Stat 201: Introduction to Statistics

Standard 25: Confidence Intervals – for Proportions

From *Naked Statistics:* Inference

- “Statistics cannot prove anything with certainty. Instead, the power of statistical inference derives from observing some pattern or outcome and then using probability to determine the most likely explanation for that outcome.”

From *Naked Statistics:* Inference

- "Of course, the most likely explanation is not always the right explanation. Extremely rare things happen. Linda Cooper is a South Carolina woman who has been struck by lightning four times. (The Federal Emergency Management Administration estimates the probability of getting hit by lightning just once as 1 in 600,000.) Linda Cooper's insurance company cannot deny her coverage simply because her injuries statistically improbable."

From *Naked Statistics:* Inference

- “Statistical inference is the process by which the data speak to us, enabling us to draw meaningful conclusions. This is the payoff!”

- **“The point of statistics is not to do myriad rigorous mathematical calculations; the point is to gain insight into meaningful social phenomena.”**

From *Naked Statistics:* Inference

- "You may recall that we can create a standard error for each of our samples - $\sqrt{\frac{\rho(1-\rho)}{n}}, \frac{\sigma}{\sqrt{n}}, \frac{\mathrm{s}}{\sqrt{n}} \ldots$ you will recall that the central limit theorem and empirical rule tell us that for 95 samples out of 100, the sample proportion or sample mean is going to lie within two standard errors of the true population mean, in either direction or the other."

From *Naked Statistics:* Inference

- “If exactly half of American adults disapprove of gay marriage, then our best guess about the attitudes of a representative sample of 1,000 Americans is that about half of them will disapprove of gay marriage.”
- “Conversely – and more important from the standpoint of polling – if we have a representative sample of 1,000 Americans who feel a certain way, such as the 46 percent who disapprove of President Obama’s job performance, then we can **infer** from that sample that the general population is likely to feel the same way.”

From *Naked Statistics:* Inference

- “ When you read that a poll has a ‘margin of error’ of +/- 3 percent, this is really just the same kind of confidence interval we’ll learn here. Our ‘95 percent confidence’ means that if we conducted 100 different polls on samples drawn from the same population, we would expect the answers we get from our sample in 95 of those polls to be within 3 percentage points in one direction or the other of the population’s true sentiment.”

From *Naked Statistics:* Inference

- "Ever the statistics guru, you point out that you cannot be certain of any result until all of the votes are counted. However, you can offer a 95 percent confidence interval instead. In this case, CNN's spinning, 3-D, multicolored graphic will be wrong, on average, only 5 times out of 100."

From *Naked Statistics:* Inference

- “When done properly, polls are uncanny instruments. According to Frank Newport, editor in chief of the Gallup Organization, a poll of 1,000 people can offer meaningful and accurate insights into the attitudes of the entire country. Statistically speaking, he’s right. But to get those meaningful and accurate results, we have to conduct a proper poll and then interpret the results correctly, both of which are much easier said than done.”

From *Naked Statistics:* Inference

- “Bad polling results do not typically stem from bad math when calculating the standard errors. Bad polling results typically stem from a biased sample, or bad questions, or both.”

From *Naked Statistics:* Inference

- **<u>Questions to ask:</u>**
- **Is this an accurate sample of the population whose opinions we are trying to measure?**
- **Have the questions been posed in a way that elicits accurate information on the topic of interest?**
- **Are respondents telling the truth?**

From *Naked Statistics:* Inference

- **Example: 3,342 adults were chosen for an American sex study**
  - 90% of couples shared the same race, religion, social class and age
  - Typical respondent was engaging in sexual activity “a few times a month” though there was wide variation. The number of sexual partners age 18 ranged from zero to over 1,000.
  - 80% had either one sexual partner in the previous year or none at all.
  - Respondents with one sexual partner were happy than those with none or with multiple partners.
  - A quarter of the married men and 10% of married women reported having extramarital sexual activity
  - Roughly 5% of men and 4 percent of women reported some sexual activity with a same gender partner

From *Naked Statistics:* Inference

- This assumes that the respondents to the survey both mirrored the population from which they were drawn (US adults) and gave accurate answers.

- Do we see any strange statistics in this sample summary?

From *Naked Statistics:* Inference

- The most suspicious thing about polling is that the opinions of so few can tell us about the opinions of so many. A proper sample will look like the population from which it is drawn. The real challenge of polling is twofold: finding and reaching that proper sample and eliciting information from that representative group in a way that accurately reflects what its members believe.

Confidence Intervals

- Often, we do not know the **population parameter, $\boldsymbol{\mu}$ or $\boldsymbol{\rho}$**

- We use our **sample statistics, $\boldsymbol{\overline{x}}$ or $\boldsymbol{\widehat{p}}$** to make **inference** on the **population parameter, $\boldsymbol{\mu}$ or $\boldsymbol{\rho}$**

Confidence Intervals

- **First,** we will consider an interval estimate which we call a confidence interval

$$\boldsymbol{point\ estimate \pm margin\ of\ error}$$

$$\boldsymbol{= point\ estimate \pm \binom{confidence}{coefficient} * \binom{St\widehat{anda}rd}{Error}}$$

Note: The margin of error is our plus/minus from Chap 1

Telling Which Parameter We’re After

- As statisticians, or data scientists, it’s our job to hear a problem and decide what we’re after
  - We call the parameter of interest the **target parameter**

