Jan 24

Chapter 7: Estimating Parameters and Determining Sample Sizes

Section 7.1: Estimating a Population Proportion Part I ← || Part 2 of Course

Stat 50

INTRO

DESCRIPTIVE STATISTICS		PROBABILITY		INFERENTIAL STATISTICS
Sample data	+	*Find a level of RELIABILITY* theoretical	=	make conclusion based off of sample data.

box plot

Prob dice / CLT

Look at the tray of beans that was brought to class today. We are going to try and estimate the percentage of black beans in the tray.

What percentage of the ~~tray~~ image's contents do you feel are black beans? Write your guess below.

Class Guesses: 0.4, 0.33, 0.4, 0.4, 0.5, 0.5, 0.42, 0.55

a. How could we find out if we are right?

* count all of them
* take a sample

b. Based on our sample, the instructor's best guess is that there is between 15 % and 20 % of black beans in the ~~tray~~ image. That's a proportion of between 0.15 and 0.2.

What level of agreement do you have with this guess? (*Circle one*)

Totally agree	Somewhat agree	Somewhat disagree	No Way!
none	none		

c. Your instructor has revised their guess! Now they believe that there is between 20 % and 32 % of black beans in the tray. That's a proportion of between 0.2 and 0.32.

What level of agreement do you now have with the instructor's new guess? (*Circle one*)

Totally agree	Somewhat agree	Somewhat disagree	No Way!
0	1	4	2

d. One last revision! Now they believe that there is between 35 % and 47 % of black beans in the tray. That's a proportion of between 0.35 and 0.47.

What level of agreement do you now have with the instructor's new guess? (*Circle one*)

Totally agree	Somewhat agree	Somewhat disagree	No Way!
2	4	0	0

e. Discuss:

- What would happen if we took another sample? Would we get the same sample proportion?

  Maybe, we expect to be close if samples nearly same size.

- Would you like to know the true proportion of black beans in the entire bag?

  Not gonna tell you!

Def ***Point Estimate*** A single value used to approximate a population parameter. ← never know

Note: *The sample proportion* $\hat{p}$ *is the point estimate of the population proportion* p.

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

SAMPLE PROPORTION: $\hat{p} = \dfrac{\textit{successes}}{\textit{total}} = \dfrac{x}{n}$

note: p (w/no hat) is pop. proportion.

Ex: I took a sample of 40 PCC students and asked them "Do you think Sir Pugsly Farnsworth Esquire III is cute or not?" The results were as follows: **24** said **YES** and **16** said **NO**.

a. What is the variable? Is it qualitative or quantitative?

↳ Y or N answer

b. This is a __binomial__ problem since there are only two outcomes.

c. What would $\hat{p}$ be?

$\hat{p} = \frac{24}{40} = \boxed{0.6}$

Reminder: proportion should always be in decimal form.

d. Which statement do you think we can make from this point estimate?

BAD → 1. ***Exactly*** __???__ (can guess 0.6) of ALL PCC students think Sir Pugsly Farnsworth Esquire III is cute. ← not stats

Good → 2. ***About*** __0.6__ of ALL PCC students think Sir Pugsly Farnsworth Esquire III is cute. ← Good

From this example, we can imagine that there may be some __error (variation)__ associated with taking a sample from a population. We will create an interval around $\hat{p}$ with the hope that the *true proportion p* lies within that interval and we state a certain level of confidence we have.

Conceptually Important!

$$\boxed{(\hat{p} \pm ERROR)}$$

p p p

0.57 0.6 0.62

key round parenthesis don't include # line

Def *A* ***confidence interval*** (abbreviated CI) is a range (or interval) of values used to estimate the true value of a population parameter.

$\hat{p} - E$ & $\hat{p} + E$

Interval Notation $(\hat{p} - E, \hat{p} + E)$ Prefered.

E *Def* ***Margin of Error*** The maximum likely difference between the observed sample proportion $\hat{p}$ and the true value of the population proportion p.

