Generally, σ^2 will not be known and must be estimated from the data by calculating the **pooled sample variance,**

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

where $s_X^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and similarly for s_Y^2. Note that s_p^2 is a weighted average of the sample variances of the X's and Y's, with the weights proportional to the degrees of freedom. This weighting is appropriate since if one sample is much larger than the other, the estimate of σ^2 from that sample is more reliable and should receive greater weight. The following theorem gives the distribution of a statistic that will be used for forming confidence intervals and performing hypothesis tests.

THEOREM **A**

Suppose that $X_1, \ldots, X_n$ are independent and normally distributed random variables with mean μ_X and variance σ^2, and that $Y_1, \ldots, Y_m$ are independent and normally distributed random variables with mean μ_Y and variance σ^2, and that the Y_i are independent of the X_i. The statistic

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

follows a t distribution with $m + n - 2$ degrees of freedom.

Proof

According to the definition of the t distribution in Section 6.2, we have to show that the statistic is the quotient of a standard normal random variable and the square root of an independent chi-square random variable divided by its $n+m-2$ degrees of freedom. First, we note from Theorem B in Section 6.3 that $(n-1)s_X^2/\sigma^2$ and $(m-1)s_Y^2/\sigma^2$ are distributed as chi-square random variables with $n-1$ and $m-1$ degrees of freedom, respectively, and are independent since the X_i and Y_i are. Their sum is thus chi-square with $m+n-2$ df. Now, we express the statistic as the ratio U/V, where

$$U = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

$$V = \sqrt{\left[\frac{(n-1)s_X^2}{\sigma^2} + \frac{(m-1)s_Y^2}{\sigma^2}\right]\frac{1}{m+n-2}}$$

U follows a standard normal distribution and from the preceding argument V has the distribution of the square root of a chi-square random variable divided by its degrees of freedom. The independence of U and V follows from Corollary A in Section 6.3. ■