1-sample Confidence Intervals (CI)

Module 8

Statistics 251: Statistical Methods

Updated 2020

Introduction

Statistical Inference: using information obtained from a proper sample to make an educated judgment about a population

Three types of inference

(1) point estimation
(2) interval estimation (aka confidence intervals (CIs))
(3) statistical tests (aka hypothesis tests)

Terms I

Point Estimation: a point estimate of a parameter (let's just say a generic parameter is called θ and its estimate (statistic) is called $\hat{\theta}$) θ is a single number that can be regarded as a sensible value for θ. It is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimate** of θ. Examples are: $\overline{X}$ estimates μ, $\hat{\pi}$ estimates π, s estimates σ, and so on.

A point estimate is just a single number and by itself provides no information about the precision and reliability of estimation; it gives no feedback on how close our estimate was to the parameter.

Terms II

Interval estimation: an alternative to reporting a sensible value for the parameter being estimated is to calculate an entire interval of plausible values, called **interval estimation**, specifically we call them **confidence intervals (CIs)**.

Select the level of confidence, it is usually 95% but others are also used often (90%, 98%, 99%). A CI with level 95% implies that 95% of samples would give an interval that contains θ, or that only 5% of samples would not contain θ.

Assumptions

Assumptions: conditions that we need to be true in order for the data to properly fit the model we are using for estimations (1) Independence: observations are independent from one another (2) Randomization: proper randomization was used (takes care of independence issue if there is one) (3) Means need an *approximate* normal distribution (if $n \geq 30$, then it is approximately normal), and proportions need to meet the S/F condition: $np \geq 5$ and $nq \geq 5$ (if $n \geq 60$, S/F condition is met)

If assumptions are violated, the results from the analyses are not as valid nor reliable

General Form

All CIs (even more complex ones) have the same form:

$$point\ estimate \pm bound$$

Where the bound on the error of estimation is $z^\star(se)$ or $t^\star(se)$ and $se_{mean} = \frac{\sigma}{\sqrt{n}}$, $se_\pi = \sqrt{\hat{\pi}(1-\pi)/n}$, or $se_{mean} = \frac{s}{\sqrt{n}}$ (the one you use depends on the situation; explanations to come)

Form for CI on μ when σ is known

$$\overline{X} \pm z^\star(se_{mean})$$

Page 443 and 445 in the textbook show examples on how to do these calculations on TI calculators.

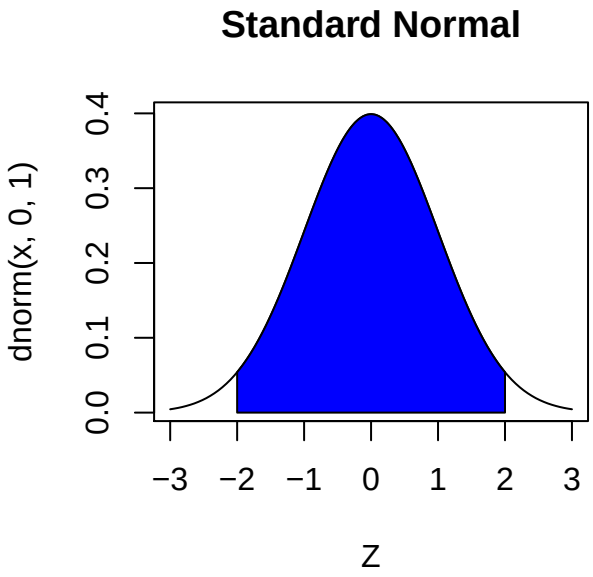
To solve for a sample size given a CL and bound:

$$n = \left(\frac{z^\star \sigma}{bound}\right)^2$$

Unlike normal rounding conventions, we *always round* ***up***

Finding $z^\star$

$z^\star$ is a critical value of z based on the confidence level. $1 - CL = \alpha$ where α is the "left-over" area