Note E changes depending on pop. parameter

MARGIN OF ERROR FORMULA: FOR POP. PROPORTION

$$\boxed{E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

def $1-\alpha$: confidence level

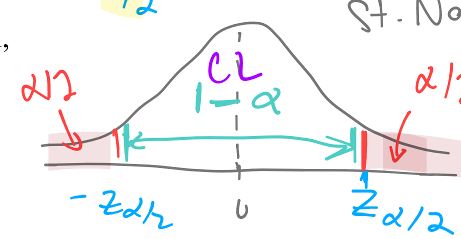
Ex: If the margin of error is 8%, then what interval would you expect the **true proportion** of PCC students to be who think Sir Pugsly Farnsworth Esquire III is cute?

assume $E = 0.08$ $\hat{p} \pm E = 0.6 \pm 0.08$

true pop. proportion is in $(0.52, 0.68)$

- $\hat{p}$ sample proportion
- $\hat{q} = 1 - \hat{p}$ (complement)
- $z_{\alpha/2}$ = critical value

St. Norm.



EFFECTS ON THE WIDTH OF YOUR CONFIDENCE INTERVAL

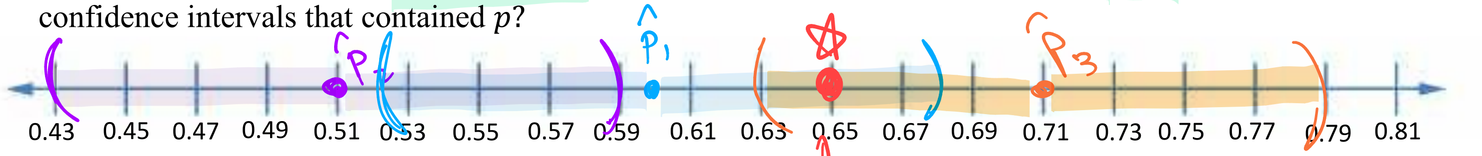
1) Sample size

- If n is small, the CI widens
- If n is large, then CI narrows

2) confidence level

- The lower the confidence level, narrower CI
- The higher the confidence level, widens CI

Let's use our first $\hat{p}$ = 0.6 and then take two more samples where the $\hat{p}$'s were .51 ($\hat{p}_2$) and then .71 ($\hat{p}_3$). Using the same margin of error, if the true proportion of PCC who thinks Sir Pugsly is cute is 0.65, then which $\hat{p}$'s had confidence intervals that contained p?



Error $E = 0.08$

$\hat{p}_1: (0.52, 0.68)$ $\hat{p}_2: (0.43, 0.59)$ $\hat{p}_3: (0.63, 0.79)$

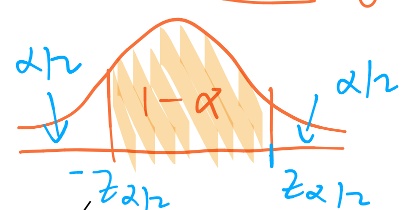
CI's for $\hat{p}_1$ & $\hat{p}_3$ contain p.

It looks like our probability of success is $2/3 = 67\%$ right now, but are there only three samples of 40 PCC students that can be taken?

Def **Confidence Level** The probability equal to the proportion of times that the confidence interval (CI) actually contains the true population parameter.

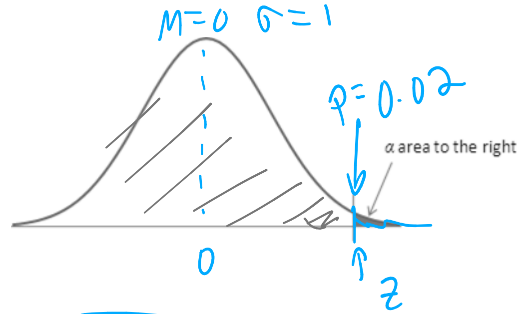
Note: If we were to repeat the estimation process a large number of times, the confidence level is the percentage of confidence intervals that would actually contain the true population proportion. GTK

Notation: The confidence level is denoted by $\boxed{1-\alpha}$



*Note: The **critical value** (denoted $z_{\alpha/2}$) is the positive z score that separates an area of $\alpha/2$ in the right tail of the standard normal distribution. It is determined by the confidence level.*

