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 *change* the reaction times of drivers when confronted with a road hazard as compared to standard radio noise.

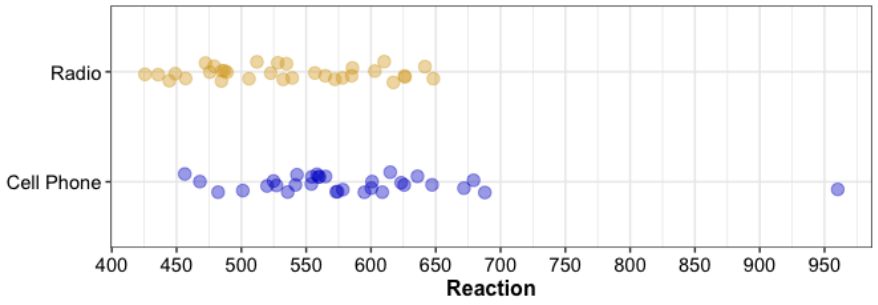
In a study with 64 people, the researchers randomly assigned 32 people to operate a simulated vehicle while holding their cell phone and having a conversation. The other 32 were randomly assigned to do the same thing, but while listening to the radio or an audio book.

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	**Radio**
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):

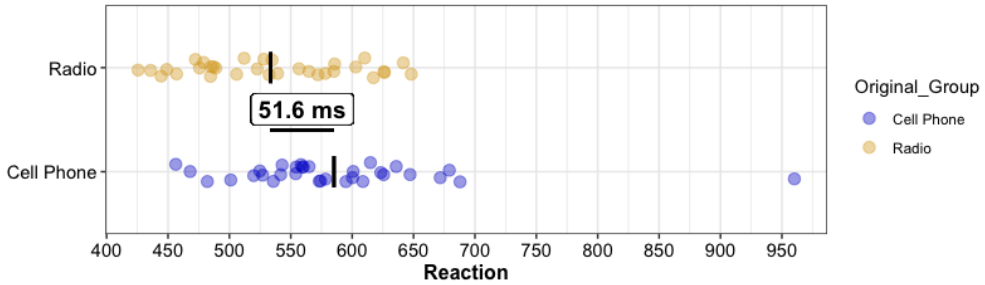
- **The Null Hypothesis:** _____.
  - Non-directionally: $\mu_1 = \mu_2$
  - Directionally: *Mirror the alternative*
- **The Alternative Hypothesis:** _____.
  - Non-directionally: $\mu_1 \neq \mu_2$
  - Directionally: $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Identify the null and alternative hypotheses for this investigation:

Let's start with a *non-parametric* approach to testing this: The **Permutation Test!** We'll be using the **Permutation Test applet** from the Art of Stat website and specifically looking at the "Reaction Times" dataset.

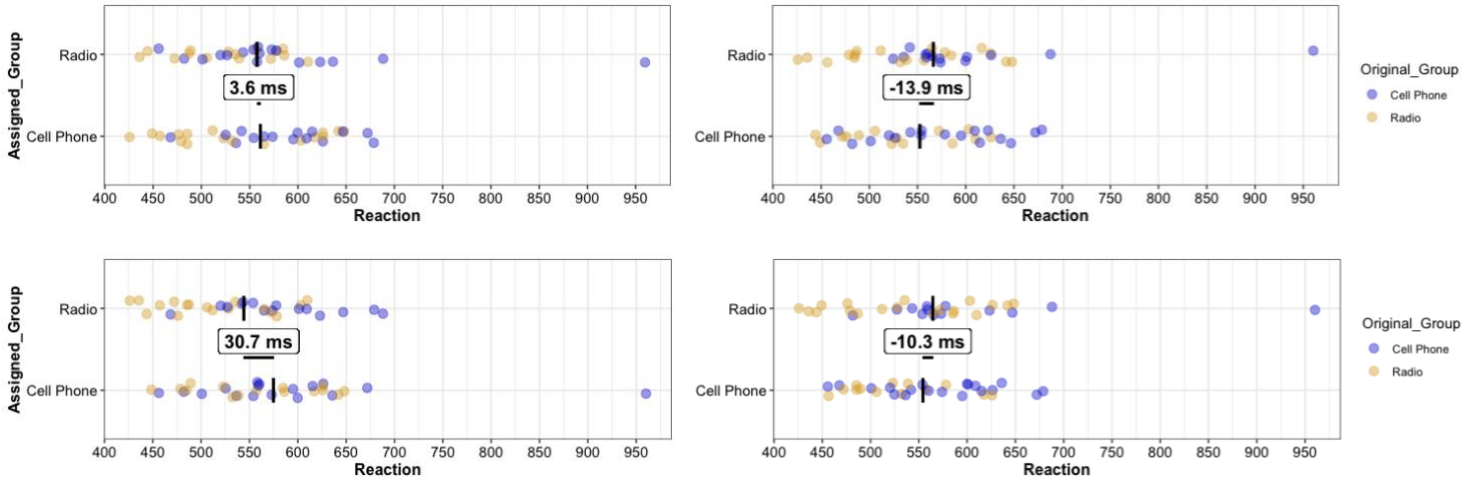
The permutation test is a simulation-based approach that takes the following process:

