Week 8: Confidence Intervals Part 2

Jessie Yeung

STA 220

Winter 2024

Announcements

- Term Test grades were released last week
  - Solutions are posted
  - Follow the steps on the syllabus if you would like to request a re-grade

Overview

- We will finish the topic of confidence intervals this week
- Topics for this week
  - Confidence Intervals for the mean
  - Robustness of confidence intervals
  - T Distribution
- This content corresponds with Module 7

Review: CIs for Proportions

Review: CIs for Proportions

- We are trying to find an estimate for the population proportion p by collecting a sample and computing $\hat{p}$
- We can then build a $(1-\alpha)\%$ confidence interval around $\hat{p}$ to give us a range of plausible values that p could be.
- The confidence interval is built by taking the estimate $\hat{p}$ and putting limits on either side, determined by the sampling error $\mathrm{SE}(\hat{p})$ (i.e. from the sampling distribution)
- We determine how many SEs away from $\hat{p}$ we are by finding $z_{\alpha/2}$, a critical value quantile from the standard normal distribution

Review: CIs for Proportions

$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

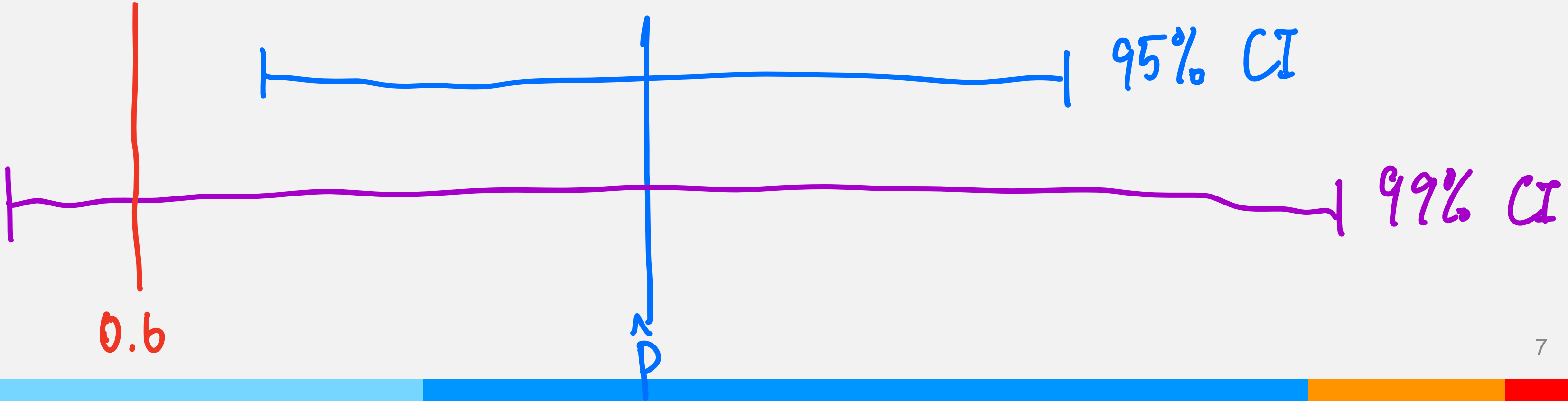
- We end up with the interval

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

- Since we don't know p, we can either use $\hat{p}$ or 0.5 in the standard error term.

If 0.6 lies inside the 99% confidence interval for the population proportion, then 0.6 is also inside the 95% confidence interval for the population proportion.

- ○ True
- ✓ False



Confidence Intervals for the mean

Confidence Intervals for Means

- So far, we have built confidence intervals around the population proportion p. This is appropriate when the population parameter we are interested in is a proportion (i.e., our data is Bernoulli).
- However, there are other cases when we are instead interested in the population mean μ.
- Recall that we are also able to use CLT to obtain the sampling distribution of the sample mean $\bar{X}$. We can use this to build a confidence interval for the population mean μ in a very similar way.

Sampling distribution of the mean

- Recall from the CLT that for a sufficiently large n,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

where μ is the population mean of the quantity measured and σ^2 is the population variance of the quantity measured.

