Chapter 9: Confidence Intervals

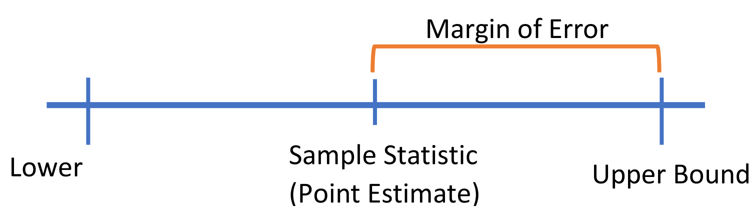
Confidence Interval Basics

- Estimation
  - Confidence intervals are used to find a range of reasonable possible values for the parameter
  - Hypothesis Testing examines one specific candidate parameter and helps us decide if it is probabilistically reasonable.
  - Both methods use the same principles, but provide insights from slightly different angles!

- Building a confidence interval
  - Our **sample statistic** is typically our **best estimate** of the **population parameter**, but due to sampling variability, it will have some error. So what we want to do is build an interval around our sample statistic that we hope will contain the true parameter.

  - Our sample statistic will be our ______________________, as it is our best guess for the parameter's value.

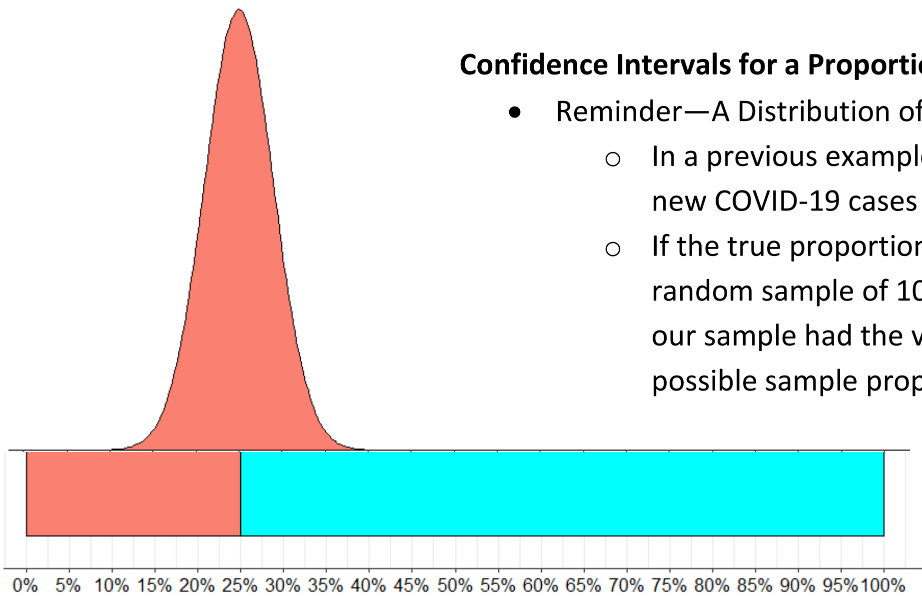
  - The distance we extend out is called the ____________________. This is how much error we are allowing from our sample statistic in order for the interval to work.



  - But how far out should we extend? What determines how large the margin of error is?

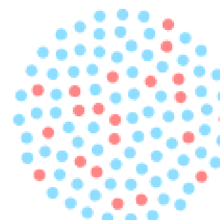
    - First, our desired _____________________! The more confident you want to be that your interval works, the farther out you need to extend the interval.

    - Second, the ___________________ of our sample statistic (which depends on both our sample size and how much variability is in the data). The larger the standard error for our sample statistic, the larger the margin of error needs to be for our confidence interval.