1. Choose an **estimator** as our measure of difference. For numeric response data, it's common to calculate the difference in our _____ _____ as an estimator for $\mu_1 - \mu_2$.



2. **Assume the Null Hypothesis** is true and $\mu_1 = \mu_2$. If that's the case, then the _____ factor is arbitrary and the difference shown above is just a result of random chance!

3. To determine how plausible it would be to get a difference at least this large by chance, we can **shuffle the reaction times** randomly to each group *many* many times. Then we can keep track of the sample mean differences that do truly happen by random chance.



4. **Create a "Permutation sampling distribution"** to represent the distribution of $\bar{x}_1 - \bar{x}_2$ under the Null.

5. **Calculate our p-value.** This will be the number of simulations that produced sample mean differences at least as large as ours by random chance.

In a two-sample context, we might ***interpret*** our p-value (non-directionally) like this:

The probability of observing a sample
mean difference at least this large

if the Null is true,

Is ____%

Reflection Questions

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

7.2. In a permutation test, how does the null hypothesis relate to the idea of shuffling the group labels around randomly?

7.3. In your own words, describe how we would get a p-value from a permutation test. For an interactive example, try the **alpaca simulation** (follow link, or google search "jwilber permutation test").

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

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* try a parametric test to do inference rather than estimate the p-value with simulations.
- **Calculating the Standard Error for $\bar{x}_1 - \bar{x}_2$**
  - In our permutation simulation, our distribution of possible sample mean differences had a standard deviation of around 20.6.
  - This is approximating the standard error for the difference in our sample means—in other words, it is how much error we should expect in $\bar{x}_1 - \bar{x}_2$ as an estimator for _____.
  - *If* the parametric assumption is true, then the standard error for our sample mean difference should be converging toward one of the formulas below:

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

- **The Pooled Method**
  - The pooled standard error calculation assumes that the two populations we're sampling from have equal standard deviations. If that assumption is true, then it is slightly more efficient to take a weighted average of the sample standard deviations and calculate this way!
  - S_p in that formula represents the *pooled standard deviation*.
- **The Unpooled Method** (sometimes referred to as "**Welch's 2-sample t-test**")
  - In many cases, our populations won't have equal standard deviations (or we may not be able to safely assume that). In which case, the unpooled calculation is the safer choice.

Practice: Calculate the standard error for $\bar{x}_1 - \bar{x}_2$ if **not** assuming equal variation in each population.

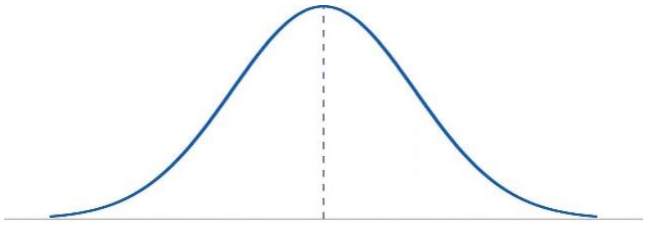
Assumptions for an independent samples z or t-test

- ✓ **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 vs. Unpooled method**
  - A pooled method *could* be used if you believe your populations have equal variance.
  - An unpooled method is not as efficient a test, but it is valid in either case.
  - In applied practice, you should typically default to using an **unpooled** method!
- ✓ When do we ***need*** a t-method adjustment?
  - If σ_1 and σ_2 known (or reasonably approximated with large sample sizes, such as when each group has 100 or more observations) then we *could* use a z-test. But as with one-sample testing, a t-test is still valid in large-sample settings too. In practice when using software, default to a t-test!

- **Test statistic and p-value**
  - 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 represents the same thing as in one-sample testing: How many _____ _____ wide is our estimate from the null hypothesized parameter?