Parameter	Point Estimate	Key Phrase	Type of Data
μ	$\bar{x}$	Mean, Average	Quantitative
p	$\hat{p}$	Proportion, percentage, fraction, rate	Qualitative (Categorical)

Confidence Intervals for Population Proportions on YouTube

- Intro:
  - https://www.youtube.com/watch?v=3ReWri_jh3M

Recall Proportion Sampling Distributions

- Recall: the mean of the sampling distribution for a sample proportion will always equal the population proportion: $\boldsymbol{\mu_{\hat{p}} = \rho}$
- The standard error, the standard deviation of the sample proportion, is:

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

Confidence Intervals

- Often, we do not know the **population proportion, p.**

- We use our **sample proportions, $\hat{p}$, ** to make **inference** on the **population parameter, p.**

Confidence Intervals: Step One

- **Assumptions:**
  1. Data must be obtained through randomization
  2. We **MUST** make sure that $n\hat{p} \geq 15$ and $n(1 - \hat{p}) \geq 15$. This ensures that $\hat{p}$ follows a bell shaped distribution
     - Recall Chapter 4 and the shape of the binomial dist.

Confidence Intervals: Step Two

- Recall: $\hat{p}$ is our **point-estimate** for the population proportion

- Recall we consider $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ when we don’t know ρ for the standard error as $\hat{p}$ can estimate the value of ρ

Confidence Intervals: Step Two

- $\hat{p}$ is our **point-estimate** for the population proportion
  - Our ‘best’ guess for the **true population proportion, $\boldsymbol{\rho}$,** is our **sample proportion, $\hat{p}$.**

Confidence Intervals: Step Two

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$ is our **margin of error**
  - $z_{1-\frac{\alpha}{2}}$ is the **confidence coefficient** and is the z value such that $P\left(Z < z_{\left(1-\frac{\alpha}{2}\right)}\right) = 1 - \frac{\alpha}{2}$
  - $\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$ is the **estimated standard deviation**

Confidence Intervals – Common Z's

- We choose z from this table based off our desired confidence level
  - **Level of confidence** = (1-∝) * 100%
  - **Error Probability** = ∝ = 1- Level of confidence

Confidence	Error Probability (∝)	$z_{\left(1-\frac{\alpha}{2}\right)}$
.9	.1	1.645
.95	.05	1.96
.99	.01	2.58

  - Our interval will get larger when the margin of error increases
    1) When we increase confidence → increase z
    2) When we decrease n

Confidence Intervals: Step Two

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$ is our **margin of error**
  - **As n increases**, the margin of error decreases causing the width of the confidence interval to narrow
  - **As n decreases**, the margin of error increases causing the width of the confidence interval to grow wider

Confidence Intervals: Margin of Error

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$ is our **margin of error**
  - **As the confidence level decreases**, z decreases causing the margin of error to decrease, causing the width of the confidence interval to narrow
  - **As the confidence level increases**, z increases causing the margin of error to increase, causing the width of the confidence interval to grow wider

Confidence Intervals – Step Two

- A fishing metaphor:
  - **As n increases** → confidence interval narrows
  - **As n decreases**→ confidence interval widens

  - Think about fishing in a pond with a net. If there are more fish you can use a smaller net to catch the fish.
  - In our case, when our sample size is larger we can use a smaller interval to catch our parameter.

Confidence Intervals – Step Two

- A fishing metaphor:
  - **Increase confidence** → confidence interval narrows
  - **Decrease confidence** → confidence interval widens

  - Think about fishing in a pond with a net. We want to be more certain that we’ll catch a fish we need a bigger net.
  - In our case, when we increase confidence to be more certain that we’ll catch the parameter, we need a bigger interval.

Confidence Intervals – Step Three

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

Lower Bound = $\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$

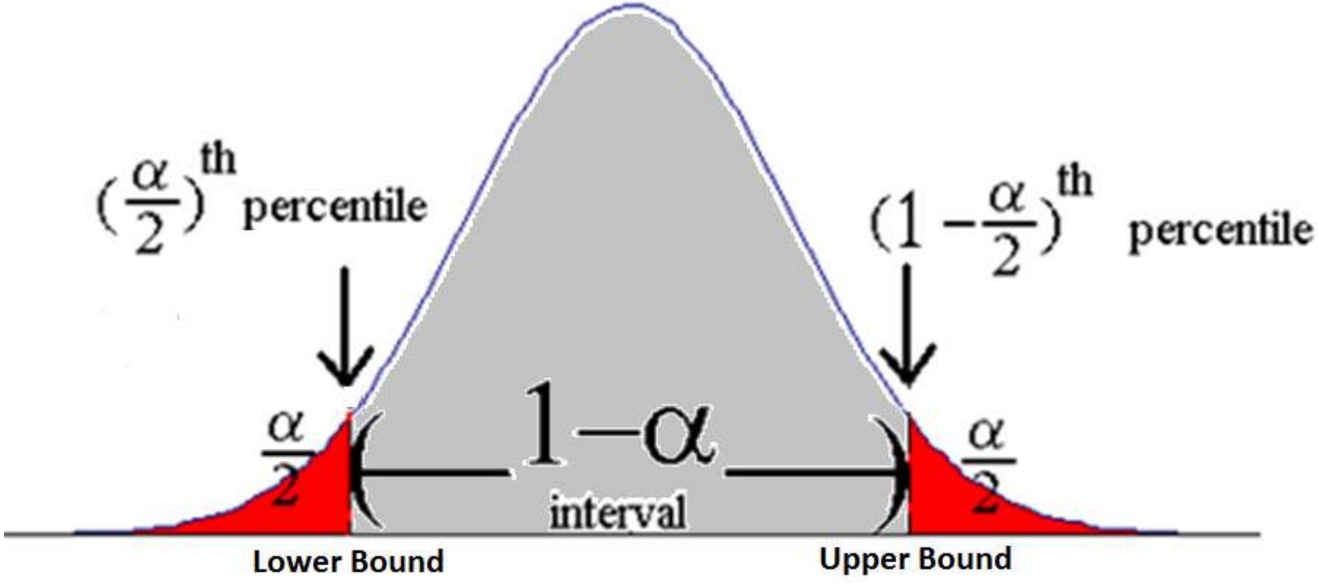
Upper Bound = $\hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\left(\hat{p}(1-\hat{p})\right)}{n}}$