Confidence Intervals for a Proportion

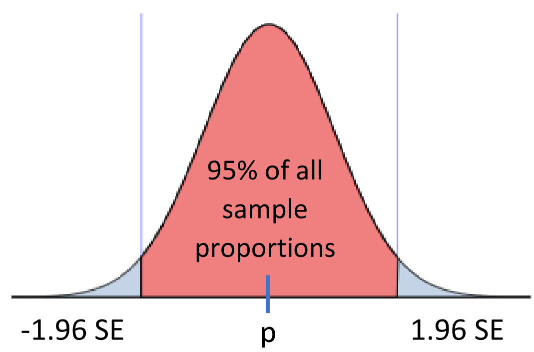
- Reminder—A Distribution of possible sample proportions
  - In a previous example, we discussed what proportion of new COVID-19 cases were detected to be a new variant.
  - If the true proportion were, say, 0.25, and we took a random sample of 100 patients to see what proportion of our sample had the variant, then there is a distribution of possible sample proportions we could get.



21%

  - But if we're doing inference, all we have is our one sample of 100 patients and one sample proportion. How does this help us *estimate* a range for the true proportion?

- Using properties of the Normal Distribution
  - We have a sample of 100, and 21 people had the variant. We know that this $\hat{p}$ value of 0.21 falls in some distribution of sample proportions, and this distribution should be normally distributed around p.
  - We know that the standard deviation of this sampling distribution will be the standard error of $\hat{p}$, which is equal to $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ *...and we can approximate it by plugging in $\hat{p}$.*
  - We can use normal distribution properties to find how often it will be within a certain number of standard errors from the true center, p.
    - When values distribute normally, 95% of those values will be within 1.96 SDs of the mean.
    - ***95% of all possible $\hat{p}$'s*** *will be within* ***1.96 standard errors*** *of p!*
    - Therefore, **the Margin of Error** for a 95% confidence interval will be **1.96*** $\mathbf{SE}_{\hat{p}}$



Confidence Interval for Population Proportion = $\hat{p} \pm Z_{\alpha/2} * SE_{\hat{p}}$

- $Z_{\alpha/2}$ is a z-score representing the number of standard errors we would need to extend for a 1-α confidence level. α in this context represents the probability that your confidence interval **does not** contain the true parameter.
- If looking at the z-table we use for this class…
  - Look at the bottom row and identify the α representing the complement of your confidence level. Example: 95% confidence → α = 0.05
  - Look at the $z_{\alpha/2}$ value to get your needed z-score.
- Identify some common confidence interval values!
  - 90% of all possible sample proportions will be within 1.645 standard errors of p
  - 95% of all possible sample proportions will be within 1.96 standard errors of p
  - 98% of all possible sample proportions will be within _____ standard errors of p
  - 99% of all possible sample proportions will be within _____ standard errors of p
  - *Use the z-table to fill in the values above!*
- This would yield the following confidence intervals. Can you fill in the missing ones?

90% Confidence Interval for Population Proportion = $\hat{p} \pm 1.65 * SE_{\hat{p}}$

95% Confidence Interval for Population Proportion = $\hat{p} \pm 1.96 * SE_{\hat{p}}$

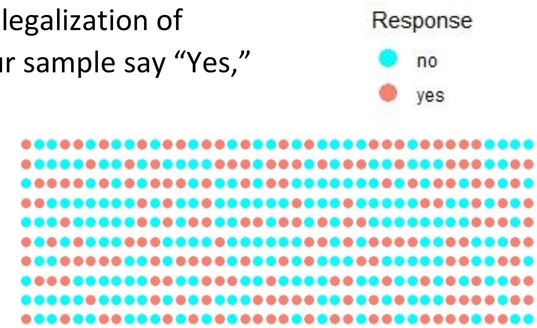
98% Confidence Interval for Population Proportion = _______________

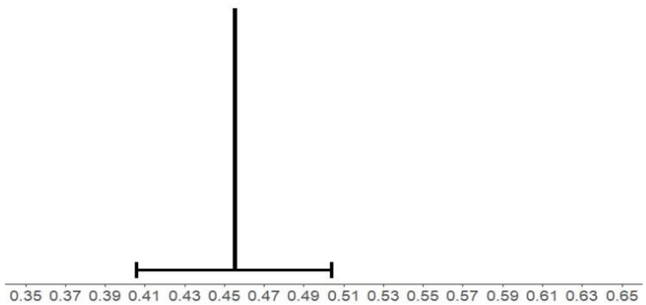
99% Confidence Interval for Population Proportion = _______________

