Chapter 5: Confidence Intervals

Confidence Interval for a Proportion: π

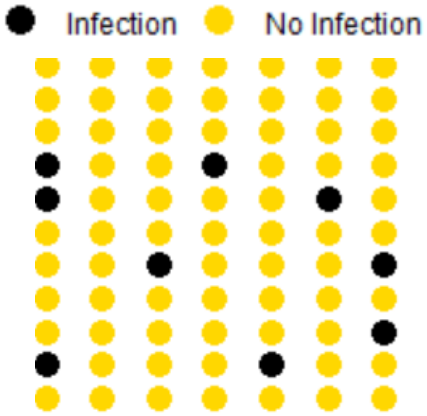
Investigation: You are part of a team of entomologists studying the presence of "Black queen cell virus" (BQCV) in honey bee colonies around the state. https://beeaware.org.au/archive-pest/black-queen-cell-virus/

After examining 84 hives that have been carefully selected across the state, you found this virus present in 10 of those hives. You'd like to use this data to estimate what proportion of Illinois honey bee hives have been infected so far this season to determine how serious this threat might be to honey bee colonies.

Population: All honey bee hives in state

Unit of observation: One hive

Variable (and type): BQCV infection present or not (categorical)



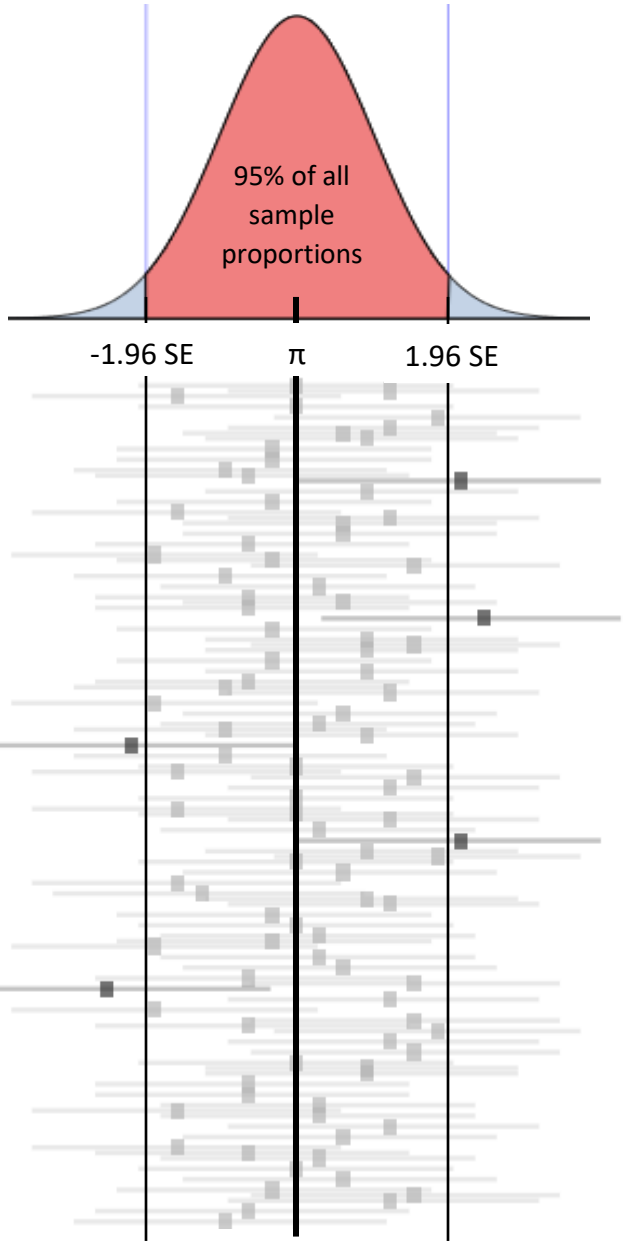
Testing vs. Estimating

- Can we test a parameter?
  - Hypothesis Testing examines one specific candidate parameter and helps us decide if it is probabilistically reasonable (e.g., Is there evidence that more than 8% of hives are infected?).
  - But in this context, we don't have a particular value to test—we just want to estimate what the true proportion might reasonably be!
- When would we estimate?
  - If there's no candidate parameter to assess, then we could instead build an interval to express where the parameter most likely is.
  - Even if there is a benchmark to test, interval estimates can *complement* p-values by showcasing how far the true parameter likely is from that benchmark.

