We Are Back! ☹ or ☺??

Unit 7 – Confidence Intervals and Sample Size

Your More-Than-Halfway-Done Professor Colton

LU 7 - Outline

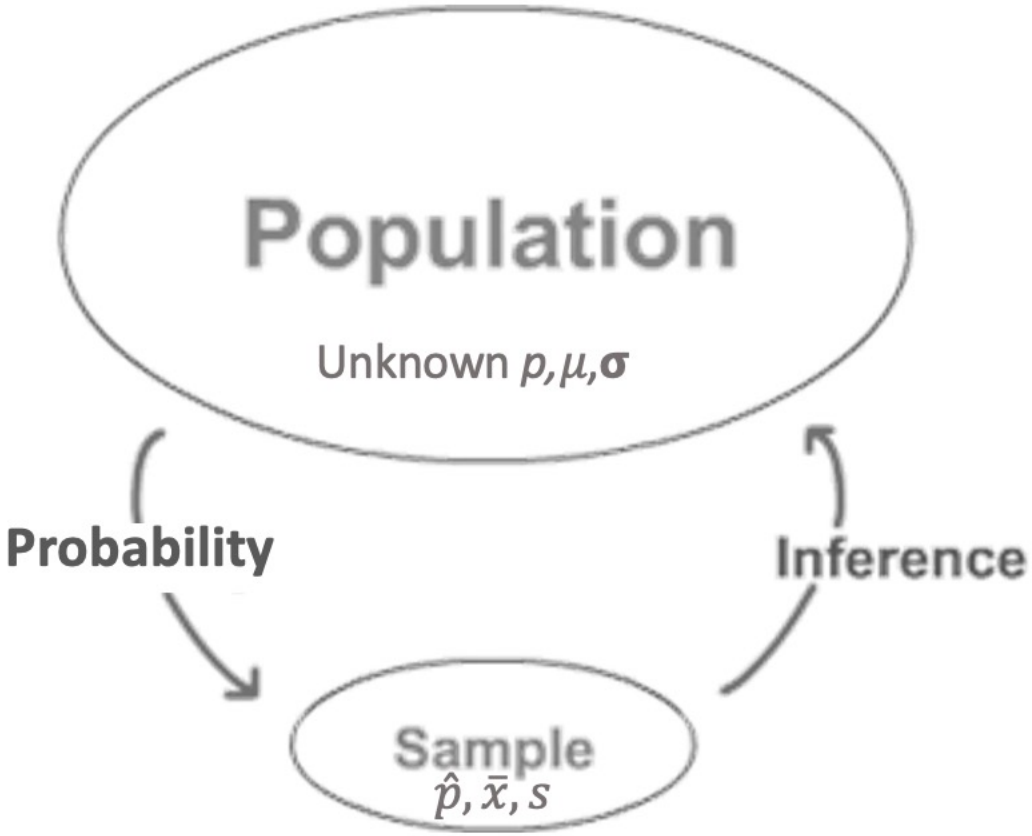
Learning Unit 7 – Confidence Intervals and Sample Size

Intro / Motivation – Why Confidence Intervals
- Point Estimate

Proportions

- Sampling Distribution of $\hat{p}$
- CLT for $\hat{p}$
- Confidence Interval Estimate
- Margin of Error
- Interpreting CI
- Minimum Sample Size

Motivation – Populations and Samples



Motivation – Why Confidence Intervals

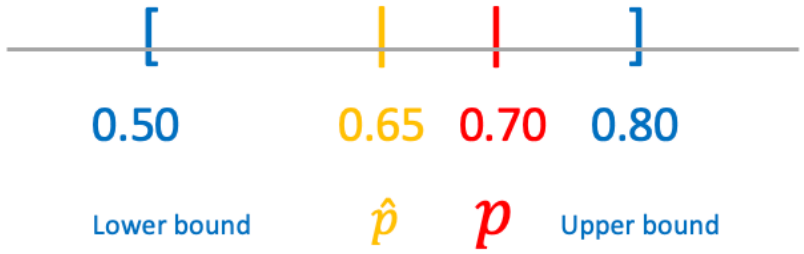
Estimating Parameters

Point Estimates

- Using a statistic to estimate a parameter (this means we use $\hat{p}$ or $\bar{x}$ to estimate *p or* μ, respectively
- It is our best guess.
- Usually the statistics do not equal the parameter (remember each sample is different; sampling variability).

Interval Estimates

- Give a range for what we think the population parameter is.
- Takes into account sampling variability.



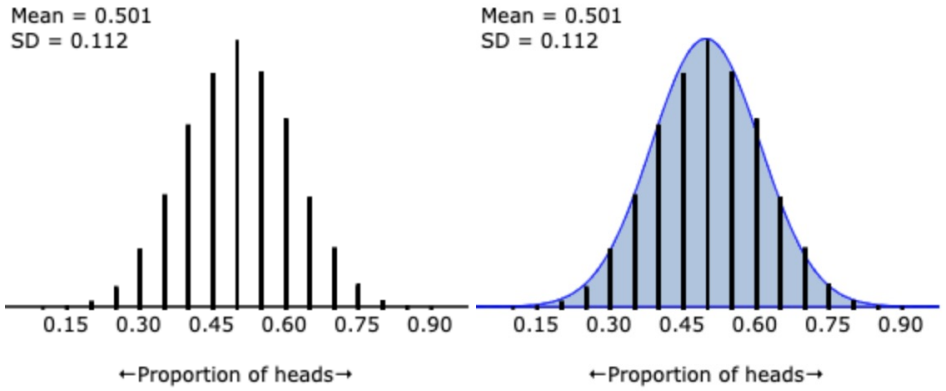
Sampling Distribution of $\hat{p}$

Sampling Distribution of Sample Proportions

- Let $\hat{p}$ be the sample proportion of successes in a random sample of size n from a population with true proportion of success p.
- The mean of $\hat{p}$ is the true population proportion; $\mu_{\hat{p}} = p$
- The standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- Again, we actually know the behavior of this distribution because of the Central Limit Theorem!

Ex) Results from 10,000 samples of 20 coin tosses
→ *p = 0.5*



Applet!!!

Central Limit Theorem for $\hat{p}$

Central Limit Theorem

- Let $\hat{p}$ be the sample proportion of successes in a random sample of size n from a population with true proportion of success p.
- If we take a large enough sample, then
  - The mean of $\hat{p}$ is equal to the population proportion, p

$$\mu_{\hat{p}} = p$$

  - The standard deviation of $\hat{p}$ is equal to

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

  - And the distribution of $\hat{p}$ is approximately Normal!

Large Enough Sample

- If both $np \geq 5$ AND $n(1-p) \geq 5$
- In other words, if we have at least 5 successes and 5 failures!

$$\hat{p} \sim Normal\left(mean = p, SD = \sqrt{\frac{p(1-p)}{n}}\right)$$

- Further,
  - As we increase the sample size, the sampling distribution of $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$ approaches a standard normal distribution ($\mu = 0, \sigma = 1$), regardless of the shape of the population distribution!

CLT Assumptions & Conditions

CLT Assumptions

- **Randomization Condition:** The data values must be sampled randomly.
- **Large Enough Sample Condition:** The sample size, n, has to be large enough to expect at least 5 successes and 5 failures;

  $np \geq 5$ **AND** $nq \geq 5$

Example: Spam

Setup

- In 2003, a major vendor of anti-spam software claimed that the proportion of email consisting of unsolicited spam was 0.40; i.e., 4 out of every 10 email messages are spam.
- Suppose 50 email messages are selected at random.

Questions

a) Carefully sketch and label the distribution of the sample proportion $\hat{p}$.

b) What is the probability the sample proportion is greater than 0.50?

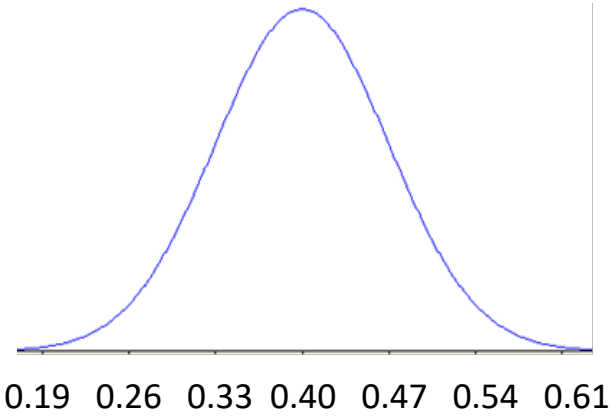
c) Find the probability the sample proportion will be between 0.32 and 0.37.