This method is known as a "**Wald Interval**." It is a common interval approach for estimating a population proportion from a sample—*but if you take more statistics courses, you can learn other methods that may work better in special situations!*

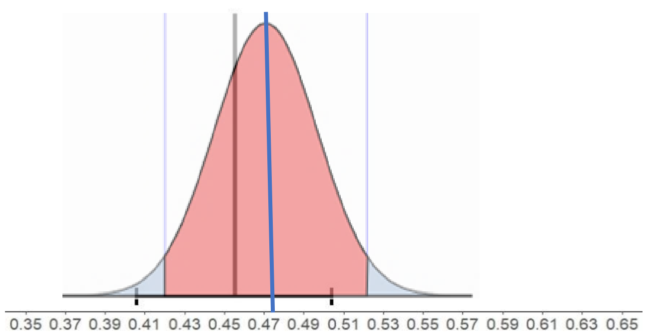
Let's think about this process from the beginning again!

We're trying to estimate the proportion of Texas voters who support legalization of recreational marijuana. We poll 400 Texans and find 182 people in our sample say "Yes," making $\hat{p}$ = 0.455

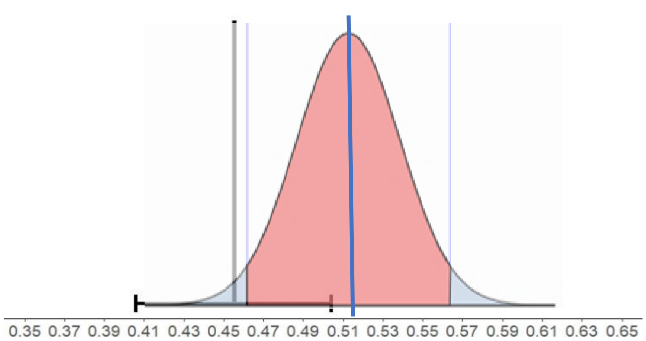




If we want to build an interval that we are 95% confident contains that true proportion, we need our interval to stretch 1.96 standard errors in each direction. That would be $1.96 * SE_{\hat{p}}$, where $SE_{\hat{p}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$. Calculate the margin of error!



Here is one of those **95% of times where our interval contains the true proportion.** The blue line represents the actual proportion, and the transparent curve in the background represents the distribution of possible sample proportions for n = 400 if this were the case. Our sample proportion is comfortably inside the 1.96 SE zone.



However, there is another story where we are in the **5% of times where our interval doesn't contain the parameter.** In this picture, our sample proportion is apparently more than 1.96 SEs away from p (something that will happen 5% of the time). Thus, our 95% interval does not quite reach the parameter.

- Confidence Intervals in the Polls
  - Many of the polls you might encounter in the news report their point estimate, along with the margin of error for that poll.
    - But beware—not all samples are perfectly representative, so even margin of error is only half the story to understanding how much polls can miss the true proportion!
    - https://www.youtube.com/watch?v=TambSayfCOE

Confidence Intervals in the News: In a sample of 5,109 people who were chosen through random sampling and weighted to be representative of the population across several factors, Pew Research estimates that approximately 59% of U.S. Americans believe abortion should be legal in "all or most" cases. *You can find full details here.* https://www.pewresearch.org/fact-tank/2021/05/06/about-six-in-ten-americans-say-abortion-should-be-legal-in-all-or-most-cases/

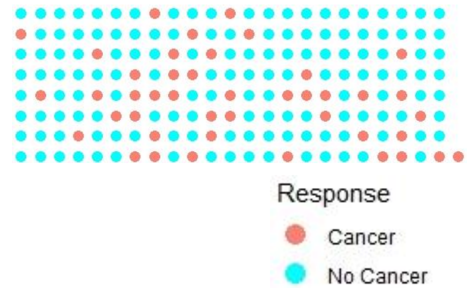
Based on the proportion and sample size, what is the 95% margin of error of this poll?