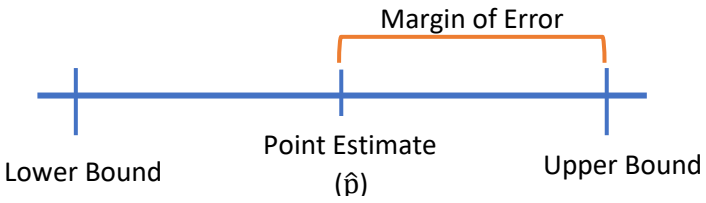
Building a "z-interval" estimate

- Let's start by thinking about a point estimate. This represents our best single estimate for the true parameter.
  - Symbolically, what parameter are we trying to estimate in the bee example? π
  - What is our point estimate for that parameter? $\hat{p}$
- Next, we'll need to think about the expected error of this sample statistic as an estimate for the parameter.
  - What is the standard error in our statistic as an estimate for this parameter?

- Finally, let's consider *how many* standard errors we should extend out from our point estimate. This choice helps us determine how confident we are that the interval contains the parameter.
- If we make a "parametric" assumption that $\hat{p}$ is normally distributed around π, then…
  - **90%** of all possible $\hat{p}$'s will be within **1.645** standard errors of π
  - **95%** of all possible $\hat{p}$'s will be within **1.960** standard errors of π
  - **98%** of all possible $\hat{p}$'s will be within **2.326** standard errors of π
  - **99%** of all possible $\hat{p}$'s will be within **2.576** standard errors of π



The actual distance we extend out from our point estimate is called the <u>margin of error</u> and will be some number of standard errors in length.



<u>Highlight point estimate, SE, and number of SE extended</u>

Confidence Interval for π: $\hat{p} \pm Z_{\alpha/2} * \sigma_{\hat{p}}$

- $Z_{\alpha/2}$ is a z-score representing the number of <u>standard errors</u> 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.
- Using the normal distribution simulator: https://istats.shinyapps.io/NormalDist/
  - Go to "Find Percentile/Quantile"
  - Take a two-tailed percentile
  - Type in your desired confidence level

Assumptions for creating a z-interval for a proportion

- Z-intervals for a proportion (also referred to as "Wald intervals") are fairly reliable, but they do depend on two assumptions.
    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 $\hat{p}$ might have some skew, or be too discrete to reasonably use a normal approximation.
    2) $\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ is a reasonable estimate of $\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
        - 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 typically "ok" as long as the 10/10 rule above still holds.

Exploring Confidence Intervals: Open the following simulation and let's make some observations about how confidence intervals behave https://istats.shinyapps.io/ExploreCoverage/

What do you notice when you increase the level of confidence for your intervals? *Which part of the equation is affected by this?*

Intervals gets wider (and more of them cover the parameter!) because we are increasing the number of SEs we are extending

What do you notice when you generate confidence intervals from larger sample sizes? *Which part of the equation is affected by this?*

Intervals get narrower (because SE is getting smaller)

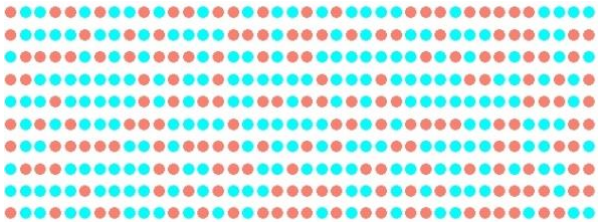
Does the accuracy/coverage of Wald Intervals (z-intervals) change when the sample size is smaller, or when the true proportion is closer to 0 or 1?

Coverage is noticeably less than 95%.

Completing our investigation: Using a z-interval, calculate a **90%** confidence interval for the true proportion of honey bee hives in the state that have a BQCV infection based on our sample data.

Investigation: Let's say that we're examining the proportion of Texas voters who support legalizing recreational marijuana. We poll 400 Texans and find 218 people in our sample say "Yes."

What is our point estimate for the true proportion of Texas voters who support legalizing recreational marijuana?



What is the standard error for our point estimate?

Would a Wald Interval be appropriate here? If so, what would be the margin of error for a **95%** confidence interval?

