

Chapter 4

Introduction to Analysis-of-Variance Procedures

Chapter Table of Contents

OVERVIEW	55
STATISTICAL DETAILS FOR ANALYSIS OF VARIANCE	56
Definitions	56
Fixed and Random Effects	56
Tests of Effects	57
General Linear Models	58
Linear Hypotheses	58
ANALYSIS OF VARIANCE FOR FIXED EFFECT MODELS	59
PROC GLM for General Linear Models	59
PROC ANOVA for Balanced Designs	59
Comparing Group Means with PROC ANOVA and PROC GLM	60
PROC TTEST for Comparing Two Groups	60
ANALYSIS OF VARIANCE FOR MIXED AND RANDOM EFFECT MODELS	61
ANALYSIS OF VARIANCE FOR CATEGORICAL DATA AND GENER- ALIZED LINEAR MODELS	61
NONPARAMETRIC ANALYSIS OF VARIANCE	62
CONSTRUCTING ANALYSIS OF VARIANCE DESIGNS	62
REFERENCES	63

Chapter 4

Introduction to Analysis-of-Variance Procedures

Overview

This chapter reviews the SAS/STAT software procedures that are used for analysis of variance: GLM, ANOVA, CATMOD, MIXED, NESTED, NPAR1WAY, TRANSREG, TTEST, and VARCOMP. Also discussed are SAS/STAT and SAS/QC software procedures for constructing analysis of variance designs: PLAN, FACTEX, and OPTEX.

The flagship analysis-of-variance procedure is the GLM procedure, which handles most standard problems. The following are descriptions of PROC GLM and other procedures that are used for more specialized situations:

ANOVA	performs analysis of variance, multivariate analysis of variance, and repeated measures analysis of variance for <i>balanced</i> designs. PROC ANOVA also performs several multiple comparison tests.
CATMOD	fits linear models and performs analysis of variance and repeated measures analysis of variance for categorical responses.
GENMOD	fits generalized linear models and performs analysis of variance in the generalized linear models framework. The methods are particularly suited for discrete response outcomes.
GLM	performs analysis of variance, regression, analysis of covariance, repeated measures analysis, and multivariate analysis of variance. PROC GLM produces several diagnostic measures, performs tests for random effects, provides contrasts and estimates for customized hypothesis tests, performs several multiple comparison tests, and provides tests for means adjusted for covariates.
MIXED	performs mixed-model analysis of variance and repeated measures analysis of variance via covariance structure modeling. Using likelihood-based or method-of-moment estimates, PROC MIXED constructs statistical tests and intervals, allows customized contrasts and estimates, and computes empirical Bayes predictions.
NESTED	performs analysis of variance and analysis of covariance for purely nested random models.
NPAR1WAY	performs nonparametric one-way analysis of rank scores.
TTEST	compares the means of two groups of observations.
TRANSREG	fits univariate and multivariate linear models, optionally with spline and other nonlinear transformations.

VARCOMP estimates variance components for random or mixed models.

The following section presents an overview of some of the fundamental features of analysis of variance. Subsequent sections describe how this analysis is performed with procedures in SAS/STAT software. For more detail, see the chapters for the individual procedures. Additional sources are described in the “References” section on page 63.

Statistical Details for Analysis of Variance

Definitions

Analysis of variance (ANOVA) is a technique for analyzing experimental data in which one or more *response* (or *dependent* or simply Y) variables are measured under various conditions identified by one or more classification variables. The combinations of levels for the classification variables form the cells of the experimental design for the data. For example, an experiment may measure weight change (the dependent variable) for men and women who participated in three different weight-loss programs. The six cells of the design are formed by the six combinations of sex (men, women) and program (A, B, C).

In an analysis of variance, the variation in the response is separated into variation attributable to differences between the classification variables and variation attributable to random error. An analysis of variance constructs tests to determine the significance of the classification effects. A typical goal in an analysis of variance is to compare means of the response variable for various combinations of the classification variables.

