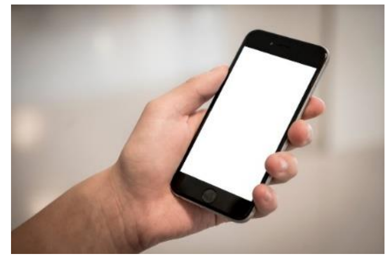
Chapter 7: Comparing Two Means

Investigation: A research study compared the reaction times of automobile drivers with and without cell phones. The goal of the study was to determine whether using a cell phone might *increase* the reaction times of drivers when confronted with a road hazard as compared to standard radio noise.

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.

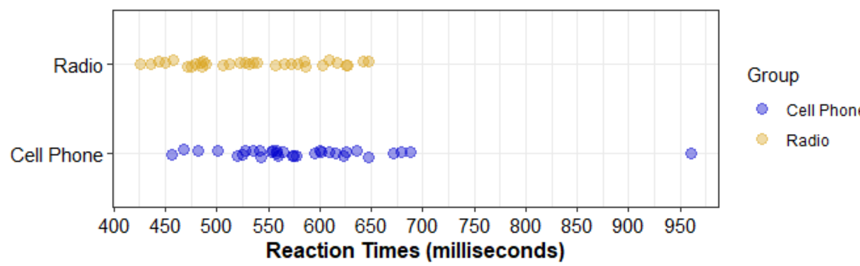


The researchers measured how many milliseconds it took drivers to hit the brake after the road hazard appeared.

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

Table 1. Summary Statistics

	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:

Response Variable (and type):

Explanatory Variable (and type):

- **The Null Hypothesis:** ____________________________.
  - Non-directionally: $\mu_1 = \mu_2$
  - Directionally: *Mirror the alternative*
- **The Alternative Hypothesis:** ____________________________.
  - 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:

- Let's start with a non-parametric approach to testing this. The **Permutation Test!**
  - To explore this testing approach conceptually, let's use this **alpaca simulation** (or google search "jwilber permutation test").
  - **Bottom line:** We can build our null model by taking our response labels and randomly shuffling around the explanatory labels.
  - This will give us a sense of how often we'd see various differences in the sample means by random chance and see how our particular sample mean difference fits in.

Open the Art of Stat Web apps and select **Permutation Test**, then choose "Reaction Times."

Exploring the Distribution of __________ as an estimator for __________

Let's "permutate" the group designations randomly across each observed response value.

Theoretically, what should the mean of our null model be?

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

How often did we observe a permuted sample mean difference at least as high as our actual sample mean difference? *Check out the "Permutation Test" tab up top.*

In a two-sample context, we might interpret our p-value like this:

The probability that one group's mean
would be *at least* this much higher than
the other group's,

if the Null were true,

Is ____%

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

- **Parametric assumption**
  - You might notice that this distribution is *approximately* ____________________________.
  - 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 $\bar{x}_1 - \bar{x}_2$**
  - The standard error for the difference in two sample means is the _____________ difference in our two sample means when assuming $\mu_1 = \mu_2$.
  - If the parametric assumption is true, then our simulation-based approach should be approximating the following calculation for the standard error.
  - **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 ___ 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 symbols; we are estimating σ with s in each formula. Due to this approximation, it would be more accurate to use a _______ 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… and a standard deviation of…

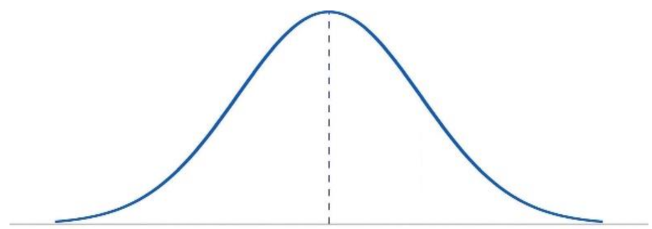
Assumptions for a "pooled" independent samples z or t-test

- ✓ **Parametric assumption:** The distribution of $\bar{x}_1 - \bar{x}_2$ is normally distributed.
  - This is met if each population is already approximately normally distributed ***OR*** the skewness in each population is mild enough for the CLT to apply.
  - When not met, we might stick with a non-parametric test (like a permutation test!)
- ✓ **Pooled method assumption**
  - Variances of each group are *reasonably* close *(we won't cover how to check that)*
  - When that's not the case, there is an **unpooled** method! Safe choice; easy with software
- ✓ Do I always need a t-method adjustment?
  - If σ_1 and σ_2 known (or reasonably approximated with large sample sizes) then a z-test is probably fine. Otherwise, stick with a t-test!

