Confidence increasing!

Unit 7 – Confidence Interval Estimates

Your ~~Mean~~ SUPER NICE! Professor Colton

Unit 7, Day 2 - Outline

Unit 7 – Confidence Interval Estimates

Sampling Distribution and CLT of $\bar{x}$ Review

Means with Known Sigma (i.e. know population standard deviation)

- Formula
- Practice

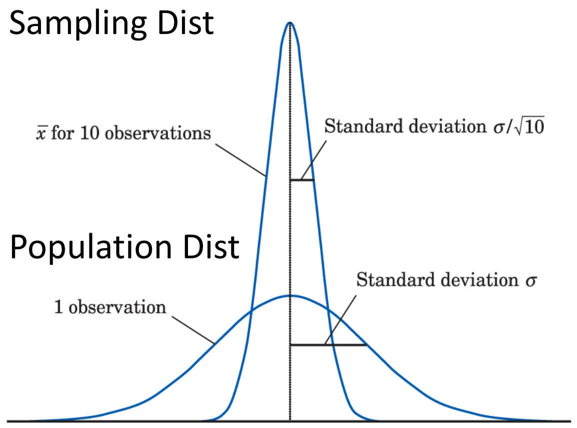
Means with unknown Sigma (i.e. only know sample standard deviation or sample data)

- Formula
- Comparison of Z and T intervals
- t-distribution
- Practice

Overall Summary

Central Limit Theorem (CLT) for $\bar{x}$

REVIEW!!



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

<u>Central Limit Theorem</u>

- Let $\bar{x}$ be the mean of observations in SRS of size n from a population with mean μ and standard deviation σ
- If we take a <u>large enough sample</u>, then
  - The mean of $\bar{X}$ is equal to the mean of the population, μ

    $\mu_{\bar{X}} = \mu$ ($\mu_{\bar{X}}$ in words = mean of the sampling distribution of x-bar)

  - The standard deviation is equal to σ divided by the square root of n

    $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ ($\sigma_{\bar{X}}$ in words = standard deviation of the sampling distribution of x-bar)

  - And this distribution will be approximately Normal!

$$\bar{X} \sim Normal(mean = \mu, SD = \frac{\sigma}{\sqrt{n}})$$

<u>*Summary*</u>

If X has mean μ and sd σ

- *(Referring to the population, selecting a sin person)*

$\bar{X}$ is Normal with mean $\mu_{\bar{X}} = \mu$ and SD $\sigma_{\bar{X}} =$

- *(Referring to the sampling distribution, selecting a group of people and summarizin*

Confidence Intervals Again!

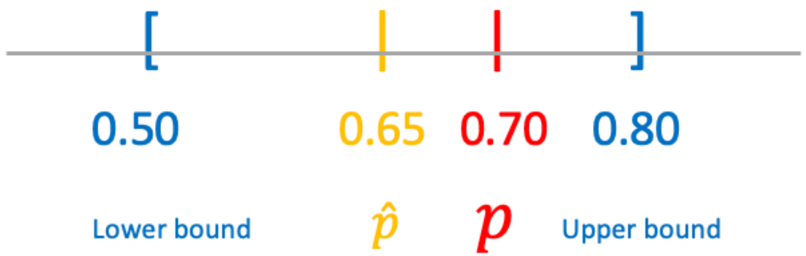
Estimating Parameters

Point Estimates

- Using a statistic to estimate a parameter (for means we use $\hat{p}$ or $\bar{x}$ to estimate p *or* μ, respectively.

Interval Estimates

- Give a range for what we think the population parameter is.

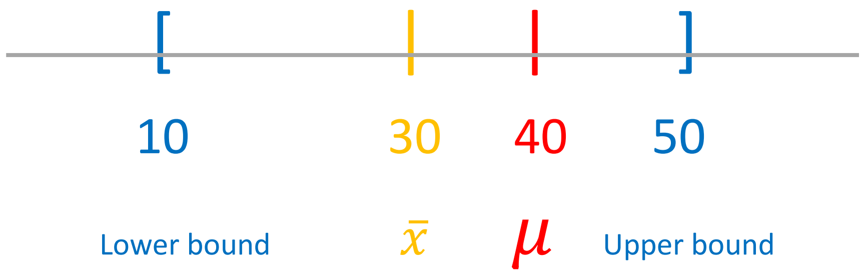


Proportions

- CI for **proportions** deal with categorical variables (yes / no).

Means

- CI for **means** deal with quantitative variables (numeric).



