Difference of Two-Means Test

Before we considered *inferences about a single sample mean*. Normally, however, we are interested in relationships. and want to compare two or more sample means between or across groups. This would allow us to explore questions like is the average amount of GNP per capita allocated by governments for military expenditures significantly greater in the richest countries compared to the poorest? Or, is the crime rate significantly greater in the largest cities compared to the smallest?

These questions require that we calculate two means and compare them to see if one is greater than the other, and by how much. To do this we have to return to a theorem derived from the central limits theorem:

> If *independent random samples* of sizes N_1 and N_2, respectively, are drawn from populations that are $Nor(m_1, s_1)$ and $Nor(m_2, s_2)$, respectively, then the sampling distribution of the difference between the two sample means $\left(\bar{X}_1 - \bar{X}_2\right)$ will be $Nor\left(m_1 - m_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)$.

That is to say: $m_{\left(\bar{X}_1 - \bar{X}_2\right)} = m_1 - m_2$

$$s_{\left(\bar{X}_1 - \bar{X}_2\right)} = \sqrt{\frac{s}{n_1} + \frac{s}{n_2}}$$

While a random selection procedure insures that each case sampled is independent of others within a sample, the selection procedures of one sample cannot influence the selection of the other sample.

So how do we proceed? There are two models that can be employed. The first assumes that the two population variances are equal ($\sigma_1=\sigma_2$). The second assumes that they are unequal ($\sigma_1\neq\sigma_2$).

Model A: ($\sigma_1=\sigma_2$)

A *t*-score is calculated in general as follows:

$$t = \frac{X - m_X}{s_X}$$

In practical terms:

$$t_{(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

The denominator, assuming equal variances, can be rewritten as:

$$s_{\left(\bar{X}_1-\bar{X}_2\right)}=\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}=s\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}=s\sqrt{\frac{N_1+N_2}{N_1N_2}}$$

However, we do not normally know the population variances, but we can create an estimate of the common variance, called a *pooled estimate*, by taking a weighted average of the sample variances divided by the degrees of freedom. This estimate is:

$$\hat{s}=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}$$

Putting the whole thing together we get:

$$\hat{s}_{(\bar{X}_1-\bar{X}_2)}=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1n_2}}$$

And finally, the whole *t*-score formula is equal to:

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1n_2}}}$$

An Example

Since we do not know the population variances, we have to use a *t*-test. We are assuming equal population variances, but this can be tested as we shall see later. For the time being, we will just assume this to be the case. We will use a significance level of 0.05 and a one-tailed test. Let's say we were given the following proposition:

P_1: The Higher the educational achievement of a man's father, the higher his educational level.

Using the 1984 GSS data we can test this. First, we can take father's education and dichotomize it. The two groups are:

Group #1: Father has some college education or more
Group #2: Father has a high school degree or less

The dependent variable is the number of years of education that each father's son possesses. From this we can derive two hypotheses, the null and alternative hypotheses:

$$H_0: \mu_1-\mu_2=0$$
$$H_1: \mu_1-\mu_2>0$$

The following information is given:

	Mean	Variance	*n*
Group #1	15.24	6.57	90
Group #2	12.57	9.82	359

Plugging this into the formula, we get:

$$t = \frac{15.24 - 12.57}{\sqrt{\frac{(90)(6.57)+(359)(9.82)}{90+359-2}}\sqrt{\frac{90+359}{(90)(359)}}} = \frac{2.67}{0.36} = 7.42$$

The critical value of *t* with *df* equal to 90+359-2 for alpha equal to 0.05 is 1.645. Clearly, we can reject the hypothesis of no difference.

Model B: ($\sigma_1 \neq \sigma_2$)

If we cannot assume that the two populations have equal standard deviations, we have to modify our procedure slightly. In this case, we cannot pool our variances from two samples together; they must be estimated separately.

The formula for the standard error is:

$$\hat{s}_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

Using the numbers from the example above, we get:

$$t = \frac{15.24 - 12.57}{\sqrt{\frac{6.57}{90-1} + \frac{9.82}{359-1}}}$$

$$= \frac{2.67}{0.3181987} = 8.39$$

The *df* for the unequal variance model is equal to:

$$df = \frac{\left(\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}\right)^2}{\left(\frac{s_1^2}{N_1 - 1}\right)^2\left(\frac{1}{N_1 + 1}\right) + \left(\frac{s_2^2}{N_2 - 1}\right)^2\left(\frac{1}{N_2 + 1}\right)} - 2$$

$$= 163.42$$

The critical value of *t* with *df* equal to 163.42 for alpha equal to 0.05 is 1.645. Clearly, we can reject the hypothesis of no difference.

Testing for Equality of Variances

So, now that we know that there are two ways to calculate the difference of means test, one for equal variances, one for unequal variances, how do we determine which one to use? To determine whether the variances are equal or not, we will take the ratio of one sample variance to the other. It turns out that this ratio has a sampling distribution with known characteristics, and is called the *F distribution*. As this ratio departs from unity, we assume, with a known probability of error, that the variances are unequal. That is to say, we reject the null hypothesis of equal variances.

The definition of the *F* distribution is:

> A random variable formed by the ratio of two independent χ^2-square distributed variables, each divided by its degrees of freedom, is said to be an *F*-ratio, and to be distributed according to the *F* probability density function.

How do you use the *F* table? Two parameters are necessary for the *F* distribution, υ_1 and υ_2 (pronounced nu). These are equal to the degrees of freedom associated with each sample where *df* is equal to *n*-1. *F* tables are normally very condensed since there are so many possible values. Furthermore, most tables only present information for a one-tailed test.

Suppose that we have two samples, one with *n*=13, and the other with *n*=20. These are associated, respectively, with s_1 and s_2. The respective degrees of freedom are 12 and 19. Assuming α=0.05, we look across the top of the *F*-table to the column labeled 12. Then we look down the rows to the one labeled 19. The critical value that the *F* ratio must exceed to reject the null hypothesis is 2.31 --the intersection of the appropriate row and column.

So, how does all this work? Take the ratio of the larger variance to the lesser variance, given as:

$$F_{h_1;h_2} = \frac{s_1^2}{s_2^2}$$

And the null and alternative hypotheses are:

$$H_0 : s_1^2 = s_2^2$$
$$H_1 : s_1^2 > s_2^2$$

In the earlier example we saw that group two had a variance equal to 9.82 and group one had a variance of 6.57 with 358 and 89 *df*, respectively. The closest we can get to these *df* in the *F*-table for α=0.05 is infinity and 60. I choose 60 as a more conservative test than 120. The critical value, then, is 1.39. The *F* ratio is equal to:

So, we would reject the null hypothesis and be forced to use the formula for unequal variances (i.e. we cannot pool the variances).