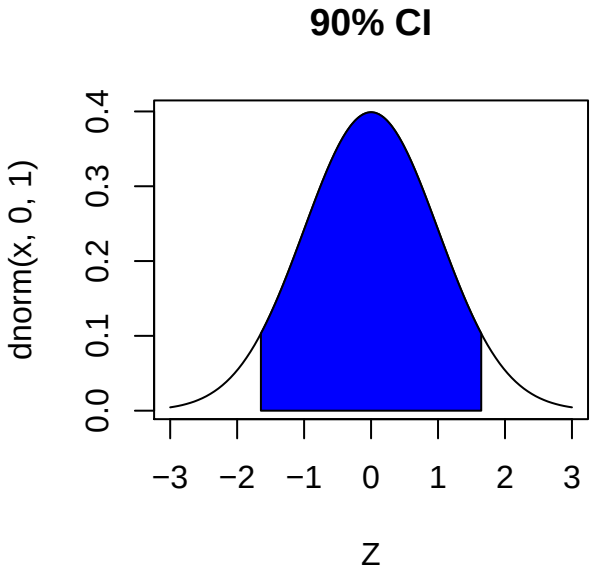


How to find $z^\star$

(1) Subtract the confidence level (CL) from 1: $1 - CL$
(2) $1 - CL = \alpha$, where α is the left over area
(3) Divide α in half: $\frac{\alpha}{2}$. **This is a probability**
(4) Using $\frac{\alpha}{2}$, look that value up *INSIDE* the body of the z table to find the corresponding value of z
(5) The z-score for a CI can be either positive or negative; *you will get the same answers either way* (I usually use positive values)

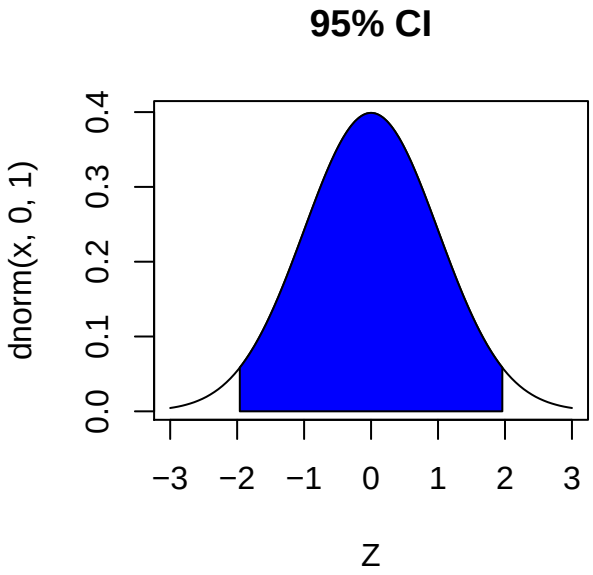
$z^\star$ for 90% CI

90%: $z_{90\%}$ $1 - 0.9 = 0.1$ $0.1/2 = 0.05$, which is in between two values so the z-score for 90% CI is: ± 1.64, ± 1.65, or ± 1.645 (1.645 is used in practice so my examples will use it)