Final Confidence Interval for μ

Now the goal is to find an estimate for the unknown population mean μ!

*** We are going to assume that the population standard deviation **σ** is known!

1 Mean Z Interval

* Same Critical Value as with a 1 Proportion Z Interval

C.I. = Point Estimate ± Margin of Error,   MOE = CV * SE

$$= \bar{x} \pm Z^* \sigma_{\bar{x}}$$

$$= \bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad \left(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + Z^* \frac{\sigma}{\sqrt{n}}\right)$$

Same interpretation too!!

I am % confident that the true/population parameter + context is between (lower bound) and (upper bound).

Using Calc!

Setup

Suppose the number of points scored by your favorite basketball team is normally distribution with known population standard deviation **σ** = 15 points.

The mean from a random sample of 35 games was 72 points. **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!

1. ZInterval
   a) Input = Stats
   b) **σ** = population standard deviation
   c) $\bar{x}$ = sample mean
   d) n = sample size
   e) C-Level = Confidence level (as a decimal or whole number, both work)

Interpret results:
??

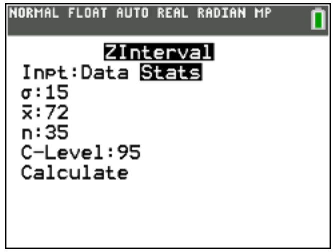
Using Calc!

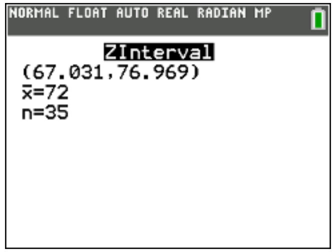
Setup

Suppose the number of points scored by your favorite basketball team is normally distribution with known population standard deviation **σ** = 15 points.

The mean from a random sample of 35 games was 72 points. **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!





1. ZInterval
   a) Input = Stats
   b) **σ** = population standard deviation
   c) $\bar{x}$ = sample mean
   d) n = sample size
   e) C-Level = Confidence level (as a decimal or whole number, both work)

*Show work: 95% CI = ZInterval(**σ** = 15, $\bar{x}$ = 72, n = 35, C-Level = 95) → (67.031, 76.969)*

Interpret results:

We are 95% confident that the true mean number of points scored by my favorite basketball team is between 67.031 and 76.969

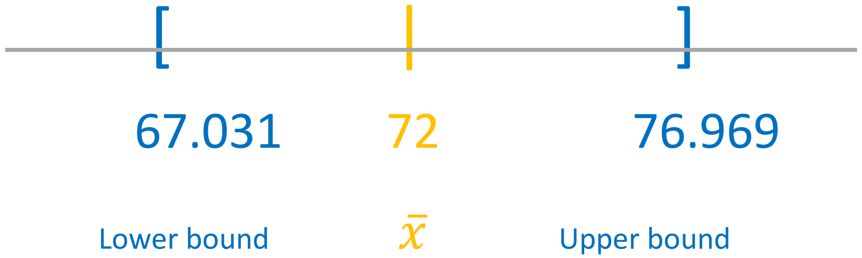
Using Calc!

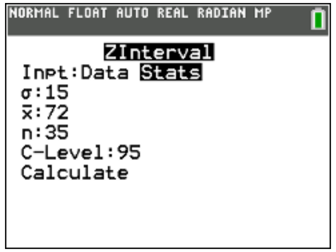
Setup

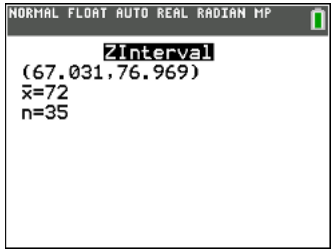
Suppose the number of points scored by your favorite basketball team is normally distribution with known population standard deviation ~~σ = 15 points.~~

The mean from a random sample of 35 games was 72 points and SAMPLE standard deviation of 12 points. **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!







1. ZInterval
   a) Input = Stats
   b) **σ** = population standard deviation
   c) $\bar{x}$ = sample mean
   d) n = sample size
   e) C-Level = Confidence level (as a decimal or whole number, both work)

Show work: *ZInterval(Input = Stats, sigma = 15, x-bar = 72, n = 35, C-Level = 0.95) = (67.031, 76.969)*

Interpret results:

a) My favorite basketball team scored an average of between 67 points and 76 points with a 95% confidence level

The true mean points scored per game is between 67 points and 76 points with a 95% confidence level

My belief my guess of where this unknown population mean is, context, the actual numbers we calculated, some confidence statement

Another LCQ

Setup: Lets assume the population of ACT scores is normally distribution with known population standard deviation **σ** = 3.5 points. From a random sample of 15 students, there was a sample mean score of 24.

