Activity #17: Two-sample mean comparisons

In the previous activity, we learned how to conduct a hypothesis test to compare a sample mean to a hypothesized value. In this activity, we'll compare the means from two samples (μ_1 versus μ_2).

Recall the logic behind hypothesis tests:

- Assume the hypothesized value for your population parameter is true
- Estimate the likelihood of your data if that hypothesized population parameter were true

Before we begin, we need to take a trip back to unit #1 in this course and recall a few facts:

1. What does it mean if two events (A and B) are independent?

2. What is an expected value? If you have two independent variables, X and Y, what is E[X – Y]?

3. Define the variance of a random variable. If you have two independent variables, X and Y, what is var[X – Y]?

4. As stated before, we're going to learn how to compare the means of two independent groups to see if those means differ by a statistically significant amount. A few (arbitrarily chosen) examples of this would be:

   Comparing the average starting salary of SAU graduates to the average for graduates of a competing school
   Comparing the average blood pressure of those who take an experimental drug to individuals who take a placebo
   Comparing the average number of chocolate chips in Chips Ahoy© cookies to a generic brand

   In each of these examples, we would have the following:

Group #1	**Group #2**
Unknown population parameter of interest = μ_1	Unknown population parameter of interest = μ_2
(Possibly) unknown population standard deviation = σ_1	(Possibly) unknown population standard deviation = σ_2
Number of subjects = n_1	Number of subjects = n_2
Observed data = $X_{11}, X_{12}, X_{13}, \ldots X_{1n}$	Observed data = $X_{21}, X_{22}, X_{23}, \ldots X_{2n}$

   Our goal would be to compare the population means. Write out the null and alternative hypotheses we would use to compare two population means. Since we only know procedures to compare a single population parameter to a hypothesized value, rewrite these hypotheses to match what we know how to do:

5. We don't know the population means, so we'll never be able to directly calculate $\mu_1 - \mu_2$. Instead, we'll need to find an estimate for this unknown parameter. What could we calculate from our sample data to estimate $\mu_1 - \mu_2$?

6. If we want to use that estimator, we'll need to know its sampling distribution. In other words, if we could repeatedly take samples of size n_1 and n_2 from our two populations and calculate our estimator for each sample, what would the distribution of all those estimates look like? Where would that distribution be centered? What would the standard error of that distribution be? Let's derive these characteristics of our sampling distribution of interest:

Sampling distribution of $\left(\bar{X}_1 - \bar{X}_2\right)$:

A) Expected value:

$E[X_1 - X_2] = E[X_1] - E[X_2] = \mu_1 - \mu_2 = 0$ (assuming our null hypothesis is true)

B) Standard error:

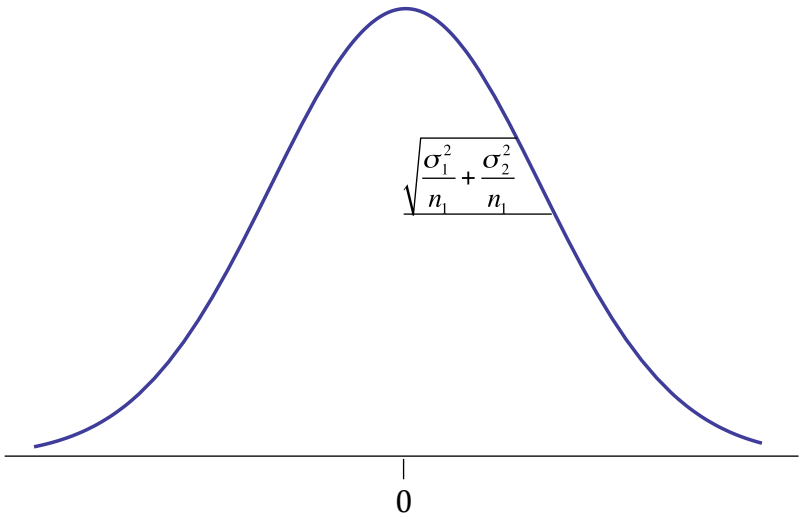
The standard error of the distribution of sample means is $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$, so the variance would be $\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$

We know: $\text{var}\left[\bar{X}_1 - \bar{X}_2\right] = \text{var}\left[\bar{X}_1\right] + \text{var}\left[\bar{X}_2\right] = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma^2_{X_1}}{n_1} + \frac{\sigma^2_{X_1}}{n_1} = \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_1}$