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

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



Identifying the degrees of freedom for a pooled t-test is rather straightforward: $df = n_1 + n_2 - 2$. However, the degrees of freedom calculation for an unpooled method is complex and typically requires software. If using software for the degrees of freedom calculation, we'd get a two-tailed p-value of around **1.1%**

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

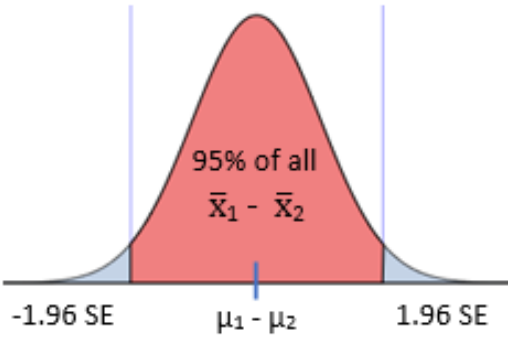
1. The probability of a randomly chosen cell phone user having a reaction time less than a randomly chosen radio user is about 1.1%.

2. The probability of cell phone users and radio users having the same mean reaction time is about 1.1%.

3. If there truly is no difference in mean reaction time between cell phone users and radio users, then we'd expect to see a difference in sample means at least this big about 1.1% 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: $\bar{x}_1 - \bar{x}_2 \pm z_{C\%} * SE_{(\bar{x}_1 - \bar{x}_2)}$

$z_{90\%} = 1.645$

$z_{95\%} = 1.960$

$z_{98\%} = 2.326$

$z_{99\%} = 2.576$

t-interval: $\bar{x}_1 - \bar{x}_2 \pm t_{C\%,df} * \sim SE_{(\bar{x}_1 - \bar{x}_2)}$

t depends on confidence *and* degrees of freedom.

In our course, t-scores for confidence intervals will always be provided as they would if using software!

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 the interval we create will include the parameter, then there is a ____ probability that this interval we create will miss 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: Given our p-value was 1.1%, what would be the minimum confidence level we'd need for our interval to just reach 0?

Reflection Questions

7.5. 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?

7.6. Describe what the standard error for $\bar{x}_1 - \bar{x}_2$ represents.

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

7.8. When completing a confidence interval for $\mu_1 - \mu_2$, what would we use as a point estimate?

7.9. 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 (Videos available in the Ch 7 module on Canvas!)

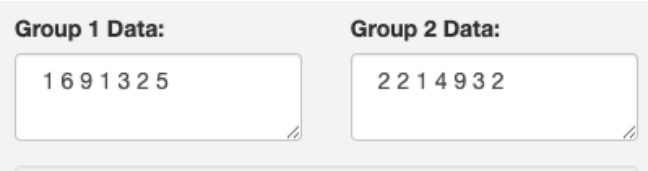
Investigation: Mario Kart 8 online allows people to compete with other players around the world in 12-person races. The youtuber "Shortcat" created a video that asked: "Which strategy is better: attack or defense?" In other words, he wanted to know whether throwing your items to attack racers vs. holding your items to defend against other racers might be a better strategy. After completing 7 races taking each strategy, he reported the following result, coming to the conclusion that defending is better than attacking. Did Shortcat collect enough data and find a large enough mean difference to declare that confidently? Let's test it!

Table 2. Summary statistics table

	Attack	**Defense**
Mean Race Placement	$\bar{x}_1 = 3.9$	$\bar{x}_2 = 3.3$
Standard Deviation	$s_1 = 2.968$	$s_2 = 2.690$
Sample Size	$n_1 = 7$	$n_2 = 7$
Pooled Standard Deviation	$s_p = 2.829$	

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

Let's proceed with a permutation test. Using the Permutation Test applet from the art of stat web apps page, we can select "provide own" for data, label our two groups, and enter our data as shown below. Generate the results and then find the appropriate p-value.



What does your p-value communicate? Would you say that Shortcat provided statistical evidence that one strategy is better than the other?

Would this be a situation where a parametric test, like an independent samples t-test, would be appropriate? Why or why not?

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

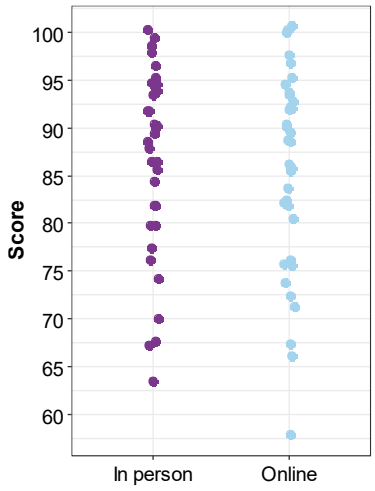
	In person	**Online**
Mean Productivity Score	$\bar{x}_1 = 86.5$	$\bar{x}_2 = 85.5$
Sample Standard Deviations	$s_1 = 9.6$	$s_2 = 10.5$
Pooled Standard Deviation	$s_p = 10.05$	

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?