Discussion 9 Review

1. The Two-sample T-Test (Normal with Equal Variances)

   The data consists of separate simple random samples from two different populations, label them 1 and 2. Let:

   μ_1 = true mean of population 1
   μ_2 = true mean of population 2
   n_1 = sample size taken from population 1
   n_2 = sample size taken from population 2
   σ_1^2 = true variance of population 1
   σ_2^2 = true variance of population 2

   We wish to test:

   $H_0 : \mu_1 - \mu_2 = \delta$
   vs.
   $H_A : \mu_1 - \mu_2 \neq \delta$

   Good numerical and graphical summaries to explore the data might include means, medians, standard deviations, side-by-side boxplots, side-by-side dotplots, stacked histograms, and normal quantile plots, among others.

   If based on our prior knowledge and after exploring the data we are willing to assume:

   - All of the data points are independent, both within and between populations (this will be true if the samples are simple random samples)
   - The two populations each follow normal distributions
   - The variances of the two populations are equal so that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

   Then the test statistic is:

   $$t = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

   Where:

   $$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

   Compare to a t distribution on $\nu = n_1 + n_2 - 2$ degrees of freedom.
   In this situation, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

   $$\bar{x}_1 - \bar{x}_2 \pm t_{\nu,\alpha/2} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

   Where $\nu = n_1 + n_2 - 2$.

2. The Welch T-Test (Normal with Unequal Variances)

The data consists of separate simple random samples from two different populations, label them 1 and 2. Let:

μ_1 = true mean of population 1
μ_2 = true mean of population 2
n_1 = sample size taken from population 1
n_2 = sample size taken from population 2
σ_1^2 = true variance of population 1
σ_2^2 = true variance of population 2

We wish to test:

$H_0 : \mu_1 - \mu_2 = \delta$
vs.
$H_A : \mu_1 - \mu_2 \neq \delta$

Good numerical and graphical summaries to explore the data might include means, medians, standard deviations, side-by-side boxplots, side-by-side dotplots, stacked histograms, and normal quantile plots, among others.

If based on our prior knowledge and after exploring the data we are willing to assume:

- All of the data points are independent, both within and between populations (this will be true if the samples are simple random samples)

- The two populations each follow normal distributions

- The variances of the two populations are not equal

Then the test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Compare to a t distribution on ν degrees of freedom, where:

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

In this situation, $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where ν is given above.

When deciding between a variance equal and variance unequal test, remember that:

- A good guideline when the sample sizes are similar in the two groups is that if $0.5 \leq \frac{s_1}{s_2} \leq 2.0$, we can feel fairly comfortable assuming the sigmas equal. The more our samples sizes differ, the more strict we need to be in having the sigmas similar

- If the variances are truly equal, but are allowed to differ, the test loses some power, but is still a good test.

- If the variances are truly different, but they are assumed equal, the test can make wildly incorrect conclusions.

3. Two-Sample Bootstrap Test

When one or both of the populations do not follow normal distributions, and n is too small to use the CLT, you can use the bootstrap to test:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$\text{vs } H_A : \mu_1 - \mu_2 \neq 0 \text{ or } H_A : \mu_1 - \mu_2 \leq 0 \text{ or } H_A : \mu_1 - \mu_2 \geq 0$$

(1) Draw a simple random sample $x_{1,1}$, $x_{1,2}$, ..., x_{1,n_1} of size n_1 from the first population and compute $\bar{x}_1$ and s_1^2. Draw a simple random sample $x_{2,1}$, $x_{2,2}$, ..., x_{2,n_2} of size n_2 from the second population and compute $\bar{x}_2$ and s_2^2. Let

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(2) Draw a simple random sample, with replacement, of size n_1, from the first sample data. Call these observations $x_{1,1}^*$, $x_{1,2}^*$, ..., x_{1,n_1}^*. Draw a simple random sample, with replacement, of size n_2, from the second sample data. Call these observations $x_{2,1}^*$, $x_{2,2}^*$, ..., x_{2,n_2}^*.

(3) Compute the means and variances of the resampled data for each group separately. Call these things $\bar{x}_1^*$ and s_1^{2*}, and $\bar{x}_2^*$ and s_2^{2*}, respectively.

(4) Compute the statistic:

$$\hat{t} = \frac{(\bar{x}_1^* - \bar{x}_2^*) - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^{2*}}{n_1} + \frac{s_2^{2*}}{n_2}}}$$

(5) Repeat steps 2-4 B times, where B is a large number, and compute $\hat{t}$ from each one. This is an approximation to the sampling distribution of $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$.

(6) For a two-sided alternative, Let mlo be the number of bootstrap samples where $\hat{t} \leq t_{obs}$, and let mup be the number of bootstrap samples where $\hat{t} \geq t_{obs}$. The bootstrap p-value is $2 * min(mlo, mup)/B$. For a one-sided less-than alternative, the p-value is mlo/B. For a one-sided greater-than alternative, the p-value is mup/B.