PART-WHOLE COMPARISONS OF MEANS

Albert Madansky
H.G.B. Alexander Professor Emeritus of Business Administration
Booth School of Business
University of Chicago

Vice President
The Analytical Group, Inc.

We consider here the situation in which we have m randomly drawn observations from population 1and n-m observations randomly drawn from population 2, and the data are drawn independently from each of the populations. We designate the mean of the two samples as $\overline{x}_1$ and $\overline{x}_2$, and the sample variances of the two data sets will be designated as s_1^2 and s_2^2. We also designate by $\overline{x}_T$ the mean of the total sample of n observations from the two populations. The object of the t-test is to test whether the mean of population 1 is different from that of the composite population consisting of populations 1 and 2.

In this note we describe the t-test used by The Analytical Group's WinCross (2012) and contrast it with the t-test recommended by the National Assessment of Educational Progress (NAEP) (2009). Finally, in the spirit of the analysis in Eberhardt and Fligner (1977), we compare the two procedures and conclude that the WinCross approach is superior.

WinCross approach

The hypothesis test is going to be based on the difference between $\overline{x}_1$ and $\overline{x}_T$, and so we need to determine the variance of that difference. Let us assume that the n observations are ordered so that the first m are from population 1 and the remaining n-m are from population 2. Then

$$\begin{aligned}
\overline{x}_1 - \overline{x}_T &= \frac{\sum_{i=1}^{m} x_i}{m} - \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{\sum_{i=1}^{m} x_i}{m} - \frac{\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i}{n} \\
&= (\frac{1}{m} - \frac{1}{n})\sum_{i=1}^{m} x_i - \frac{\sum_{i=m+1}^{n} x_i}{n} \\
&= (1 - \frac{m}{n})\overline{x}_1 - (1 - \frac{m}{n})\overline{x}_2 \\
&= (1 - \frac{m}{n})(\overline{x}_1 - \overline{x}_2)
\end{aligned}$$

Therefore the variance of $\overline{x}_1 - \overline{x}_T$ is given by

$$V[\bar{x}_1 - \bar{x}_T] = (1-\frac{m}{n})^2[V[\bar{x}_1] + V[\bar{x}_2]]$$
$$= (1-\frac{m}{n})^2[\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n-m}]$$

which is estimated by

$$(1-\frac{m}{n})^2[\frac{s_1^2}{m} + \frac{s_2^2}{n-m}]$$

That is, if p=m/n is the fraction of the n observations that come from population 1, then the variance of $\bar{x}_1 - \bar{x}_T$ is given by $(1\text{-}p)^2$ times the sum of squares of the standard errors of population 1 and population 2. The t statistic for testing the hypothesis that the mean of population 1 is different from that of the composite population is given by

$$t = \frac{\bar{x}_1 - \bar{x}_T}{(1-p)\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n-m}}}$$

When the variances of the two populations are identical, the common variance may be estimated by s_T^2 and so this reduces to

$$t = \frac{\bar{x}_1 - \bar{x}_T}{(1-p)\sqrt{\frac{s_T^2}{m} + \frac{s_T^2}{n-m}}}$$
$$= \frac{\bar{x}_1 - \bar{x}_T}{s_T(1-p)\sqrt{\frac{1}{m} + \frac{1}{n-m}}}$$

The denominator of t can be rewritten as

$$s_T(1-p)\sqrt{\frac{1}{np} + \frac{1}{n(1-p)}} = \frac{s_T}{\sqrt{n}}\sqrt{\frac{1-p}{p}}$$

As p gets close to 1 the denominator of t approaches 0, so that t approaches infinity. In this case the test will find significant differences whenever the "part" is almost equal to the "whole." As a precaution, I recommend not performing this test when p > 0.9. As p gets close to 0 the denominator of t approaches infinity, so that t approaches 0. In this case the test will find no significant differences whenever the "part" is an infinitesimal part of the "whole." As a precaution, I recommend not performing this test when p < 0.1.

NAEP approach

Alternatively,

$$V[\bar{x}_1 - \bar{x}_T] = V[\bar{x}_1] + V[\bar{x}_T] - 2C[\bar{x}_1, \bar{x}_T]$$

and so we must determine C[$\bar{x}_1, \bar{x}_T$], the covariance of $\bar{x}_1$ and $\bar{x}_T$. This covariance is

$$C[\bar{x}_1, \bar{x}_T] = C[\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{n} x_i}{n}]$$
$$= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} C[x_i, x_j]$$
$$= \frac{1}{mn} m\sigma_1^2$$

and so

$$V[\bar{x}_1 - \bar{x}_T] = V[\bar{x}_1] + V[\bar{x}_T] - 2C[\bar{x}_1, \bar{x}_T]$$
$$= \frac{\sigma_1^2}{m} + \frac{\sigma_T^2}{n} - 2\frac{\sigma_1^2}{n}$$
$$= \frac{\sigma_T^2}{n} + \frac{\sigma_1^2}{m}(1-2p)$$

which is the expression given by NAEP.