- This can be standardized so that

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

- We can use this to construct a confidence interval just like before

Deriving the confidence interval

$$P(-1.96 < Z < 1.96) = 0.95$$

$Z \sim N(0, 1)$

Sub in $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} < 1.96\right) = 0.95$$

multiply by $\sqrt{\frac{\sigma^2}{n}}$

$$P\left(-1.96\sqrt{\sigma^2/n} < \bar{X} - \mu < 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

subtract $\bar{X}$

$$P\left(-\bar{X} - 1.96\sqrt{\sigma^2/n} < -\mu < -\bar{X} + 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

$$P\left(-\bar{X} - 1.96\sqrt{\sigma^2/n} < -\mu < -\bar{X} + 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

multiply by -1

$$P\left(\bar{X} + 1.96\sqrt{\sigma^2/n} > \mu > \bar{X} - 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

Rearrange

$$P\left(\bar{X} - 1.96\sqrt{\sigma^2/n} < \mu < \bar{X} + 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

Therefore, the 95% confidence interval for μ is

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Any level of confidence

- Similar to before, the level of confidence can be anything you want it to be.
- In general, for a $(1-\alpha)\%$ confidence interval, we will need to find the critical value $z_{\alpha/2}$ such that $P(Z < -z_{\alpha/2}) = \alpha/2$.
- This gives a confidence interval of

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- Also can be written as

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Unknown σ

- The formula is not in terms of μ which is good since that's the unknown population parameter that we're trying to learn about
- The trouble here is that, just like we don't know the true population mean or proportion, we also don't know what the population SD σ is.
- Similar to one of the options with CIs for proportions, we can use a value computed from our sample:
  - In this case, it makes sense to use the **sample standard deviation s**
  - This is the standard deviation that we can compute based on our data
  - We use $\frac{s}{\sqrt{n}}$ to <u>estimate</u> the standard deviation of the sampling distribution $\frac{\sigma}{\sqrt{n}}$

Sample Standard Deviation

- Recall from Week 1 when we were looking at summary statistics
- Sample Variance = $s^2 = \frac{\Sigma_{\mathrm{i}=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Sample standard deviation = s = $\sqrt{s^2} = \sqrt{\frac{\Sigma_{\mathrm{i}=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- In R, we can calculate the sample variance and sample ~~variance~~ std. deviation using the var() and sd() commands

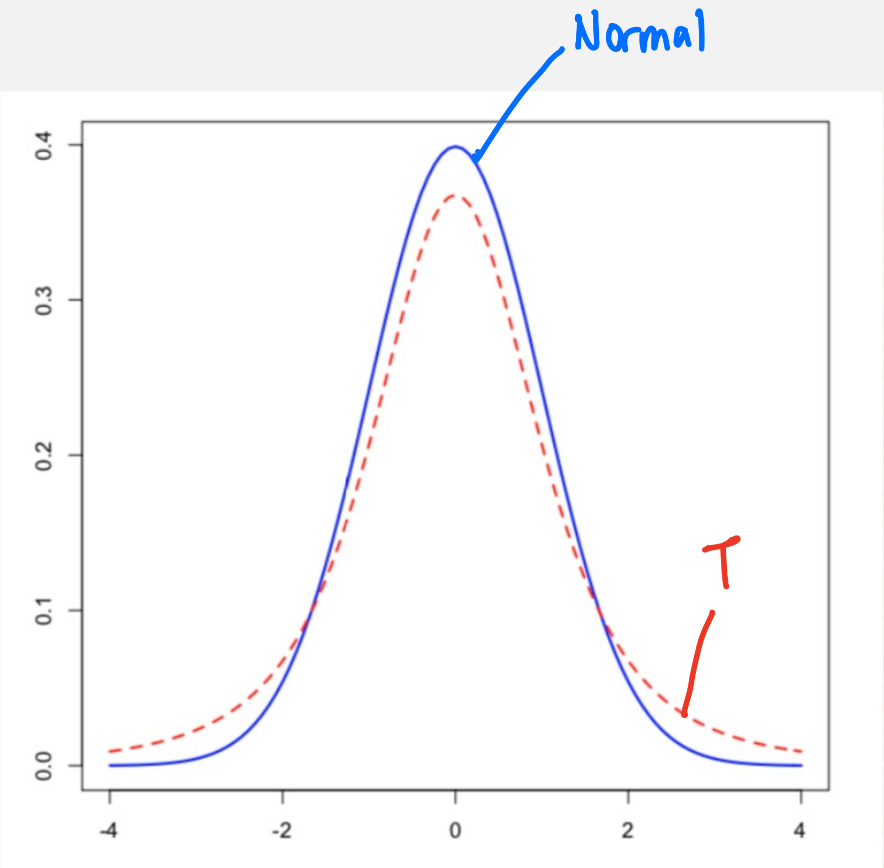
Extra Variation

- We encounter another snag though when working with means:
  - Our interval is now using two pieces of information that we are computing from our sample data: $\bar{x}$ and s
  - Every time we estimate something, we are introducing more uncertainty/variability
  - Our formula for the CI of the mean only accounts for the variability of using $\bar{x}$ to estimate μ. It doesn't account for the extra variability of using s to estimate σ
  - Then, even though we say we are building a $(1-\alpha)\%$ interval, that may not be what we actually get!
  - So our interval may not actually have the **coverage** that we want.
- Solution: Instead of using N(0,1) to find the critical values, we will need to use another distribution.