$\mu = 0$ $\sigma = 1$



Ex: Find $z_{0.02}$

$\alpha/2 = 0.02$

$\alpha = 0.04$

CL $1-\alpha = 1 - 0.04 = 0.96$

$z_{0.02} = \text{invNorm}(0.02, 0, 1, \text{RIGHT}) = \boxed{2.05}$ (prob. μ σ tail)

$z = \text{invNorm}(0.98, 0, 1)$

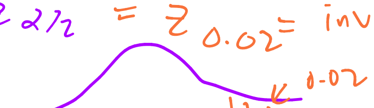
Ex: Find the margin of error if $\hat{p} = 0.48$, $n = 200$, and the confidence level is as follows:

a. 96% confidence level

$E = ?$

$E = z_{\alpha/2} \cdot \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

- $CL = 0.96$
- $\alpha = 1 - CL = 1 - 0.96 = 0.04$
- $\alpha/2 = 0.02$
- $\hat{p} = 0.48$ $\hat{q} = 1 - 0.48 = 0.52$
- $z_{\alpha/2} = z_{0.02} = \text{invNorm}(0.02, 0, 1, \text{RIGHT})$



$z_{\alpha/2} = 2.05$

$E = 2.05 * \sqrt{\dfrac{(0.48)(0.52)}{200}}$

$\boxed{E = 0.072}$

rounded to 3 decimal places

b. 80% confidence level

$E = 0.0452\ldots$

$\boxed{E = 0.045}$

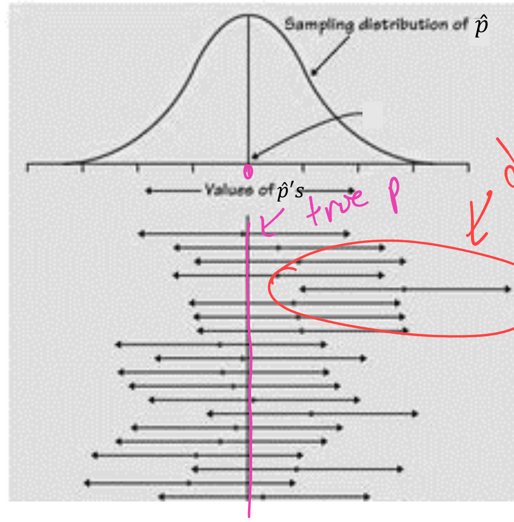
$CL = 0.80$

$\alpha = 1 - 0.8 = 0.2$

$\alpha/2 = 0.1$

$z_{\alpha/2} = 1.28$

Ex: In the example to the right we took twenty different samples of size n, and found the corresponding confidence intervals with a 95% level of confidence.



true p

doesn't have true p

A 95% confidence interval indicates that 19 out of **20** samples from the same population will produce confidence intervals that contain p.

$20 * 0.95 = 19$

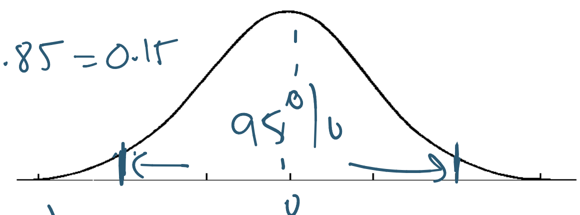
1 CI shouldn't have true p

Ex: In a 2018 random survey of 160 American Democrats, 136 said that they support Medicare-for-all, also known as single-payer healthcare. Find the 95% confidence interval estimate for the true proportion of Democrats in America who support Medicare-for-all.