Assumptions for using a Wald Interval for Proportions

1) The distribution of possible sample proportions is normally distributed
   - **The 10/10 Rule**: This is a reasonable assumption to make if our sample has **at least 10 of each response (e.g., >=10 "yes" and >=10 "no" responses)**
   - When this isn't true, the distribution of sample proportions might be skewed if it is centered too close to 0% or 100% and Wald intervals are unreliable.
2) $\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ is a reasonable estimate of $SE_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
   - When we have a fairly large sample size (n >100), then this is no problem. The difference due to using $\hat{p}$ will be very tiny.
   - When n is smaller, then Wald Intervals are probably "ok" as long as the 10/10 rule above still holds. Though there are corrections you can make with software!

- **For our class, we will <u>only</u> use the 10/10 rule to decide if a Wald Interval is appropriate!**

Practice: A particular variant of the BRCA1 gene has been linked to breast cancer. Researchers are trying to estimate the true risk of breast cancer among women with this gene variant. 185 women with this gene variant were studied over time, and 43 of them developed breast cancer by Age 60. Let's assume that these 185 women constitute a representative sample of women in the population with this gene variant.



Is it appropriate to create a Wald Interval in this context?

What proportion of women in our sample developed breast cancer?

Calculate the standard error (expected error) for our sample proportion

Find a 95% confidence interval (using a Wald-Interval) for the proportion of women with the BRCA1 gene variant who will eventually develop breast cancer by age 60.

If we were reporting our sample proportion to a journal as an estimate of the risk of breast cancer among women in this population, what would be the 95% margin of error we report alongside?

A news article sees this study and makes the following claim: "Researchers are 95% confident that women with this BRCA1 variant will develop breast cancer by age 60." Is this a correct interpretation? How might we write a correct interpretation?

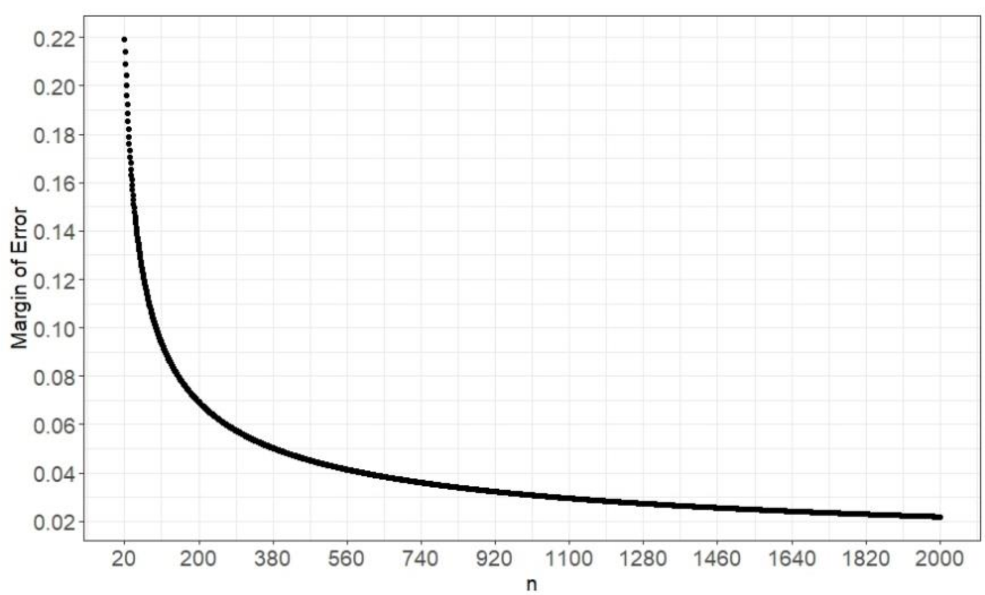
- Choosing a confidence level
  - The more confident you want to be, the more SEs you need to stretch out and the more the margin of error ________________. More confidence is great, but be careful that it doesn't make your interval so wide that it becomes useless!

Practice: If we wanted to be 99% sure that our intervals contains the true risk of breast cancer for women with the BRCA1 variant, what would the margin of error be then?