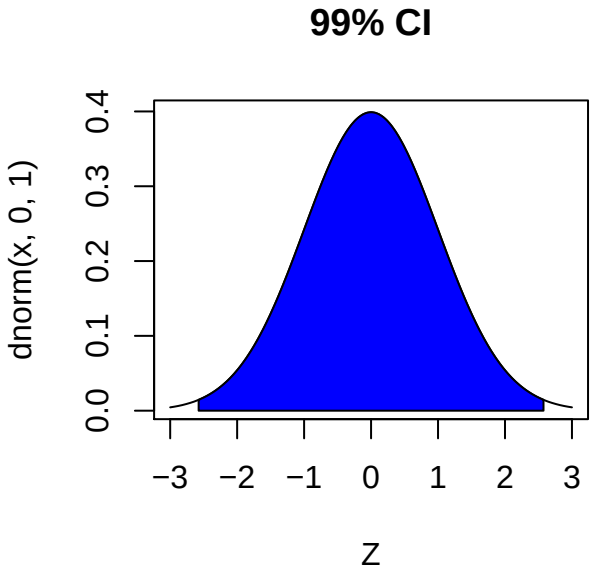
$z^\star$ for 95% CI

95%: $z_{95\%}$ $1 - 0.95 = 0.05$ $0.05/2 = 0.025$ $z = \pm 1.96$



$z^\star$ for 99% CI

99%: $z_{99\%}$ $1 - 0.99 = 0.01$ $0.01/2 = 0.005$, which is in between two values so the z-score for 99% CI is: ± 2.57, ± 2.58, or ± 2.575



Generic Interpretation of CI

To interpret the CI, the statement we use discusses the level of confidence, the parameter we are trying to estimate, and the values of the interval.

> "We are (CL)% confident the true (mean, proportion, etc.) of (insert context) is between (lower number) and (upper number) (units of measurement)."

Estimating μ when σ is known

A consumer testing agency wants to evaluate the claim made by a manufacturer of discount tires; the claim is that its tires can be driven at least 35,000 miles before wearing out. Assume that these tires have a normal distribution and its standard deviation is 5,000 miles.
To determine the average number of miles that can be obtained from the tires, the agency randomly selects 60 tires from the manufacturer's warehouse and places the tires on 15 similar cars driven by test drivers on a 2-mile oval track, with sample mean of 31.47. (The mean was scaled down from 31,470 miles for ease of calculations but will NOT change the answer). $\overline{X} \sim N\left(31.47, \frac{5}{\sqrt{60}}\right) = \overline{X} \sim N(31.47, 0.646)$ (we could do it with the values in thousands, $\overline{X} \sim N(31470, 645.5)$, and other than rounding errors, there would be no difference)

Examples for μ with known σ

(1) Estimate μ, the true average lifetime miles of these tires with 90% confidence, interpret
(2) Estimate μ, the true average lifetime miles of these tires with 95% confidence, interpret
(3) Estimate μ, the true average lifetime miles of these tires with 98% confidence, interpret
(4) Calculate the necessary sample size for a new study that has 95% CL and a bound of 1.05 miles

90% CI for μ example (1)

$$31.47 \pm (1.645)(0.646) = 31.47 \pm 1.063 = (30.41, 32.53)$$

Interpretation: We are 90% confident the true average lifetime of these tires is between 30,410 and 32,530 miles.

95% CI for μ example (2)

$$31.47 \pm (1.96)(0.646) = 31.47 \pm 1.266 = (30.2, 32.74)$$

Interpretation: We are 95% confident the true average lifetime of these tires is between 30,200 and 32,740 miles.

99% CI for μ example (3)

$$31.47 \pm (2.575)(0.646) = 31.47 \pm 1.663 = (29.81, 33.13)$$

Interpretation: We are 99% confident the true average lifetime of these tires is between 29,810 and 33,130 miles.

Calculate the necessary sample size for a new study that has 95% CL and a bound of 1.05

$bound = 1.05$, $z^\star = 1.96$, and $\sigma = 5$

$$n = \left(\frac{z^\star \sigma}{bound}\right)^2 = \left(\frac{(1.96)(5)}{1.05}\right)^2 = 87.11 \to 88$$

CI for π, proportion

$$\hat{\pi} \pm bound = \hat{\pi} \pm z^\star(se_{\hat{\pi}})$$

Where $bound = z^\star(se)$ and $se_{\hat{\pi}} = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, $z^\star$ comes from the table the same way as previous example with μ CI

To solve for a sample size given a CL and bound:

$$n = [\hat{\pi}(1-\hat{\pi})]\left(\frac{z^\star}{bound}\right)^2$$

When no previous information about the proportion should be, use $\hat{\pi} = 0.5$. Again, unlike normal rounding conventions, we *always round* ***up***

Estimating π

In a given town, a random sample of 892 voters contained 500 people who favored a particular bond issue.