↳ give a decimal

1. *Identify point estimate (sample proportion)*

$\hat{p} = \frac{136}{160} = \boxed{0.85}$ $\hat{q} = 1 - \hat{p} = 1 - 0.85 = 0.15$



2. *Determine critical value* $z_{\alpha/2}$

$z_{\alpha/2} = z_{0.025} = \text{invNorm}(0.025, 0, 1, \text{RIGHT}) = \boxed{1.96}$

- $CL = 0.95$
- $\alpha = 1 - 0.95 = 0.05$
- $\alpha/2 = 0.05/2 = 0.025$

3. *Find margin of error*

$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 * \sqrt{\frac{(0.85)(0.15)}{160}} = \boxed{0.0553}$

$n = 160$

Advice keep 4 decimal places

4. *Construct confidence interval* CI

$CI: (\hat{p} - E, \hat{p} + E)$

to 3 decimal places

$CI: (0.85 - 0.0553, 0.85 + 0.0553) \approx (0.795, 0.905)$ or

$0.795 < p < 0.905$

5. *Interpretation of CI*

(M→E) "We are 95% confident that the true proportion of American Democrats who support Medicare-for-All is between 0.795 and 0.905."

Does the proportion of Democrats who support Medicare-for-all appear to be substantially different than the 70% rate for the general population? If so, why do you think that is?

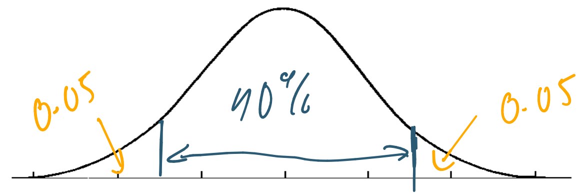
- Democrats: 79.5% to 90.5%
- General pop: 70%
- Yes, it appears substantially different (almost 10%!) because it includes other parties w/ different views.

Ex: In a 2018 survey on climate change, 626 out of 1278 Americans (18 years or older) were "extremely" or "very sure" it is happening. Construct a 90% confidence interval for the true percentage of Americans who are "extremely" or "very sure" climate change is happening.

note %

1 *Identify point estimate (sample proportion)*

$$\hat{p} = \frac{626}{1278} = \boxed{0.490}$$



CL = 0.90

$\alpha = 1 - CL = 1 - 0.9 = 0.1$

$\alpha/2 = 0.1/2 = 0.05$

2 *Determine critical value* $z_{\alpha/2}$

$$z_{\alpha/2} = z_{0.05} = \text{invNorm}(0.05, 0, 1, \text{RIGHT}) = \boxed{1.64}$$

Note keep more accurate on calc using $\boxed{\text{STO}\to}$

3 *Find margin of error*

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.64^{*}\sqrt{\frac{(0.49)(0.51)}{1278}} = 0.0230009\ldots = \boxed{0.023}\ E$$

4 *Construct confidence interval*

$$CI: (\hat{p} - E, \hat{p} + E) = (0.467, 0.513)$$

CI: or $0.467 < p < 0.513$

5 * *Interpretation of CI* * (MSE)

"We are 90% confident that the true percentage of Americans (18 years or older) who are "extremely" or "very sure" climate change is happening is between 46.7% and 51.3%."

GRAPHING CALCULATOR (TI-83 OR 84)

Instructions: (a) STAT $\Rightarrow$ TESTS $\Rightarrow$ 1-PropZInt One proportion z-score interval (conf.)

(b) Enter $\begin{cases} x = \text{number of successes} \\ n = \text{number of trials} \\ C-Level = \text{confidence level} \end{cases}$

Funky name be careful to pick correct one.

don't know $\hat{p} \to \hat{p} = 0.5$
$\hat{q} = 0.5$
$\hat{p} * \hat{q} = (0.5)^2 = 0.25 = \frac{1}{4}$

DETERMINING SAMPLE SIZE

★ Always ROUND UP

Point Estimate $\hat{p}$ *Known*	*Point Estimate* $\hat{p}$ *Unknown*
$n = \dfrac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$	$n = \dfrac{[z_{\alpha/2}]^2}{4E^2}$

"n people counting"
$n = 67.1$
$n = 68$

Ex: An economist wants to know if the proportion of the children in the United States that speak a language other than English at home has changed since 2016, when 22% of children did. How many children would need to be surveyed if the economist wants to be within 2 percentage points of the true proportion with 92% confidence? (M→E)

