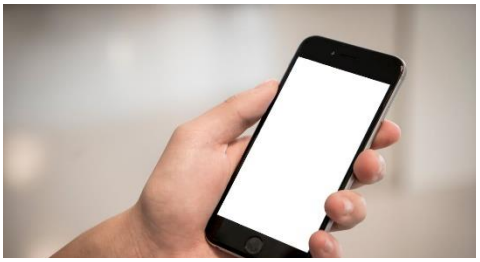
Chapter 7: Comparing Two Means

Investigation: A research study compared the reaction times of automobile drivers with and without cell phones. The goal of the study was to determine if (and by how much) using a cell phone might influence the reaction times of drivers when confronted with a road hazard. https://doi.org/10.1111/1467-9280.00386

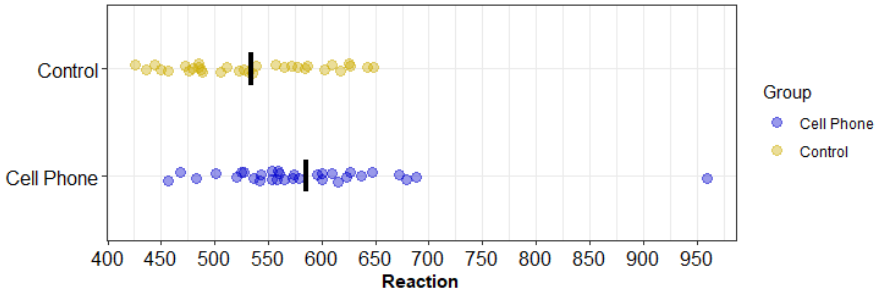
In a study with 64 people, the researchers randomly assigned 32 people to operate a simulated vehicle while holding their cell phone and having a conversation. The other 32 were randomly assigned to do the same thing, but while listening to the radio or an audio book.



The researchers measured how many milliseconds it took drivers to hit the brake after the road hazard appeared.

Is there evidence that drivers' reaction times when on a cell phone is different than it is without?

	Phone	**Control**
Mean Reaction	$\bar{x}_1 = 585.4$	$\bar{x}_2 = 533.8$
SD	$s_1 = 89.6$	$s_2 = 65.4$
Sample Size	$n_1 = 32$	$n_2 = 32$



Unit of Observation: <u>One person/driver</u>

Response Variable (and type): <u>Reaction time (numeric)</u>

Explanatory Variable (and type): <u>Cell phone vs. radio/standard distraction (categorical)</u>

- **The Null Hypothesis:** <u>There is **no** difference</u>
  - Non-directionally: $\mu_1 = \mu_2$
  - Directionally: *Mirror the alternative*
- **The Alternative Hypothesis:** <u>There is a difference</u>
  - Non-directionally: $\mu_1 \neq \mu_2$
  - Directionally: $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Identify the null and alternative hypotheses for this investigation

- Exploring this investigation through a **Permutation Test**
    - IF the null is true and there is no underlying difference between these two populations on average**…**we can "shuffle" the group designation randomly and simulate sample mean differences we should expect under the null hypothesis.
    - Go to this site and choose "Reaction Times" https://istats.shinyapps.io/PermDist_2samples/

Exploring the Distribution of $\bar{X}_1 - \bar{X}_2$ as an estimator for $\mu_1 - \mu_2$

Let's "permutate" the group designations randomly across each observed response value. Notice the distribution of sample mean differences that emerge from these random permutations and draw it below.

What is the mean of this distribution? *(what is it converging towards as we complete more permutations?)*

What is the standard deviation of this distribution? *This would be the standard error of* $\bar{x}_1 - \bar{x}_2$

How often do we observe a permuted sample mean difference as large or larger than our actual sample mean difference? *Check out the "Permutation Test" tab up top.*

Do we have evidence for a difference? Or could the difference observed in our sample be reasonably explained as random chance?

Exploring this investigation through an **Independent Samples z or t-test**

- **Parametric assumption**
  - You might notice that this distribution is *approximately* normally distributed
  - For that reason, we *could* use a parametric testing approach to do inference rather than estimate the p-value with a finite number of simulations.
- **Calculating the Standard Error for $\bar{x}_1 - \bar{x}_2$**
  - We can estimate the standard error of the sample mean difference using our simulation, but if the assumptions for this method are true (see below), then it is converging toward this value.
  - **Pooling Assumption:** If we can assume that each population has approximately the same variance, we can use a "pooled" method to calculate this value. If there is a large discrepancy, we might choose to allow each group to have a different variance.

$$\text{(Pooled) } SE_{(\bar{x}_1 - \bar{x}_2)} \approx s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{(Unpooled) } SE_{(\bar{x}_1 - \bar{x}_2)} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

…Where s_p represents the pooled standard deviation. This, more or less, averages the standard deviation of each group separately, and weights them by their sample sizes.

  - *Notice the approximation symbol, since we are estimating σ_p with s_p. Due to this approximation, we will likely need to use a t-test rather than a z-test.*

Practice: Calculate the standard error for $\bar{x}_1 - \bar{x}_2$. We **won't** assume equal variances.