Solution

Check the Conditions

- ✓ Randomization Condition
  - The 50 emails are a random sample from the population
- ✓ Large Enough Sample Condition
  - $np \geq 5$ AND $nq \geq 5$
  - $50(0.4) = 20 \geq 5$ AND $50(1 - 0.4) = 30 \geq 5$

- Conditions for CLT are met, so we can use the results to find the probabilities of interest!

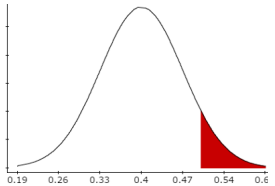


Solve

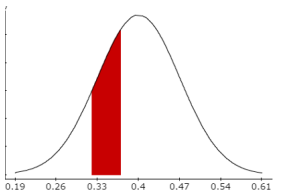
a) Find / define the distribution of sample proportions of spam emails.

- Based on CLT, we know:
  - $\mu_{\hat{p}} = p = 0.4$ and $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.4(0.6)}{50}} = 0.069 \approx 0.07$.
  - So…

$$\hat{p} \sim Normal(mean = 0.4, SD \approx 0.07)$$



b) $P(\hat{p} \geq 0.50) = normalcdf\left(lower = 0.5, upper = 10000, \mu = 0.4, \sigma = \sqrt{\frac{0.4(0.6)}{50}}\right) = 0.0766$



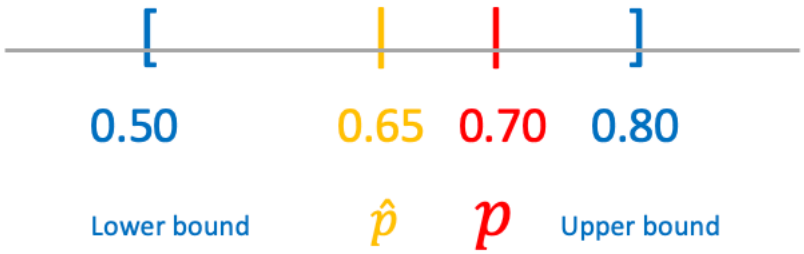
c) $P(0.32 \leq \hat{p} \leq 0.37) = normalcdf\left(lower = 0.32, upper = 0.37, \mu = 0.4, \sigma = \sqrt{\frac{0.4(0.6)}{50}}\right) = 0.2076$

Now we are ready!

How do we build this interval?

What are the different pieces that make up a confidence interval?

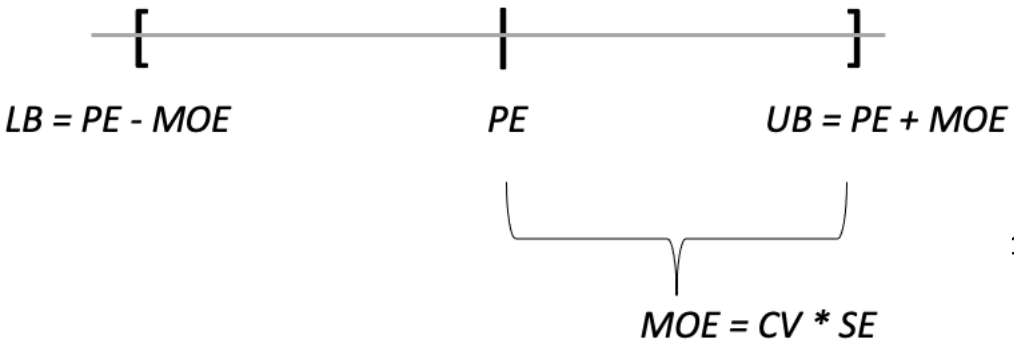
How do we interpret the final interval? What does it mean?



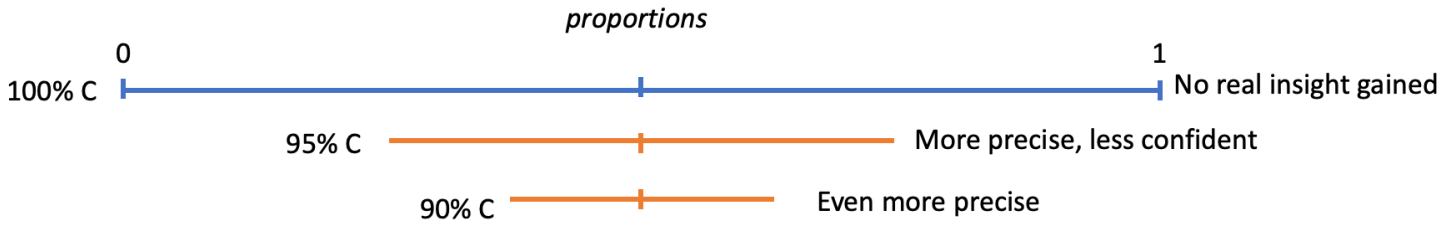
How to Build a Confidence Interval

C.I. = Point Estimate ± Margin of Error

- Point Estimate is your best guess; at the center of the interval.
- Margin of Error (MOE) = Critical Value (CV) * Standard Error (SE).
  - SE (standard deviation of your statistic) measures sampling error.
  - % Confident is determined by confidence level set and incorporated via the Critical Value (CV).
- Smaller MOE, more precise your estimate is.
  - The more confident, the wider your interval is (if everything else stays the same)



Tradeoff between Precision and Confidence



Margin of Error

Margin of Error

MOE = Critical Value (CV) * Standard Error (SE)

$= Z^* \sigma_{\hat{p}}$

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

Now let's breakdown the two pieces!

Standard Error

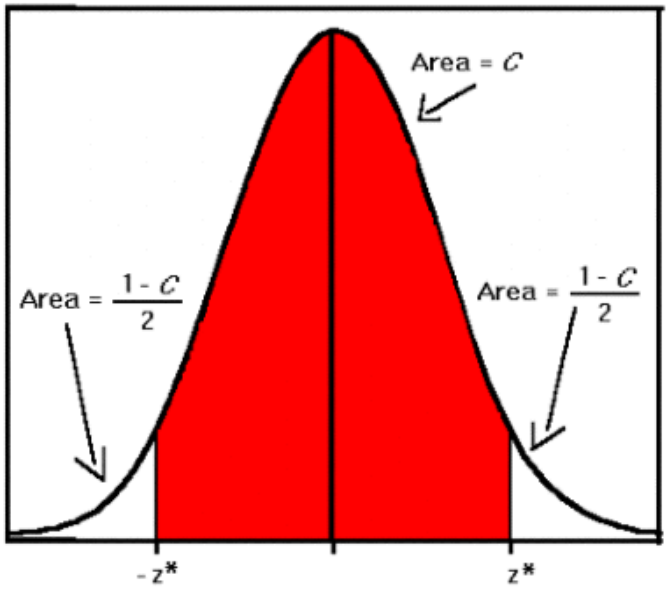
- Measures sampling error.
- The *standard deviation of the sampling distribution* $\hat{p}$ is
  - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- When we don't know *p* (i.e. when making CIs), we substitute $\hat{p}$ in for *p* and it becomes
  - $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- This quantity is now referred to as the *standard error*!