Calculate the lower and upper bound for a **95%** confidence interval.

Let's say that in an upcoming election, 60% of voters would need to approve for this to become legal. Assuming we have a truly random sample, would you say that 60% is a plausible possibility for the parameter? *What is the smallest confidence level you can find that would include 0.60?*

https://istats.shinyapps.io/Inference_prop/

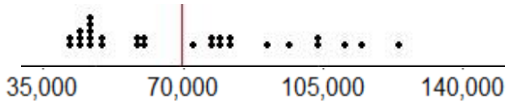
Note our 95% CI does not overlap 0.60. But if extend to about a 97.3% confidence level, we will hit 0.6 at the upper bound.

This should correspond to a two-sided p-value of around 2.7% (slight discrepancy since it used 0.60 rather than 0.545 in the SE calculation)

Confidence Intervals for a Mean: μ

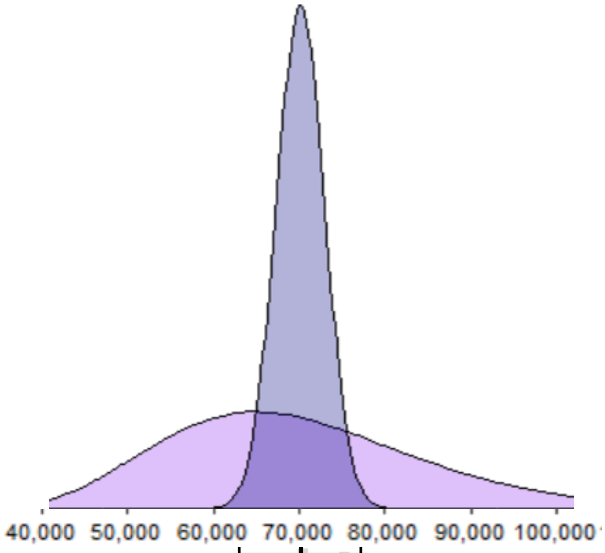
Investigation: We would like to examine the starting yearly income for students who have graduated from the U of I with a Bachelor's degree in Statistics. Let's say that this distribution is well represented by the density curve below. This is a positively skewed distribution with μ = $70,000 and σ = $23,000.

Now, let's imagine we sampled 30 recent graduates randomly from this population.



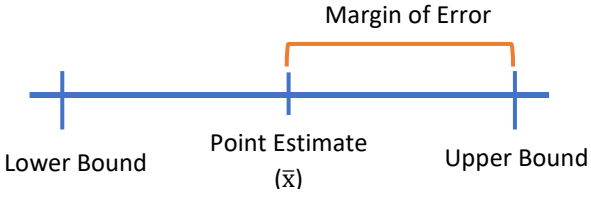
Let's consider how might follow a similar process to estimate μ using our sample information!

- Our **point estimate** represents our best single estimate for the parameter.
  - *Symbolically,* what is our point estimate for μ? $\bar{x}$
- The **standard error** represents the expected error in our point estimate.
  - What is the standard error in our statistic as an estimate for this parameter?



- Finally, let's consider *how many* standard errors we should extend out from our point estimate. This will help us determine the **margin of error** for our interval.
  - What would be our margin of error if want 95% confidence in capturing μ? *Assume $\bar{x}$ is distributing normally.*

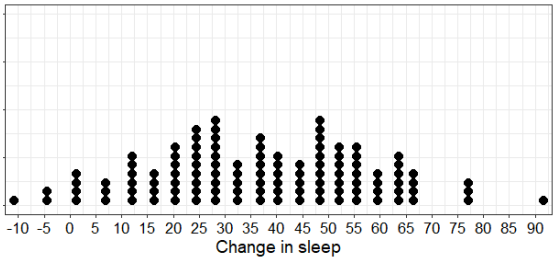
Consider this statement: *"If we choose a graduate from this population at random, we are 95% confident that their starting salary will be within this margin of error from $70,000."* Is this a correct interpretation?



Investigation: 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 36.14 minutes more sleep a night while on the sleep aid.
- The standard deviation for change in sleep duration is 18.00 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.



Identify our point estimate for μ

Identify the expected error in our point estimate

Calculate the margin of error we will need if we want to be 98% confident in capturing the true mean change in sleep as a result of this medication. If done correctly, we should find the following interval.