Confidence Intervals – Step Three

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

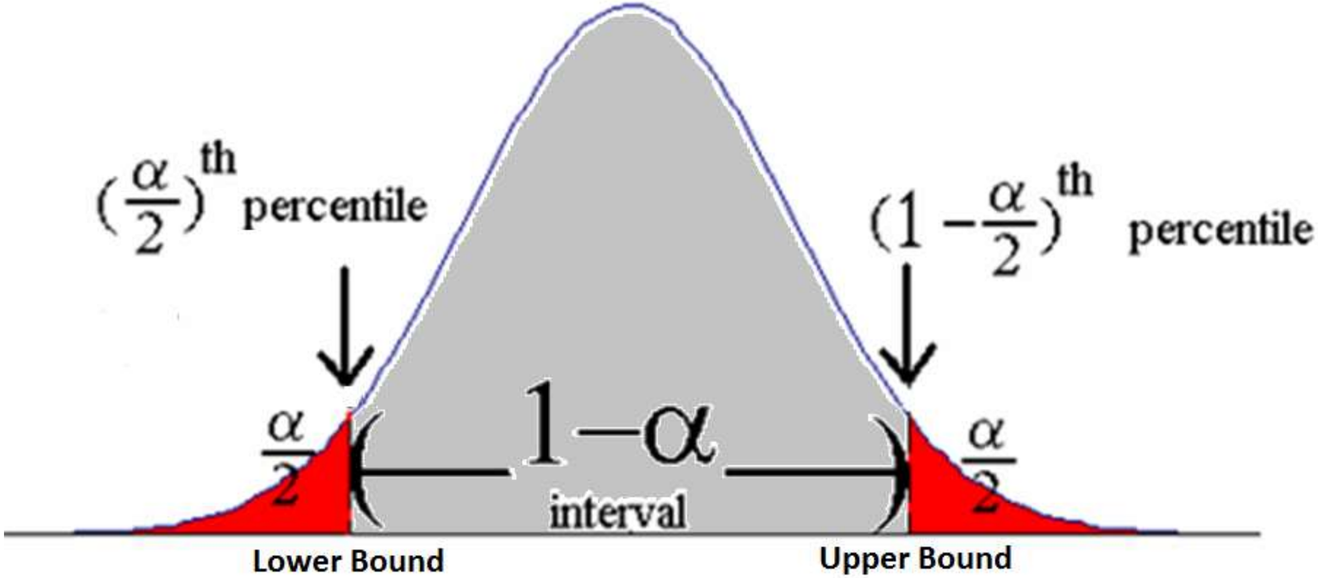
“We are <u>--%</u> confident that the true population proportion, *ρ*, is between the <u>lower bound</u> and <u>upper bound</u>.”

Confidence Intervals



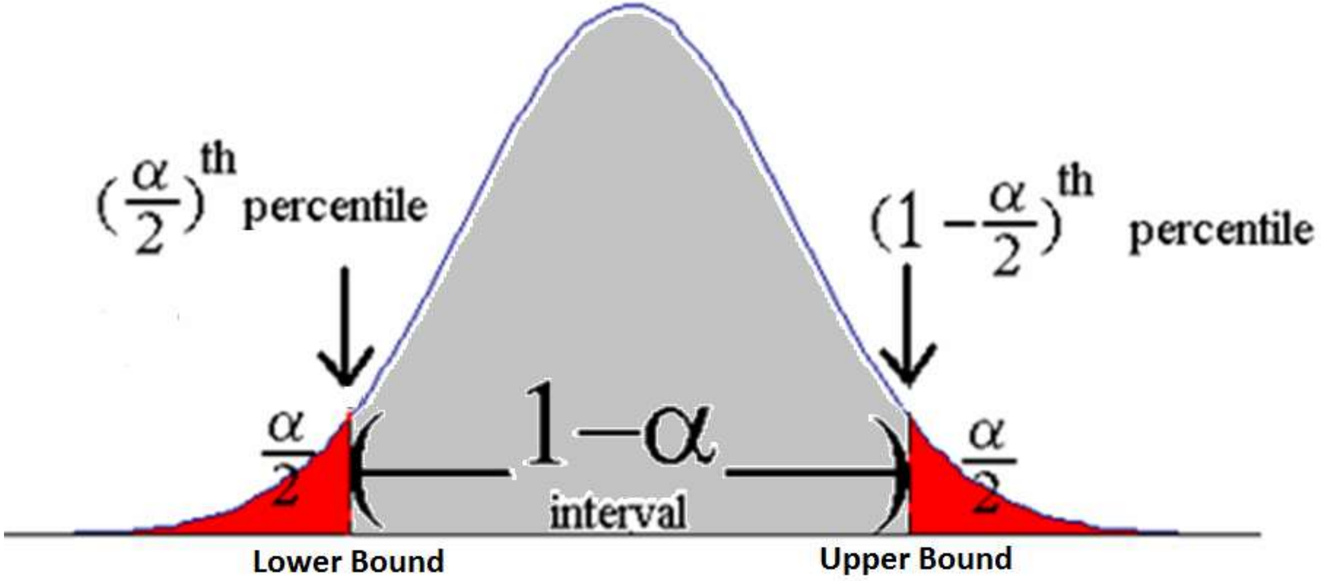
- We choose our values such that
  - Our **point estimate** is the mean, the 50th percentile
  - Our **lower bound** is the $\frac{\alpha}{2}^{\text{th}}$ percentile
  - Our **upper bound** is the 1-$\frac{\alpha}{2}^{\text{th}}$ percentile