GIVENS
- $\hat{p} = 0.22$
- $E = 2\% = 0.02$
- $CL = 0.92$

WANT
$n = ?$

Can Find
- $\hat{q} = 0.78$
- $z_{\alpha/2}$

Do we know $\hat{p}$? Yes! → Use $n = \dfrac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$

92

$CL = 0.92$
$\alpha = 1 - 0.92 = 0.08$
$\alpha/2 = 0.04$

$z_{0.04} = \text{invNorm}(0.04, 0, 1, R) = 1.75^*$

$n = \dfrac{[1.75^*]^2 * (0.22) * (0.78)}{(0.02)^2}$

$n = 1314.84\ldots$ ⇧ Round

$\boxed{n = 1315}$

(M→E) "We would need 1315 children to survey to get a 92% confidence w/ 2% Error."

Ex: Before US presidential elections potential candidates have their exploratory teams work to determine the percentage of people who will vote for their candidate. Kamala Harris has such a team looking into what percentage of people would vote for her. If they want to construct a 98% confidence interval with a 3% margin of error, how many people do they need to poll to achieve this?

"proportion of voters who would vote for her."

Givens
- $CL = 0.98$
- $E = 0.03$

Wants
- $n = ?$

Do we know $\hat{p}$? NO!

Use $n = \dfrac{[z_{\alpha/2}]^2}{4E^2}$

Can Find
- $z_{\alpha/2}$

$CL = 0.98$
$\alpha = 1 - CL = 0.02$
$\alpha/2 = 0.01$

$z_{0.01} = \text{invNorm}(0.01, 0, 1, R) = 2.33^*$

$n = 1503.304\ldots$ ⇧

$\boxed{n = 1504}$

(M→E)
"Harris' Team would need 1504 people to poll to construct a 98% confidence interval with a 3% margin of error w/o knowing a sample

μ – population mean σ = pop st-dev
$\bar{x}$ – sample mean s – sample st.dev

Section 7.2: Estimating a Population Mean

CONFIDENCE INTERVAL FOR THE POPULATION MEAN

Alternative Forms: $\bar{x} - E < \mu < \bar{x} + E$ or $\bar{x} \pm E$ or $(\bar{x} - E, \bar{x} + E)$

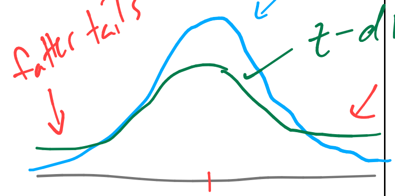
Requirements

1. The sample is a simple random sample. SRS
2. The value of the population standard deviation σ is not known.
3. Either or both of the given conditions are satisfied: { The population is normally distributed or $n > 30$ } (CLT) just like Sampling Dist.

To create these confidence intervals, we won't have σ but we can get a sample standard deviation $\underline{s}$ (little s). The larger the sample, the closer the sample s will get you to σ. We will need to use a different distribution. Student's t-dist.

STUDENT T DISTRIBUTION

STUDENT'S t- DISTRIBUTION RULES	
	• $\underline{\text{bell}}$ - shaped but the tails will be $\underline{\text{fatter}}$ than a standard $\underline{\text{dist}}$ normal curve. • Centered around $\mu_t = \underline{0}$ and $\sigma_t = \frac{s}{\sqrt{n}}$ • Characterized by the *degree of freedom* FORMULA: $df = n - 1$ • We change $z_{\frac{\alpha}{2}}$ to * $\underline{t_{\alpha/2}}$



calc: $t_{\alpha/2} = \text{invT}\left(\frac{\alpha}{2}, df\right)$ abs value! OR $= \text{invT}\left(1 - \frac{\alpha}{2}, df\right)$

Ex: Find the critical t –value that corresponds to 98% confidence with 15 degrees of freedom.
↳ sample size is $n=16$.