- **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 the t distribution simulator to get a p-value using the *unpooled*, independent samples t-test.

Below is **one sensible** p-value interpretation and **two impostors**! Can you sort out which is which?

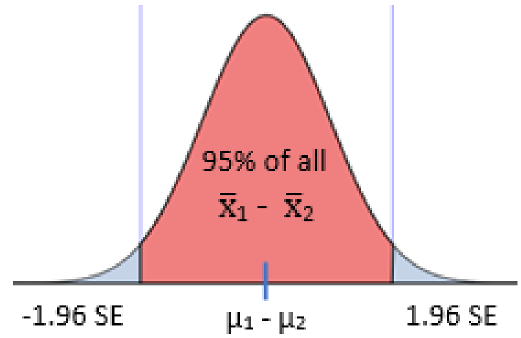
1. The probability of a cell phone user having a reaction time less than the mean reaction time of radio users is about 0.55%.

2. We have strong evidence that the cell phone users have reaction times at least 50 milliseconds longer on average than the radio users.

3. If there truly is no difference in mean reaction time between cell phone users and radio users, then we'd expect to see the cell phone group's sample mean this much higher about 0.55% of the time.

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…



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:

Margin of Error:

Interval bounds:

Confidence Intervals and p-values

- Remember that confidence levels correspond to significance levels.
  - If there is a 95% probability that this interval includes the parameter, then there is a 5% probability 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.

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

Chapter 7 Reflection Questions

When completing a test to compare two means, the response variable will be (categorical or numeric?) and the explanatory variable is (categorical or numeric?).

In a permutation test, why would we shuffle group labels randomly to create a null model?

In your own words, describe how we would get a p-value from a permutation test. Perhaps follow along with the **alpaca simulation** (follow link, or google search "jwilber permutation test").

If $\mu_1 = \mu_2$, the distribution of $\bar{x}_1 - \bar{x}_2$ should have a mean of what value?

Independent samples t-tests and z-tests are parametric tests. What is the parametric assumption we need to be true for these tests to be valid?

In words, how might we describe what the standard error for $\bar{x}_1 - \bar{x}_2$ represents?

What should be true in order for a *pooled* method to be appropriate when completing an independent samples t-test?

When completing a confidence interval for $\mu_1 - \mu_2$, our point estimate would be what?

If a 98% t-interval for $\mu_1 - \mu_2$ does **not** include 0, then an independent samples t-test with 0 as the null hypothesis should yield a p-value less than what?

Chapter 7 Additional Practice (if needed!)

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.

Table 2. Summary statistics table

	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.

Table 3. Summary Statistics Table

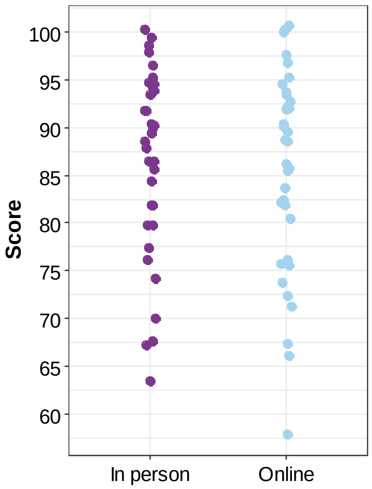
	In person	Online
Mean Productivity Score	$\bar{x}_1 = 86.5$	$\bar{x}_2 = 85.5$
Sample Standard Deviations	$s_1 = 9.6$	$s_2 = 10.5$
Pooled Standard Deviation	$s_p = 10.05$	

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?