(1) Find the true proportion of voters who favor the particular bond with 90% confidence
(2) If another sample is to be taken later, how many people should be sampled to be within 0.02 of the estimate?
(3) Suppose for another sample, there is no previous information about the proprotion of people who would favor the bond issue. How many people should be sampled to be within 0.02 of the estimate?

90% CI for π example

$\hat{\pi} = \frac{X}{n}$ where X is the count of successes and n is the sample size. $\frac{500}{892} = 0.561$

$se_{\hat{\pi}} = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{(0.561)(1-0.561)/892} = 0.0166$

$$0.561 \pm (1.645)(0.0166) = 0.561 \pm 0.0273 = (0.5337, 0.5883)$$

Interpretation: We are 90% confident the true proportion of voters who favor the particular bond issue is between 53.37% and 58.83%.

Calculate new sample size with 90% CL and bound of 0.02

$bound = 0.02$, $z^\star = 1.645$, and $\hat{\pi} = 0.561$

$$n = \hat{\pi}(1-\hat{\pi})\left(\frac{z^\star}{bound}\right)^2 = (0.561)(1-0.561)\left(\frac{(1.645)}{0.02}\right)^2$$

$$= 1666.093 \rightarrow 1667$$

Calculate new sample size with 90% CL and bound of 0.02, no information about π

$bound = 0.02$, $z^\star = 1.645$, and with no known information about the proportion, use $\hat{\pi} = 0.5$

$$n = \hat{\pi}(1-\hat{\pi})\left(\frac{z^\star}{bound}\right)^2 = (0.5)(1-0.5)\left(\frac{(1.645)}{0.02}\right)^2$$

$$= 1691.266 \rightarrow 1692$$

Student's t Distribution

What happens when we do not know the standard deviation (or variance)? Sometimes the z distribution may not be conservative enough. There is a distribution that is similar to the z distribution but made more for distributions with heavier tails (more area in the tails than z).

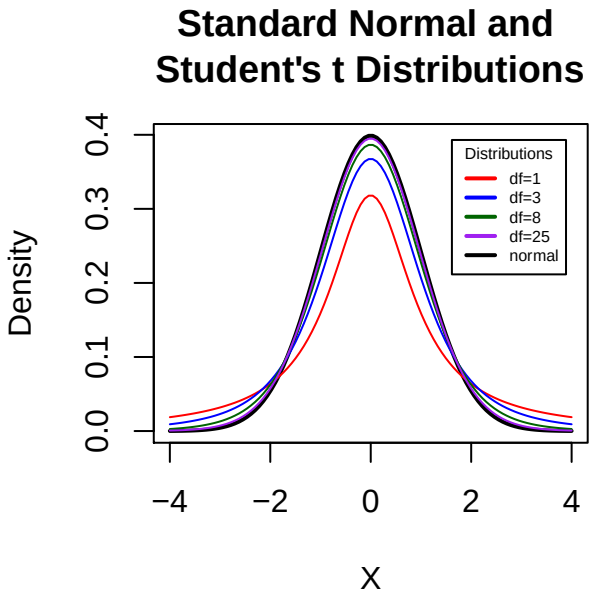
It is called Student's t distribution. It was created by a quality control manager at Guinness Beer named Gosset in 1908. He worked in quality control and worked with small samples. He couldn't publish the results of his study under his true name because of his work contract – no company secrets to be given out.

