Combined Variance of Two Groups with Equal Numbers of Observations

Sara E. Burke

Created 2014-01-12, Last updated 2021-06-11

Special Thanks to Rachel Nolan

If we have two groups of n observations each, and we know the mean and sample variance of each group separately, how can we calculate the mean and sample variance of the combination of the two groups?

Let's refer to the groups as X and Y. The elements in the first group are $x_1 \dots x_n$ and the elements in the second group are $y_1 \dots y_n$. We are given four values:

$$\begin{aligned} \text{Mean of } X &= \bar{x} \\ \text{Variance of } X &= V_x \\ \text{Mean of } Y &= \bar{y} \\ \text{Variance of } Y &= V_y \end{aligned}$$

And we are expected to calculate two values:

$$\begin{aligned} \text{Overall Mean} &= M \\ \text{Overall Variance} &= V \end{aligned}$$

The overall mean is easy to calculate. It is the sum of all the observations divided by the total number of observations $2n$. Because each group has equally many observations, the overall mean is just the midpoint of the two means.

$$M = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{2n} = \frac{n\bar{x} + n\bar{y}}{2n} = \frac{\bar{x} + \bar{y}}{2}$$

The overall variance is not so simple.

The variance of X can be written as follows. As shown, from now on all sums across indices i from 1 to n will be written using only the $\sum$ symbol.

$$V_x = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The overall variance follows the same structure, but we have to sum across all of the X observations and all of the Y observations.

$$V = \frac{\sum(x_i - M)^2 + \sum(y_i - M)^2}{2n - 1}$$

Let's take the X portion of the numerator and expand it as shown below.

$$\begin{aligned}
\sum(x_i - M)^2 &= \sum(x_i - \bar{x} + \bar{x} - M)^2 \\
&= \sum((x_i - \bar{x}) + (\bar{x} - M))^2 \\
&= \sum((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - M) + (\bar{x} - M)^2) \\
&= \sum(x_i - \bar{x})^2 + 2\sum(x_i\bar{x} - \bar{x}^2 - x_iM + \bar{x}M) + \sum(\bar{x} - M)^2 \\
&= \sum(x_i - \bar{x})^2 + 2\bar{x}\sum x_i - 2n\bar{x}^2 - 2M\sum x_i + 2n\bar{x}M + n(\bar{x} - M)^2
\end{aligned}$$

Note that $\bar{x} = \dfrac{\sum x_i}{n}$ so $\sum x_i = n\bar{x}$. Continuing from above,

$$\begin{aligned}
\sum(x_i - M)^2 &= \sum(x_i - \bar{x})^2 + 2n\bar{x}^2 - 2n\bar{x}^2 - (\bar{x} + \bar{y})n\bar{x} + (\bar{x} + \bar{y})n\bar{x} + n(\bar{x}^2 - 2\bar{x}M + M^2) \\
&= \sum(x_i - \bar{x})^2 + n\bar{x}^2 - 2n\bar{x}M + nM^2 \\
&= \sum(x_i - \bar{x})^2 + n\bar{x}^2 - 2n\bar{x}\left(\frac{\bar{x} + \bar{y}}{2}\right) + n\left(\frac{\bar{x} + \bar{y}}{2}\right)^2 \\
&= \sum(x_i - \bar{x})^2 + n\bar{x}^2 - (n\bar{x}^2 + n\bar{x}\bar{y}) + n\left(\frac{\bar{x}^2 + 2\bar{x}\bar{y} + \bar{y}^2}{4}\right) \\
&= \sum(x_i - \bar{x})^2 + n\bar{x}^2 - n\bar{x}^2 - n\bar{x}\bar{y} + \frac{n}{4}\bar{x}^2 + \frac{n}{2}\bar{x}\bar{y} + \frac{n}{4}\bar{y}^2 \\
&= \sum(x_i - \bar{x})^2 + \frac{n}{4}\bar{x}^2 - \frac{n}{2}\bar{x}\bar{y} + \frac{n}{4}\bar{y}^2 \\
&= \sum(x_i - \bar{x})^2 + \frac{n}{4}(\bar{x}^2 - 2\bar{x}\bar{y} + \bar{y}^2) \\
&= \sum(x_i - \bar{x})^2 + \frac{n}{4}(\bar{x} - \bar{y})^2
\end{aligned}$$

The $\sum(y_i - M)^2$ portion of the numerator of V can be similarly expanded, so V can be expanded as follows.

$$\begin{aligned} V &= \frac{\sum(x_i - M)^2 + \sum(y_i - M)^2}{2n-1} \\ &= \frac{\sum(x_i - \bar{x})^2 + \frac{n}{4}(\bar{x} - \bar{y})^2 + \sum(y_i - \bar{y})^2 + \frac{n}{4}(\bar{x} - \bar{y})^2}{2n-1} \\ &= \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 + \frac{n}{2}(\bar{x} - \bar{y})^2}{2n-1} \end{aligned}$$

Note that $V_x = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$ so $\sum(x_i - \bar{x})^2 = (n-1)V_x$. The same applies to Y.

$$V = \frac{(n-1)V_x + (n-1)V_y + \frac{n}{2}(\bar{x} - \bar{y})^2}{2n-1}$$

$$V = \frac{(n-1)(V_x + V_y) + \frac{n}{2}(\bar{x} - \bar{y})^2}{2n-1}$$

This is the simplest form.

Note that, as the sample sizes increases, the variance of the combined sample approaches the mean of the two variances plus the square of half the distance between the two means.

$$\begin{aligned} \lim_{n\to\infty}(V) &= \lim_{n\to\infty}\left(\frac{(n-1)(V_x + V_y)}{2n-1} + \frac{\frac{n}{2}(\bar{x} - \bar{y})^2}{2n-1}\right) \\ &= \frac{V_x + V_y}{2} + \frac{(\bar{x} - \bar{y})^2}{4} \\ &= \frac{V_x + V_y}{2} + \left(\frac{\bar{x} - \bar{y}}{2}\right)^2 \end{aligned}$$