Margin of Error

Critical Value – Z Star

- % Confident determined by confidence level set and incorporated via the Critical Value (CV), Z*.
- The Z-scores that mark the middle %C of the standard normal curve!
- Values are symmetric, so can just find one!

$$Z^* = invNorm(area = \frac{1-C}{2}, \mu = 0, \sigma = 1)$$



http://www.stat.yale.edu/Courses/1997-98/101/confint.htm

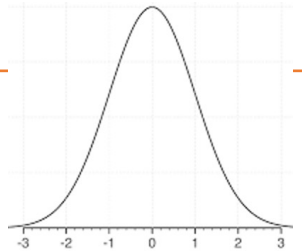
Mini LCQ

Find the Critical Values for the following Confidence Levels:

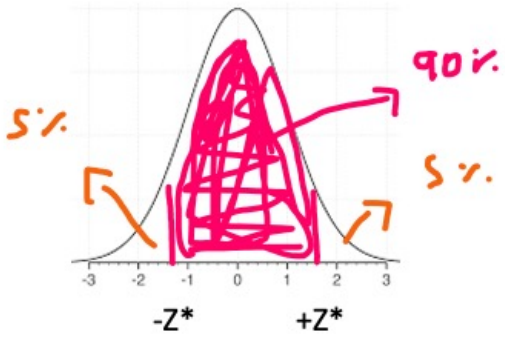
a) 90% Confident
b) 95% Confident

Find the % Confidence based on the following Critical Value

c) Z* = 1.281



LCQ Solution

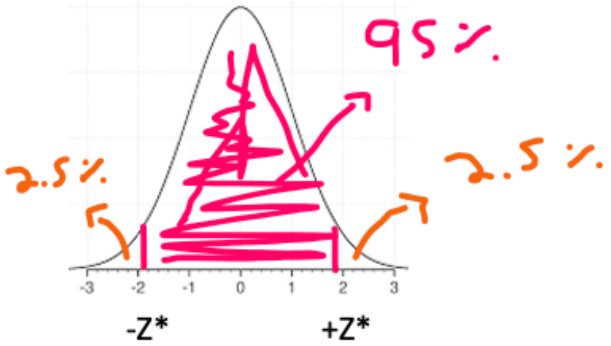


Mini LCQ

Find the Critical Values for the following Confidence Levels:

a) 90% Confident -> *Z* = invNorm(area = 0.05, mean = 0, sd = 1) = -1.645*

$area = \frac{1-C}{2} = \frac{1-0.9}{2} = 0.05$

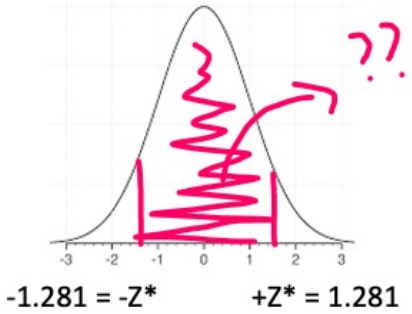


a) 95% Confident -> *Z* = invNorm(area = 0.95+0.025, mean = 0, sd = 1) = 1.96*

(because they are symmetric, could find the upper one as well)

Find the % Confidence based on the following Critical Value

c) Z* = 1.281 -> *normalcdf(lower = -1.281, upper = 1.281, mean = 0, sd = 1) = 0.80 -> 80% Confident*



Final Confidence Interval for *p*

1 Proportion Z Interval

Recall: Our point estimate is the sample proportion $\hat{p} = \frac{x}{n}$, which represents the number of success divided by the sample size.

C.I. = Point Estimate ± Margin of Error

$$= \hat{p} \pm Z^* \sigma_{\hat{p}}$$

$$= \hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \rightarrow \quad \left(\hat{p} - Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Don't forget to **Check the Conditions**

✓Randomization Condition

Need to have a random sample

✓Large Enough Sample Condition

$n\hat{p} \geq 5$AND $n(1-\hat{p}) = n\hat{q} \geq 5$ OR

AT LEAST 5 successes and 5 failures from the sample

Interpreting Confidence Intervals

General Structure

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

Example

95% CI = (0.05, 0.25)

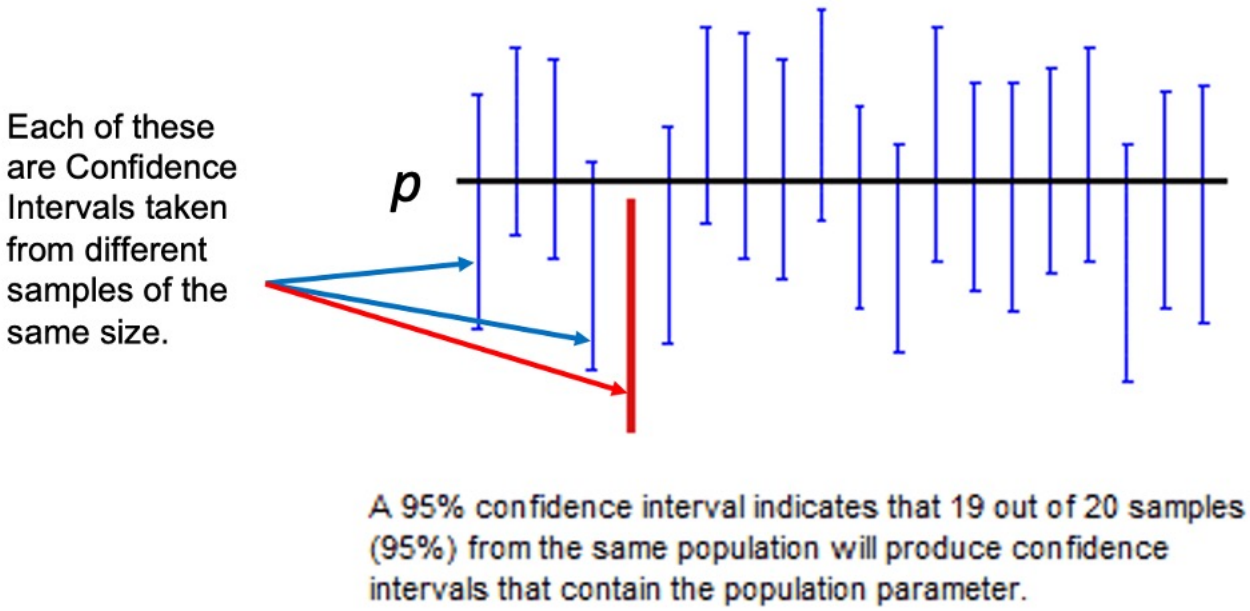
- We are 95% confident that the true (population) proportion of all Columbus residents who enjoy running is between 0.05 and 0.25.

3 Pieces

1. 95% Confident: This is a Confidence Statement
   - Tells us what percent off ALL possible samples result in a CI that captures the true proportion.
2. Parameter + Context: We are talking about population proportions.
   - But what population proportion??? We ALWAYS need context.
3. Interval: The range of plausible values!
   - Uses our sample statistic and MOE.

Interpreting Confidence Intervals

Confidence Interval Interpretation Visualized



Dope applet!

Very Important!

- The confidence level is NOT the probability the parameter is in the interval.
- It refers to the long run capture rate (i.e. over many, many intervals constructed in the same way).
- Either the interval contains the parameter or it does not.

Summarizing LCQ!

Setup

A NatGeo Poll interviewed 1200 hiking enthusiasts and asked “Are you more afraid of spiders or snakes???” Out of the 1200 people, 768 responded “Ewww, snakes….”. **Calculate** and **interpret** the corresponding *95% confidence interval*!

Solution

- $p =$
- $\hat{p} =$
- $CV =$
- $SE =$
- $MOE =$
- $95\%\ CI =$
- *Interpretation:*

Summarizing LCQ!

Setup

A NatGeo Poll interviewed 1200 hiking enthusiasts and asked "Are you more afraid of spiders or snakes???" Out of the 1200 people, 768 responded "Ewww, snakes….". **Calculate** and **interpret** the corresponding *95% confidence interval*!

Solution

- *p = What is the context?? In words, p represents the true proportion of hikers that are more afraid of snakes*
- $\hat{p} = \frac{x}{n} = \frac{768}{1200} = 0.64$
- *CV = Z^* = 1.96*
- $SE = \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.64\ (1-0.64)}{1200}} = 0.0139$
- $MOE = Z^*\sigma_{\hat{p}}$ *= 1.96 * 0.0139 = 0.0272*
- $95\%\ CI = \hat{p} \pm MOE = 0.064 \pm 0.0272 = (0.6128,\ 0.6672)$
- *Interpretation: We are 95% confident that the true proportion of hiking peeps that are more afraid of snakes is between 0.6128 and 0.6672.*

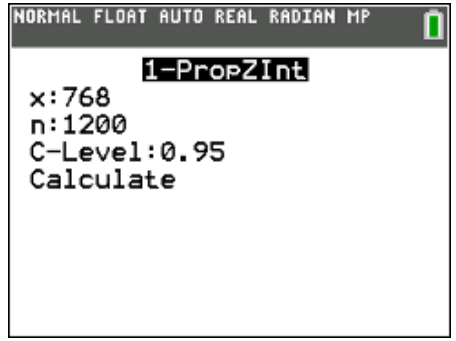
Using Calc!