  - You should balance finding a width that is narrow enough to be helpful and confidence that is strong enough to be meaningful. *Many researchers find 95% to be a nice balance!*
- Are there any other ways to narrow the interval (i.e., decrease the margin of error)?
  - Besides just losing confidence, we could also increase our sample size.

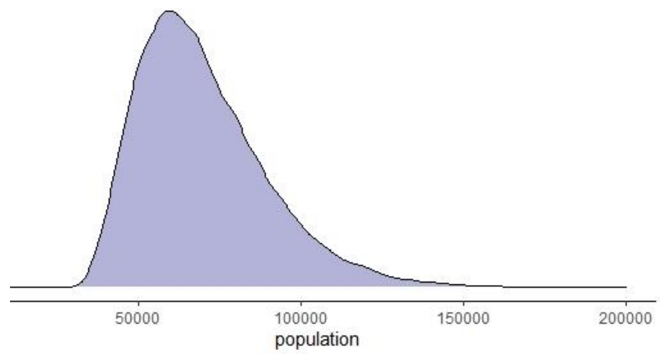
Practice: What if in the breast cancer study, we increased the sample size to 300? Keep the sample proportion the same value, and calculate the new margin of error.

  - There will be diminishing returns in sample size in sample size though. Check out the picture below to see the trend *(these would be the exact values if p = 0.5).*

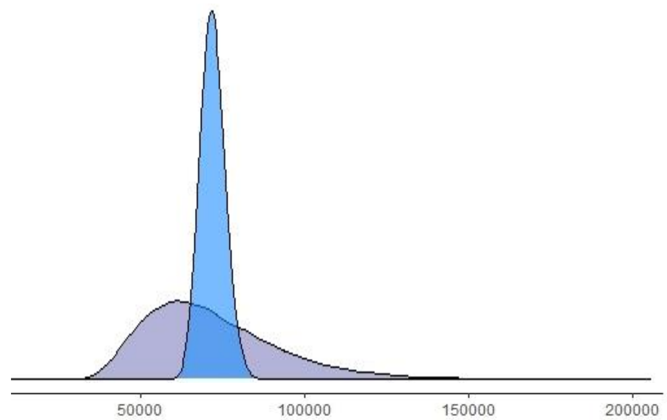


Confidence Intervals for Population Means

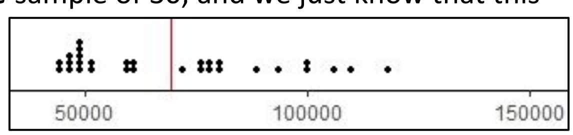
- Basically the same process!
  - When working with binary data, it makes sense to estimate a proportion. But when working with numeric data (especially continuous data), we're often working with a measure like the mean.



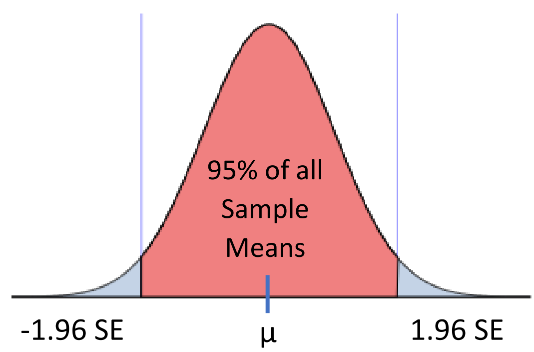
  - Unlike binary data, numeric data has a distribution that we can represent with a density curve. Consider some population to the right, representing the *(fictional)* salaries of individuals who graduate with a Bachelors degree in statistics. This is a positively skewed distribution with μ = \$70,000 and σ = \$20,000.