An analysis of variance may be written as a linear model. Analysis of variance procedures in SAS/STAT software use the model to predict the response for each observation. The difference between the actual and predicted response is the *residual error*. Most of the procedures fit model parameters that minimize the sum of squares of residual errors. Thus, the method is called *least squares regression*. The variance due to the random error, σ^2 , is estimated by the mean squared error (MSE or s^2).

Fixed and Random Effects

The explanatory classification variables in an ANOVA design may represent fixed or random effects. The levels of a classification variable for a fixed effect give all the levels of interest, while the levels of a classification variable for a random effect are typically a subset of levels selected from a population of levels. The following are examples.

- In a large drug trial, the levels that correspond to types of drugs are usually considered to comprise a fixed effect, but the levels corresponding to the various clinics where the drugs are administered comprise a random effect.

- In agricultural experiments, it is common to declare locations (or plots) as random because the levels are chosen randomly from a large population of locations and you assume fertility to vary normally across locations.
- In repeated-measures experiments with people or animals as subjects, subjects are declared random because they are selected from the larger population to which you want to generalize.

A typical assumption is that random effects have values drawn from a normally distributed random process with mean zero and common variance. Effects are declared random when the levels are randomly selected from a large population of possible levels. Inferences are made using only a few levels but can be generalized across the whole population of random effects levels.

The consequence of having random effects in your model is that some observations are no longer uncorrelated but instead have a covariance that depends on the variance of the random effect. In fact, a more general approach to random effect models is to model the covariance between observations.

Tests of Effects

Analysis of variance tests are constructed by comparing independent mean squares. To test a particular null hypothesis, you compute the ratio of two mean squares that have the same expected value under that hypothesis; if the ratio is much larger than 1, then that constitutes significant evidence against the null. In particular, in an analysis-of-variance model with fixed effects only, the expected value of each mean square has two components: quadratic functions of fixed parameters and random variation. For example, for a fixed effect called A, the expected value of its mean square is

$$E(\text{MS}(A)) = Q(\beta) + \sigma_e^2$$

Under the null hypothesis of no A effect, the fixed portion $Q(\beta)$ of the expected mean square is zero. This mean square is then compared to another mean square, say $\text{MS}(E)$, that is independent of the first and has expected value σ_e^2 . The ratio of the two mean squares

$$F = \frac{\text{MS}(A)}{\text{MS}(E)}$$

has the F distribution under the null hypothesis. When the null hypothesis is false, the numerator term has a larger expected value, but the expected value of the denominator remains the same. Thus, large F values lead to rejection of the null hypothesis. The probability of getting an F value at least as large as the one observed given that the null hypothesis is true is called the *significance probability value* (or the p -value). A p -value of less than 0.05, for example, indicates that data with *no* real A effect will yield F values as large as the one observed less than 5% of the time. This is usually considered moderate evidence that there *is* a real A effect. Smaller p -values constitute even stronger evidence. Larger p -values indicate that the effect of interest

is less than random noise. In this case, you can conclude either that there is no effect at all or that you do not have enough data to detect the differences being tested.

General Linear Models

An analysis-of-variance model can be written as a linear model, which is an equation that predicts the response as a linear function of parameters and design variables. In general,

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad i = 1, 2, \dots, n$$

where y_i is the response for the i th observation, β_k are unknown parameters to be estimated, and x_{ij} are design variables. Design variables for analysis of variance are indicator variables; that is, they are always either 0 or 1.

The simplest model is to fit a single mean to all observations. In this case there is only one parameter, β_0 , and one design variable, x_{0i} , which always has the value of 1:

$$\begin{aligned} y_i &= \beta_0 x_{0i} + \epsilon_i \\ &= \beta_0 + \epsilon_i \end{aligned}$$

The least-squares estimator of β_0 is the mean of the y_i . This simple model underlies all more complex models, and all larger models are compared to this simple mean model. In writing the parameterization of a linear model, β_0 is usually referred to as the *intercept*.