T Distribution

T Distribution

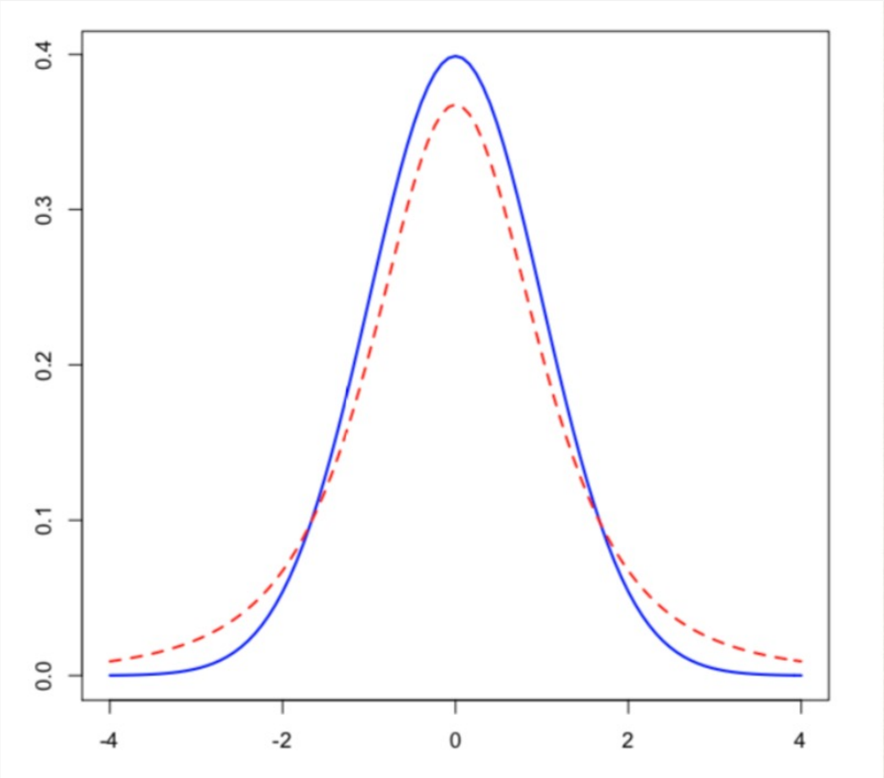
- The T distribution is a probability distribution that is also symmetric, unimodal, and bell-shaped
- It is always centered at 0
- The only difference from N(0,1) is the probability of getting values far from the mean (i.e. in the tails of the distribution)
- The T distribution has heavier tails, so bigger chance of being far from the mean.



Eg. $X \sim T_3$

T Distribution

- The T distribution only has one parameter that controls its shape, called the degrees of freedom (df)
- This is written as $X \sim T_{df}$
- To fully define the T distribution, you need to specify the df
- Turns out that as the df gets larger and larger, the T distribution more closely resembles N(0,1)