(32.48, 39.80)

Interpreting Confidence Intervals

Which statement below correctly interprets this confidence interval in context?

I am 98% confident that an individual who takes this medication will experience between a 32.48 and 39.80 minute increase in their sleep.

<u>Nope. Individuals may vary quite a bit more.</u>

I am 98% confident that the average change in sleep duration among all individuals who may take this medication is between 32.48 and 39.80 minutes.

<u>Yes! This is describing μ</u>

I am 98% confident that the average change in sleep duration among these 131 individuals is between 32.48 and 39.80 minutes.

<u>Nope. This is describing $\bar{x}$. We already know what our sample mean is.</u>

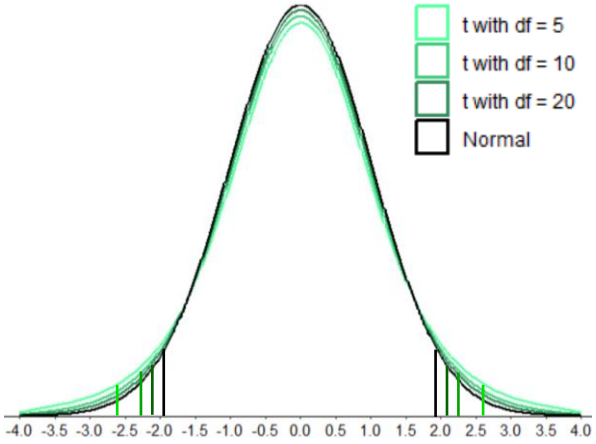
Confidence intervals are designed to identify the position of a <u>parameter</u>.

- Accounting for the estimation error when using s in place of σ
  - When we calculate the standard error for the mean, we typically need to calculate it using the standard deviation of our sample (s) since we likely won't know σ.

$$\sigma_{\bar{x}} \approx s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

  - For large sample sizes, this discrepancy should be negligible, but at smaller sample sizes, we will want to make an appropriate adjustment to our method!

- **Calculating t-intervals using the t-distribution**
  - Our standard error calculations will now *vary* (potentially quite a bit!) with each sample.
  - If I still want to be, say, 95% confident that my interval includes the parameter, then I might need to extend slightly more than 1.96 standard errors out from $\bar{x}$
  - The number of approximated standard errors we now extend will be referred to as a **t-score** rather than a z-score.



  - This t-score will depend both on the <u>confidence level</u> and on the <u>degrees of freedom</u> associated with our $s_{\bar{x}}$ estimation.

t-interval for the Population Mean = $\bar{x} \pm t_{\alpha/2} * s_{\bar{x}}$ (use especially when $n \leq 100$)

z-interval for the Population Mean = $\bar{x} \pm z_{\alpha/2} * \sigma_{\bar{x}}$ (also ok with $s_{\bar{x}}$ when n is large)

  - As an example, $t_{0.025} > 1.96$, but the magnitude of difference depends on how large a sample size we have.
    - For df = 5, $t_{0.025} = 2.571$. For df = 10, $t_{0.025} = 2.228$. For df = 20, $t_{0.025} = 2.086$
    - …as n increases, $t_{0.025}$ will approach 1.96 as "s" converges toward "σ"
  - Using the t-simulator - https://istats.shinyapps.io/tdist/

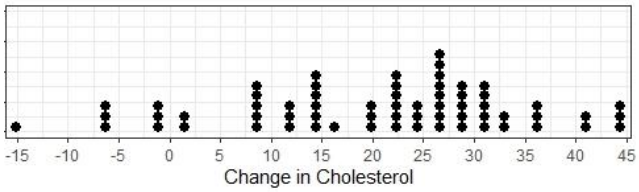
Find the t-score needed to create a 95% confidence interval for a mean using a sample of size 25	Find the t-score needed to create a 99% confidence interval for a mean using a sample of size 51

Assumptions for Creating z or t-intervals for a <u>Mean</u>

- ✓ z-intervals are appropriate if…
  - Sample size is large, like n>100 (*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 >100, then CLT will likely apply except in very highly skewed populations
- ✓ 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 >100, it would be reasonable to do a z-interval!*

Investigation: 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. *Use t = 1.998*

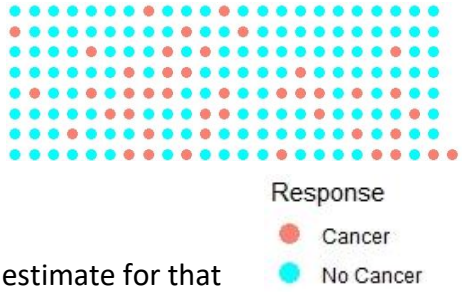
If you had a patient over 50 with high cholesterol, could you use this interval to say that you are 95% sure *HIS* cholesterol drop would be between these two values?