A one-way model is written by introducing an indicator variable for each level of the classification variable. Suppose that a variable A has four levels, with two observations per level. The indicator variables are created as follows:

Intercept	A1	A2	A3	A4
1	1	0	0	0
1	1	0	0	0
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	1	0
1	0	0	0	1
1	0	0	0	1

The linear model for this example is

$$y_i = \beta_0 + \beta_1 A1_i + \beta_2 A2_i + \beta_3 A3_i + \beta_4 A4_i$$

To construct crossed and nested effects, you can simply multiply out all combinations of the main-effect columns. This is described in detail in “Specification of Effects” in Chapter 28, “The GLM Procedure.”

Linear Hypotheses

When models are expressed in the framework of linear models, hypothesis tests are expressed in terms of a linear function of the parameters. For example, you may want to test that $\beta_2 - \beta_3 = 0$. In general, the coefficients for linear hypotheses are some set of L s:

$$H_0: L_0\beta_0 + L_1\beta_1 + \cdots + L_k\beta_k = 0$$

Several of these linear functions can be combined to make one joint test. These tests can be expressed in one matrix equation:

$$H_0: \mathbf{L}\boldsymbol{\beta} = 0$$

For each linear hypothesis, a sum of squares (SS) due to that hypothesis can be constructed. These sums of squares can be calculated either as a quadratic form of the estimates

$$SS(\mathbf{L}\boldsymbol{\beta} = 0) = (\mathbf{L}\mathbf{b})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\mathbf{b})$$

or, equivalently, as the increase in sums of squares for error (SSE) for the model constrained by the null hypothesis

$$SS(\mathbf{L}\boldsymbol{\beta} = 0) = SSE(\text{constrained}) - SSE(\text{full})$$

This SS is then divided by appropriate degrees of freedom and used as a numerator of an F statistic.

Analysis of Variance for Fixed Effect Models

PROC GLM for General Linear Models

The GLM procedure is the flagship tool for analysis of variance in SAS/STAT software. It performs analysis of variance by using least squares regression to fit general linear models, as described in the section “General Linear Models” on page 58. Among the statistical methods available in PROC GLM are regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation.

While PROC GLM can handle most common analysis of variance problems, other procedures are more efficient or have more features than PROC GLM for certain specialized analyses, or they can handle specialized models that PROC GLM cannot. Much of the rest of this chapter is concerned with comparing PROC GLM to other procedures.

PROC ANOVA for Balanced Designs

When you design an experiment, you choose how many experimental units to assign to each combination of levels (or cells) in the classification. In order to achieve good statistical properties and simplify the computations, you typically attempt to assign the same number of units to every cell in the design. Such designs are called *balanced designs*.

In SAS/STAT software, you can use the ANOVA procedure to perform analysis of variance for balanced data. The ANOVA procedure performs computations for analysis of variance that assume the balanced nature of the data. These computations are simpler and more efficient than the corresponding general computations performed by PROC GLM. Note that PROC ANOVA can be applied to certain designs that are not balanced in the strict sense of equal numbers of observations for all cells. These additional designs include all one-way models, regardless of how unbalanced the cell counts are, as well as Latin squares, which do not have data in all cells. In general, however, the ANOVA procedure is recommended only for balanced data. **If you use ANOVA to analyze a design that is not balanced, you must assume responsibility for the validity of the output.** You are responsible for recognizing incorrect results, which may include negative values reported for the sums of squares. If you are not certain that your data fit into a balanced design, then you probably need the framework of general linear models in the GLM procedure.

Comparing Group Means with PROC ANOVA and PROC GLM

When you have more than two means to compare, an F test in PROC ANOVA or PROC GLM tells you whether the means are significantly different from each other, but it does not tell you which means differ from which other means.

If you have specific comparisons in mind, you can use the CONTRAST statement in PROC GLM to make these comparisons. However, if you make many comparisons using some given significance level (0.05, for example), you are more likely to make a type 1 error (incorrectly rejecting a hypothesis that the means are equal) simply because you have more chances to make the error.

Multiple comparison methods give you more detailed information about the differences among the means and enables you to control error rates for a multitude of comparisons. A variety of multiple comparison methods are available with the MEANS statement in both the ANOVA and GLM procedures, as well as the LSMEANS statement in PROC GLM. These are described in detail in “Multiple Comparisons” in Chapter 28, “The GLM Procedure.”

