ANALYSIS OF VARIANCE AND ANALYSIS OF COVARIANCE

V. Nollau
Institute of Mathematical Stochastics, Technical University of Dresden, Germany

Keywords: Analysis of variance, least squares method, models with fixed effects, models with random effects, mixed models, one-way layout, higher way layouts, partitioning a sum of squares, analysis of covariance.

Contents

Summary

The analysis of variance (ANOVA) models have become one of the most widely applied tools of statistics for investigating multifactor data. It is a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to find out which kinds of effects are important and to estimate these effects. Such theory of analyzing measurements is naturally very useful in experiment design. The analysis of covariance is a generalization of regression analysis as well as analysis of variance for analyzing simultaneously the qualitative effects of some factors and the quantitative effects of the other factors.

1. Analysis of Variance

The analysis of variance is a statistical technique for analyzing measurements, quantitative results of observations and experiments depending on several kinds of effects (factors) operating simultaneously, to decide which kinds of effects are important and to estimate the effects (factors). A suitable theory of analyzing results of measurements, observations, and experiments naturally has useful application in experiment design.

The methodology of analysis of variance, which was originally developed by Sir Ronald A. Fisher, is concerned with the investigation of the factors probably having significant effects, by suitable choice of experiments. The main technique consists in an isolating procedure of the variations associated with different factors or defined sources. Mathematically this is a suitable partitioning of the total sample variance that means a partitioning of the total sum of squares.

In this way the procedure involves division of total observed variations in the data into individual components attributable to various factors and those due to random effects, and tests of significance to find out which factors influence the experimental results.

2.1 Fixed Models

1.3.1. One-Way Classification

The one-way classification (also called one-way layout) refers to the comparison of the means of several (univariate) populations. Considering an experiment having m treatment groups or m different levels of a single factor A one supposes n_i observations have been made at the i-th level giving a total of $N = \sum_{i=1}^{m} n_i$ observations (cf. Experiment Design (I)). If y_{ij} is the observed score corresponding to the j-th observation at the i-th level or treatment group, the analysis of variance model for such an experiment is given as

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad (i = 1, ..., m; \quad j = 1, ..., n_i) \qquad (1)$$

where μ (μ-real number) is a common (or general) mean common to all the observations, α_i is the special effect due to the i-th level of the considered factor and e_{ij} is the realization of a random error associated with the j-th observation, at the i-th level or treatment group. Without loss of generality one assumes $\sum_{i=1}^{m} n_i \alpha_i = 0$ (in the other case one transforms $\mu + \sum_{i=1}^{m} n_i \alpha_i \to \mu$).

Experiment Design (I)					
j	$\rightarrow$	replicated experiments			
i	$\downarrow$	1	2	$\cdots$	n_i
levels of factor A (treatment groups)	1	y_{11}	y_{12}	$\cdots$	y_{1n_1}
	2	y_{21}	y_{22}	$\cdots$	y_{2n_2}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$

	m	y_{m1}	y_{m2}	$\cdots$	y_{mn_m}

Table 1. Experiment Design (I)

The $N(=n_1+n_2+....+n_m)$ observed or measured values y_{ij} are considered as realizations of m mathematical samples

$$(Y_{11},Y_{12},....,Y_{1n_1}), \qquad (Y_{21},Y_{22},....,Y_{2n_2}),.....,(Y_{m1},Y_{m2},....,Y_{mn_m}).$$

Thus (1) implies

$$Y_{ij} = \mu + \sigma_i + E_{ij} \qquad (i=1,....,m; \quad j=1,.....,n_i) \tag{2}$$

where μ and α_i $(i=1,....,m)$ are (generally unknown) real parameters. Moreover, it is assumed that E_{ij} are random normally distributed independent variables with the expected value $\mathbb{E}(E_{ij})=0$ (that means, the random influence is a non-systematic one) and the variance $\mathrm{var}(E_{ij})=\sigma^2$ $(\sigma>0)$ (that means, the variability of the random influences is constant).