How We Found the Common Z's: 90%



- For a 90% confidence interval lower bound, we need to find the z with a percentile of
$$\frac{\alpha}{2} = \frac{1 - confidence}{2} = \frac{1 - .90}{2} = \frac{.10}{2} = .0500$$
- If we look this up in the z-table we see that a z-score of -1.65 or -1.64 gives us a value very close to .0500

How We Found the Common Z's: 90%



- For a 90% confidence interval upper bound, we need to find the z with a percentile of

$$1-\frac{\alpha}{2}=1-\frac{1-confidence}{2}=1-\frac{1-.90}{2}=1-\frac{.10}{2}=.9500$$

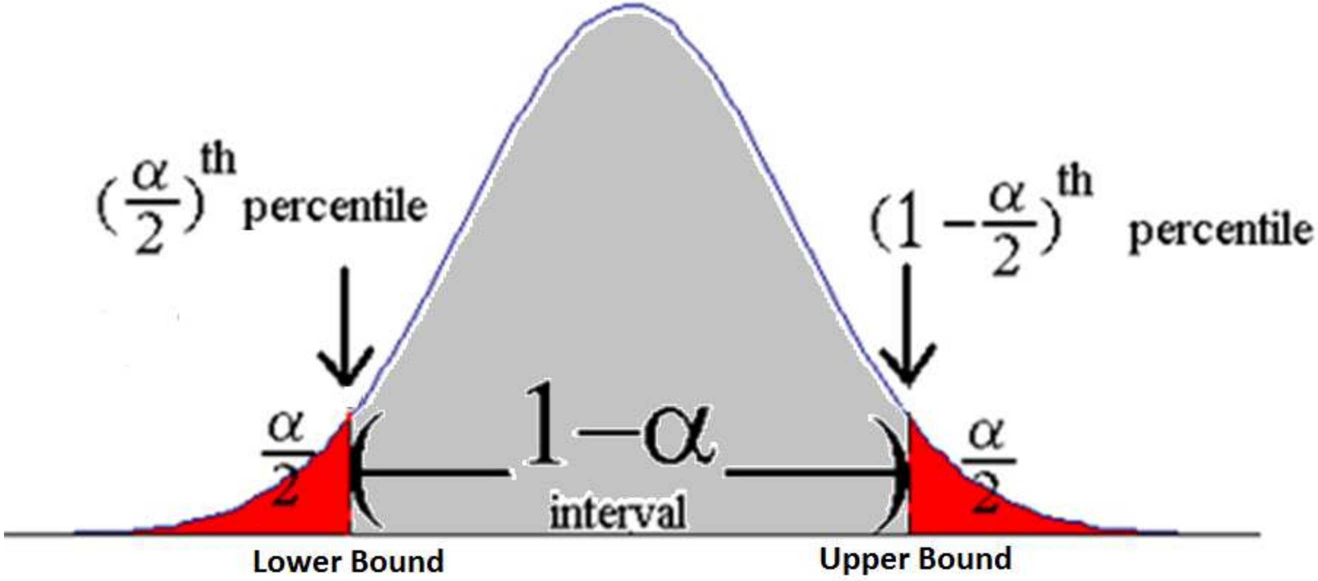
- If we look this up in the z-table we see that a z-score of 1.64 or 1.65 gives us a value very close to .9500

How We Found the Common Z's: 90%

- **Lower Bound:** If we look this up in the z-table we see that a z-score of -1.65 or -1.64 gives us a value very close to .0500
- **Upper Bound:** If we look this up in the z-table we see that a z-score of 1.65 or 1.64 gives us a value very close to .9500
- Since it's in the middle we average 1.64 and 1.65