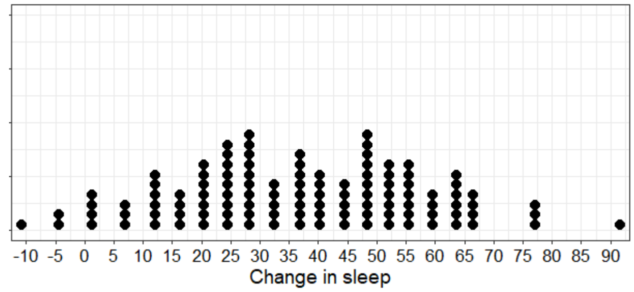
  - The second picture superimposes the distribution of possible $\bar{x}$'s for samples of size 30. In other words, if we randomly sample 30 people from this population, take their average salary, and repeat this process many many times.
  - As we learned with the Central Limit Theorem, the sample mean usually distributes normally around the true population mean, with a standard deviation of $\frac{\sigma}{\sqrt{n}}$ (which we often have to estimate with the formula $\frac{s}{\sqrt{n}}$)
  - BUT…if we're doing inference, all we have is just ***one*** sample of 30, and we just know that this sample mean falls *somewhere* in that normal distribution around the true parameter.



  - We can again use properties of the normal distribution to estimate how far we should be from μ a certain percent of the time.
  - When values distribute normally, 95% of those values will be within 1.96 SDs of the mean. That means that ***95% of all possible $\bar{x}$'s** will be within* ***1.96 standard errors** of the mean!*



Practice: Researchers would like to estimate the average change in sleep duration for adults trying a new experimental sleep aid. The researchers find that over the length of the study, the 131 participants report an average of 36 minutes more sleep a night while on the sleep aid. The standard deviation for change in sleep duration is 18 minutes. *Assuming this is a representative sample of adults*, use this information to **build a <u>98%</u> confidence interval** for the average change in sleep duration for adults on this sleep aid.



First, calculate the approximated $\mathrm{SE}_{\bar{x}}$

Now calculate the 98% confidence interval for the average change in sleep duration in this population as a result of taking this sleep aid.

Consider this statement—"this study finds that 98% of **individuals** on this sleep aid had an increase in sleep within the confidence interval range reported above." Is this a correct interpretation?

Practice: A geriatrics researcher reported findings from her study that residents over the age of 80 who live alone spend, on average, only 24 minutes a day talking to another human being. She also reported that the 95% confidence margin of error is 4.8 minutes. What would be the 95% confidence interval for the average amount of talking time that solo-living seniors seem to be getting?

If she wanted to have a smaller margin of error, what *could* this researcher do? (What two things affect the size of the margin of error?)

Practice: In a study to try to find the average commute time for workers in Chicago, researchers took a representative sample of 182 people who worked in the city limits. They used this information to construct a 95% confidence interval for the mean commute time for all Chicago workers in the city limits. The interval was computed to be (34.12, 38.60) minutes.

Which of the following would be a correct interpretation?

i) If I pick a random Chicago worker in the city limits, we are 95% confident that their commute time is between 34.12 and 38.60 minutes.

ii) We are 95% confident that the average commute time for all Chicago workers in city limits is between 34.12 and 38.60 minutes.

iii) 95% of the 182 workers sampled take between 34.12 and 38.60 minutes for their commute.

iv) We are 95% confident that the average commute time for my 182 workers is between 34.12 and 38.60 minutes.

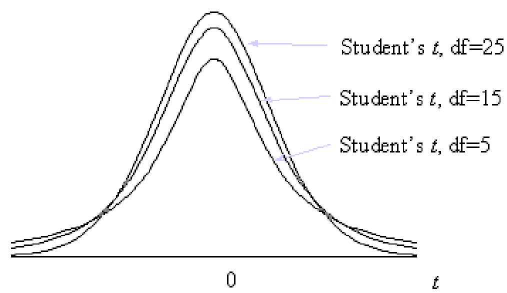
- Interpret carefully!
  - Confidence intervals tell you about the position of a _________________, but they do not necessarily tell you about the position of individual data points.
    - Be careful with means—a mean is not a measure for an individual!
    - This is in contrast to proportions—a proportion can also be used to represent the probability that an individual will fit the category described.

  - Confidence intervals *shouldn't* be used to make claims about our _____________. We already know our sample information completely—it's the population parameter that we are estimating!
- Accounting for the estimation error in the Standard Error
  - When we calculate the standard error for the mean, we will almost always need to calculate it using the standard deviation of our sample (s) since we likely won't know σ.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad SE_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

  - When our sample is quite large (in our class, we'll say n > 120), then this isn't much of a problem as the error will be negligible, but when n is smaller, it can make a difference.
  - We again need to turn to the t-distribution to adjust our calculations.