Our **null model** is approximately normally distributed with…

- a mean of… 0
- and a standard deviation of… SE (calculated above)

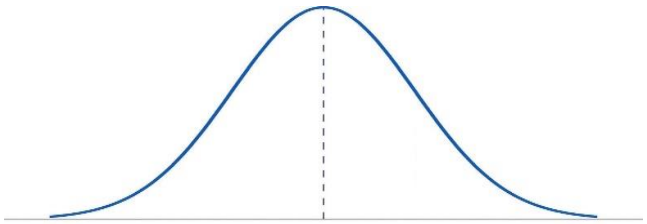
Assumptions for a "pooled" independent samples z or t-test

- ✓ **Parametric assumption:** The distribution of $\bar{x}_1 - \bar{x}_2$ is normally distributed
  - Each population is already approximately normally distributed ***OR*** the Central Limit Theorem would apply (e.g., each sample size > 30 and no long tail/large skew).
  - When not met, we might stick with a permutation method.
- ✓ **Pooled method assumption**
  - Variances of each group are *reasonably* close *(could check with "F-test")*
  - When that's not the case, there is an **unpooled** method! Leave it to software.
- ✓ Do I always need a t-method adjustment?
  - If σ_1 and σ_2 known, or reasonably approximated with large sample sizes (n_1 and $n_2 > 30$ and $n_1 + n_2 > 100$), then both methods should yield about the same result!

- **Test statistic and p-value**
  - Once we have identified our null, we can find a standardized value to identify where our sample result falls on this null model.
  - If sample sizes are not very large, we will need to use an **"independent samples t-test"** to account for standard deviation estimates.
    - Note that an independent samples **z-test** would be reasonable if our **sample sizes were large.** A t-test is a safer option, and even in larger sample cases, a t-test won't be inaccurate. *It's computationally more complex, but easy with software!*
  - Either way, our test statistic will have the same form: How many standard errors wide is our discrepancy from the null hypothesis?

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sim SE_{(\bar{x}_1 - \bar{x}_2)}} \qquad z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

Calculate your test statistic, then let's label it on the t distribution.



Let's use this simulation to get a p-value using this unpooled, independent samples t-test approach.
https://istats.shinyapps.io/2sample_mean/

Making Conclusions. Which statement best describes what we have found from this hypothesis test result?

1. About 99% of drivers using a cell phone have a longer reaction time as compared to those who didn't use a cell phone.

   This is a statement about individuals, but we're testing the position of $\mu_1 - \mu_2$

2. We have strong evidence that the average reaction time of those using cell phones is several seconds higher than those who don't.

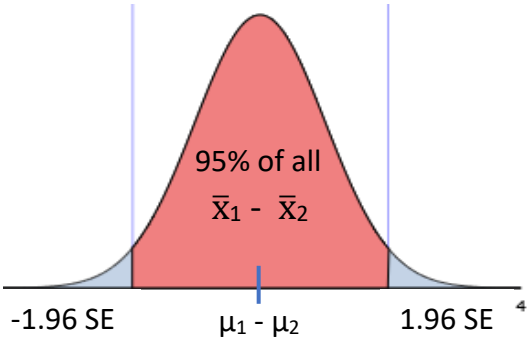
   Not quite. A p-value doesn't tell us how large the difference is.

3. If there is no difference in reaction time on average due to using a cell phone, then we'd see this much difference in our sample result about 1% of the time by random chance.

   This is exactly what a p-value represents!

Confidence Interval for $\mu_1 - \mu_2$

- P-values help us determine how confident we are in *any* departure from the null. However, they alone cannot tell us how large that difference is or whether we should care.
- We can also estimate the parameter $\mu_1 - \mu_2$ using a confidence interval.
- Our point estimate for this parameter is… $\bar{x}_1 - \bar{x}_2$



z-interval: $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} * SE_{(\bar{x}_1 - \bar{x}_2)}$

90%: $z_{0.05} = 1.645$
95%: $z_{0.025} = 1.960$
98%: $z_{0.01} = 2.326$
99%: $z_{0.005} = 2.576$

t-interval: $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} * \sim SE_{(\bar{x}_1 - \bar{x}_2)}$

t depends on confidence *and* degrees of freedom

In our course, t-scores for confidence intervals will always be provided, or we will use software to find it!

Practice: Calculate a 95% t-interval for the true difference in average reaction time between those using cell phones and those who don't while driving. *Use a t-score of 2.003.*

Point Estimate:

Standard Error:

Margin of Error:

Confidence Intervals and p-values

- Remember that confidence levels correspond to significance levels
  - We are 95% confident this interval includes the parameter implies we are 5% confident that this interval misses the parameter.
- If our 95% confidence interval does not include 0, that implies that a hypothesis test with 0 as the null hypothesis would yield a p-value above / below 0.05.
- Why?
  - In our recent 95% confidence interval, our margin of error extends 2.003*SE.
  - When doing a hypothesis test with 0 as the null, our test statistic was more / less than 2.003
  - If we extend to a higher level of confidence to eventually reach to 0 with our interval, then a two-sided test p-value should be the **complement** of that confidence level.