- This is why we have plus or minus z=1.645 for a 90% confidence interval

How We Found the Common Z's: 95%

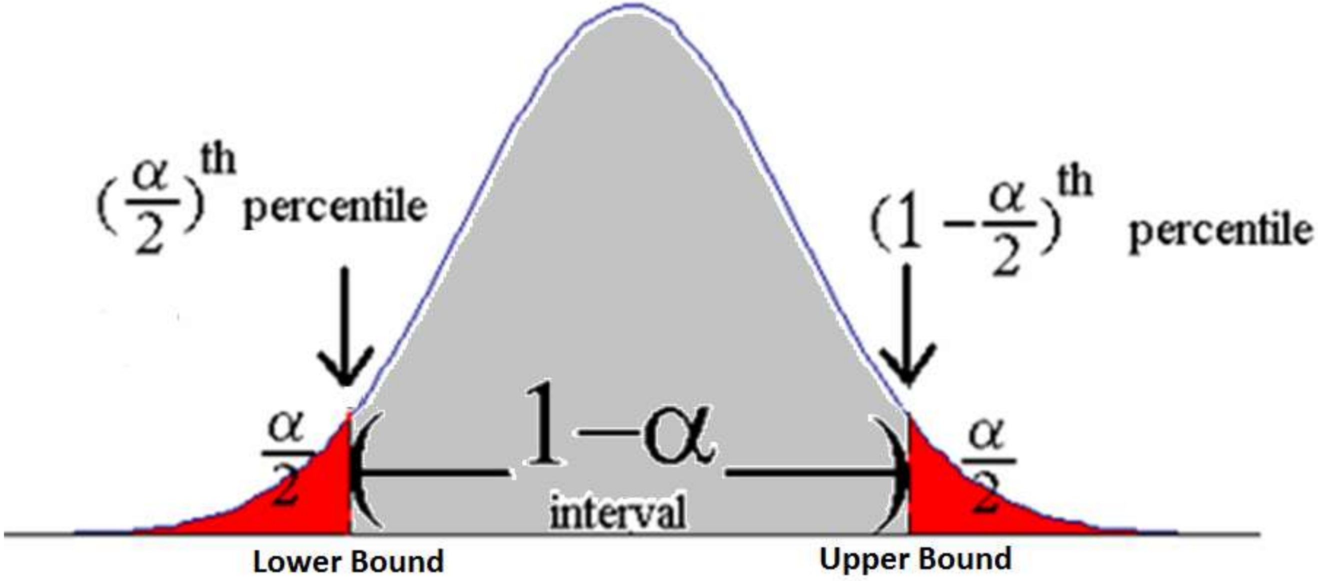


- For a 95% confidence interval lower bound, we need to find the z with a percentile of

$$\frac{\alpha}{2} = \frac{1 - confidence}{2} = \frac{1 - .95}{2} = .00250$$

- If we look this up in the z-table we see that a z-score of -1.96 gives us a value very close to .0250

How We Found the Common Z’s: 95%



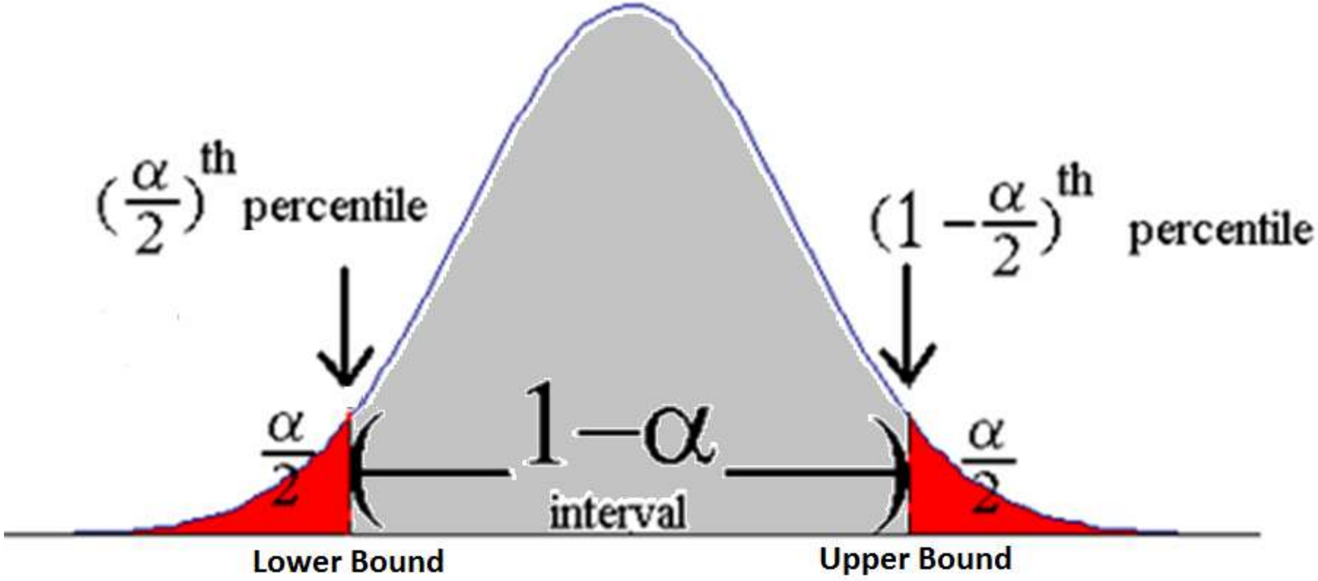
- For a 95% confidence interval upper bound, we need to find the z with a percentile of $1-\frac{\alpha}{2}=1-\frac{1-confidence}{2}=1-\frac{1-.95}{2}=1-\frac{.05}{2}=.9750$
- If we look this up in the z-table we see that a z-score of 1.96 gives us a value very close to .9750

