



*Analysis of variance one and two
factors*

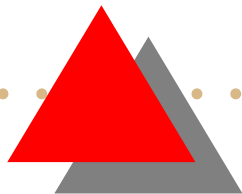
(STATS545.4: Analyses of variance and covariance)

Gunnar Stefansson



Factors and levels

A factor is a classification (categorical) variable such as a farm, gender, color and so forth. The possible values which a factor can take on are called levels. For example color may be red, blue, green and so forth.



Classification variables - two groups

When comparing two means the basic model is

$$y_i = \beta_1 + e_i, i = 1, \dots, n$$

$$y_i = \beta_2 + e_i, i = n + 1 \dots m$$

Note that the \mathbf{X} -matrix can be of arbitrary form. In particular one can define classification variables:

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & n \\ 0 & 1 & n + 1 \\ 0 & 1 & n + 2 \\ \vdots & \vdots & \vdots \\ 0 & 1 & n + m \end{bmatrix}$$

i.e. $y = \mathbf{X}\beta + \mathbf{e}$ is equivalent to the above model, which concerns estimation or comparisons of two

Classification variables - another representation

One could also write

$$y_i = \mu + e_i \quad 1 \leq i \leq n$$

$$y_i = \mu + \beta + e_i \quad n + 1 \leq i \leq n + m$$

and $H_0 : \mu_1 = \mu_2$ becomes $H_0 : \beta = 0$.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & 0 \\ 1 & 1 \\ \vdots & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Simple analysis of variance

Several groups

$$y_{1j} = \mu_1 + e_{1j} \quad j = 1, \dots, J_1$$

$$y_{2j} = \mu_2 + e_{2j} \quad j = 1, \dots, J_2$$

\vdots

$$y_{Ij} = \mu_I + e_{Ij} \quad j = 1, \dots, J_I,$$

with a total of $n = J_1 + \dots + J_I$ measurements.

In addition to simple comparisons of two means, i.e.

tests of $H_0 : \mu_1 = \mu_2$ with data of the form

$$y_i = \mu_1 + e_i \quad i = 1, \dots, n$$

$$y_i = \mu_2 + e_i \quad i = n + 1, \dots, n + m$$

it is also of interest to compare several means.

Thus we want to consider data from several (I) groups.

Developing matrix notation

Want

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

-prefer independent columns...

The models are set up using matrix notation,

- usually omit those columns in \mathbf{X} which would make them linearly dependent (also set the corresponding elements of the $\boldsymbol{\beta}$ -vector to zero without further estimation).

Different versions of the same model

The model can be written in different ways, e.g.

$$y_{1j} = \mu + \alpha_1 + e_{1j}, \quad j = 1, \dots, J_1$$

$$y_{2j} = \mu + \alpha_2 + e_{2j}, \quad j = 1, \dots, J_2$$

⋮

$$y_{Ij} = \mu + \alpha_I + e_{Ij}, \quad j = 1, \dots, J_I.$$

Here, μ is an overall mean but α_i is the deviance of each group from the overall mean.

Deviations from overall mean in matrix form

This model can be written using matrix notation as:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J_2} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & & 1 \\ 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \end{bmatrix} + \mathbf{e}$$



Null hypotheses, several means

The null hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_J$$

is the same as

$$H_0 : \alpha_1 = \dots = \alpha_I = 0.$$

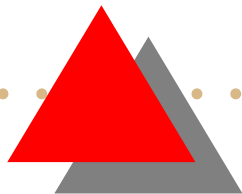
The alternative hypothesis H_a is simply that H_0 is not correct.





Dependent column vectors of X

Note now that the columns of X are dependent so that $(X'X)^{-1}$ does not exist. Therefore columns must be dropped or some other conditions set in order to find a solution.



Point estimates

One solution...

$$\mu_i = \mu + \alpha_i$$

$$\sum_i \alpha_i = 0$$

$$J_i = J$$

$$\hat{\mu}_i = \bar{y}_{i.}$$

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

The sum of squares is well-defined

$$SSE = \sum_{ij} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^I \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$$

where

$$\bar{y}_{i.} = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}.$$

We also know that

$$SSTOT = \sum_{i=1}^I \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$$

so the following variation is explained by the model

$$SSR = SSTOT - SSE = \dots = \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_i J_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

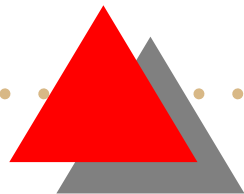


Components of sums of squares

The residuals add up and so do the sums of squares:

$$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 = \sum_{ij} (y_{ij} - \bar{y}_{i.})^2 + \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2$$



One-way anova

The ANOVA table becomes

	df	SS	MS	F
Model	$I - 1$	$SSR = \sum_{i=1}^I J_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$MSR = SSR / (I - 1)$	$F = MSR / MSE$
Error	$n - I$	$SSE = \sum_{i=1}^I \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$	$MSE = SSE / (n - I)$	
Total	$n - 1$	$SSTOT = \sum_{i=1}^I \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$		

We will reject H_0 if $F > F_{I-1, n-I, 1-\alpha}$