Unknown σ

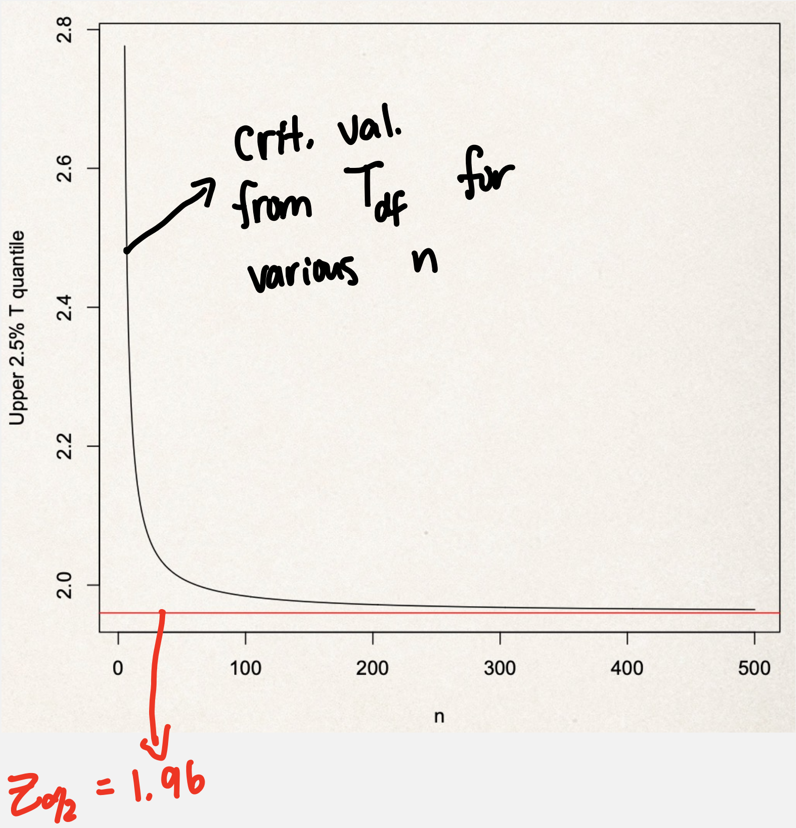
- So why does this help us when we don't know σ?
  - If we need to estimate σ with s, we need to account for the fact that both our estimates for σ and μ will vary with each different sample we take.
  - more variability in two places means higher chance of getting extreme sample values
  - To have the right coverage for the interval, we need to use a distribution that also has higher chance of getting extreme values
- In order to find the critical values, we will want to use the T distribution with $df = n - 1$ where n is the sample size

Finding the critical value

- Finding the critical value is the same as before
- We are interested in a value $t_{\frac{\alpha}{2},n-1}$ such that $P\left(T_{n-1} < -t_{\frac{\alpha}{2},n-1}\right) = \frac{\alpha}{2}$

Finding the critical value for 95% CI

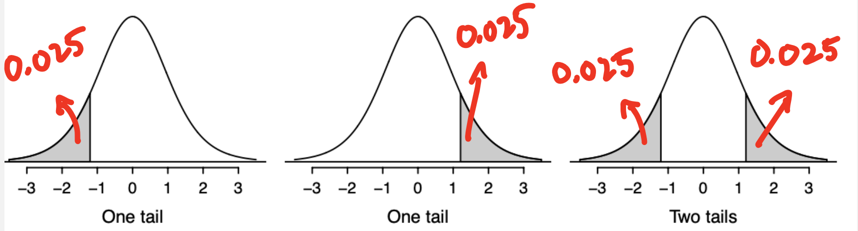
- Say we want to build a 95% confidence interval around the mean.
- We then need quantiles for $\alpha = 0.05$
- Since the T distribution parameter is based on the sample size n, we can look at how larger sample sizes get us closer to a N(0, 1).
  - we know $z_{\alpha/2}$ = 1.96 (red line)
  - by the time we have a sample of size of 100, the T distribution quantile for the same $t_{df=99}$ is already 1.98
  - So, very close to the Normal one



T distribution table

- Similar to how there is a table for the standard normal distribution, there is one for the T distribution
- For confidence intervals, should use the two tails part of the table
    - we are taking our α tail probability and splitting it over 2 tails.
    - If we want α = 0.05 over 2 tails, that is the same as 0.025 in one tail.
- Find the correct $df = n - 1$ for your sample.
- Where the tail probability column and df row intersect, that's your critical value for your CI

t distribution probability table



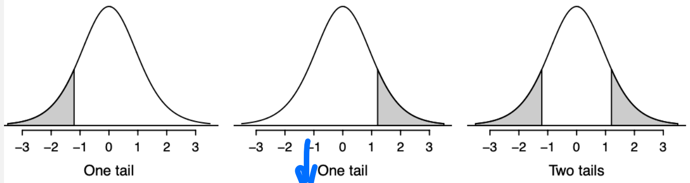
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95

Example

Find the critical value for a 90% CI using the T-distribution where $n = 15$

df = 14

t distribution probability table



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95

T Distribution and CIs

- By using the T distribution, especially when sample size is small(ish), we can account for the fact that $\bar{x}$ and s may not be good estimates
  - use a heavier tailed distribution to see higher chance of extreme values
  - the T distribution will make our CI wider because the quantile is bigger than the Normal
- Just like sample proportions and means, s gets better (and closer to σ) when sample size goes up.
  - then T distribution gets closer to Normal, and we no longer need to worry about extreme estimates