Therefore, the standard error would be: $SD\left[\bar{X}_1 - \bar{X}_2\right] = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_1}}$

C) If we know the population standard deviations, σ_1 and σ_2, we could calculate z-scores from the sampling distr.:

$$Z = \frac{\text{(observed)} - \text{(hypothesized)}}{\text{(standard error)}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_1}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_1}}}$$



7. The previous sampling distribution only applies if we know the population standard deviations, σ_1 and σ_2. If we don't know those population standard deviations, logic would dictate that we substitute their sample estimates, s_1 and s_2, and use the t-distribution:

$$\text{Logical Idea} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}} = \text{t-distribution}$$

Unfortunately, that won't work. We can only use the t-distribution if we replace **one** population variance with a sample variance. We replaced **two** parameters with their estimates. If we can't replace both variances with their estimates and use the t-distribution, what can we do?

Option A: If we can safely assume the two groups have equal population variances ($\sigma_1^2 = \sigma_2^2$), then we can substitute a single estimate (s^2) for that single parameter (σ^2). If we do this, then we will have a t-distribution.

Remember, we'll never know the population variances. How could we possibly know if they're equal if we don't know their true values?

In this class, we'll use the eyeball method. If the variances we calculate for our samples look approximately equal ($s^2 \approx s^2$), then we'll work under the assumption that the population variances are equal ($\sigma_1 = \sigma_2$). If you take MATH 301, you'll learn a few more sophisticated (and defensible) methods for testing if variances are equal.

Suppose we sample data from two independent groups and calculate s_1=8 and s_2=10. These values look fairly close to one another. Would we assume the population variances are equal? Note that the variances for our samples would be 64 and 100.

Even if we assume, in this example, that the population variances are equal, what value would we use for that population variance? If the two groups had equal sample sizes, it might make sense to assume the population standard deviation is 9. If the groups differ in sample sizes, then we should use a weighted average:

If $s_1^2 = \frac{\sum\left(X_{i1} - \bar{X}_1\right)^2}{n_1 - 1}$ and $s_2^2 = \frac{\sum\left(X_{i2} - \bar{X}_2\right)^2}{n_2 - 1}$, then the weighted average would be:

$$s_{\text{pooled}}^2 = \frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}$$

Therefore, if we assume population variances are equal, the standard deviation would be:

$$s_{\text{pooled}} = \sqrt{\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}}$$

To convert that standard deviation to a standard error, we need to do something similar to dividing by the square root of our sample size:

$$SE_{\text{pooled}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}}$$

We can now calculate our test statistic:

$$t_{n_1+n_2-2} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{\left(n_1-1\right)s_1^2 + \left(n_2-1\right)s_2^2}{\left(n_1-1\right)+\left(n_2-1\right)}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{\left(n_1-1\right)s_1^2 + \left(n_2-1\right)s_2^2}{\left(n_1-1\right)+\left(n_2-1\right)}}}$$

We can also calculate a confidence interval:

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm \left(t_{n_1+n_2-2}\right)\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{\left(n_1-1\right)s_1^2 + \left(n_2-1\right)s_2^2}{\left(n_1-1\right)+\left(n_2-1\right)}}\right)$$

Option B: If we cannot reasonably assume the two groups have equal population variances ($\sigma_1^2 \neq \sigma_2^2$), then we cannot use Option A. If we really want to use the t-distribution, we could use the Welch-Satterthwaite Method. In this method, we need to modify the degrees of freedom of our t-statistic, as follows:

$$df^* = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2/n_1}{n_1-1}\right)+\left(\frac{s_2^2/n_2}{n_2-1}\right)}$$

I'm going to assume you could use that formula if I forced you to. We won't ever calculate this by hand in-class. If you wanted to use this method (or if you wanted to conduct any hypothesis testing procedures), I'd recommend using a computer.

Option C: If you cannot assume equal population variances, you could also try another hypothesis test method. For example, you could run a nonparametric test, like the Mann-Whitney U (a.k.a. Wilcoxon rank-sum test), a randomization-based test (like a permutation test), or use bootstrap methods.

We've discussed randomization-based tests throughout this semester (including running a permutation test the first day of class). We've seen bootstrap methods when we were constructing confidence intervals. If you take MATH 301, we'll introduce ourselves with additional nonparametric tests.

8. When we were deriving the mean and standard error of the sampling distribution for μ1 – μ2, what assumptions did we make?