

**Example 6.6 (The sum of squares decomposition for univariate ANOVA)**

Consider the following independent samples.

Population 1: 9, 6, 9

Population 2: 0, 2

Population 3: 3, 1, 2

Since, for example,  $\bar{x}_3 = (3 + 1 + 2)/3 = 2$  and  $\bar{x} = (9 + 6 + 9 + 0 + 2 + 3 + 1 + 2)/8 = 4$ , we find that

$$\begin{aligned} 3 &= x_{31} = \bar{x} + (\bar{x}_3 - \bar{x}) + (x_{31} - \bar{x}_3) \\ &= 4 + (2 - 4) + (3 - 2) \\ &= 4 + (-2) + 1 \end{aligned}$$

Repeating this operation for each observation, we obtain the arrays

$$\begin{pmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{pmatrix}$$

$$\begin{array}{ccccc} \text{observation} & = & \text{mean} & + & \text{treatment effect} & + & \text{residual} \\ (x_{\ell j}) & & (\bar{x}) & & (\bar{x}_\ell - \bar{x}) & & (x_{\ell j} - \bar{x}_\ell) \end{array}$$

The question of equality of means is answered by assessing whether the contribution of the treatment array is large relative to the residuals. (Our estimates  $\hat{\tau}_\ell = \bar{x}_\ell - \bar{x}$  of  $\tau_\ell$  always satisfy  $\sum_{\ell=1}^g n_\ell \hat{\tau}_\ell = 0$ . Under  $H_0$ , each  $\hat{\tau}_\ell$  is an estimate of zero.) If the treatment contribution is large,  $H_0$  should be rejected. The size of an array is quantified by stringing the rows of the array out into a vector and calculating its squared length. This quantity is called the *sum of squares* (SS). For the observations, we construct the vector  $\mathbf{y}' = [9, 6, 9, 0, 2, 3, 1, 2]$ . Its squared length is

$$SS_{\text{obs}} = 9^2 + 6^2 + 9^2 + 0^2 + 2^2 + 3^2 + 1^2 + 2^2 = 216$$

Similarly

$$SS_{\text{mean}} = 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 = 8(4^2) = 128$$

$$\begin{aligned} SS_{\text{tr}} &= 4^2 + 4^2 + 4^2 + (-3)^2 + (-3)^2 + (-2)^2 + (-2)^2 + (-2)^2 \\ &= 3(4^2) + 2(-3)^2 + 3(-2)^2 = 78 \end{aligned}$$

and the residual sum of squares is

$$SS_{\text{res}} = 1^2 + (-2)^2 + 1^2 + (-1)^2 + 1^2 + 1^2 + (-1)^2 + 0^2 = 10$$

The sums of squares satisfy the same decomposition, (6-30), as the observations. Consequently,

$$SS_{\text{obs}} = SS_{\text{mean}} + SS_{\text{tr}} + SS_{\text{res}}$$

or  $216 = 128 + 78 + 10$ . The breakup into sums of squares apportions variability in the combined samples into mean, treatment, and residual (error) components. An analysis of variance proceeds by comparing the relative sizes of  $SS_{\text{tr}}$  and  $SS_{\text{res}}$ . If  $H_0$  is true, variances computed from  $SS_{\text{tr}}$  and  $SS_{\text{res}}$  should be approximately equal. ■

The sum of squares decomposition illustrated numerically in Example 6.6 is so basic that the algebraic equivalent will now be developed.

Subtracting  $\bar{x}$  from both sides of (6-30) and squaring gives

$$(x_{\ell j} - \bar{x})^2 = (\bar{x}_{\ell} - \bar{x})^2 + (x_{\ell j} - \bar{x}_{\ell})^2 + 2(\bar{x}_{\ell} - \bar{x})(x_{\ell j} - \bar{x}_{\ell})$$

We can sum both sides over  $j$ , note that  $\sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell}) = 0$ , and obtain

$$\sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x})^2 = n_{\ell}(\bar{x}_{\ell} - \bar{x})^2 + \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})^2$$

Next, summing both sides over  $\ell$  we get

$$\sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x})^2 = \sum_{\ell=1}^g n_{\ell}(\bar{x}_{\ell} - \bar{x})^2 + \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})^2 \quad (6-31)$$

$$\left( \begin{array}{c} SS_{\text{cor}} \\ \text{total (corrected) SS} \end{array} \right) = \left( \begin{array}{c} SS_{\text{tr}} \\ \text{between (samples) SS} \end{array} \right) + \left( \begin{array}{c} SS_{\text{res}} \\ \text{within (samples) SS} \end{array} \right)$$

or

$$\begin{aligned} \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} x_{\ell j}^2 &= (n_1 + n_2 + \cdots + n_g)\bar{x}^2 + \sum_{\ell=1}^g n_{\ell}(\bar{x}_{\ell} - \bar{x})^2 + \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})^2 \\ (SS_{\text{obs}}) &= (SS_{\text{mean}}) + (SS_{\text{tr}}) + (SS_{\text{res}}) \end{aligned} \quad (6-32)$$

## ANOVA TABLE FOR COMPARING UNIVARIATE POPULATION MEANS

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)
Treatments	$SS_{tr} = \sum_{\ell=1}^g n_{\ell} (\bar{x}_{\ell} - \bar{x})^2$	$g - 1$
Residual (Error)	$SS_{res} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x}_{\ell})^2$	$\sum_{\ell=1}^g n_{\ell} - g$
Total (corrected for the mean)	$SS_{cor} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell j} - \bar{x})^2$	$\sum_{\ell=1}^g n_{\ell} - 1$

The usual  $F$ -test rejects  $H_0: \tau_1 = \tau_2 = \dots = \tau_g = 0$ , at level  $\alpha$  if

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/\left(\sum_{\ell=1}^g n_{\ell} - g\right)} > F_{g-1, \sum n_{\ell} - g}(\alpha)$$

where  $F_{g-1, \sum n_{\ell} - g}(\alpha)$  is the upper  $(100\alpha)$ th percentile of the  $F$ -distribution with  $g-1$  and  $\sum n_{\ell} - g$  degrees of freedom. This is equivalent to rejecting  $H_0$  for large values of  $SS_{tr}/SS_{res}$  or for large values of  $1 + SS_{tr}/SS_{res}$ . The statistic appropriate for a multivariate generalization rejects  $H_0$  for *small* values of the reciprocal

$$\frac{1}{1 + SS_{tr}/SS_{res}} = \frac{SS_{res}}{SS_{res} + SS_{tr}} \quad (6-33)$$

**Example 6.7 (A univariate ANOVA table and  $F$ -test for treatment effects)**

Using the information in Example 6.6, we have the following ANOVA table:

Source of variation	Sum of squares	Degrees of freedom
Treatments	$SS_{tr} = 78$	$g - 1 = 3 - 1 = 2$
Residual	$SS_{res} = 10$	$\sum_{\ell=1}^g n_{\ell} - g = (3 + 2 + 3) - 3 = 5$
Total (corrected)	$SS_{cor} = 88$	$\sum_{\ell=1}^g n_{\ell} - 1 = 7$