t logistics

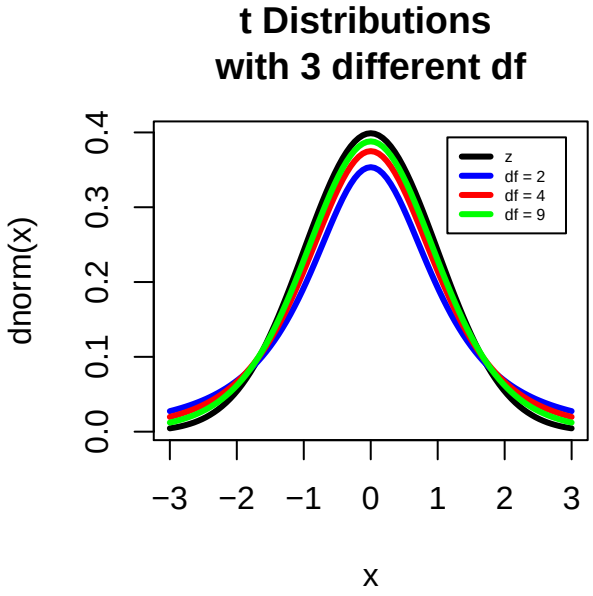
t needs two things: (1) df, degrees of freedom ($df = n - 1$), and (2) CL or $\alpha/2$

Degrees of freedom arises from the fact that we are now using an estimate, $\overline{X}$, for the true mean, μ. We have lost a bit of information, and now have one less degree of freedom. For more detailed description, see Lecture link for this module on class website (links provided there)

Student's t graphs



z and t with varying df



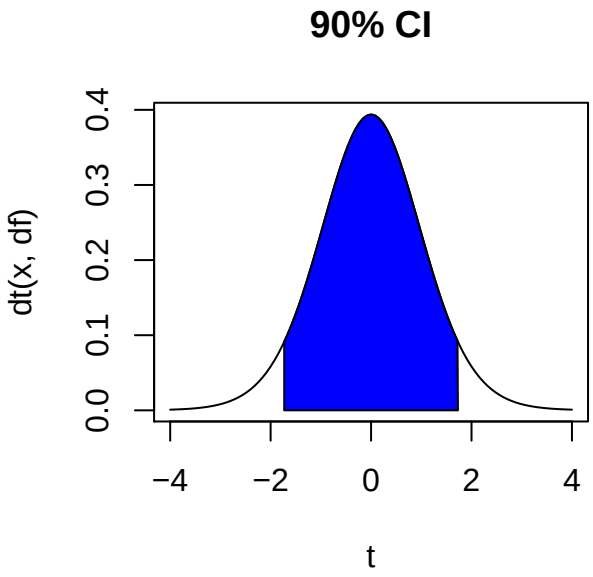
How to find $t^\star$

(1) Subtract the confidence level (CL) from 1: $1 - CL$
(2) $1 - CL = \alpha$, where α is the left over area
(3) Divide α in half: $\frac{\alpha}{2}$. **This is a probability**
(4) Using $\frac{\alpha}{2}$, find the column that says that number
(5) The rows are the df, go to the appropriate row and find where that row and the α column meet
(6) t is symmetric so that the values can be negative for the other side of the curve (below the mean)

$t^\star$ for 90% CI

90%: $t_{90\%,df}$ $1 - 0.9 = 0.1$ $0.1/2 = 0.05$, now line up the column for α (or CL) with the appropriate row, where the rows are the $df = n - 1$. The graph shows a $t^\star$ with $n = 20$ so $df = 19$