Chapter 5 Additional Practice

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 by age 60.

185 women with this gene variant were studied over time, and 43 of them developed breast cancer by Age 60. Assume these 185 women are a representative sample of women with this gene variant.

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



Symbolically, what parameter are we trying to estimate? What is our point estimate for that parameter?

What is the expected error (standard error) in our point estimate as an estimate for the parameter?

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

In drafting an abstract for a conference, your co-author writes “We are 95% confident that women with this BRCA1 variant will develop breast cancer by age 60.” What’s wrong with this statement? How could you rewrite it?

In related research, someone finds a 95% confidence interval for the risk of ovarian cancer based on a different gene variant. The interval is (0.084, 0.131). What is the point estimate and margin of error for this interval?

Practice: How long does it take semi-truck drivers to deliver supplies to a Champaign store when driving from a Chicago warehouse?

Let's say that we take a sample of 40 trips and recorded the amount of time in minutes it took them to complete the drive. The average drive time for that sample was 168 minutes with a *sample standard deviation* 9 minutes.

Create a 95% confidence interval for the true average drive time. *According to the t-simulator, we'll need to use a t-score of 2.023.*

Now consider if we had created **98%** confidence interval with the same data.

First, **predict**. Which do you think will be wider, this 98% confidence interval or the 95% confidence interval from before? *Think conceptually or mathematically about what is changing here!*

Now, calculate the 98% confidence interval for the sample of 40. *Use t = 2.426*

Now let's say that we just took a sample of **20 trips**. Their average drive time was also 168 minutes with sample standard deviation 9 minutes.

First, **predict.** If we created a 95% confidence interval for the true average drive time, would it be larger or smaller than the interval using 40 trips? *Think conceptually or mathematically about what is changing here!*

Now check your work by calculating this new confidence interval. *Use t = 2.093*

Chapter 5 Learning Goals

After this chapter, you should be able to…

- Recognize interval estimation and hypothesis testing as two different approaches for inference
    - Testing can judge the plausibility of a specific value as a possible parameter
    - Estimation can provide an interval of reasonable guesses for the parameter
- Identify $\hat{p}$ as the point estimate for π and as the basis for building an interval
- Identify $z^*\sigma_{\hat{p}}$ as the margin of error for a confidence interval for π, where z represents how many standard errors needed to achieve a specified level of confidence
- Calculate a z-interval for a proportion based on provided count data and z-score.
- Describe how the width of a confidence interval will be affected by changes in desired confidence level or sample size.
- Recognize situations when z-intervals for a proportion (Wald intervals) would **not** be appropriate
    - When the distribution of $\hat{p}$ is not well approximated by a normal distribution—which we will judge using the 10/10 rule in this course.
    - Recognize that other interval methods exist with better coverage in these cases.
- Identify $\bar{x}$ as the point estimate for μ and as the basis for building an interval
- Identify $z^*\sigma_{\bar{x}}$ or $t^*s_{\bar{x}}$ as the margin of error for a confidence interval for μ, where z or t represents how many standard errors needed to achieve a specified level of confidence
- Calculate a z-interval or t-interval for a mean based on sample data and a provided z or t-score
- Recognize that t-intervals are needed when estimating σ with s, especially in smaller sample contexts like $n \leq 100$
- Recognize situations when z or t intervals for a mean would **not** be appropriate
    - When the distribution of $\bar{x}$ is **not** well approximated by a normal distribution. This happens when population distribution is skewed and sample size is too small for the CLT to apply.
- Interpret a confidence interval, in context, as an estimate for the position of a parameter
- Convert between values flexibly, such as identifying the point estimate or margin of error for a confidence interval based on interval bounds provided.