Inferences Involving Two Populations

When comparing two populations we will need two samples, one from each populations. Two basic kinds of samples can be used, dependent and independent samples. The dependance or independence of the data depends on the source of the data.

The **source** of the data is a person, object or anything else that yields a piece of data.

If the same set of sources or related sets are used to obtain the data representing both populations, we have **dependent samples**.

If two (unrelated) sets of sources are used, one set from each population, we have **independent samples**.

Let's look at some specific examples:

Randomly select 50 participants from the list of those enrolled in BZS at Dawson and give them a pretest. At the end of the course randomly select 50 participants and give them a posttest.

Randomly select 50 participants from the list of those enrolled in BZS at Dawson and give them a pretest. Give the same set of 50 students the posttest when they complete the course.

In before versus after studies, a dependent sample is usually used.

Let's try it again:

A sample of cars will be selected randomly, equipped with brand A tires, and driven for a month. Another sample of cars will be selected, equipped with brand B tires and driven for a month. The wear on the tires will be compared.

A sample of cars will be selected randomly, equipped with one brand A tire and one brand B tire and driven for a month (the other two tires are not part of the test). The wear on the tires will be compared.

This is, of course, ignoring other factors such as age, weight and mechanical condition of the car, driving habits of the drivers, location of the dire on the car and how much the car is driven.

Inferences Concerning the Mean Difference Using Two Dependent Samples

In both the dependent and independent case we will want to draw conclusions about the difference between the means of the two populations, $\mu_1 - \mu_2$.

In the dependent case, however, the data from the two populations comes from the same sources so we have the same sample size in both samples and we can pair the data points that come from the same source. Say, x_1 is a data point from the first population and x_2 is a data point from the second population from the **same source**.

To determined whether or not this student improved we can look at the random variable $d = x_1 - x_2$ called the **paired difference**.

To determine if the class improved, we might want to look at the mean of the paired differences (the mean of the improvements) $\mu_d = \overline{x_1 - x_2}$.

But notice that

Written in words, the difference between population means is the same as the mean of paired difference for dependent samples.

The good news is that if the populations are normally distributed so is the population for paired differences, d.

Information for the population of mean differences

- $\overline{d}$ is the mean of the sample differences

(it is also the difference in sample means)

- $\mu_{\overline{d}} = \mu_d$ (unbiased statistic)

- s_d is the standard deviation of the sample differences

- To find a confidence interval estimate for μ_d we use

(don't forget that $t(\alpha/2)$ depends on the degrees of freedom $df = n - 1$)

- To do a hypothesis test about μ_d we use the test statistic (t-value)

(again, when using the t-tables don't forget about $df = n - 1$)

Example: A sample of six cars was selected randomly, equipped with one brand A tire and one brand B tire and driven for a month (the other two tires are not part of the test). The wear on the tires (in thousandths per inch) is given below.

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	56	103	37	102	115

Construct a 95% C.I.E. for the mean difference in the paired data.

Example: Twenty-six patients were randomly selected from a large pool of potential subjects, and their pulse rates were recorded. A calcium channel blocker was administered to each patient for a fixed period of time, and then each patient's pulse rate was again determined. The tow resulting data sets appear to have normal distributions and it was found that $\overline{d} = 1.07$ and $s_d = 1.74$. Does the sample provide sufficient evidence to show that the calcium channel blocker lowered the pulse rate with 5% significance?

Solution: (classical approach)

Solution: (p-value approach)

Notice that although we are able to make hypothesis about μ_d with respect to any value, we will most often make hypothesis about whether there is any difference between these two means (in other words with respect to 0). This is useful if we want to check if there was any change to a population mean after an experiment.