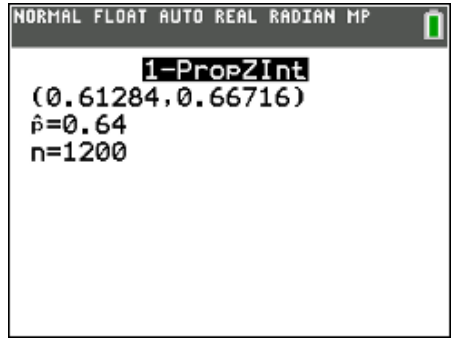
Setup

A NatGeo Poll interviewed 1200 hiking enthusiasts and asked "Are you more afraid of spiders or snakes???" Out of the 1200 people, 768 responded "Ewww, snakes….". **Calculate** and **interpret** the corresponding *95% confidence interval*!

GOAL: Find the Confidence Interval!

1. 1-PropZInt
   a) x = # of successes (people that said yes)
   b) n = sample size
   c) C-Level = Confidence level (as a decimal or whole number, both work)





Interpret results:

- *We are 95% confident that the true proportion of hiking peeps that are more afraid of snakes is between 0.61284 and 0.66716.*

Another LCQ

Setup: 15 out of 23 people from a random sample said their National Championship team is still remaining in their NCAA March Madness Bracket.

1) Check the conditions for a 1-Proportion Z Interval.

2) Calculate the 90% Confidence Interval.

3) Interpret this interval.

Another LCQ

Setup: 15 out of 23 people from a random sample said their National Championship team is still remaining in their NCAA March Madness Bracket.

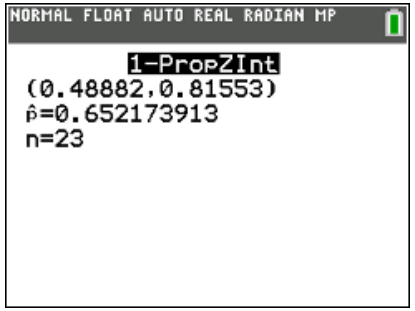
1) Check the conditions for a 1-Proportion Z Interval.

- *Randomization -> Yes, a random sample was taken*
- *Large enough sample -> Yes, there were 15 successes and 8 failures, both are greater than 5.*
  - *Very long way (just for illustration):*

$\hat{p} = \frac{x}{n} = \frac{15}{23} \rightarrow n\widehat{p} = 23\left(\frac{15}{23}\right) = 15 \geq 5$ *AND* $n\hat{q} = 23\left(\frac{23-15}{23}\right) = 23\left(\frac{8}{23}\right) = 8 \geq 5$

2) Calculate the 90% Confidence Interval.

90% CI = (0.489, 0.816)



3) Interpret this interval.

We are 90% confident that the true proportion of people who still have their national championship team remaining is between 0.489 and 0.815.

One more LCQ

Setup: From a random sample 500 people, 64% said they prefer to vacation at the beach compared to the mountains.

1) Calculate the 85% Confidence Interval.

2) If I increase the sample size to 600 people and all else remains constant, what will happen to the new confidence interval?

3) If I change the Interval from Question 1 to be 90% Confident, what will happen to the new confidence interval?

One more LCQ

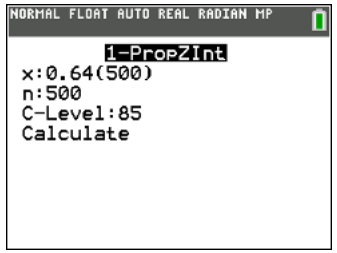
Setup: From a random sample 500 people, 64% said they prefer to vacation at the beach compared to the mountains.

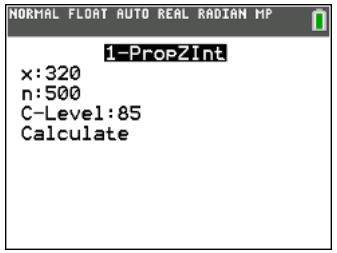
1) Calculate the 85% Confidence Interval.

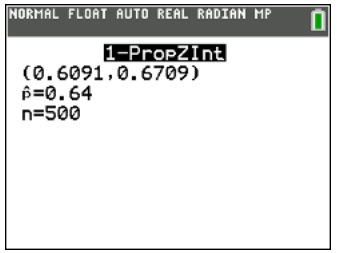
85% CI = (0.6091, 0.6709)

Have to type in x, but weren't given it directly

- *So need to calculate it using $\hat{p}$ and n*







2) If I increase the sample size to 600 people and all else remains constant, what will happen to the new confidence interval?

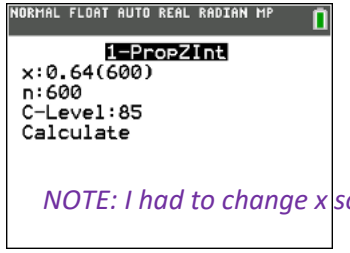
Interval becomes narrower!

This is because of the MOE, and specifically the standard error!

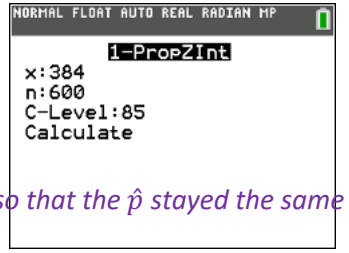
MOE = $Z^*\sigma_{\hat{p}}$

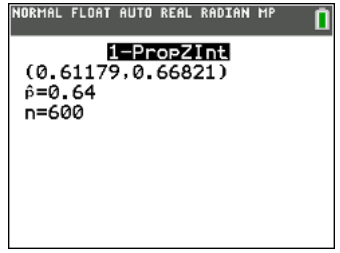
= $Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- *If the confidence level stays the same, then the Z* value doesn't change*
- *BUT with a larger sample size, in the standard error equation we are dividing by a larger number.*
- *Thus making that overall quantity smaller and the MOE smaller -> and confidence interval becomes narrower!*



NOTE: I had to change x so that the $\hat{p}$ stayed the same





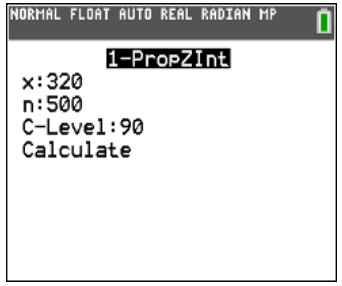
3) If I change the Interval from Question 1 to be 90% Confident, what will happen to the new confidence interval?

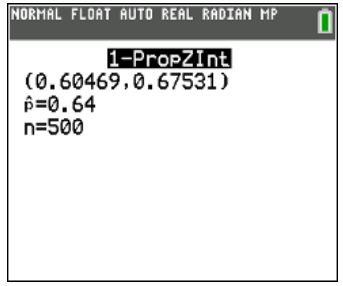
It becomes wider.

Same idea as number 2, BUT now the standard error remains the same (because of the same n and $\hat{p}$).

And the Z value changes because of the new confidence level!*

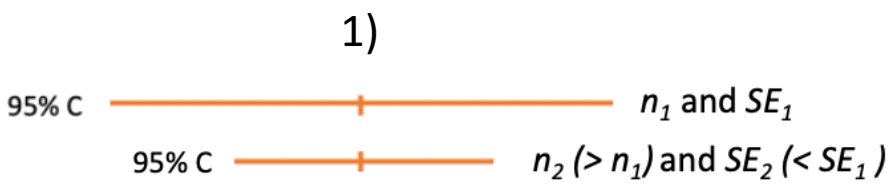
- *To be more confident, we need to cover more values.*
  - *So have a larger critical value*
- *This makes the MOE increase and our CI become wider!*

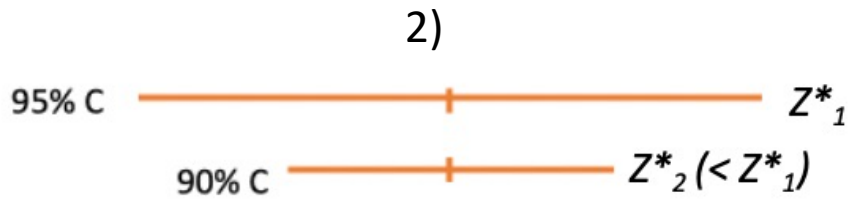




Summary of ideas from previous LCQ

- There are <u>two ways</u> to get a **more precise (narrower) confidence interval**!
  - (Assuming everything else remains the same)
  - <u>Increase the sample size</u>!
    - *This decreases the standard error and as a result the MOE.*
  - <u>Decrease the confidence level</u>!
    - *This decreases the critical value and as a result the MOE.*

