σ is known;
- $z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, if σ is not known.

t statistic as:

- $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

Assumption:

- Population approx normal

Confidence Intervals for Population Mean: t-statistic

When the sample is small, we have two major problems:

1. CLT is not longer valid, since n is small. Thus, we can no longer be sure that the shape of the distribution of $\overline{x}$ is normal. In reality, now the shape of the distribution of $\overline{x}$ depends on the distribution of the population.
   - **How to solve it.** A theorem tells us that if the random sample of size n is selected from a normal population, the distribution of $\overline{x}$ will be normal.
   - We now require the Assumption of Normality of the Population.
2. The st.dev. is not known and, often, s provides a "poor" approximation for σ when n is small.
   - **How to solve it.** Rather than using $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$, we now use $t = \frac{\overline{x}-\mu}{s/\sqrt{n}}$, and we assume that t follows a t distribution;
   - What is the difference between a normal distribution and a t distribution?

Nonparametric Statistics

If the assumption of normality of the population does not hold, use a nonparametrical statistical test.

THE *t* DISTRIBUTION

William Gosset (Source: Wikipedia)

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t* DISTRIBUTION.

The t distribution

William Gosset (Source: Wikipedia)

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.

THE t DISTRIBUTION

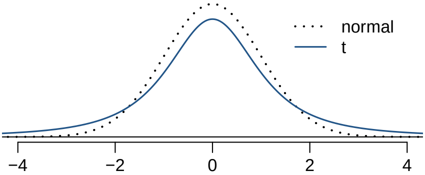
William Gosset (Source: Wikipedia)

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t DISTRIBUTION.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

The t distribution

William Gosset (Source: Wikipedia)

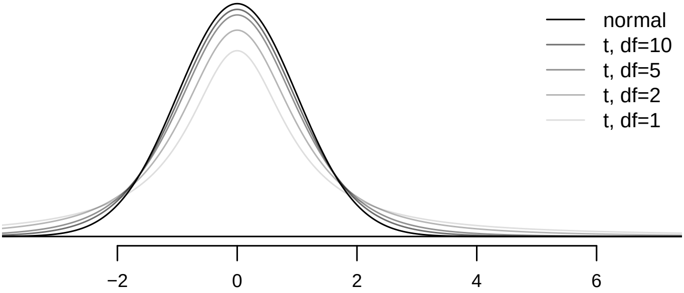
- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t DISTRIBUTION.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)



-

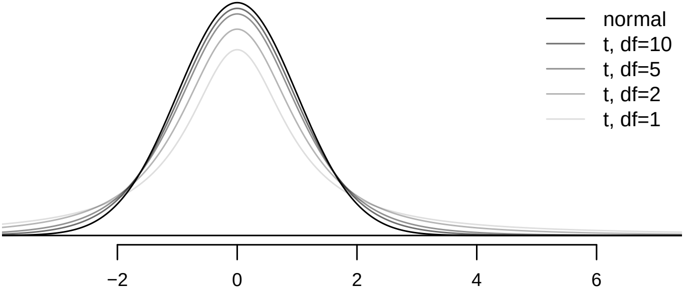
The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: degrees of freedom (df).



The t distribution (cont.)

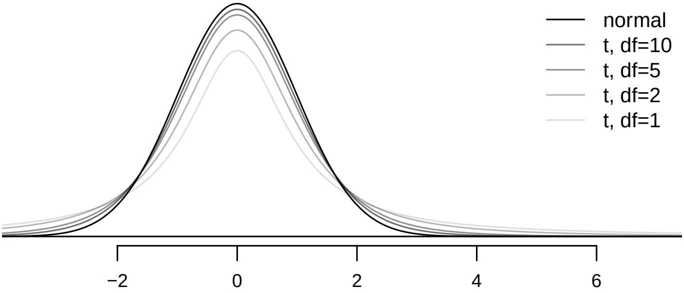
- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: DEGREES OF FREEDOM (df).



What happens to shape of the t distribution as df increases?

The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: degrees of freedom (df).



What happens to shape of the t distribution as df increases?

Approaches normal.

Confidence Intervals for Population Mean: t-statistic

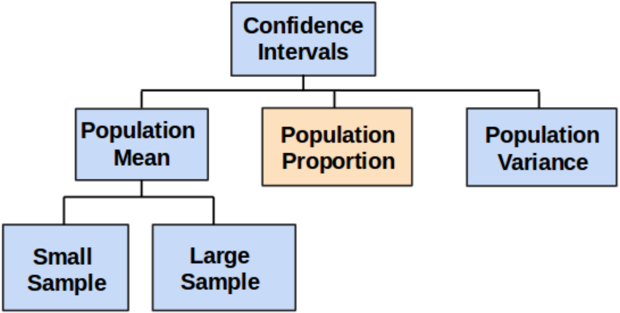
Confidence Interval for Exam Grade

A sample of 12 students from a large group obtain an average of 56.9% in an exam, and the sample variance is 25%. Give a 95% confidence interval for the mean exam grade for the whole group, assuming that exam grades have a normal distribution.