1) Calculate the 85% Confidence Interval

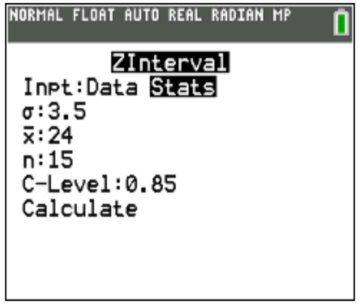
2) Interpret this interval.

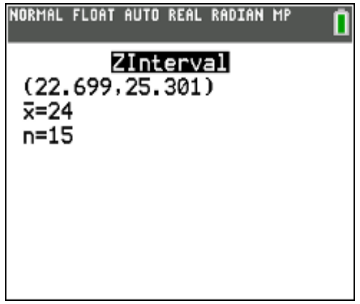
Another LCQ

Setup: Lets assume the population of ACT scores is normally distribution with known population standard deviation **σ** = 3.5 points. From a random sample of 15 students, there was a sample mean score of 24.

1) Calculate the 85% Confidence Interval = *ZInterval(**σ** = 3.5, $\bar{x}$ = 24, n = 15, C-Level = 85) → (22.699, 25.301)*

Don't forget to include the name of the calculator function!





2) Interpret this interval.

Common answers:

a) *We are 85% confident the ACT score of students is between 22.699 and 25.301 → MISSING parameter (true mean)*

b) *We are 85% confident that the true population mean is 22.699 and 25.301 → MISSING context (ACT score of students)*

c) *We are 85% confident that population mean of students scores on the ACT is between 22.69 and 25.301*

Option C is PERFECT!! Has all the parts we need!

This is also correct: We are 85% confident that the true mean ACT score is between 22.699 and 25.301 → as long as we get enough context, which this does!

There is NO NEED TO BE CREATIVE!!
Use the general format, that is what I am

One more LCQ

Setup: Suppose an instructor wishes to predict the average time needed to take the final exam. A random sample of 45 students shows a mean of 1.4 hours. The population standard deviation is known to be 0.25 hours.

1) Calculate and interpret the 97% Confidence Interval.

2) If I increase the sample size to 60, find the margin of error of the new CI. Is it smaller or larger than the MOE in (1)?

3) If we were mistaken and the actual population standard deviation is 0.75 hours. With a sample size of 45, will a 97% CI be wider or narrower than the result in (1)? Find the new interval.

One more LCQ

Setup: Suppose an instructor wishes to predict the average time needed to take the final exam. A random sample of 45 students shows a mean of 1.4 hours. The population standard deviation is known to be 0.25 hours.

SHOW YOUR WORK! For calculating an interval, that means writing the name of the procedure and the inputs!

1) Calculate and interpret the 97% Confidence Interval.

Calculation: Zinterval(input = Stats, st dev = 0.25, mean = 1.4, n = 45, C-Level: 0.97) → (1.3191, 1.4809)

a) I'm 97% confident that the mean of the time it takes to take the exam is between 1.3191 and 1.4809 hours → MISSING true

b) We are 97% confident that the sample mean of the time to take the test is between 1.3191 and 1.4809 hour → WRONG! We are trying estimate the POPULATION mean, we already know what the SAMPLE mean is (DON'T write SAMPLE!)

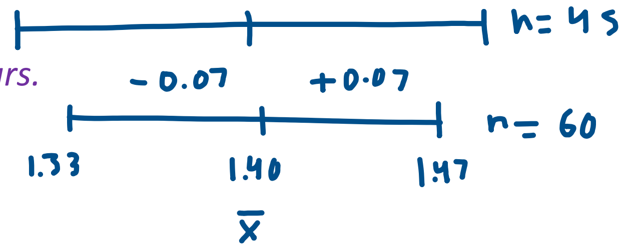
c) CORRECT!!! I'm 97% confident that the true mean of the time it takes to take the exam is between 1.3191 and 1.4809 hours

2) If I increase the sample size to 60, find the margin of error of the new CI. Is it smaller or larger than the MOE in (1)?

After increasing our sample size to 60, 97% CI = (1.33, 1.47).

