11.5: Standard Error and Pooled Variance

Recall that the standard error is the average distance between any given sample mean and the center of its corresponding sampling distribution, and it is a function of the standard deviation of the population (either given or estimated) and the sample size. This definition and interpretation hold true for our independent samples t-test as well, but because we are working with two samples drawn from two populations, we have to first combine their estimates of standard deviation – or, more accurately, their estimates of variance – into a single value that we can then use to calculate our standard error.

The combined estimate of variance using the information from each sample is called the pooled variance and is denoted s_p^2; the subscript p serves as a reminder indicating that it is the pooled variance. The term "pooled variance" is a literal name because we are simply pooling or combining the information on variance – the Sum of Squares and Degrees of Freedom – from both of our samples into a single number. The result is a weighted average of the observed sample variances, the weight for each being determined by the sample size, and will always fall between the two observed variances. The computational formula for the pooled variance is:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{11.5.1}$$

This formula can look daunting at first, but it is in fact just a weighted average. Even more conveniently, some simple algebra can be employed to greatly reduce the complexity of the calculation. The simpler and more appropriate formula to use when calculating pooled variance is:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} \tag{11.5.2}$$

Using this formula, it's very simple to see that we are just adding together the same pieces of information we have been calculating since chapter 3. Thus, when we use this formula, the pooled variance is not nearly as intimidating as it might have originally seemed.

Once we have our pooled variance calculated, we can drop it into the equation for our standard error:

$$S_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \tag{11.5.3}$$

Once again, although this formula may seem different than it was before, in reality it is just a different way of writing the same thing. An alternative but mathematically equivalent way of writing our old standard error is:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \tag{11.5.4}$$

Looking at that, we can now see that, once again, we are simply adding together two pieces of information: no new logic or interpretation required. Once the standard error is calculated, it goes in the denominator of our test statistic, as shown above and as was the case in all previous chapters. Thus, the only additional step to calculating an independent samples t-statistic is computing the pooled variance. Let's see an example in action.