Comparing Two Populations

OPRE 6301

Introduction ...

In many applications, we are interested in hypotheses concerning differences between the means of two populations. For example, we may wish to decide on the basis of suitable samples whether men can perform a certain task as fast as women, or we may want to decide on the basis of a survey whether the average weekly food expenditures of families in one city exceeds that of families in another city.

We will summarize several basic tests that can be used in such scenarios. These are:

- Comparison of two means with two independent samples
- Comparison of two means with paired samples
- Comparison of two proportions with two independent samples

Two Means, Independent Samples ...

Suppose two **independent** samples of sizes n_1 and n_2 are taken from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Let $\bar{X}_1$ and $\bar{X}_2$ be the respective sample means. Then, the central limit theorem tells us that for large n_1 and n_2, $\bar{X}_1 - \bar{X}_2$ is approximately normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

If the population variances are *known*, this immediately suggests that we can use the statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{1}$$

to test the hypotheses pair

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 \text{ (or } \mu_1 - \mu_2 = 0) \\ H_1 : &\quad \mu_1 \neq \mu_2 \text{ (or } \mu_1 - \mu_2 \neq 0) \end{aligned}$$

Such a test, however, is hardly useful because in most applications the population variances are *unknown*.

As in the case of a single population, the natural approach is to replace the unknown variances by their sample estimates s_1^2 and s_2^2. Doing this results in the following statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \tag{2}$$

There are two cases to consider when working with the above statistic:

— The unknown population variances are *equal*.

— The unknown population variances are *not equal*.

Equal Variances

The idea here is that if the two variances are the same, we can pool data for both samples together to produce a **pooled** variance estimate, defined as:

$$s_p^2 \equiv \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{3}$$

Note that the ratios $(n_1 - 1)/(n_1 + n_2 - 2)$ and $(n_2 - 1)/(n_1 + n_2 - 2)$ can be viewed as weights assigned to s_1^2 and s_2^2, since they sum up to 1.

With s_p^2 replacing both s_1^2 and s_2^2 in (2), we then have

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{4}$$

It can be shown that if both populations are normally distributed, then the statistic in (4) follows the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Our test procedure is therefore as follows.

— Compute from two independent samples the t statistic in (4).

— Reject H_0 if $|t| > t_{\alpha/2}$ with $\nu = n_1 + n_2 - 2$.

The Excel Data Analysis tool

"t-Test: Two-Sample Assuming Equal Variances"

can be used to conduct this test.

Unequal Variances

When the two population variances are not the same, the statistic in (2) is neither normally nor t distributed. (In fact, the exact sampling distribution in this case has not yet been found!) However, it has been shown that the sampling distribution can be well approximated by a t distribution with the following degrees of freedom:

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} \,. \tag{5}$$

Our test procedure is therefore as follows.

— Compute from two independent samples the t statistic in (2).

— Reject H_0 if $|t| > t_{\alpha/2}$ with ν given by (5).

The Excel Data Analysis tool

"t-Test: Two-Sample Assuming Unequal Variances"

can be used to conduct this test.

For truly large n_1 and n_2, the distribution of the statistic in (2) can of course also be approximated by the standard normal distribution.

A Recommendation

Unless there is sufficient evidence that the two variances are not the same, the equal-variances t test should be used. This is because ν in (5) is bounded from above by n_1+n_2-2, implying that the equal-variances test is more *powerful* (greater "sample size" yields lower β).

Two Means, Paired Samples . . .

In many applications, the two populations are naturally **paired** or **coupled**. Some simple examples are:

— Decide on the basis of weights "before and after" whether a certain diet is really effective.

— Decide on the basis of salaries "before and after" receiving an MBA whether that degree contributes to financial well being.

— Decide whether an observed difference between the average I.Q.'s of husbands and their wives is really significant.

— Decide whether or not a new industrial safety program is effective on the basis of data taken from 12 factories.

Even if two populations are not "naturally" coupled, it is often desirable to *design* experiments that are "controlled" by some factors. The intent is to reduce "noise" and hence variability in the sample differences. A good example of this is that when testing the performance of a new piece of equipment, the same operator should test both the new machine and the old machine. This serves to reduce potential noise coming from variability in operator skills, and is often referred to as "controlling for the operator" (or for whatever).