New MOE = Width / 2 = (UB – LB) / 2 = (1.47 – 1.33) / 2 = 0.055, which is a smaller interval than the original MOE. We knew it would be smaller even before calculating the new interval!

Bonus, interpretation: We are 97% confident the true mean of the time to take the exam is between 1.33 and 1.47 hours.



3) If we were mistaken and the actual population standard deviation is 0.75 hours. With a sample size of 45, will a 97% CI be wider or narrower than the result in (1)? Find the new interval.

Wider! $MOE = Z^* \frac{\sigma}{\sqrt{n}}$*, increasing the numerator of the SE (and keeping the sample size and CL the same) will increase the MOE and this a wider interval!*

New Calculation: Zinterval(input = Stats, st dev = 0.75, mean = 1.4, n = 45, C-Level: 0.97) → (1.1574, 1.6426)

A New CI for a New Unknown

- We've looked at how to create a CI for a population mean:

$$\text{C.I.} = \bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

- In doing so, what did we have to assume?? Population standard deviation **σ** was KNOWN!
- Is this practical?? NO!! It's a veryyy big assumption, unlikely to be known.

- What do we do when we don't have this information? Any ideas??? $\cancel{\sigma} \longrightarrow s$
  - If we known the value of **σ**, we can **estimate** it with our **sample standard deviation**!!
  - This intuitively makes sense! It is essentially the same thing we are doing with the population mean too. We use our sample mean as a point estimator!
    - And the population mean is ultimately what we are after, so we then give an interval around the point estimator, the sample mean.
    - With the population standard deviation, we aren't necessarily interested in this quantity, but we still need it! So just having a single point estimator suffices!

Final Confidence Interval for μ, unknown $\boldsymbol{\sigma}$!

The goal is still to find an estimate for the unknown population mean μ!

*** But now we have to **estimate** the <u>population standard deviation</u> $\boldsymbol{\sigma}$ is with our <u>sample standard deviation</u> s!

- So we have a new procedure! And specifically, a new <u>critical value</u> and <u>standard error</u>!
- It is now based on the ***t*-distribution** rather than the <u>standard normal distribution Z</u>.

<u>1 Mean T Interval</u>

This is wayyy more common than a Z Interval!

C.I. = Point Estimate ± Margin of Error,   MOE = CV * SE

$= \bar{x} \pm t^* \sigma_{\bar{x}}$

$= \bar{x} \pm t^* \frac{s}{\sqrt{n}}, \quad df = n - 1$

$\rightarrow \left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$

$\cancel{\sigma} \longrightarrow s$

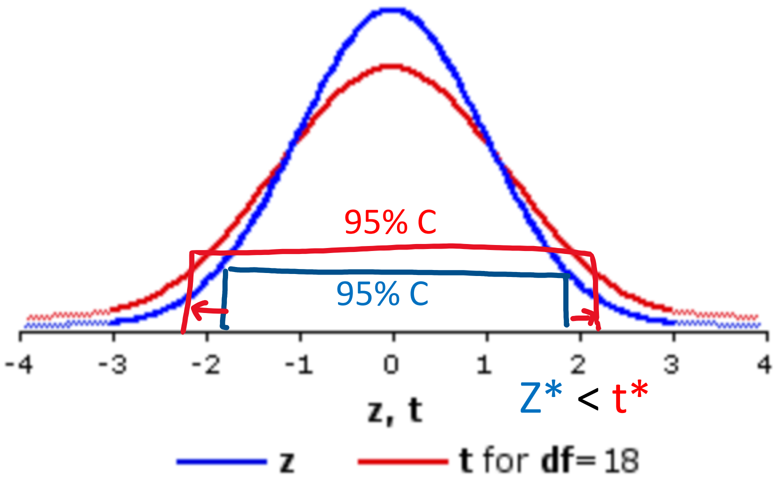
df = Degrees of freedom

Still same interpretation!!

I am <u>% confident</u> that the true/population <u>parameter + context</u> is between <u>(lower bound)</u> and <u>(upper bound)</u>.

CIs for Known vs Unknown **σ**

- Because we have to **estimate** an extra parameter, there is inherently more variability in a *t*-interval!
    - This makes sense! Sample means vary from sample to sample, so do the standard deviations!
- So to account for that extra level of uncertainty, ***t*-intervals** produce wider intervals compared to Z-intervals (for the same confidence level and sample size).