How We Found the Common Z’s: 95%

- **Lower Bound:** If we look this up in the z-table we see that a z-score of -1.96 gives us a value very close to .0250
- **Upper Bound:** If we look this up in the z-table we see that a z-score of 1.96 gives us a value very close to .9750

- This is why we have plus or minus z=1.96 for a 95% confidence interval

How We Found the Common Z’s: 99%

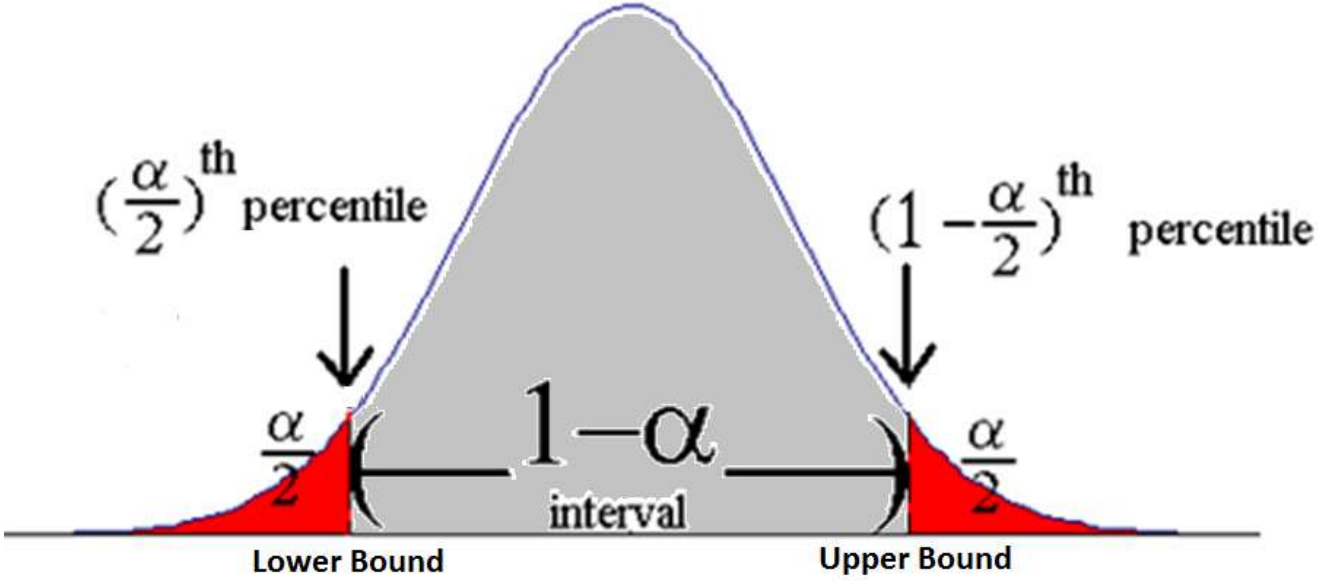


- For a 99% confidence interval lower bound, we need to find the z with a percentile of

$$\frac{\alpha}{2} = \frac{1 - confidence}{2} = \frac{1 - .99}{2} = .0050$$

- If we look this up in the z-table we see that a z-score of -2.58 gives us a value very close to .0050

How We Found the Common Z’s: 99%



- For a 99% confidence interval upper bound, we need to find the z with a percentile of

$$1-\frac{\alpha}{2}=1-\frac{1-confidence}{2}=1-\frac{1-.99}{2}=1-\frac{.01}{2}=.9950$$

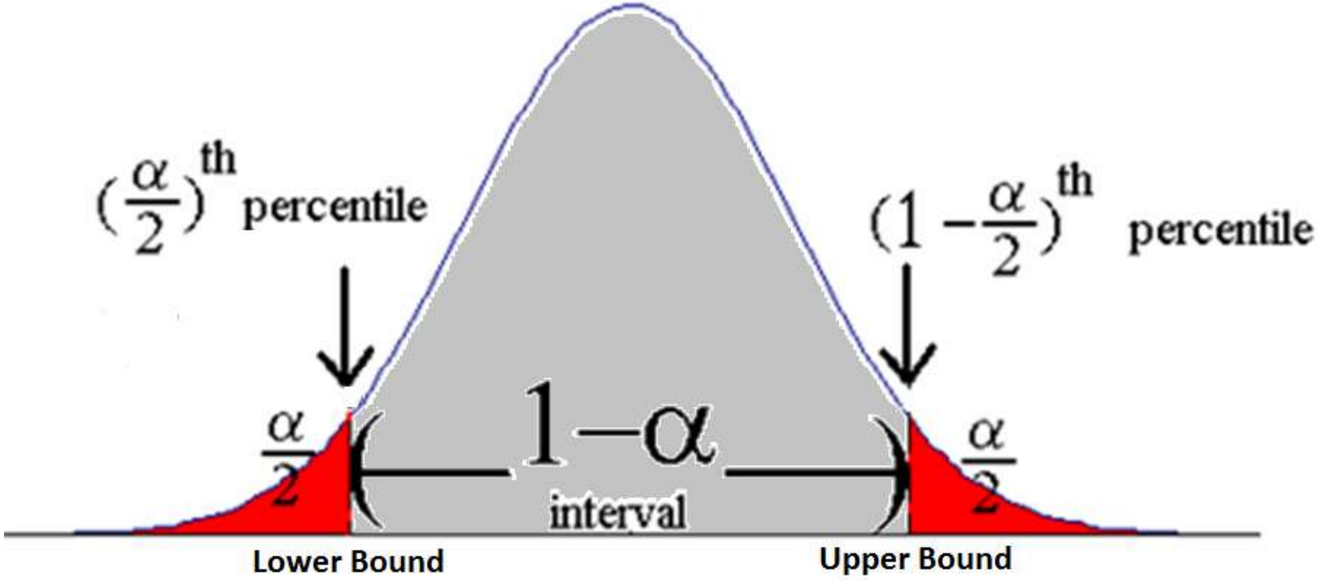
- If we look this up in the z-table we see that a z-score of 2.58 gives us a value very close to .9950