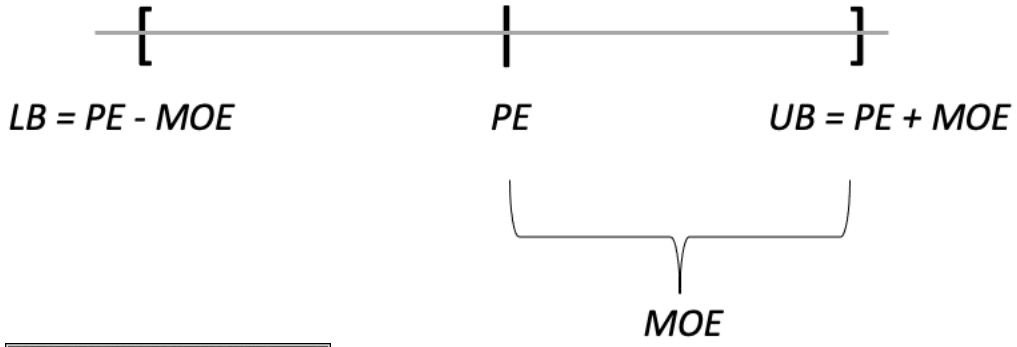




Margin of Error Revisited

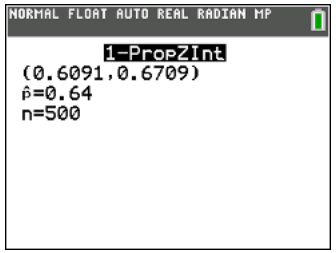
Recall: **MOE** is what you add and subtract from your point estimate to get your upper bound (UB) and lower bound (LB) of your confidence interval.

- If you are given an interval, your **margin of error** is the following:
  - Margin of Error = $\frac{UB - LB}{2} = \frac{Width}{2}$
- **Width** of an interval = 2 * MOE



Example:

Width = UB – LB = 0.6709 – 0.6091 = 0.0618
MOE = Width / 2 = 0.0618 / 2 = 0.0309



Finding the Minimum Sample Size - Motivation

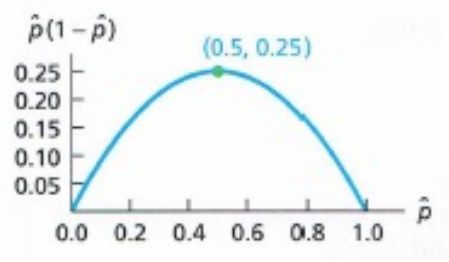
- In practice, and important step when planning a study is determining how large of a sample is needed.
    - If *n* is too large, it is a waste of resources (studies are expensive, time and $$$).
    - If *n* is too small, they are less confident in the results (i.e. too imprecise).
- Researchers try to figure out how large their sample needs to be to yield a confidence interval with a predetermined width.
    - In doing so, they are controlling the precision!

Finding the Minimum Sample Size - Calculation

- Start with the formula for Margin of Error and rearrange to solve for *n*:

$$MOE = Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \frac{\hat{p}(1-\hat{p})(Z^*)^2}{MOE^2}$$

- We know ahead of time what we want the MOE (or width) to be and confidence level (so Z*), the only other unknown is $\hat{p}$.
    - We have 2 options for this!
    1. Set $\hat{p}$ based on previous research or experience. This will have to be given to us.
    2. If no prior information is available, set $\hat{p} = 0.5$.
        - This is because it results in the largest *n* for a specific MOE. We want to be safe!



- Solve for *n*!
    - Since this formula is for the *minimum* sample size, if any decimal occurs, ALWAYS round up to the next largest whole number.

Example - Choosing Sample Size for Proportions

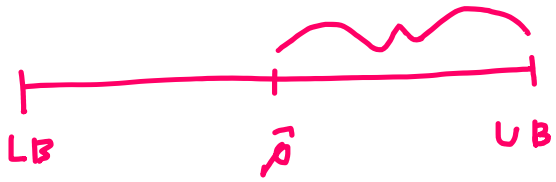
Setup: How large of a sample of apples do we need to estimate the proportion of defective apples within 2% of the true proportion of defective apples with 95% confidence?

1) Option 1: no guess for $\hat{p}$, use 0.5

2) Option 2: Based on previous experience, we estimate $\hat{p}$ = 0.08

Option 1 Solution

MOE = 0.02, trying to guarantee this!



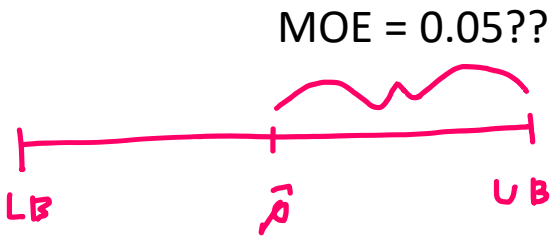
- Margin of error, MOE = 2%, or 0.02
- Since we want a 95% confidence interval, z^* = 1.96
- We don't have an estimate for $\hat{p}$, so use $\hat{p}$ = 0.5

$$n = \frac{\hat{p}(1-\hat{p})(Z^*)^2}{MOE^2} = \frac{0.5(1-0.5)1.96^2}{0.02^2} = 2401$$

- We need at least 2401 apples to estimate the proportion of defective apples within 2% of the true proportion of defective apples with 95% confidence

Option 2 Solution

With the wrong $\hat{p}$, the *n* we find could result in a MOE larger than wanted…. ☹



- Margin of error, MOE = 2%, or 0.02
- Since we want a 95% confidence interval, z^* = 1.96
- Our estimate for $\hat{p}$ = 0.08

$$n = \frac{\hat{p}(1-\hat{p})(Z^*)^2}{MOE^2} = \frac{0.08(1-0.08)1.96^2}{0.02^2} = 706.8 \ldots \rightarrow 707$$

- **Always round UP to the nearest whole number**
- We need at least 707 apples to estimate the proportion of defective apples within 2% of the true proportion of defective apples with 95% confidence

Problem Session!!!

Problem #1

An investment website can tell what devices are used to access the site. The site managers wonder whether they should enhance the facilities for trading via “smart phones” so they want to estimate the proportion of users who access the site that way (even if they also use their computers sometimes). They draw a random sample of 200 investors from their customers. Suppose that the true proportion of smart phone users was 36%.

a) What would you expect the shape of the sampling distribution for the sample proportion to be?

b) What would be the mean of this sampling distribution?

c) If the sample size were increased to 500, would your answers change? Explain.

Problem #1 Solution

a) np = 200(0.36) = 72 and nq = 200(1 – 0.36) = 128; Normal

b) $\mu_{\hat{p}} = 0.36$

c) $\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{200}} = 0.0339$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{500}} = 0.0215$$

No, only the standard deviation of the sampling distribution would change; it would decrease from 0.0339 to 0.0215.

Problem #3

The investment website of Exercise 1 draws a random sample of 200 investors from their customers. Suppose that the true proportion of smart phone users is 36%.

a) What would the standard deviation of the sampling distribution of the proportion of smart phone users be?
b) What is the probability that the sample proportion of smart phone users is greater than 0.36?
c) What is the probability that the sample proportion of smart phone users is between 0.30 and 0.40?
d) What is the probability that the sample proportion of smart phone users is less than 0.28?
e) What is the probability that the sample proportion of smart phone users is greater than 0.42?

Problem #3 Solution

a) $\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{200}} = 0.0339$

b) P($\hat{p}$ > 0.36) = 0.50

c) P(0.30 < $\hat{p}$ < 0.40) = 0.8426

d) P($\hat{p}$ < 0.28) = 0.0091

e) P($\hat{p}$ > 0.42) = 0.0384

Problem #5

A real estate agent wants to know how many owners of homes worth over $1,000,000 might be considering putting their home on the market in the next 12 months. He surveys 40 of them and finds that 10 of them are considering such a move. Are all the assumptions and conditions for finding the sampling distribution of the proportion satisfied? Explain.