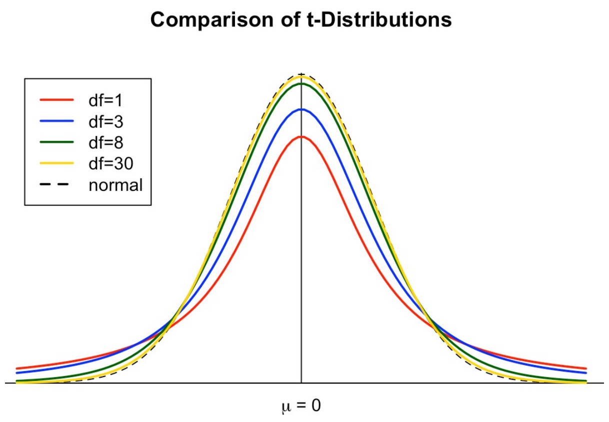
http://vassarstats.net/textbook/ch9pt3.html

- Z-interval is the GOLD standard (more precise, which is great!), so we want to do this procedure IF WE CAN
- But if not and we only have the SAMPLE STANDARD DEVIATION, then we have to do T-Interval

t-Distribution

We are just going to briefly mention some ideas about the t-Distribution.

https://www.geo.fu-berlin.de



Shape and Parameters

- Symmetric and unimodal distribution (slightly resembles a bell shape).
- The t-distribution has heavier tails than the Z distribution!
    - This means there is more probability near edges and likewise less probability in the middle!
- Indexed by the degrees of freedom, *df = n – 1 = sample size – 1*
    - This tells us exactly which t-distribution we are talking about!

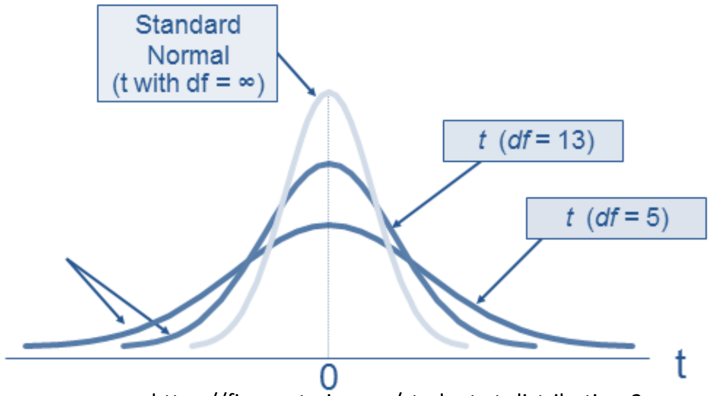
Why do we need to use t??

- Recall, as the sample size *n increases*, the sampling distribution

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \rightarrow Z = \text{Normal}(\mu = 0, \sigma = 1)$$

- Well now, because we have to substitute *s* for, this becomes

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



https://financetrain.com/students-t-distribution-2

Interesting Tidbit

- As the df goes towards infinity, t becomes Z!

Doper applet!

- **USE T WHEN DON'T KNOW SIGMA**
- PRODUCES WIDER INTERVALS THAN Z FOR THE SAME CONFIDENCE

- FOCUSING ON APPLICATION
    - KNOWING WHEN TO USE WHAT AND HOW TO LET CALC DO IT

Using Calc!

Setup

Lets assume the population of SAT scores is normally distribution with unknown population standard deviation.

From a random sample of 6 students, there was a sample mean score of 1190 and sample standard deviation of 205.91 points. **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!

TInterval

- Option 1) Input = Stats

a) $\bar{x}$ = sample mean
b) Sx = sample standard deviation
c) n = sample size
d) C-Level = Confidence level (as a decimal or whole number, both work)

- Option 2) Input = Data
  - Enter raw data in L_1

a) List = L_1
b) Freq = 1
c) C-Level = Confidence level (as a decimal or whole number, both work)

Score
1300
1200
1190
1050
1500
900
Mean = 1190
SD = 205.91

Interpret results:
??

Using Calc!

Setup

Lets assume the population of SAT scores is normally distribution with unknown population standard deviation.

From a random sample of 6 students, there was a sample mean score of 1190 and sample standard deviation of 205.91 points. **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!