How We Found the Common Z’s: 99%

- **Lower Bound:** If we look this up in the z-table we see that a z-score of -2.58 gives us a value very close to .0050
- **Upper Bound:** If we look this up in the z-table we see that a z-score of 2.58 gives us a value very close to .9950

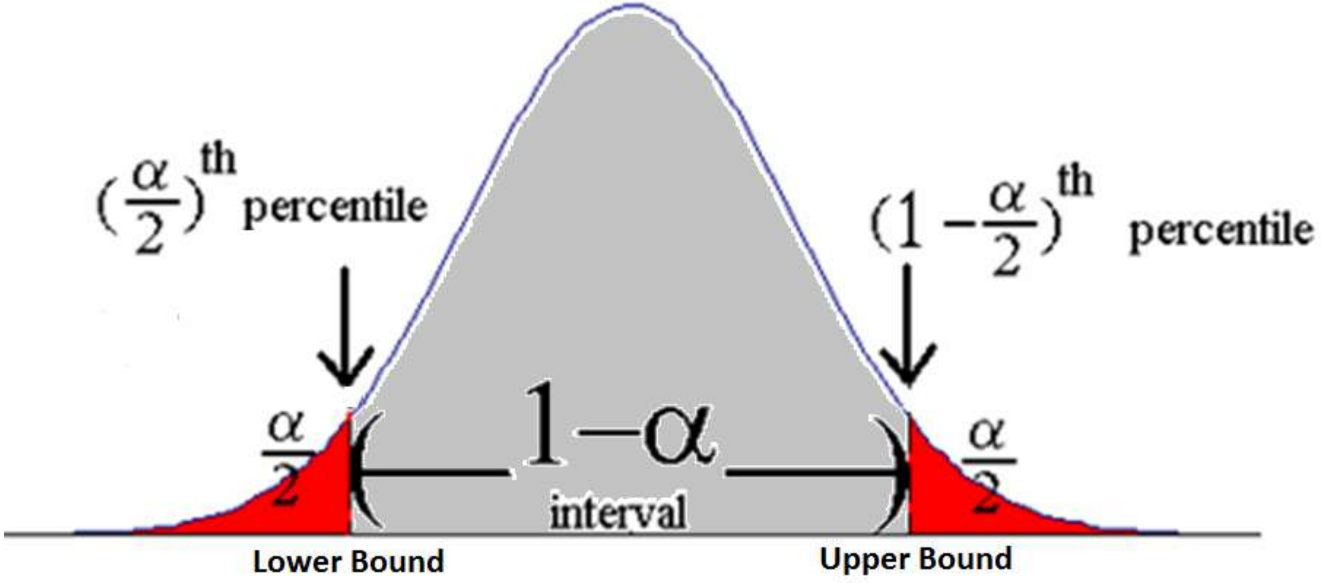
- This is why we have plus or minus z=2.58 for a 99% confidence interval

How We Find an Uncommon Z: 98%



- For a 98% confidence interval lower bound, we need to find the z with a percentile of $\frac{\alpha}{2} = \frac{1 - confidence}{2} = \frac{1-.98}{2} = \frac{.02}{2} = .0100 \rightarrow 1\%$
- If we look this up in the z-table we see that a z-score of -2.33 gives us a value very close to .0100

How We Found the Common Z's: 98%



- For a 98% confidence interval upper bound, we need to find the z with a percentile of

$$1-\frac{\alpha}{2}=1-\frac{1-confidence}{2}=1-\frac{1-.98}{2}=1-\frac{.02}{2}=.9900$$

- If we look this up in the z-table we see that a z-score of 2.33 gives us a value very close to .9900

How We Found the Common Z's: 98%

- **Lower Bound:** If we look this up in the z-table we see that a z-score of -2.33 gives us a value very close to .0100
- **Upper Bound:** If we look this up in the z-table we see that a z-score of 2.33 gives us a value very close to .9900

- This is why we have plus or minus z=2.33 for a 98% confidence interval

Examples

Example

- A random sample of MLB home games showed that the home teams **won 1335 of 2429 games.**
- Our **sample proportion** = $\hat{p} = \frac{1335}{2429} = .5496$
- We should know this is a proportion problem because we’re considering it as a percentage, not an average.
- **Find the 95% confidence interval for the population proportion**

Example

- **Step One:**
- Check Assumptions:
    - $n * \hat{p} = 2429 * .5496 = 1334.9784 \geq 15$
    - $n * (1 - \hat{p}) = 2429 * .4504 = 1094.0216 \geq 15$
    - It is safe to assume the distribution of $\hat{p}$ has a bell shaped distribution
    - The data is from a random sample

Example

- **Step Two:**
- 95% CI:

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

$$.5496 \pm (1.96) \sqrt{\frac{.5496(.4504)}{2429}} = (.5298, .5694)$$

- We are 95% confident that the **true population proportion** of home team wins **is between** 52.98 and 56.94 percent.