What's the smallest confidence level we can choose that would extend to include 0?

Chapter 7 Additional Practice

Investigation: How does coffee correlate with daily calorie consumption? A dietician collects data using a weight-loss app, where 3,682 regular users grant consent to share their data. These users record their caloric intake for the day, along with specific items they consumed. A dietician recorded whether the participant included "coffee" as one of their items for the day. This dietician then compares the caloric intake of coffee drinkers to non-coffee drinkers.

	Coffee	**No Coffee**
Mean Caloric Intake	$\bar{x}_1 = 1{,}765$	$\bar{x}_2 = 1{,}691$
Standard Deviation	$s_1 = 289$	$s_2 = 313$
Sample Size	$n_1 = 2{,}075$	$n_2 = 1{,}607$
Pooled Standard Deviation	$s_p = 299.6$	

Let's assume this is a representative sample of coffee and non-coffee drinkers who use this app and who are trying to lose weight.

Population (who are we generalizing to?)

Response variable (and type):

Explanatory variable (and type):

What is the Null and Alternative hypothesis in this investigation? *Is this directional or non directional?*

Let's assume that the variance in caloric intake of each population is about the same. What would be the expected error in our sample mean difference as an estimate for the true mean difference?

Since the sample size is large, let's complete a z-test. Calculate the z-score for our sample mean difference within the null model.

Our sample mean is _____ standard errors below / above the null hypothesized mean difference of ___.

The p-value should come out to be quite small…<0.0001. Do we have evidence of a difference in mean caloric intake among coffee and non-coffee drinkers?

Complete the appropriate calculation to fill this in: "We are 98% confident that the true average difference in caloric intake between coffee drinkers and non-coffee drinkers using this app is between _____ and _____ in favor of _____ drinkers being higher."

Investigation: Consider an investigation to determine if there is a difference in mean exam scores among students who are enrolled in a section with an in-person peer tutoring program versus students enrolled in a section with an online peer tutoring program. We obviously can't study every student's experience who might ever take it, but we can compare the 35 students who took each section this semester.

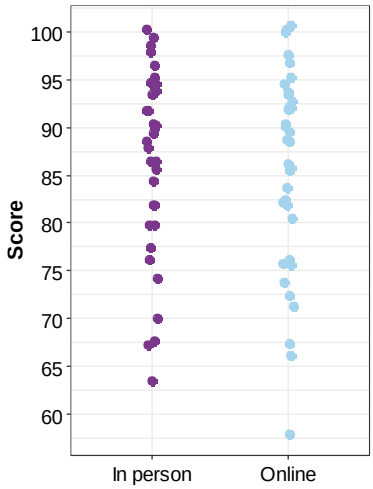
	In person	**Online**
Mean Productivity Score	$\bar{x}_1 = 86.5$	$\bar{x}_2 = 85.5$
Sample Standard Deviations	$s_1 = 9.6$	$s_2 = 10.5$
Pooled Standard Deviation	$s_p = 10.05$	

Population:

Unit of Observation:

Response variable:

Explanatory variable:



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

Calculate a **95%** confidence interval to estimate the true average difference in exam score between each section. *Assume the variances are equal and that the score distributions are not highly skewed. Use* ***t=1.995****.*

Does the interval include 0? Based on this, what would you expect to find if you completed a t-test with 0 as the null hypothesized mean difference—would you expect the p-value to be above 0.05 or below?

Let's say that we extended this study into an additional semester and doubled the sample size for our study to about 70 students per section. Assuming the exam score variability remains around 10.05, how might that affect the width of our confidence interval?

Chapter 7 Learning Goals

After this chapter, you should be able to…

- Identify the null and alternative hypotheses in a two mean comparison context
- Understand random permutations as a way to simulate the distribution of $\bar{x}_1 - \bar{x}_2$, and use this simulated distribution to assess whether it's reasonable that $\mu_1 = \mu_2$
- Recognize a parametric test (z or t-test) as appropriate when the distribution of $\bar{x}_1 - \bar{x}_2$ is likely normally distributed, and a permutation test as reasonable otherwise.
- Calculate (and distinguish) the standard error for $\bar{x}_1 - \bar{x}_2$ using a pooled and unpooled method and calculate an appropriate test statistic (t-score or z-score)
- With the aid of an app or provided p-value, complete an independent samples z or t-test and make an appropriate conclusion
- Recognize $\bar{x}_1 - \bar{x}_2$ as a point estimate for $\mu_1 - \mu_2$
- Complete a z or t-interval for the difference in two means and make an appropriate interpretation.
- Recognize that a 95% confidence interval for $\mu_1 - \mu_2$ that does not include 0 would suggest a strong evidence (p-value < 0.05) that 0 is not the true mean difference, and a 95% interval including 0 would suggest lack of strong evidence (p-value > 0.05)