

FORMULAS FOR THE POOLED STANDARD DEVIATION

This arises in the two-sample problem. There are two samples of data $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$.

Here we've noted the random variables and X 's and Y 's. You'll see this also identified as "group 1" and "group 2" or perhaps "group A" and "group B."

The usual assumption is that these come from normal populations with the same standard deviation σ but with (possibly) different means μ_x and μ_y . For both hypothesis testing and for confidence intervals there is a need to estimate σ .

The estimate of σ is the pooled standard deviation, denote s_p . You will find many formulas for s_p or, equivalently, the pooled variance s_p^2 .

Then $s_p =$ pooled standard deviation $=$

$$\begin{aligned}
 & \sqrt{\frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2}{m + n - 2}} \\
 &= \sqrt{\frac{\left\{ \sum_{i=1}^m x_i^2 - \frac{\left(\sum_{i=1}^m x_i \right)^2}{m} \right\} + \left\{ \sum_{j=1}^n y_j^2 - \frac{\left(\sum_{j=1}^n y_j \right)^2}{n} \right\}}{m + n - 2}} \\
 &= \sqrt{\frac{\left\{ \sum_{i=1}^m x_i^2 - m \bar{x}^2 \right\} + \left\{ \sum_{j=1}^n y_j^2 - n \bar{y}^2 \right\}}{m + n - 2}} \\
 &= \sqrt{\frac{S_{xx} + S_{yy}}{m + n - 2}} = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m + n - 2}}
 \end{aligned}$$

The value of s_p will lie between s_x and s_y . If m and n are not too far apart, then s_p is surprisingly close to halfway between s_x and s_y .