Confidence Intervals for Means, Unknown σ

- Now that we understand why we need to use the T distribution when dealing with means and their confidence intervals, we can look at the interval itself.
- Just like with proportions, we can specify whatever confidence level $(1-\alpha)\%$ we want
  - For some α, find the corresponding quantile of the T distribution, $t_{\frac{\alpha}{2},n-1}$ using a T table (like the Normal table)
  - Estimate both $\bar{x}$ and $s = \sqrt{\frac{\Sigma_i (x_i - \bar{x})^2}{n-1}}$ from your sample
  - Compute the $(1-\alpha)\%$ CI: $\left(\bar{X} - t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}\right)$
  - In rare cases where σ is known, you can use the CI based on the Normal distribution

Sample Sizes

Size of a Confidence Interval

- Similar to before, the margin of error is the distance between the sample mean and the end of the confidence interval
- ME = critical value $\times \frac{\sigma}{\sqrt{n}}$ or ME = critical value $\times \frac{s}{\sqrt{n}}$
- General format of the CI is: sample estimate $\pm$ margin of error
- We can see the width of the CI is impacted by several things:
  - The confidence level $(1 - \alpha)$
    - Higher confidence means wider CI
  - Standard deviation σ or sample standard deviation s
    - Higher SD means wider CI
  - Sample size n
    - Larger sample size means narrower CI

Calculating the required sample size

- We can re-arrange the formula for the margin of error to get the following:
  - If σ is known, then $n = \left(\frac{z_{\alpha/2} \times \sigma}{ME}\right)^2$
  - If σ is NOT known, then $n = \left(\frac{t_{\alpha/2,n-1} \times s}{ME}\right)^2$
    - But we have a problem here! The formula requires a critical value with df=n-1. Which means that the critical value depends on n.
    - So, let's just use the normal distribution
    - Therefore, you can calculate the required sample size by using $n = \left(\frac{z_{\alpha/2} \times s}{ME}\right)^2$ and then rounding up!

Summary

Summary of Notation

- Population Parameters: Unknown characteristics of whole population
  - Population proportion: p
  - Population mean: μ
  - Population standard deviation: σ
  - Population variance: σ^2
- Sample Statistics: Summary measures of sample data
  - Sample proportion: $\hat{p}$
  - Sample mean: $\bar{X} = \hat{\mu}$
    - Note: $\bar{X}$ refers to the sample mean as a random variable, $\bar{x}$ refers to a computed value based on a specific dataset
  - Sample standard deviation: $s = \hat{\sigma}$
  - Sample variance: $s^2 = \hat{\sigma}^2$

Summary of CI for p

- A $(1-\alpha)\%$ CI can be expressed as:

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

- Use either 0.5 or $\hat{p}$ in place of p in the formula

Summary of CI for μ

- When σ is known, a $(1-\alpha)\%$ CI can be expressed as:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- When σ is NOT known, a $(1-\alpha)\%$ CI can be expressed as:

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

Example: Parking Fees

A city builds a new parking structure in the central business district. The city plans to pay for the structure through parking fees. During a 44 day period, the daily parking fees collected were on average \$126 with a standard deviation of \$15.

a) Find a 90% confidence interval for the mean daily income this parking structure will generate.

b) The consultant who advised the city thinks the parking structure fees will generate an average of \$130 of revenue for the city. Based on your interval, do you think his guess is correct?

c) Recalculate the 90% confidence interval with the additional information that the true standard deviation of parking fees collected is \$15.

d) Compare the answers in part a and c. Do they make sense?

a) We are interested in μ, which is the average parking fees collected in a day.

$n = 44$ $\bar{x} = 126$ $s = 15$

$t_{\alpha/2, n-1} = 1.68$ from t-distribution table

Alternatively, we can use qt(p= 0.05, df = 43) in R.

90% CI: $\bar{x} \pm t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}$

$\Rightarrow 126 \pm 1.68 \times \frac{15}{\sqrt{44}}$

$\Rightarrow (122.20, \ 129.80)$

We are 90% confident that the daily average parking fees collected is between \$122.20 and \$129.80

b) No, based on the 90% CI, it doesn't support the estimate that the city can generate \$130 per day.

c) $n = 44$ $\bar{x} = 126$ $\sigma = 15$ $z_{\alpha/2} = 1.64$

90% CI: $\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

$\Rightarrow 126 \pm 1.64 \times \frac{15}{\sqrt{44}}$

$\Rightarrow (122.29, \ 129.71)$

d) Yes they make sense — interval in part a is wider due to extra variability from using s to estimate σ

Practice Problems

- In this lecture, we covered all of Module 7: Confidence Intervals Part 2
- Practice problems are posted