In fact, given a choice, pairing the samples is preferred over collecting independent samples.

Interestingly, our test procedure actually turns out to be easier. The idea is simply to look at the pairwise differences of the observed data.

Formally, ...

Consider n pairs of observations (X_{1i}, X_{2i}) for $i = 1, 2, \ldots, n$. Since the data is paired, we can define for each pair:

$$D_i = X_{1i} - X_{2i} \tag{6}$$

(the order is irrelevant).

The key fact is that, *regardless of whether* X_{1i} *and* X_{2i} *are independent*, if it is reasonable to assume that the D_is are normally distributed with mean μ_D and variance σ_D^2, then the one-sample t test discussed before *directly* applies! One only needs to note that $\mu_D = \mu_1 - \mu_2$, the difference in the two separate population means.

It follows that our hypotheses are:

$$\begin{aligned} H_0 : &\quad \mu_D = 0 \\ H_1 : &\quad \mu_D \neq 0 \end{aligned}$$

Now, define the test statistic:

$$t = \frac{\bar{D} - 0}{\sqrt{s_D^2/n}}, \tag{7}$$

where

$$\bar{D} \equiv \frac{1}{n} \sum_{i=1}^{n} D_i \tag{8}$$

and

$$s_D^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2. \tag{9}$$

Then, the statistic t follows the t distribution with $\nu = n - 1$ degrees of freedom.

The test procedure is therefore:

— Compute t from the paired sample data.

— Reject H_0 if $|t| > t_{\alpha/2}$ with $\nu = n - 1$.

The Excel tool "t-Test: Paired Two Sample for Means" can be used to conduct this test.

Two Proportions ...

Suppose the Human Resources Department of a company does a study on male and female's salaries. They find that on average female's salaries are substantially lower than those of men. When brought to the attention of senior management, one vice-president points out that he thinks that on average more men have graduate degrees (MBAs) and thus would tend to have higher salaries then women.

The director of Human Resources then conducts two random samples. One of male employees and one of female employees. The results are:

	Males	Females
Number with MBA	20	8
Sample Size	100	75
Proportion	0.2	0.106667

Thus, 20% of the men have an MBA while only 10.7% of the women have an MBA. Does this indicate that the proportions *differ* in the populations of all male and all female employees? This motivates the following ...

Consider the "generic framework":

$X_i =$ Number of "successes" in a sample from population i, $i = 1, 2$

$n_i =$ Size of the sample from population i, $i = 1, 2$

$\hat{p}_i =$ Proportion of successes in the sample from population i, $i = 1, 2$; that is, X_i/n_i

$p_i =$ Proportion of successes in population i, $i = 1, 2$

We wish to test the hypotheses pair:

$$\begin{aligned} H_0 : &\quad p_1 = p_2 \text{ (or } p_1 - p_2 = 0) \\ H_1 : &\quad p_1 \neq p_2 \text{ (or } p_1 - p_2 \neq 0) \end{aligned}$$

Clearly, the natural test statistic is: $\hat{p}_1 - \hat{p}_2$. As usual, we will standardize this.

It can be shown that if the two samples are taken independently, then the variance of our sample statistic is

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \qquad (10)$$

However, since p_1 and p_2 are unknown, we will estimate this variance by substituting $\hat{p}_1$ and $\hat{p}_2$ into the right-hand side of (10). This results in the standardized statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}, \qquad (11)$$

which, when both $n_1\hat{p}_1$ and $n_2\hat{p}_2$ are greater than 5, can be shown to approximately follow the standard normal distribution.

The test procedure is therefore:

— Compute z according to (11).

— Reject H_0 if $|z| > z_{\alpha/2}$.

Let $\alpha = 0.05$. It is easily shown that in our salary example, we have $z = 1.741$. Since this is less than $z_{0.025} = 1.96$, we conclude that there is insufficient evidence to reject H_0, even though, *in this particular sample*, the MBA proportion for males is almost twice as large as that for females. That is, what we observed in this sample is not considered a rare event.