TInterval

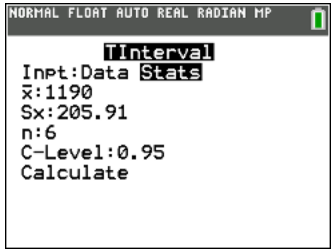
- Option 1) Input = Stats

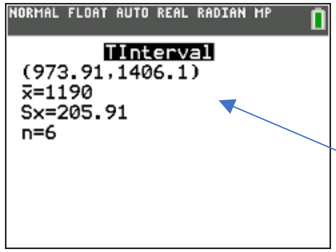
a) $\bar{x}$ = sample mean

b) Sx = sample standard deviation

c) n = sample size

d) C-Level = Confidence level (as a decimal or whole number, both work)





Same intervals! (maybe a little roundoff error from $\bar{x}$ *or s)*

TInterval(input = Stats, $\bar{x}$ = 1190, Sx = 205.91, n = 6, C-Level: 0.97) → (973.91, 1406.1)

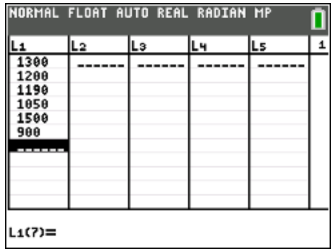
- Option 2) Input = Data
  - Enter raw data in L_1

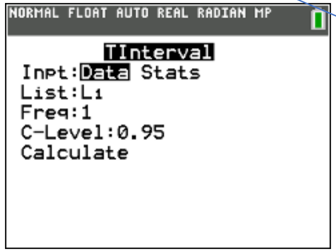
a) List = L_1

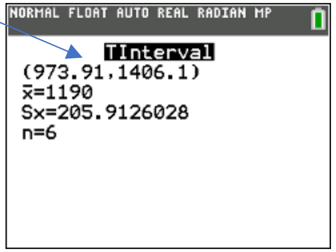
b) Freq = 1

c) C-Level = Confidence level (as a decimal or whole number, both work)

Score
1300
1200
1190
1050
1500
900
Mean = 1190
SD = 205.91







Interpret results:

We are 95% Confident that the true average SAT scores for students is between 973.91 and 1406.1 points.

TInterval(input = Data, List = L_1, Freq = 1, C-Level: 0.97) → (973.91, 14

Summarizing LCQ

Setup: We know that number credit hours students at CSCC is normally distributed. From a random sample of 20 students, there was a sample mean 8.5 credit hours and sample sd = 2 credit hours.

1) (Only) Calculate the 90% Confidence Interval.

7.73, 9.27

Tinterval(sd = 2, x-bar 8.5, sample size 20 confidence level 95) = 7.7267, 9.2733

2) Admissions department tells us the true (population) standard deviation is 1.25 credit hours. (Only) Calculate the 90% Confidence Interval.

T –Interval(X-bar - 8.5, SD - 1.25, N=20) -> (LB = 8.0167, UB = 8.9833)

Tinterval (sd =1.25, x-bar 8.5, sample size 20 confidence level 90) = 8.0167, 8.9833

T -Interval = LB - 8.0167, UB - 8.9833) , X-bar - 8.5, SD - 1.25, N=20, C-LEVEL = 90

EVERYONE IS WRONG……………

Z-INTERVAL!!!!! *ZInterval -> Stats, ~~Sx~~ σ = 1.25 , $\bar{x}$ = 8.5, n = 20, C-Level = 0.9 = (8.0402, 8.9598)*

3) Out of the 20 students, 8 said they were taking a Statistics course. YES/ NO variable -> categorical (Only) Calculate the 85% Confidence Interval.

SHOW WORK AND FINAL RESUTLS

~~tinterval, sd =1.25, x-bar 8 sample size 20 confidence level 90 = 7.58, 8.4193~~ 0 points!!!

1-PropZInt → x = 8, n = 20, C-Level = 0.9 = (0.24231, 0.55769)

~~TRUE MEAN~~

TRUE PROPORTION

Summarizing LCQ

Setup: We know that number credit hours students at CSCC is normally distributed. From a random sample of 20 students, there was a sample mean 8.5 credit hours and sample sd = 2 credit hours.

1) (Only) Calculate the 90% Confidence Interval.

2) Admissions department tells us the true (population) standard deviation is 1.25 credit hours. (Only) Calculate the 90% Confidence Interval.

3) Out of the 20 students, 8 said they were taking a Statistics course. (Only) Calculate the 85% Confidence Interval.