Problem #5 Solution

Yes, assuming the survey is random, they should be independent. We don't know the true proportion, so we cannot check to see if $np \geq 5$ and $nq \geq 5$; but we have observed 10 successes and 30 failures, which is sufficient.

Problem #7

A market researcher for a provider of iPod accessories wants to know the proportion of customers who own cars to assess the market for a new iPod car charger. A survey of 500 customers indicates that 76% own cars.

a) What is the standard deviation of the sampling distribution of the proportion?

b) If she wants to reduce the standard deviation by half, how large a sample would she need?

Problem #7 Solution

a) $\sigma_{\hat{p}} = \sqrt{\frac{0.15(1-0.15)}{100}} = 0.0357$

b) 400

Problem #9

For each situation below identify the population and the sample and identify p and $\hat{p}$ if appropriate and what the value of $\hat{p}$ is. Would you trust a confidence interval for the true proportion based on these data? Explain briefly why or why not.

a) As concert goers enter a stadium, a security guard randomly inspects their backpacks for alcoholic beverages. Of 130 backpacks checked so far, 17 contained alcoholic beverages of some kind. The guards want to estimate the percentage of all backpacks of concertgoers at this concert that contain alcoholic beverages.

b) The website of the English newspaper *The Guardian* asked visitors to the site to say whether they approved of the recent "bossnapping" actions by British workers who were outraged over being fired. Of those who responded, 49.2% said "Yes. Desperate times, desperate measures."

c) An airline wants to know the weight of carry-on baggage that customers take on their international routes, so they take a random sample of 50 bags and find that the average weight is 17.3 pounds.

Problem #9 Solution

a) Population: backpacks of concertgoers, sample: the 130 backpacks that were searched, p = proportion of backpacks that contain alcoholic beverages, and $\hat{p}$ = sample proportion of backpacks that contain alcoholic beverages = 17/130 = 0.1308. Yes, the sample is random, independent, and we have at least 5 successes and failures.

b) Population: British people, sample: those who responded to *The Guardian's* survey, p = proportion of those who approve of bossnapping, and $\hat{p}$ = sample proportion of those who approve of bossnapping = 0.492. No, this sample is not random and is probably biased.

c) Population: carry-on baggage for international flights, sample = 50 bags, μ = mean weight of carry-on bags on international flights, $\bar{x}$ = 17.3. You could use this random sample to construct a confidence interval to estimate the mean weight as the sample size is large. Since this is about means and not proportions, the methods of this chapter are not appropriate.

Sampling Distribution Example: Sample Proportion

Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

Sampling Distribution Example: Sample Proportion

Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

Is it appropriate to use the CLT?

Sampling Distribution Example: Sample Proportion

Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

Is it appropriate to use the CLT?

- Randomization? Yes, ***randomly*** selected 80 players.
- Independent? Yes, one player doesn't affect another.
- 10%? Yes, our sample size of 80 is not more than 10% of the population (896)
- Sample Size? Yes because…
  - n*p = 80*0.2467 = 19.736 is at least 5 successes.
  - n*(1-p) = 80*0.7533 = 60.264 is at last 5 failures.

It is appropriate!

Sampling Distribution Example: Sample Proportion

Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

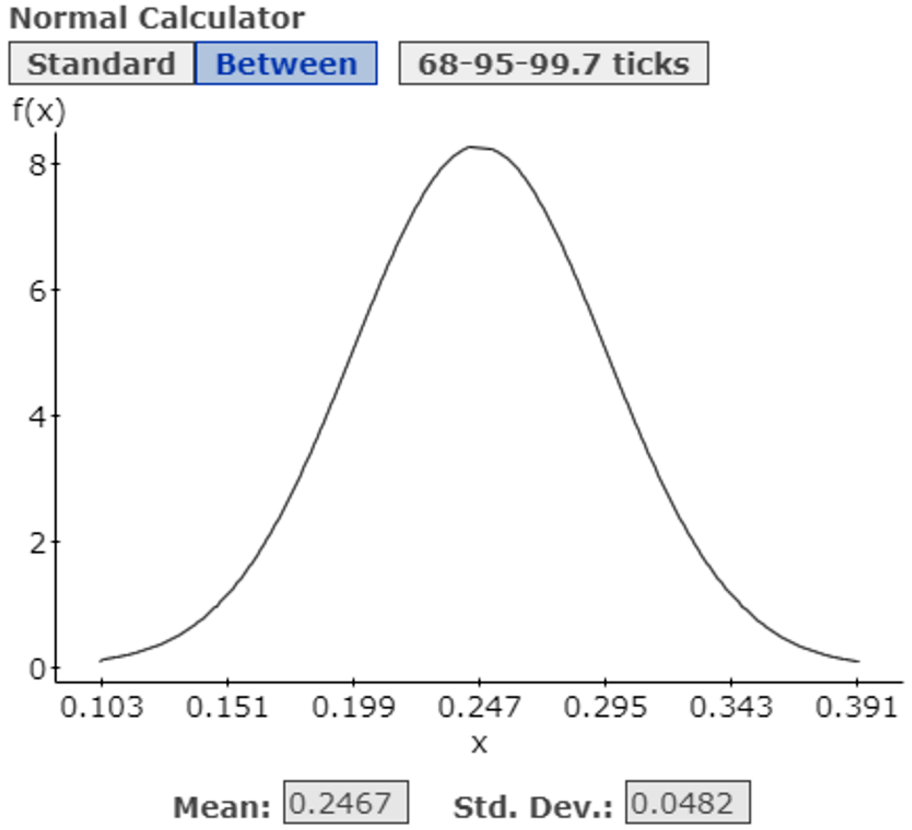
What is the name of the sampling distribution? Be sure to include the appropriate measures of its center and variability.

Sampling Distribution Example: Sample Proportion

What is the name of the sampling distribution? Be sure to include the appropriate measures of its center and variability.

Because the assumptions of the CLT hold up, we know the **sampling distribution of the sample proportions follows a Normal Distributio with a mean of 0.2467 and a standard deviation of 0.0482.**

$$\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$



Sampling Distribution Example: Sample Proportion

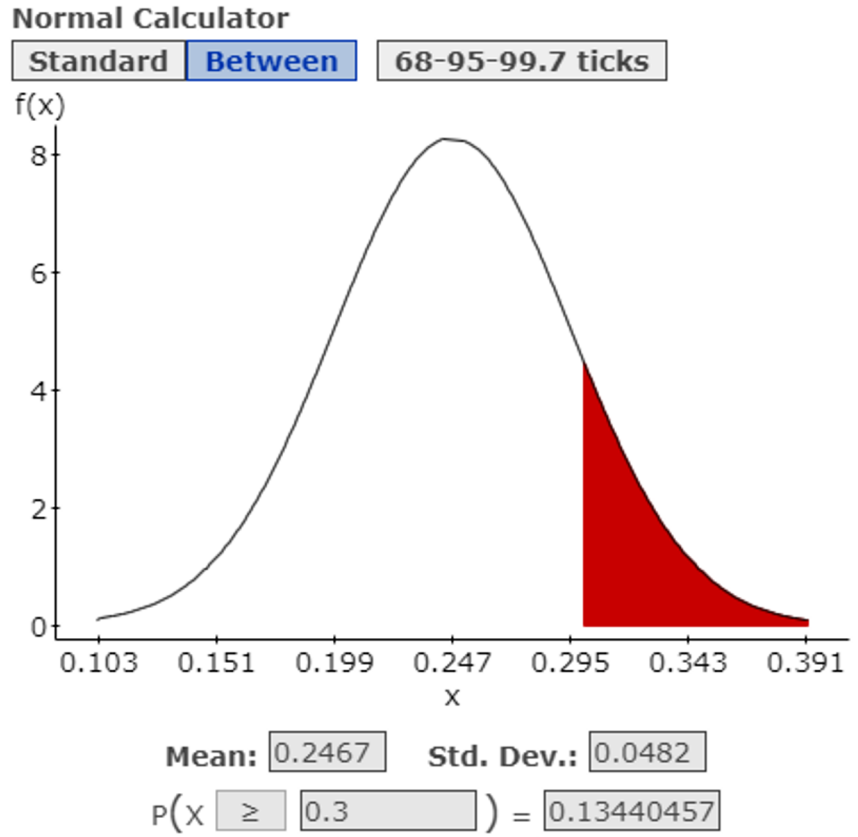
Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

What is the probability we get a sample proportion of American players of greater than 30%?