- **Calculating t-intervals using the t-distribution**
  - Think of the t-distribution as a "corrected" normal distribution. It is slightly wider than a normal distribution, so t-intervals create slightly wider confidence interval than a z-intervals.
  - Updated Equation!
    - With our intervals before, we plugged in exact values (like 1.96 or 2.33) to signal how many standard errors we need to go out, but now, we're going to substitute in $t_{\alpha/2}$ to represent how many <u>*approximated* standard errors</u> we need to go out given our sample size and desired confidence level.



Confidence Interval for the Population Mean = $\bar{x} \pm t_{\alpha/2} * SE_{\bar{x}}$

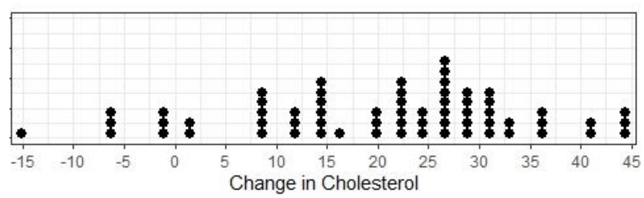
    - As an example, $t_{0.025} > 1.96$, but the magnitude of difference depends on how large a sample size we have. If n = 15, t = 2.145. If n = 50, t = 2.01…as n increases, $t_{0.025}$ will approach 1.96 because the error correction will become unnecessary.
  - Using the t-table
    - Note that you will need to divide the α value in half to find the correct column. Then look up the row by degrees of freedom.

Find the t-score needed to create a 95% confidence interval for a mean using our sample of 25	Find the t-score needed to create a 99% confidence interval for a mean using our sample of 51

Assumptions for Creating z or t-intervals for a Mean

- ✓ z-intervals are appropriate if…
  - Sample size is large, like n>120 (*or in the rare situation that σ is known)*
  - The sampling distribution for the mean is approximately normally distributed. *We need one of these situations to be true.*
    - Population distribution is approximately normally distributed
    - n >30 and population is **not** highly skewed (no long tail)
    - If n >120, then normal sampling distribution is generally safe to assume
- ✓ t-intervals are appropriate if…
  - σ is being approximated by s
  - The sampling distribution for the mean is approximately normally distributed. *We need one of these situations to be true.*
    - Population distribution is approximately normally distributed
    - n >=30 and population is **not** highly skewed (no long tail)
    - *If n >120, it would be reasonable to do a z-interval!*

Practice: 64 patients who are considered to be representative of the high-cholesterol population over 50 are trying new medication for cholesterol. They had an average cholesterol drop of 19.0 mg/dL with a standard deviation of 14.3 mg/dL.



What type of confidence interval should we create?
Are assumptions met?

Find a 95% confidence interval for the average cholesterol drop of patients over 50 with high cholesterol

If you had a patient over 50 with high cholesterol, could you use this interval to say that you are 95% sure *HIS* cholesterol drop would be between these two values?

What type of design does this study seem to use? Does this study have good internal validity to make claims about the effectiveness of this drug?

Practice: How long does it take to drive from Champaign to Chicago O'Hare Airport?

Let's say that we took a random, independent sample of 40 Champaign drivers and recorded the amount of time in minutes it took them to complete the drive. The average drive time for that sample of 40 was 158 minutes with a *sample standard deviation* 8 minutes.

Create a 95% confidence interval for the true average drive time. *You'll need to use a t-interval.*

Now create a **98%** confidence interval with the same data. Which do you think will be wider, this 98% confidence interval or the 95% confidence interval from before? Explain your prediction.

Now, calculate the 98% confidence interval for the sample of 40.

Now let's say that we just took a sample of **20 drivers**. Their average drive time was also 158 minutes with sample standard deviation 8 minutes. If we created a 95% confidence interval for the true average drive time, would it be larger or smaller than the interval using 40 drivers? Explain your prediction.

Now calculate the 95% confidence interval using this data.