$CL = 0.98$

$\alpha = 1 - CL = 0.02$

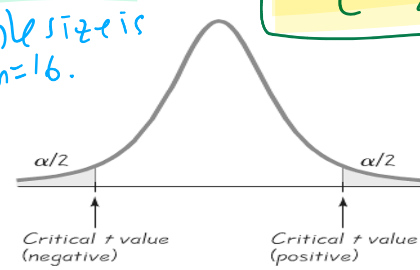
• $\alpha/2 = 0.01$

• $df = 15$

• $t_{\alpha/2} = t_{0.01}$

$= \text{invT}(\underbrace{1 - 0.01}_{1-\alpha/2}, \underbrace{15}_{df})$

$= \boxed{2.60}$



Def ***Point Estimate*** A single value used to approximate a population parameter. (same def)

Note: The sample mean $\bar{x}$ is the point estimate of the population mean μ.

SAMPLE MEAN: $\bar{x} = \frac{\sum x}{n}$

Def ***Margin of Error*** The maximum likely difference between the observed sample mean $\bar{x}$ and the true value of the population mean μ.

MARGIN OF ERROR FORMULA: $\boxed{E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}}$

"MR. T IS MEAN"

Ex: A 2017 sample of 34 cities in the US found that the average ($\bar{x}$) cost of a wedding was \$36,000 with a standard deviation of \$14,114. Determine a 99% confidence interval for the corresponding population mean cost of a US wedding.

↳ μ mean → use t

$n = 34$

$df = 33$

1. *Identify point estimate (sample mean)*

↳ sample mean $\boxed{\bar{x} = \$36{,}000}$

2. *Determine critical value $t_{\alpha/2}$*

$CL = 0.99$

$\alpha = 1 - 0.99 = 0.01$

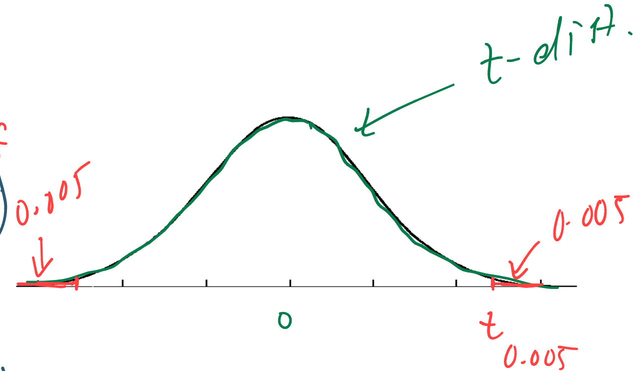
$\alpha/2 = 0.01/2 = 0.005$

$t_{\alpha/2} = t_{0.005}$

$= \text{invT}(\underbrace{1-0.005}_{1-\alpha/2}, \underbrace{33}_{df})$

$\boxed{t_{0.005} = 2.73}$

(stored or calc)



3. *Find margin of error* E

$E = t_{\alpha/2} * \frac{s}{\sqrt{n}}$

$s = 14{,}114$ (given)

$= \frac{2.73 * 14114}{\sqrt{34}} = 6615.9808\ldots$

$\boxed{E = \$6615.98}$

4. *Construct confidence interval*

$CI: (\bar{x} - E, \bar{x} + E) = (36000 - 6615.98, 36000 + 6615.98)$

$CI = (\$29384.02, \$42615.98)$

5. *Interpretation of CI*

(M→E) "I am 99% confident that the average cost of a wedding in US cities in 2017 is between \$29384.02 and \$42,615.98"

↑ notice: mean do have units!

GRAPHING CALCULATOR (TI-83 OR 84)

Instructions: (a) STAT ⇒ TESTS ⇒ TInterval

(b) Enter { Data → if individual values are entered in list; Stats → if using summary statistics

Ex: A Statistics instructor wanted to find out the mean amount of time PCC students spent on social media per day. She randomly sampled 12 students and found out how many minutes per day they used social media. Those results are below: $n = 12$ $\quad df = 11$

sample { 80 112 95 72 150 120 85 67 92 102 50 115 } { to get $\bar{x}$ use stat → calc → 1-VARSTATS

Assuming that such times for all PCC students are normally distributed, make a 95% confidence interval for the corresponding population mean amount of time spent on social media for all students at PCC.