PROC TTEST for Comparing Two Groups

If you want to perform an analysis of variance and have only one classification variable with two levels, you can use PROC TTEST. In this special case, the results generated by PROC TTEST are equivalent to the results generated by PROC ANOVA

or PROC GLM. In addition to testing for differences between two groups, PROC TTEST performs a test for unequal variances. You can use PROC TTEST with balanced or unbalanced groups. The PROC NPAR1WAY procedure performs nonparametric analogues to t tests. See Chapter 12, “Nonparametric Tests,” for an overview and Chapter 42 for details on PROC NPAR1WAY.

Analysis of Variance for Mixed and Random Effect Models

Just as PROC GLM is the flagship procedure for fixed effect analysis of variance models, PROC MIXED is the flagship procedure for random and mixed effect models. The MIXED procedure fits a variety of mixed linear models to data and enables you to use these fitted models to make statistical inferences about the data. The default fitting method maximizes the restricted likelihood of the data under the assumption that the data are normally distributed and any missing data are missing at random. This general framework accommodates many common correlated-data methods, including variance component models and repeated measures analyses.

A few other procedures in SAS/STAT software offer limited mixed-model capabilities. PROC GLM fits some random-effects and repeated-measures models, although its methods are based on method-of-moments estimation and a portion of the output applies only to the fixed-effects model. PROC NESTED fits special nested designs and may be useful for large data sets because of its customized algorithms. PROC VARCOMP estimates variance components models, but all of its methods are now available in PROC MIXED. PROC LATTICE fits special balanced lattice designs, but again, the same models are available in PROC MIXED. In general, PROC MIXED is recommended for nearly all of your mixed-model applications.

Analysis of Variance for Categorical Data and Generalized Linear Models

A *categorical variable* is defined as one that can assume only a limited number of values. For example, a person’s sex is a categorical variable that can assume one of two values. Variables with levels that simply name a group are said to be measured on a *nominal scale*. Categorical variables can also be measured using an *ordinal scale*, which means that the levels of the variable are ordered in some way. For example, responses to an opinion poll are usually measured on an ordinal scale, with levels ranging from “strongly disagree” to “no opinion” to “strongly agree.”

For two categorical variables, one measured on an ordinal scale and one measured on a nominal scale, you may assign scores to the levels of the ordinal variable and test whether the mean scores for the different levels of the nominal variable are significantly different. This process is analogous to performing an analysis of variance on continuous data, which can be performed by PROC CATMOD. If there are n nominal variables, rather than 1, then PROC CATMOD can do an n -way analysis of variance of the mean scores.

For two categorical variables measured on a nominal scale, you can test whether the distribution of the first variable is significantly different for the levels of the second variable. This process is an analysis of variance of proportions, rather than means, and can be performed by PROC CATMOD. The corresponding n -way analysis of variance can also be performed by PROC CATMOD.

See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” and Chapter 20, “The CATMOD Procedure,” for more information.

GENMOD uses maximum likelihood estimation to fit generalized linear models. This family includes models for categorical data such as logistic, probit, and complementary log-log regression for binomial data and Poisson regression for count data, as well as continuous models such as ordinary linear regression, gamma and inverse Gaussian regression models. GENMOD performs analysis of variance through likelihood ratio and Wald tests of fixed effects in generalized linear models, and provides contrasts and estimates for customized hypothesis tests. It performs analysis of repeated measures data with generalized estimating equation (GEE) methods.

See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” and Chapter 27, “The GENMOD Procedure,” for more information.

Nonparametric Analysis of Variance

Analysis of variance is sensitive to the distribution of the error term. If the error term is not normally distributed, the statistics based on normality can be misleading. The traditional test statistics are called *parametric tests* because they depend on the specification of a certain probability distribution except for a set of free parameters. Parametric tests are said to depend on distributional assumptions. Nonparametric methods perform the tests without making any strict distributional assumptions. Even if the data are distributed normally, nonparametric methods are often almost as powerful as parametric methods.