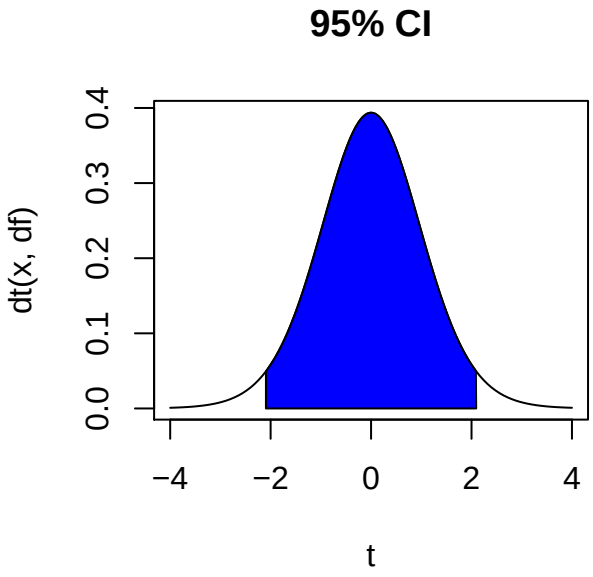
$t^\star = t_{90\%,19} = 1.729$



$t^\star$ for 95% CI

95%: $t_{95\%,df}$ $1 - 0.95 = 0.05$ $0.05/2 = 0.025$, now line up the column for α (or CL) with the appropriate row, where the rows are the $df = n - 1$. The graph shows a $t^\star$ with $n = 20$ so $df = 19$

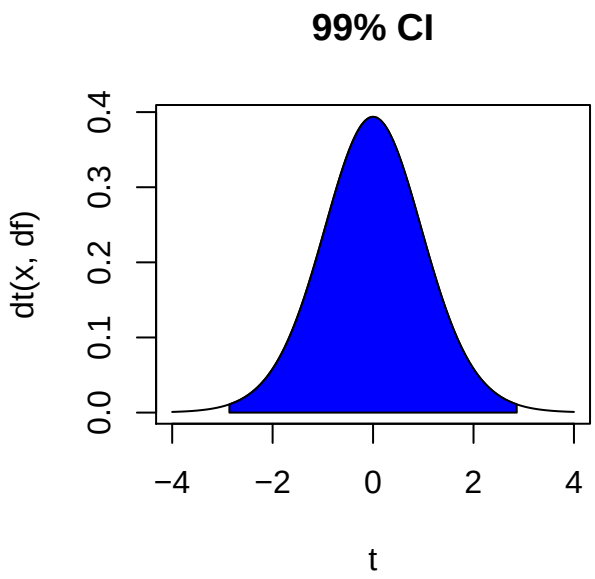
$t^\star = t_{95\%,19} = 2.093$



$t^\star$ for 99% CI

99%: $t_{99\%,df}$ $1 - 0.99 = 0.01$ $0.01/2 = 0.005$, now line up the column for α (or CL) with the appropriate row, where the rows are the $df = n - 1$. The graph shows a $t^\star$ with $n = 20$ so $df = 19$

$t^\star = t_{99\%,19} = 2.861$



Form for CI on μ when σ is not known (we use s)

$$\overline{X} \pm t^\star(se_{mean})$$

Where $t^\star = t_{\alpha/2,df}$, $df = n - 1$, and $se_{mean} = \frac{s}{\sqrt{n}}$

Page 453+ in the textbook show examples on how to do these calculations on TI calculators.

Estimating μ when σ is not known, using s

A product comes in cans labeled "38 oz", and a random sample of 10 cans had the following weights:
{34.6,39.65,34.75,40,39.5,38.9,34.25,36.8,39,37.2}

```
rbind(mean,sd)
```

```
         [,1]
mean 37.46500
sd    2.26532
```

Estimate μ, the true average weight of the product, with 98% confidence, interpret

95% CI for μ when σ is unknown

$t^\star = t_{98\%,df}$ where $df = n - 1 = 10 - 1 = 9$ so $t_{98\%,9} = t_{0.02/2,9} = 2.821$, $se_{mean} = \frac{s}{\sqrt{n}} = \frac{2.265}{\sqrt{10}} = 0.716$

$$37.465 \pm (2.821)(0.716) = 37.465 \pm 2.0198 = (35.45, 39.48)$$

Interpretation: We are 98% confident the true average weight of product is between 35.45 and 39.48 ounces.

Reality

In practice, the use of Z happens seldomly. You have to know a lot about the population and such ahead of time and that does not happen often. Statisticians primarily use t since it is good with data samples, small sample sizes, and if the sample size happens to be large enough, t will eventually converge to z so use of t is

just better. It is a little more conservative so you may not get as many false positives that way (which is good).