Combining Results

Two Samples and No Original Data

By Dean V. Neubauer

Q How can I combine the results of two samples that are statistically alike based on hypothesis testing when the original data are now unavailable?

A It is not uncommon for a researcher (scientist or engineer) to have two samples that have been determined to be statistically alike in terms of variability, based on an *F*-test, and in terms of averages, based on a *t*-test. Both of these tests are described in E2586, Practice for Calculating and Using Basic Statistics, as well as any elementary statistics text.

However, the researcher often may no longer have access to the original raw data due to various reasons. In such situations, the logical next step may be simply to pool the sample variances that an insignificant *F*-test would allow you to do. Furthermore, the means of the two samples can be weighted by their sample sizes to produce a combined mean. These are reasonable things to do.

Let's assume that our first sample has n_1 observations, a mean of $\bar{x}_1$ and a standard deviation of s_1. Likewise, our second sample has n_2 observations, a mean of $\bar{x}_2$ and a standard deviation of s_2. The combined mean, as described above, is simply a weighted average of the two means using the sample sizes as the weights as in the following formula:

$$\bar{x}_{both\ samples} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \quad (1)$$

Assume that after performing an *F*-test on the sample variances, s_1^2 and s_2^2, we find that they are statistically the same, i.e., both samples came from the same parent population. Then, as shown in many elementary statistics texts and E2586, the sample variances are pooled together using the degrees of freedom for each sample as the weights as seen below:

$$s^2_{both\ samples} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (2)$$

Now, while this pooled variance is statistically acceptable, we see that the degrees of freedom show a loss of two degrees of freedom in the denominator. This is understandable as we lose a single degree of freedom for each sample since we calculate its mean from the data. But, if we combine the two samples we are estimating only one mean, and we would want to have a pooled overall sample variance estimate showing the loss of only one degree of freedom representing the combined mean.

While this would be a simple matter if the original data are available for both samples, it becomes a more difficult problem to compute the overall variance estimate for both samples when the data are unavailable. Fortunately, the relationship between this overall variance estimate and the means $\bar{x}_1$ and $\bar{x}_2$, standard deviations s_1 and s_2, and sample sizes n_1 and n_2 of both samples was determined by Baker and Nissim[1]:

$$s^2_{both\ samples} = \frac{1}{n_1 + n_2 - 1}\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2\right] \quad (3)$$

So, we can see in Equation 3 that the denominator is now $n_1 + n_2$ -1 degrees of freedom.

Suppose that we have two samples and the original data are no longer available to us. For the first sample of $n_1 = 30$ observations, we found that $\bar{x}_1 = 13.5$ and $s_1 = 0.8$. For the second sample of $n_2 = 70$ observations, we found that $\bar{x}_2 = 13.7$ and $s_2 = 0.9$. Thus, using Equation 1, we find the combined mean of both samples to be:

$$\bar{x}_{both\ samples} = \frac{30(13.5) + 70(13.7)}{30 + 70} = \frac{1364}{100} = 13.64$$

We apply the *F*-test to our sample variances as:

$$F = \frac{s_2^2}{s_1^2} = \frac{(0.9)^2}{(0.8)^2} = \frac{0.81}{0.64} = 1.266$$

which is compared to a critical value of $F_{.95,69,29} = 1.74$ (you can use the Excel function F.INV(0.95,69,29) to get this critical value). Since the computed F value of 1.27 is less than the critical value of 1.74 we say there is insufficient evidence to say the sample variances differ, so we can pool the sample variances as shown in Equation 2 as:

$$s^2_{both\ samples} = \frac{(30-1)(0.8)^2 + (70-1)(0.9)^2}{30+70-2} = \frac{74.45}{98} = 0.7597$$

and the combined standard deviation estimate for both samples is:

$$s_{both\ samples} = \sqrt{0.7597} = 0.8716$$

which has 98 degrees of freedom. However, if we no longer have access to the original data to produce an estimate of the sample variance with 99 degrees of freedom, then we must use Equation 3 to produce the desired result as:

$$\begin{aligned} s^2_{both\ samples} &= \frac{1}{30+70-1}\left[(30-1)(0.8)^2 + (70-1)(0.9)^2 + \frac{(30)(70)}{30+70}(13.5-13.7)^2\right] \\ &= \frac{1}{99}\left[(29)(0.64) + (69)(0.81) + \frac{2100}{100}(-0.2)^2\right] \\ &= \frac{75.29}{99} \\ &= 0.7605 \end{aligned}$$

and the combined standard deviation estimate for both samples is:

$$s_{both\ samples} = \sqrt{0.7605} = 0.8721$$

which has 99 degrees of freedom. What has been demonstrated here is that if the samples do not differ statistically, then you do not need the original data to compute an overall estimate of the mean and variability of the combined samples.

Here's an interesting scenario as well. Suppose that the researcher is getting the data continuously, but slowly one at a time. Wouldn't it be desirable to update both the mean and the standard deviation estimate after you obtain each value? If we have an initial sample of size n with a mean $\bar{x}_1$ and standard deviation s_n, then we obtain a new observation of size 1. This new observation has a mean equal to itself, call it x_{n+1}. We can compute the new combined mean using a formula similar to Equation 1 as:

$$\bar{x}_{new} = \frac{n\bar{x}_n + (1)x_{n+1}}{n+1} \qquad (1a)$$

The new combined standard deviation will look similar to Equation 3 with the form:

$$s^2_{new} = \left(\frac{n-1}{n}\right)s^2_n + \frac{(\bar{x}_n - x_{n+1})^2}{n+1} \qquad (3a)$$

with n degrees of freedom. Of course, $s_{new} = \sqrt{s^2_{new}}$.

Let's illustrate Equations 1a and 3a using a simple example. Suppose that we have already collected $n = 5$ observations: 1.2, 1.4, 1.1, 1.2 and 1.6. The sample mean and standard deviation estimates for this sample are $\bar{x}_5 = 1.30$ and $s = 0.20$, respectively. Now, suppose that we obtain the next observation, $x_6 = 1.8$. The new mean, using Equation 1a, becomes:

$$\bar{x}_{new} = \frac{5(1.30) + (1)(1.8)}{5+1} = \frac{8.3}{6} = 1.38$$

and the new variance estimate, using Equation 3a, is:

$$s^2_{new} = \left(\frac{5-1}{5}\right)(0.20)^2 + \frac{(1.30-1.8)^2}{5+1} = 0.074$$

and the new standard estimate is $s_{new} = \sqrt{0.074} = 0.271$ with 5 degrees of freedom.

Hopefully, the equations provided in this article will prove useful to you if you are in the situations discussed here. Having the ability to combine data to generate overall summary statistics when all you have are the summary statistics for two samples, but no raw data, is nice to have handy.

REFERENCE

1. Baker, R. W. R. and J. A. Nissim, "Expressions for Combining Standard Errors of Two Groups and for Sequential Standard Error," *Nature*, Vol. 198, June 8, 1963, p. 1020.

Dean V. Neubauer, Corning Inc., Corning, New York, is an ASTM International fellow, is chairman of E11.90.03 on Publications and coordinator of the Data Points column; he is immediate past chairman of Committee E11 on Quality and Statistics.