With these propositions the model (2) has the following matrix structure

$$\mathbf{Y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{E} \tag{3}$$

with

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{m1} \\ \vdots \\ Y_{mn_m} \end{pmatrix}, \quad \mathbf{X}^{\mathrm{T}} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix},$$

and

$$\mathbf{E} = \begin{pmatrix} E_{11} \\ \vdots \\ E_{1n_1} \\ E_{21} \\ \vdots \\ E_{2n_2} \\ \vdots \\ E_{m_1} \\ \vdots \\ E_{mn_m} \end{pmatrix} .$$

Using the least squares method, that means

$$\left| \mathbf{Y} - \mathbf{X}^T \boldsymbol{\beta} \right|^2 \to \min_{\boldsymbol{\beta} \in \mathbb{R}^{m+1}} \tag{4}$$

or component-wise

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2 \to \min_{\mu, \alpha_1, \ldots, \alpha_m \in \mathbb{R}}, \tag{5}$$

with the condition $\sum_{i=1}^{m} n_i \alpha_i = 0$ one obtains the following estimates $\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_m$ of $\mu, \alpha_1, \ldots, \alpha_m$ respectively:

$$\begin{aligned} \hat{\mu} &= \overline{Y}_{\bullet\bullet} \\ \hat{\alpha}_1 &= \overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet} \\ &\vdots \\ \hat{\alpha}_m &= \overline{Y}_{m\bullet} - \overline{Y}_{\bullet\bullet} \end{aligned} \tag{6}$$

with

$$\overline{Y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij} \tag{7}$$

$$\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \qquad (i = 1, \ldots, m).$$

The so-called analysis of variances table has the following form (Table 2)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean square	Expected Mean Square
Between treatment groups	$m-1$	SQA	$\text{MQA} = \frac{\text{SQA}}{m-1}$	$\sigma^2 + \frac{\sum_{i=1}^{m} n_i \alpha_i^2}{m-1}$
Within treatment groups	$N-m$	SQR	$\text{MQR} = \frac{\text{SQR}}{N-m}$	σ^2
Total	$N-1$	SQG	-	-

Table 2. Analysis of Variances

with

$$\text{SQA} = \sum_{i=1}^{m} n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \tag{8}$$

$$\text{SQR} = \sum_{i=1}^{m} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{i\bullet})^2 \tag{9}$$

$$\text{SQG} = \sum_{i=1}^{m} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{\bullet\bullet})^2 \tag{10}$$

and

$$\text{SQG} = \text{SQA} + \text{SQR}. \tag{11}$$

For testing the hypothesis $H_0 : \alpha_1 = ... = \alpha_m = 0$ one uses the quotient $\frac{\text{MQA}}{\text{MQR}}$ as a (test) statistic. If the hypothesis H_0 is true, then the statistic is F-distributed with $(m-1, N-M)$ degrees of freedom. That means the identity of the m treatment means $\alpha_1, ..., \alpha_m$ can be statistically verified by a test based on the comparison of the mean square between the treatment groups and the mean square within the treatment groups.

-
-
-

Bibliography

Hocking R.R. (1996). *Methods and Applications of Linear Models*. New York: John Wiley & Sons, Inc. [This book presents a thorough treatment of the concepts and methods of linear models analysis in statistics and illustrates them with numerical and conceptual examples].

Kshirsagar A.M. (1983). *A Course in Linear Models*. New York: Marcel Dekker, Inc. [This is a text book, which describes the main aspects of general linear models of statistics including ANOVA (models I and II) as well as analysis of covariance].

Müller P.H. (ed.) (1991). *Lexikon der Stochastik*. 5. Auflage. Berlin: Akademie-Verlag. [This is a dictionary for all fields of Stochastics with comprehensive description of ANOVA].

Nollau V. (1979) *Statistische Analysen*. 2. Auflage. Basel und Stuttgart: Birkhäuser. [This book presents all statistical methods based on general linear models of statistics including ANOVA (models I and II)].

Rao C.R. (1973). *Linear Statistical Inference and Its Application*. New York: John Wiley & Sons, Inc. [This is an excellent book on linear models in statistics].

Sahai H. and Ageel M.I. (2000). *The Analysis of Variance*. Boston: Birkhäuser. [The authors consider the analysis of variance on models I, II, and III. In this way this book is a very modern dictionary of ANOVA].

Scheffé H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons, Inc. [Since more than 40 years this book is a standard one in the fields of ANOVA].

Biographical Sketch

V. Nollau was born in 1941. After studying at secondary school he entered Technical University of Dresden (Germany) to study mathematics and theoretical physics. He graduated in 1964, obtaining doctorate there in 1966 and 1971 (Dr. habil.). From 1969 he was assistant professor at TU Dresden. His main research topics were operator theory, stochastic processes and random search. In 1972 he made the first contributions to stochastic optimization and decision processes theory. Since 1990 he is professor for stochastic analysis and control. He wrote several text works including "Statistische Analysen" (Linear Models in Statistics). He is now dean of the faculty of mathematics in Dresden.