Sampling Distribution Example: Sample Proportion

Using our NHL Data as our example, we know for the population that 24.67% of NHL players are Americans. We take a random sample of size 80 from the population.

What is the probability we get a sample proportion of American players of greater than 30%? **0.1344**



Tossing Coins

- Let's say you toss a coin 40 times and it lands on heads 16 times.
- What is the point estimate?
- Find the mean, $\mu_{\hat{p}}$, and standard error, $\sigma_{\hat{p}}$, based on our point estimate.
- Find and interpret a 90% confidence interval
- Find and interpret a 95% confidence interval
- Find and interpret a 99% confidence interval

Tossing Coins Solution

- Point estimate is $\hat{p} = \frac{\#\ successes}{Total} = \frac{16}{40} = 0.40$
- Find the mean, $\mu_{\hat{p}}$, and standard error, $\sigma_{\hat{p}}$, based on our point estimate
- $\mu_{\hat{p}} = 0.40$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.40(0.60)}{40}} = 0.077$

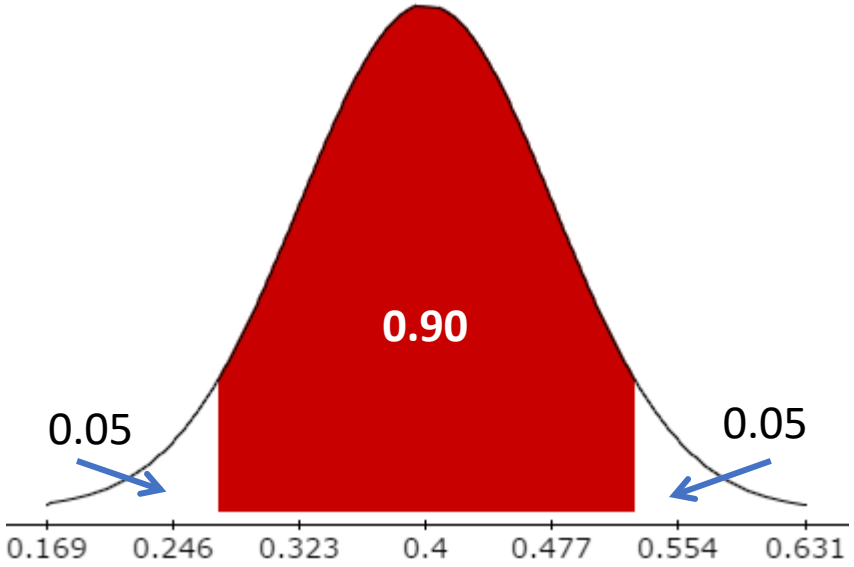
Check the Assumptions

- We can assume that the 40 tosses of the coin are a random sample of all coin tosses
- $n\hat{p} = 40(0.40) = 16 \geq 10$
- $n(\hat{q}) = 40(0.60) = 24 \geq 10$

Tossing Coins Solution, p. 2

Using calc:

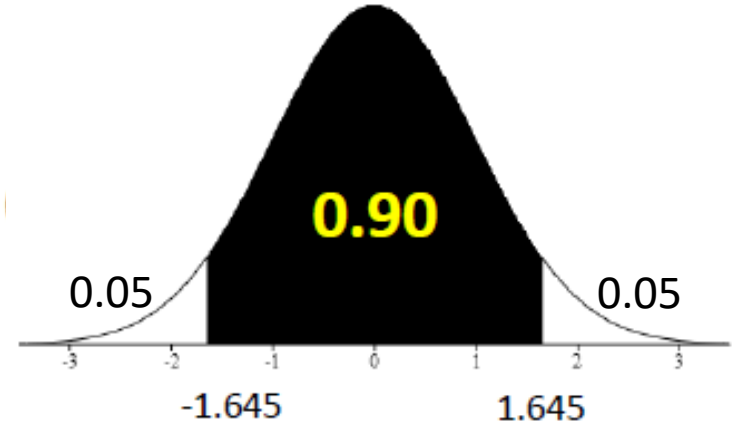
- 90% CI = (0.273, 0.527)



Tossing Coins Solution, p. 3

Long way solution:

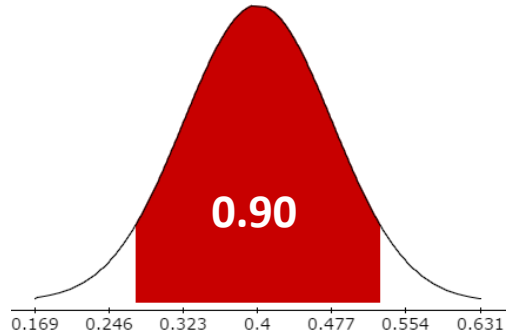
- $\mu_{\hat{p}} = 0.40$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.40(0.60)}{40}} = 0.077$
- Find z*: $\mu = 0$ and $\sigma = 1$
- P(z ≤ -1.645) = 0.05
- P(z ≥ 1.645) = 0.05

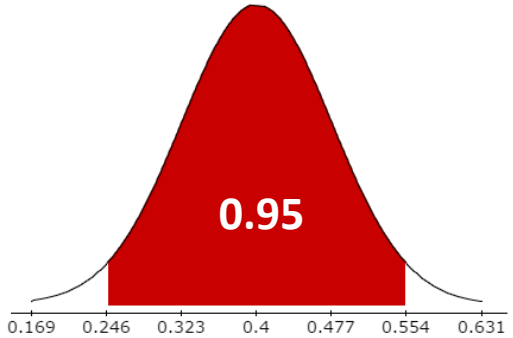


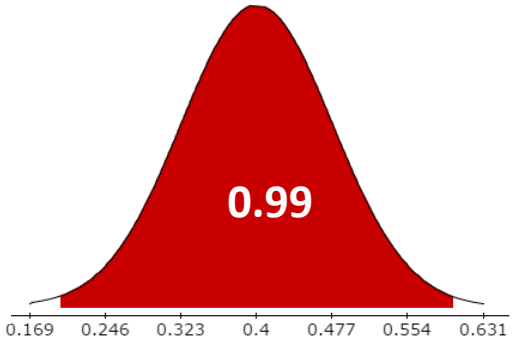
- $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(\hat{q})}{n}} = 0.40 \pm 1.645(0.077)$
- 90% CI = (0.273, 0.527)

Comparing Widths of CIs

- 90% CI = (0.273, 0.527)
  - The width of the 90% CI is .254
- 95% CI = (0.248, 0.552)
  - The width of the 95% CI is .304
- 99% CI = (0.200, 0.600)
  - The width of the 99% CI is .400
- If the sample size remains the same, as the confidence level increases, so does the width of the CI







Problem #11

A survey of 200 students is selected randomly on a large university campus. They are asked if they use a laptop in class to take notes. The result of the survey is that 70 of the 200 students responded “yes.”

a) What is the value of the sample proportion of $\hat{p}$. What is the standard error of the sample proportion?

b) Construct an approximate 95% confidence interval for the true proportion of p by taking ±2 SEs from the sample proportion.

Problem #11 Solution

a) $\hat{p}$ = 70/200 = 0.35

b) $\sigma_{\hat{p}} = \sqrt{\frac{0.35(1-0.35)}{200}} = 0.0337$

c) 0.35 ± 2(0.0337) = (0.2826, 0.4174)

I am 95% confident that the true proportion of university students who use a laptop to take notes is between 0.283 and 0.417.

Problem #15

From the survey in Exercise 11,

a) How would the confidence interval change if the confidence level had been 90% instead of 95%?

b) How would the confidence interval change if the sample size had been 300 instead of 200? (Assume the same sample proportion.)

c) How would the confidence interval change if the confidence had been 99% instead of 95%?

d) How large would the sample size have to be to make the margin of error half as big in the 95% confidence interval?

Problem #15 Solution

a) The confidence interval would be more narrow.
b) The confidence interval would be more narrow as the standard deviation would be equal to 0.0275 instead of 0.0337 → (.295, .405)
c) The confidence interval would be wider
d) The sample size would need to be 4 times as large; 800

Problem #17

Suppose you want to estimate the proportion of the traditional college students on your campus who own their own car. You have no preconceived idea of what your proportion might be.

a) What sample size is needed if you wish to be 95% confident that your estimate is within 0.02 of the true proportion?

b) What sample size is needed if you wish to be 99% confident that you estimate is within 0.02 of the true proportion?

c) What sample size is need if you wish to be 95% confident that you estimate is within 0.05 of the true proportion?