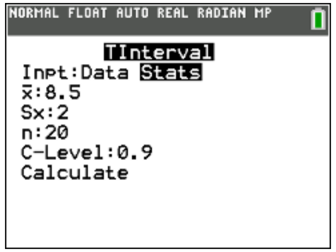
LCQ

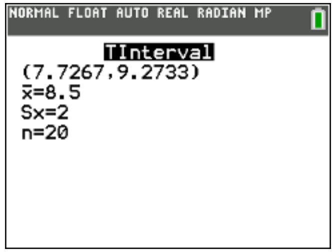
*** **Need to be able to recognize which type of interval to make**! To help, <u>first think about the response variable</u> (or the <u>parameter of interest</u>)!
- Mean μ or proportion p?? (Quantitative or Qualitative)
- Then think about what information you have, what you need, and what to type in calc!

Setup: We know that number credit hours students at CSCC is normally distributed. From a random sample of 20 students, there was a sample mean 8.5 credit hours and sample sd = 2 credit hours.

1) (Only) Calculate the 90% Confidence Interval.

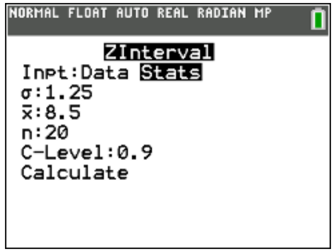
- *Tinterval(input = Stats, $\bar{x}$ = 8.5, Sx = 2, n = 20, C-Level= 0.9 → (7.7267, 9.2733)*

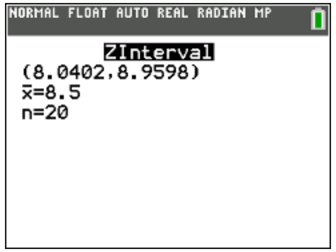




2) Admissions department tells us the true (population) standard deviation is 1.25 credit hours. (Only) Calculate the 90% Confidence Interval.

- ***NOT CORRECT**: TInterval → Stats, , $\bar{x}$ = 8.5, Sx = 1.25, n = 20, C-Level = 0.9 → (8.0167, 8.9833)*
- *We KNOW what the value of $\boldsymbol{\sigma}$ is now!! So we <u>don't have to estimate it</u> anymore and can therefore use a **Z Interval**!!*
    - *This will give us <u>a more precise (better) interval</u> for the same confidence level! So this would be the CORRECT type of interval*
- *Zinterval(input = Stats, $\boldsymbol{\sigma}$ = 1.25 , $\bar{x}$ = 8.5, n = 20, C-Level = 0.9 = (8.0402, 8.9598)*



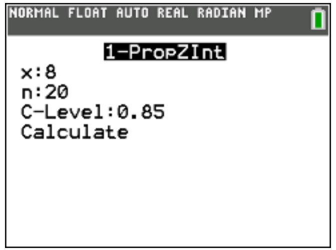


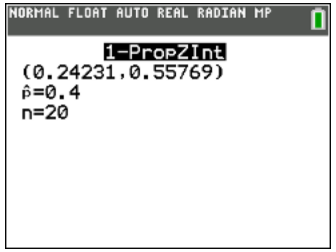
3) Out of the 20 students, 8 said they were taking a Statistics course. (Only) Calculate the 85% Confidence Interval.

- *We were given are **sample proportion** → NEED TO FIND PROPORTIONS INTERVAL!!!!!!*
- *The <u>response variable has changed</u>! We NO LONGER are <u>after the amount of credit hours</u>, rather <u>whether or not students are taking statistics</u>!*
    - *This is a <u>YES/NO variable</u>, i.e. **CATEGORICAL**!!*
- *So we CANT use <u>mean</u> formulas, we need to use **PROPORTIONS and 1PROPZINT**!!!*