Note that when m=n/2 this variance reduces to that of $\bar{x}_T$. (Note also that when m=n this reduces to 0, as it should, because $\bar{x}_T$ is identical to $\bar{x}_1$.)

Comparison

A way of seeing the equivalence of these expressions for the variance of the difference of the part and whole means is by expressing $\bar{x}_T$ as

$$\bar{x}_T = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{m\bar{x}_1 + (n-m)\bar{x}_2}{n} = p\bar{x}_1 + (1-p)\bar{x}_2$$

where p=m/n. Then

$$V\bar{x}_T = \frac{\sigma_T^2}{n} = p^2 \frac{\sigma_1^2}{m} + (1-p)^2 \frac{\sigma_2^2}{n-m}$$

and so

$$V[\bar{x}_1 - \bar{x}_T] = \frac{\sigma_T^2}{n} + \frac{\sigma_1^2}{m}(1-2p)$$
$$= p^2 \frac{\sigma_1^2}{m} + (1-p)^2 \frac{\sigma_2^2}{n-m} + \frac{\sigma_1^2}{m}(1-2p)$$
$$= (1-p)^2 \frac{\sigma_1^2}{m} + (1-p)^2 \frac{\sigma_2^2}{n-m}$$
$$= (1-p)^2 [\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n-m}]$$

Unfortunately, though in the population

$$\frac{\sigma_T^2}{n} + \frac{\sigma_1^2}{m}(1-2p) = (1-p)^2 [\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n-m}]$$

this equality does not hold for the sample counterparts:

$$d_1 = \frac{s_T^2}{n} + \frac{s_1^2}{m}(1-2p), d_2 = (1-p)^2[\frac{s_1^2}{m} + \frac{s_2^2}{n-m}]$$

To see the relationship between these sample variances, we note that

$$(n-1)s_T^2 = \sum_{i=1}^{n}(x_i - \overline{x}_T)^2 = \sum_{i=1}^{m}(x_i - \overline{x}_1 + \overline{x}_1 - \overline{x}_T)^2 + \sum_{i=m+1}^{n}(x_i - \overline{x}_2 + \overline{x}_2 - \overline{x}_T)^2$$

$$= \sum_{i=1}^{m}(x_i - \overline{x}_1)^2 + m(\overline{x}_1 - \overline{x}_T)^2 + \sum_{i=m+1}^{n}(x_i - \overline{x}_2)^2 + (n-m)(\overline{x}_2 - \overline{x}_T)^2$$

$$= (m-1)s_1^2 + (n-m-1)s_2^2 + m(\overline{x}_1 - \overline{x}_T)^2 + (n-m)(\overline{x}_2 - \overline{x}_T)^2$$

Then

$$d_1 = \frac{(m-1)s_1^2}{n(n-1)} + \frac{(n-m-1)s_2^2}{n(n-1)} + \frac{m}{n(n-1)}(\overline{x}_1 - \overline{x}_T)^2 + \frac{n-m}{n(n-1)}(\overline{x}_2 - \overline{x}_T)^2 + \frac{s_1^2}{m}(1-2p)$$

As we let m and n approach infinity but with m/n=p, nd_1 approaches

$$p\sigma_1^2 + (1-p)\sigma_2^2 + p(\mu_1 - \mu_T)^2 + (1-p)(\mu_2 - \mu_T)^2 + (1/p - 2)o_1^2$$

$$= \frac{1-p^2}{p}\sigma_1^2 + (1-p)\sigma_2^2 + p(\mu_1 - \mu_T)^2 + (1-p)(\mu_2 - \mu_T)^2$$

and nd_2 approaches

$$(1-p)^2[\frac{\sigma_1^2}{p} + \frac{\sigma_2^2}{1-p}] = \frac{(1-p)^2}{p}\sigma_1^2 + (1-p)\sigma_2^2$$

Under the null hypothesis the terms in nd_1 involving the means are both equal to 0, leaving us with a comparison of

$$\frac{1-p^2}{p}\sigma_1^2 + (1-p)\sigma_2^2$$

and

$$\frac{(1-p)^2}{p}\sigma_1^2 + (1-p)\sigma_2^2$$

Since $1-p^2 .> (1-p)^2$, we see that even when the null hypothesis is true d_1 is larger than d_2. And when the null hypothesis is false d_1 is much greater than d_2 because of the added terms involving the means.

In summary, the NAEP formula is based on a larger standard error for the difference between the part and whole mean than that used by WinCross, thereby leading to fewer significant differences detected than are actually to be found in the data.

References

Eberhardt, K.A. and Fligner, M.A. (1977), "A Comparison of Two Tests for the Equality of Two Proportions," *The American Statistician*, 31, 151-5.

National Assessment of Educational Progress (NAEP) (2009), Technical Documentation: http://nces.ed.gov/nationsreportcard/tdw/analysis/2004_2005/infer_compare2_overlap.aspx

The Analytical Group (2012), Wincross: http://www.analyticalgroup.com/wincross.html