Problem #17 Solution

- Solve the margin of error formula for n, $\mathrm{MO}E = z\sqrt{\frac{pq}{n}} \rightarrow n = \frac{pqz^2}{MOE^2}$

$a)\ n = \frac{(0.5)(0.5)(1.96)^2}{(0.02)^2}$ = 2401

$b)\ n = \frac{(0.5)(0.5)(2.576)^2}{(0.02)^2}$ = 4147.36 → 4148

$c)\ n = \frac{(0.5)(0.5)(1.96)^2}{(0.05)^2}$ = 384.16 → 385

Problem #45

A catalog sales company promises to deliver orders placed on the Internet within 3 days. Follow-up calls to a few randomly selected customers show that a 95% confidence interval for the proportion of all orders that arrive on time is 88% +/- 6%. What does this mean? Are the conclusions in parts a-e correct? Explain.

a) Between 82% and 94% of all orders arrive on time.

b) 95% of all random samples of customers will show that 88% of orders arrive on time.

c) 95% of all random samples of customers will show that 82% to 94% of all orders arrive on time.

d) The company is 95% sure that between 82% and 94% of the orders placed by the customers in this sample arrived on time.

e) On 95% of the days, between 82% and 94% of the orders will arrive on time.

Problem #45 Solution

88% ± 6% → (82%, 94%)

a) No, we are 95% confident that the true proportion of orders that arrive on time is between 82% and 94%.

b) No, it is not about the % of random samples that arrive on time, but how confident we are that the true proportion falls between 0.82 and 0.94.

c) No, it is not about the percentage of random samples that fall between the two values, but about the unknown parameter.

d) No, it is not a probability and does not describe the sample, but is used to estimate the population proportion. In this sample we know 88% arrived on time.

e) No, it is not a probability, it is about the parameter, not about the days.

Problem #47

Several factors are involved in the creation of a confidence interval. Among them are the sample size, the level of confidence, and the margin of error. Which statements are true?

a) For a given sample size, higher confidence means a smaller margin of error.

b) For a specified confidence level, larger samples provide smaller margins of error.

c) For a fixed margin of error, larger samples provide greater confidence.

d) For a given confidence level, halving the margin of error requires a sample twice as large.

Problem #47 Solution

a) False, higher confidence means larger margin of error
b) True, as sample size increases, standard error and margin of error decrease
c) True
d) False, halving the margin of error requires a sample size 4 times as large

Problem #49

A student is considering publishing a new magazine aimed directly at owners of Japanese automobiles. He wanted to estimate the fraction of cars in the United States that are made in Japan. The computer output summarizes the results of a random sample of 50 autos. Explain carefully what it tells you.

z-interval for proportion

With 90.00% confidence

0.29938661 < p(japan) < 0.46984416

Problem #49 Solution

We are 90% confident that the true proportion of cars in the US that are made in Japan is between 0.299 and 0.470.

Problem #53

An insurance company checks police records on 582 accidents selected at random and notes that teenagers were at the wheel in 91 of them.

a) Create a 95% confidence interval for the percentage of all auto accidents that involve teenage drivers.

b) Explain what your interval means.

c) Explain what “95% confidence” means.

d) A politician urging tighter restrictions on drivers’ licenses issued to teens says, “In one of every five auto accidents, a teenager is behind the wheel.” Does your confidence interval support or contradict this statement? Explain.

Problem #53 Solution

a) $0.156 \pm 1.96(0.015) = (0.1266, 0.1854)$

b) I am 95% confident that the true percentage of all auto accidents that involve teenage drivers is between 12.7% and 18.5%.

c) If I were to construct confidence intervals for all possible samples of size 582, 95% of them would capture the true population percentage of teenage drivers involved in auto accidents.

d) No, the confidence interval does not support this statement, 20% is above the confidence interval.

Problem #75

A state's environmental agency worries that a large percentage of cars may be violating clean air emissions standards. The agency hopes to check a sample of vehicles in order to estimate that percentage with a margin of error of 3% and 90% confidence. To gauge the size of the problem, the agency first picks 60 cars and finds 9 with faulty emissions systems. How many should be sampled for a full investigation?

Problem #75 Solution

- Solve the margin of error formula for n, $MOE = z\sqrt{\frac{pq}{n}} \rightarrow n = \frac{pqz^2}{MOE^2}$
- $\hat{p} = \frac{9}{60} = 0.15$
- $n = \frac{(0.15)(0.85)(1.645)^2}{(0.03)^2}$ = 383.35 → 384

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

A total of 567 observations in the population.

Am I ultimately interested in estimating the statistic or parameter?

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

A total of 567 observations in the population.

Am I ultimately interested in estimating the statistic or parameter?

A **parameter** since I am interested in the **population proportion** of freshman in STAT 1450.

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

I want ***an*** estimate of my population proportion. What should I report based on my sample?

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

I want an estimate my population proportion. What should I report based on my sample?

The **sample proportion** of my sample of 50 students 18/50!

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

Is the Population proportion equal our sample proportion?

So how do we take into account sampling variability?

Example: STA 261 Proportion of Freshman

- Assumptions
  - Random Sample?
    - Yes, I randomly selected 50 students.
  - Enough successes and failures?
    - 18 success and 32 failures in our sample
    - The one time we look at sample values, why? We don't know population proportion P.

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

Is the Population proportion equal our sample proportion? No, the statistic is not likely to equal the population proportion. Each sample will give a different point estimate (sample proportion) due to sampling variability.

So how do we take into account sampling variability? Confidence Interval!

Example: STA 261 Proportion of Freshman

I am interested in knowing what the true proportion of STAT 1450 students that are freshman.

We can do CI by hand, or use our calc. Let's use our calc.

We use 1-PropZint.

Now let's generate a 95% Confidence Interval for the Population Proportion of STAT 1450 students that are freshman.

???? (you can do it ☺)

95% C.I. How do I interpret?

???? (you can do it ☺)

Sample Size Example

What sample size is needed if we wanted a 90% confidence interval with a width no larger than 10% (0.1) if we have no prior information? (0.5)

???? (you can do it ☺)

What if we had previous information that the proportion was 30% (0.3)?

???? (you can do it ☺)

Proportion Confidence Interval

The Miami Student wants to conduct a survey to see how many Miami University students can drive a manual car. A random sample of 27 students found that 5 could drive a manual car. Find a 99% confidence interval for the population proportion of Miami students that can drive a manual transmission car.

Proportion Confidence Interval

The Miami Student wants to conduct a survey to see how many Miami University students can drive a manual car. A random sample of 27 students found that 5 could drive a manual car. Find a 99% confidence interval for the population proportion of Miami students that can drive a manual transmission car.

What parameter are we trying to estimate?

Proportion Confidence Interval

The Miami Student wants to conduct a survey to see how many Miami University students can drive a manual car. A random sample of 27 students found that 5 could drive a manual car. Find a 99% confidence interval for the population proportion of Miami students that can drive a manual transmission car.

What parameter are we trying to estimate?

The **population proportion** of Miami students that can drive a manual car.

Proportion Confidence Interval

The Miami Student wants to conduct a survey to see how many Miami University students can drive a manual car. A random sample of 31 students found that 7 could drive a manual car. Find a 99% confidence interval for the population proportion of Miami students that can drive a manual transmission car.

Do our assumptions hold up?

Proportion Confidence Interval

The Miami Student wants to conduct a survey to see how many Miami University students can drive a manual car. A random sample of 31 students found that 7 could drive a manual car. Find a 99% confidence interval for the population proportion of Miami students that can drive a manual transmission car.

Do our assumptions hold up? Yes!

Our assumption of 5 successes and 5 failures is met! And we have a random sample!

We can continue to making interval!

1-Proportion Z-interval

One sample proportion summary confidence interval:
p : Proportion of successes
Method: Standard-Wald

99% confidence interval results:

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	7	31	0.22580645	0.075095186	0.03237407	0.41923883

Interpretation:

We are 99% confident that the **population proportion** of Miami students that can drive a manual car is between 3.24% and 41.92%.