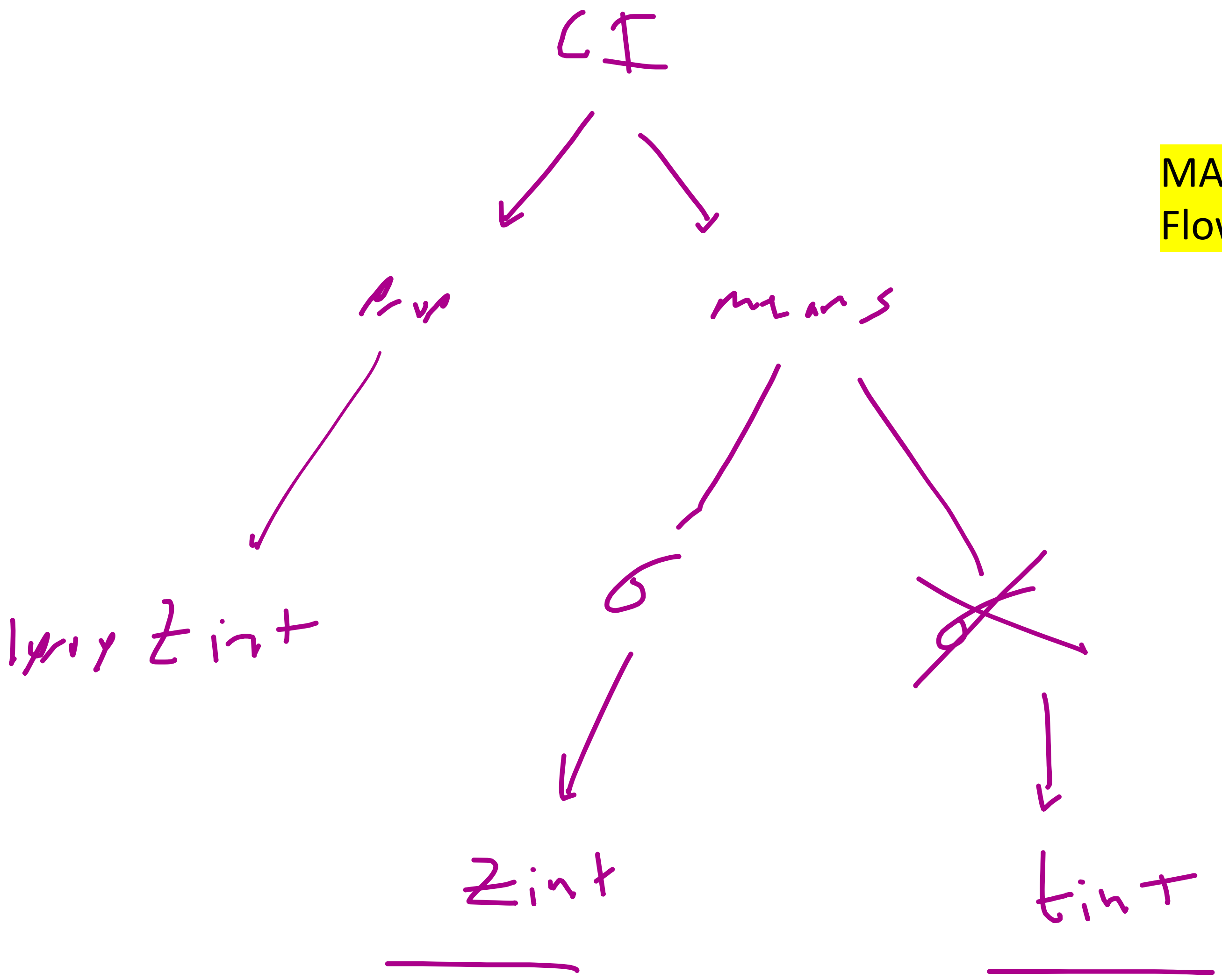
1-PropZInt(x = 8, n = 20, C-Level = 0.9 = (8.0402, 8.9598)

1PropZInt!!! → (0.24231, 0.55769) true proportion of CSCC students taking a stats course!





MAKE THIS NICER LATER!!
Flow chart decision process



When do I use which? *p* vs *Z* vs. *t*

Use **1 Proportion Z Interval** when…

- You are asked to find a Confidence interval for a Population **Proportion**
  - Dealing with Proportions (i.e. a success or failure)

Use **T Interval** when...

- You are asked to find a Confidence interval for a Population **Mean**
  - This applies to estimating the population mean, not proportion
  - There is **no** “success” or “failure” being measured.
  - Generally always the population standard deviation will be unknown!

Use **Z Interval** only for…

- (Very uncommon) Finding a CI for a Population mean when the population standard deviation is known.

As always, think about what is being measured for each observation/individual!

USE THE CORRECT NOTATION AND FORMULAS!!!!

Means vs. Proportions: When do I use which formulas?

Looking at a categorical variable that has a **"success" and "failure"**? (sample proportions)

Sampling Distributions:

- Mean: $\mu_{\hat{p}} = p$ aka the population proportion
- Standard Deviation: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Confidence Interval Procedure: **1-PropZInt**

Formula: $\hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Looking at a **quantitative** variable?
(sample means)

Sampling Distributions:

- Mean: $\mu_{\bar{X}} = \mu$ aka the mean of the population
- Standard Deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Confidence Intervals Procedures and Formulas:

ZInterval $\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}$

TInterval $\bar{x} \pm t^* \frac{s}{\sqrt{n}}, df = n - 1$