Example

- A random sample of MLB home games showed that the home teams won 1335 of 2429 games.
- 95% CI:

$$(.5298, .5694)$$

- We see here that there is a small home field advantage because all of the values in our 95% CI are above 0.5.
    - We know that 0.5 is interesting because it means **more than half the time** or **most**

Example

- A random sample of MLB home games showed that the home teams won 1335 of 2429 games.
- **99% CI:**

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

$$.549 \pm (\mathbf{2.58})\sqrt{\frac{.549(.451)}{2429}} = (.5236, .5756)$$

- We are **99% confident** that the true population proportion of home team wins is between 52.36 and 57.56 percent.

Example

- A random sample of MLB home games showed that the home teams won 1335 of 2429 games.
- 99% CI:

$$(.5236, .5756)$$

- Still, we see here that there is a small home field advantage but we note the interval is larger

Confidence Intervals for Proportions on your TI Calculator

- Confidence Intervals for proportions TI83/84
  - https://www.youtube.com/watch?v=e3HZ6Xv-plk

Confidence Intervals for Proportions on your TI Calculator

- **INPUT:**
  1. Press STAT
  2. Press → to TESTS
  3. Scroll down using ↓ to highlight ‘A: 1-PropZInt’
  4. Press ENTER
  5. Enter the number of the total that had the behavior we’re looking for next to ‘x:’
  6. Enter the total number observations next to ‘n:’
  7. Enter the desired confidence level next to ‘C-Level:’
  8. Highlight ‘Calculate’
  9. Press ENTER

Confidence Intervals for Proportions on your TI Calculator

- **OUTPUT:**
  - (lower bound, upper bound) is our confidence interval
  - $\hat{p}$ is the sample proportion for the problem
  - n is the sample size and should match the number you entered in step 6 above

Confidence Intervals for Proportions

- **StatCrunch Commands w/ data**
  - Stat→Proportion Stats→One Sample →with data (if you have the a list of data)→Choose the column→type the success value into the success box→ choose confidence interval→enter the significance level → Compute

- **StatCrunch Commands w/ summaries**
  - Stat→Proportion Stats→One Sample →with summary (if you have the count) → enter the number of success and total observations→ choose confidence interval→enter the significance level → Compute

Confidence Intervals for Proportions

- **StatCrunch Commands w/ data**
  - Stat→Proportion Stats→One Sample →with data (if you have the a list of data)→Choose the column→type the success value into the success box→ choose hypothesis→ enter the correct hypothesis→ Compute

- **StatCrunch Commands w/ summaries**
  - Stat→Proportion Stats→One Sample →with summary (if you have the count) → enter the number of success and total observations→ enter the correct hypothesis→ Compute

Confidence Intervals

Assumptions	Point Estimate	Margin of Error	Margin of Error
1. Random Sample *2.* $n\hat{p} \geq 15$ And $n(1-\hat{p}) \geq 15$	$\hat{p}$	$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- We are --% confident that the true population proportion lays on the confidence interval.

Extra Topic!

Wilson’s Adjustment for Estimating ρ

- Wilson’s Adjustment is a nice trick to ‘correct’ our confidence interval when n isn’t extremely large and performs poorly when ρ is near 0 or 1

$$\tilde{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\left(\tilde{p}(1-\tilde{p})\right)}{n}}$$

- Where $\tilde{p} = \frac{\mathrm{x+2}}{\mathrm{n+4}}$ is the adjusted proportion of observations

Example

- Let’s complete our previous example about MLB home games with Wilson’s Adjustment this time
  - The only difference here will be how we calculate the sample proportion: $\tilde{p} = \frac{x+2}{n+4}$ instead of $\hat{p} = \frac{x}{n}$
  - **Note:** we shouldn’t see a drastic change because we aren’t in the case where n isn’t extremely large and performs poorly when ρ is near 0 or 1

Example

- A random sample of MLB home games showed that the home teams **won 1335 of 2429 games.**
- Our **sample proportion** =$\tilde{p} = \frac{1335+2}{2429+4} = \ .5495$
- We should know this is a proportion problem because we’re considering a qualitative (categorical) random variable
- **Find the 95% confidence interval for the population proportion**

Example

- **Step One:**
- Check Assumptions:
    - $n * \hat{p} = 2429 * .5496 = 1334.9784 \geq 15$
    - $n * (1 - \hat{p}) = 2429 * .4504 = 1094.0216 \geq 15$
    - Thus, it is safe to assume the distribution of $\hat{p}$ has a bell shaped distribution
    - The data is from a random sample

Example

- **Step Two:**
- 95% CI:

$$\tilde{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\left(\tilde{p}(1-\tilde{p})\right)}{n}}$$

$$.5495 \pm (1.96)\sqrt{\frac{.5495(.4505)}{2429}}$$

$$= (.5297, .5693)$$

- We are 95% confident that the **true population proportion** of home team wins **is between** 52.97 and 56.93 percent.

Example

- A random sample of MLB home games showed that the home teams won 1335 of 2429 games.
- 95% CI:

$$(.5297, .5693)$$

- We see here that there is a small home field advantage because all of the values in our 95% CI are above 0.5.
  - We know that 0.5 is interesting because it means **more than half the time** or **most**