Solution:

- Sample size is small ⇒ use a t distribution
- $(1-\alpha) = 0.95 \Rightarrow \alpha = 0.05$
- To obtain the t value, we use the Excel function `TINV(error, df)`, where `error` is α and `df` is the degrees of freedom (n-1)
- $t_{\alpha/2} = t_{0.025} = 2.201$ `TINV(`α`, df) = TINV(0.05,11)`
- The interval is

$$56.9 \pm 2.201\sqrt{\frac{25}{12}} \Rightarrow (53.72, 60.08)$$


Confidence Intervals
Population Mean
Population Proportion
Population Variance
Small Sample
Large Sample

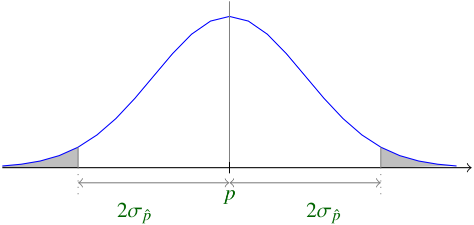
Confidence Interval for Population Proportion

Customers Proportion

We want to estimate the true, *i.e.*, population, proportion of customers buying a certain brand of cereals. We collect a sample of 1000 observations, of which, $x = 313$ bought the specific brand of cereals:

$$\hat{p} = \frac{313}{1000} = 0.313$$

- Consider buying/not buying as an experiment in which buying is “success”
- Statistic $\hat{p}$ represents the average number of successes out of n trials
- CLT tells us how $\hat{p}$ is distributed (*no matter* what the underlying population is), as long as the sample size is sufficiently large



Sampling Distribution of $\hat{p}$

- Mean of sampling distribution of $\hat{p}$ is p
- St.dev. of sampling distribution of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{pq/n}$
- For large samples, $\hat{p}$ follows a normal distribution, as long as:
  - $\hat{p}n \geq 15$, and
  - $\hat{q}n \geq 15$
- Confidence Interval

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{pq}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

Customers Proportion

Build a 95% CI for the mean proportion of customers buying that specific brand of cereals.

Confidence Interval for Population Variance

- We want to build a confidence interval for the population variance σ^2
- The confidence interval is based on the sample variance s^2
- We assume that the population is normally distributed
- The random variable

  $$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

  follows a chi-square distribution with $n-1$ degrees of freedom
- The $(1-\alpha)\%$ confidence interval for the population variance is given by:

$$LCL = \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \tag{1}$$

$$UCL = \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \tag{2}$$

Confidence Interval for Population Variance

Confidence Interval for Speed of Computer Processors

You are testing the speed of a batch of computer processors. You collect the following data (in Mhz):

- sample size = 17
- sample mean = 3004
- sample stdev = 74

Assume the population is normal. Determine the 95% confidence interval for σ^2.

Confidence Interval for Population Variance

Confidence Interval for Speed of Computer Processors

You are testing the speed of a batch of computer processors. You collect the following data (in Mhz):

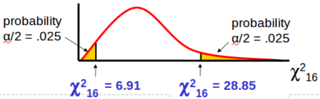
- sample size = 17
- sample mean = 3004
- sample stdev = 74

Assume the population is normal. Determine the 95% confidence interval for σ^2.

- Let us find the χ^2 value, using $df = 16$, and $\alpha = 0.05$

$$\chi^2_{n-1,\alpha/2} = \chi^2_{16,0.025} = 28.85$$

$$\chi^2_{n-1,1-\alpha/2} = \chi^2_{16,0.975} = 6.91$$



- The 95% CI is:

$$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \rightarrow \frac{(17-1)(74)^2}{28.85} < \sigma^2 < \frac{(17-1)(74)^2}{6.91}$$

$$3037 < \sigma^2 < 12680$$

Determining the Sample Size

Sample Size

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

Determining the Sample Size

Sample Size

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

- We know that the critical value associated with the 96% confidence level: $z^\star = 2.05$.

Determining the Sample Size

Sample Size

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

- We know that the critical value associated with the 96% confidence level: $z^{\star} = 2.05$.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

Determining the Sample Size

Sample Size

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

- We know that the critical value associated with the 96% confidence level: $z^\star = 2.05$.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

- The minimum number of children required to attain the desired margin of error is 85.1. Since we can't sample 0.1 of a child, we must sample at least 86 children (round up, since rounding down to 85 would yield a slightly larger margin of error than desired).