1. *Find point estimate (sample mean)* → μ so sample mean $\bar{x}$

$$\boxed{\bar{x} = 95 \text{ minutes per day}}$$

2. *Determine critical value $t_{\alpha/2}$*

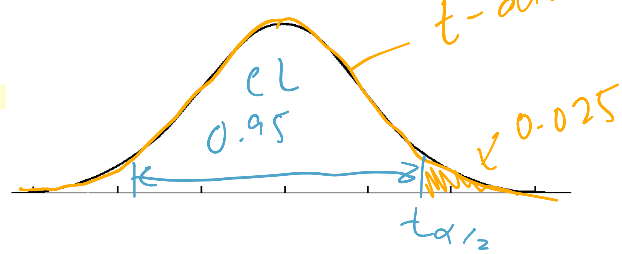
$CL = 0.95$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$$t_{\alpha/2} = t_{0.025} = \boxed{2.20}$$

$$= \text{invT}(1 - 0.025, 11)$$



3. *Find margin of error*

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = \frac{2.20^{*} \cdot (27.06893)}{\sqrt{12}} = 17.19877745\ldots = 17.1988$$

$$\boxed{E = 17.1988 \text{ min/day}}$$

4. *Construct confidence interval*

$$CI: (\bar{x} - E, \bar{x} + E) : (95 - 17.1988, 95 + 17.1988)$$

$$CI: (77.8012, 112.1988)$$

$$CI: (77.8, 112.2)$$

5. ☆ *Interpretation of CI*

(M→E)

"We are 95% confident that all PCC students spend between 77.8 to 112.2 minutes per day on their social media."

* on average

key don't forget!

(M->E)

Ex: The last example produced an error within 17.2 min/day, and with the fact that we were 95% confident, we believed that the true mean (average) number of minutes spent on social media by PCC (all) students is between 77.8 minutes and 112.2 minutes. per day.

new

Let's say we want to be 95% confident that the true mean (code for population) will be within 5 minutes ($E=5$). What sample size is needed? ($\to n$)

Remember: $E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$. Need to solve for n.

Solve for n:

$$\sqrt{n}\,(E) = \left(t_{\alpha/2} \cdot \frac{s}{\cancel{\sqrt{n}}}\right) \cdot \cancel{\sqrt{n}}$$

$$(\sqrt{n} \cdot E)^2 = (t_{\alpha/2} \cdot s)^2$$

$$\frac{n \cdot \cancel{E^2}}{\cancel{E^2}} = \frac{(t_{\alpha/2})^2 \cdot s^2}{E^2}$$

$$n = \left(\frac{t_{\alpha/2} \cdot s}{E}\right)^2$$

DETERMINING SAMPLE SIZE

Population Standard Deviation σ Known
$n = \left(\dfrac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$

• why use $z_{\alpha/2}$? it is a best guess for $t_{\alpha/2}$ b/c don't have n thus df

Ex: The president of PCC Erika Endrijonas is concerned about the amount of time that her students spend on jobs. She would like to estimate the mean number of hours worked per week by these students. She knows that the standard deviation of the times spent per week on such jobs by all students is 2.5 hours. What sample size should she choose so that she can be 90% confident that the estimate is within .75 hours of the population mean?

Givens

$\sigma = 2.5$ hrs/wk

$CL = 0.9$

$E = 0.75$

Want

sample size $n = ?$

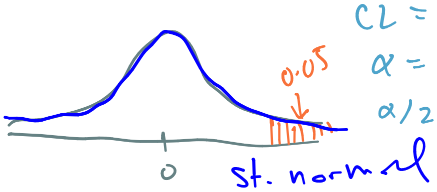
(also want μ)

Want n → do we know σ? Yes!

→ don't know σ?

Can find

$z_{\alpha/2}$ = invNorm(0.05, 0, 1, RIGHT) = 1.64*



$CL = 0.9$

$\alpha = 1 - 0.9 = 0.1$

$\alpha/2 = 0.05$

$$n = \left(\frac{1.64^* \ast 2.5}{0.75}\right)^2$$

$= 30.065\ldots$ people, round ↑

$\boxed{n = 31 \text{ PCC students}}$