Most nonparametric methods are based on taking the ranks of a variable and analyzing these ranks (or transformations of them) instead of the original values. The NPAR1WAY procedure performs a nonparametric one-way analysis of variance. Other nonparametric tests can be performed by taking ranks of the data (using the RANK procedure) and using a regular parametric procedure (such as GLM or ANOVA) to perform the analysis. Some of these techniques are outlined in the description of PROC RANK in the *SAS Procedures Guide* and in Conover and Iman (1981).

Constructing Analysis of Variance Designs

Analysis of variance is most often used for data from designed experiments. You can use the PLAN procedure to construct designs for many experiments. For exam-

ple, PROC PLAN constructs designs for completely randomized experiments, randomized blocks, Latin squares, factorial experiments, and balanced incomplete block designs.

Randomization, or randomly assigning experimental units to cells in a design and to treatments within a cell, is another important aspect of experimental design. For either a new or an existing design, you can use PROC PLAN to randomize the experimental plan.

Additional features for design of experiments are available in SAS/QC software. The FACTEX and OPTTEX procedures can construct a wide variety of designs, including factorials, fractional factorials, and D-optimal or A-optimal designs. These procedures, as well as the ADX Interface, provide features for randomizing and replicating designs; saving the design in an output data set; and interactively changing the design by changing its size, use of blocking, or the search strategies used. For more information, see *SAS/QC Software: Reference*.

References

Analysis of variance was pioneered by R.A. Fisher (1925). For a general introduction to analysis of variance, see an intermediate statistical methods textbook such as Steel and Torrie (1980), Snedecor and Cochran (1980), Milliken and Johnson (1984), Mendenhall (1968), John (1971), Ott (1977), or Kirk (1968). A classic source is Scheffe (1959). Freund, Littell, and Spector (1991) bring together a treatment of these statistical methods and SAS/STAT software procedures. Schlotzhauer and Littell (1997) cover how to perform *t* tests and one-way analysis of variance with SAS/STAT procedures. Texts on linear models include Searle (1971), Graybill (1976), and Hocking (1984). Kennedy and Gentle (1980) survey the computing aspects.

Conover, W.J. and Iman, R.L. (1981), "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124–129.

Fisher, R.A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Freund, R.J., Littell, R.C., and Spector, P.C. (1991), *SAS System for Linear Models*, Cary, NC: SAS Institute Inc.

Graybill, F.A. (1976), *Theory and Applications of the Linear Model*, North Scituate, MA: Duxbury Press.

Hocking, R.R. (1984), *Analysis of Linear Models*, Monterey, CA: Brooks-Cole Publishing Co.

John, P. (1971), *Statistical Design and Analysis of Experiments*, New York: Macmillan Publishing Co.

Kennedy, W.J., Jr. and Gentle, J.E. (1980), *Statistical Computing*, New York: Marcel Dekker, Inc.

Kirk, R.E. (1968), *Experimental Design: Procedures for the Behavioral Sciences*,

Monterey, CA: Brooks-Cole Publishing Co.

Mendenhall, W. (1968), *Introduction to Linear Models and the Design and Analysis of Experiments*, Belmont, CA: Duxbury Press.

Milliken, G.A. and Johnson, D.E. (1984), *Analysis of Messy Data Volume I: Designed Experiments*, Belmont, CA: Lifetime Learning Publications.

Ott, L. (1977), *Introduction to Statistical Methods and Data Analysis*, Second Edition, Belmont, CA: Duxbury Press.

Scheffe, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons, Inc.

Schlotzhauer, S.D. and Littell, R.C. (1997), *SAS System for Elementary Statistical Analysis*, Cary, NC: SAS Institute Inc.

Searle, S.R. (1971), *Linear Models*, New York: John Wiley & Sons, Inc.

Snedecor, G.W. and Cochran, W.G. (1980), *Statistical Methods*, Seventh Edition, Ames, IA: Iowa State University Press.

Steel R.G.D. and Torrie, J.H. (1980), *Principles and Procedures of Statistics*, Second Edition, New York: McGraw-Hill Book Co.