Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Introduction

Analysis of Variance (ANOVA) is one of the most widely used techniques in data analysis. For example, this next data set which is contained in the test file `anova.tf1` is for six replicate estimates for strontium concentrations (mg/ml) in five different locations, and it is wished to test if there are significant differences between the population means based on the sample means as listed in the last row.

	28.2	39.6	46.3	41.0	56.3
	33.2	40.8	42.1	44.1	54.1
	36.4	37.9	43.5	46.4	59.4
	34.6	37.1	48.8	40.2	62.7
	29.1	43.6	43.7	38.6	60.0
	31.0	42.4	40.1	36.3	57.3
Means	32.1	40.2	44.1	41.1	58.3

In the subsequent discussion concerning ANOVA it will be assumed that the reader is familiar with the normal, chi-square, and F distributions, and statistical tests based on them as described in the appropriate SimFIT tutorial documents. In particular, the Shapiro-Wilks test for normality and the Bartlett or Levene tests for homogeneity of variance could be used by purists determined to check if ANOVA is justified, because it should be pointed out that ANOVA is often used uncritically where better techniques may be more appropriate.

The basic theory

In studying the distribution of the variance estimate from a sample of size n from a normal distribution with mean μ and variance σ^2, you will have encountered the following decomposition of a sum of squares

$$\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma}\right)^2+\left(\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}\right)^2$$

into independent chi-square variables with $n-1$ and 1 degrees of freedom respectively. Analysis of variance is an extension of this procedure based on linear models, assuming normality and constant variance, then partitioning of chi-square variables into two or more independent components, invoking Cochran's theorem and comparing the ratios to F variables with the appropriate degrees of freedom for variance ratio tests. It can be used, for instance, when you have a set of samples (column vectors) that come from normal distributions with the same variance and wish to test if all the samples have the same mean. Due to the widespread use of this technique, many people use it even though the original data are not normally distributed with the same variance, by first applying variance stabilizing transformations, like the square root with counts, which can sometimes transform non-normal data into transformed data that are approximately normally distributed. Note that you should never make the common mistake of supposing that ANOVA is model free: ANOVA is always based upon data collected as replicates and organized into groups, where it is assumed that all the data are normally distributed with the same variance but with mean values that differ from cell to cell according to an assumed linear model.

Variance stabilizing transformations

A number of transformations are in use that attempt to create new data that is more approximately normally distributed than the original data, or at least has more constant variance, as the two aims can not usually both be achieved. If the distribution of a random variable X is known, then the variance of a function of X can in

some cases be calculated explicitly. However, to a very crude first approximation, if a random variable X is transformed by $Y = f(X)$, then the variances are related by the differential equation

$$V(Y) \approx \left(\frac{dY}{dX}\right)^2 V(X)$$

which yields $f(.)$ on integration, e.g. for constant variance where $V(Y) = k$ for some constant k would be required, given $V(X)$.

Note that SIMFIT provides the ability to explore the commonly used transformations, to be discussed next, whenever ANOVA or tests for homogeneity of variance are used.

The angular transformation

This arcsine transformation is sometimes used for binomial data with parameters N and p, e.g., for X successes in N trials, when

$$\begin{aligned} X &\sim b(N, p) \\ Y &= \arcsin(\sqrt{X/N}) \\ E(Y) &\simeq \arcsin(\sqrt{p}) \\ V(Y) &\simeq 1/(4N) \text{ (using radial measure).} \end{aligned}$$

However, note that the variance of the transformed data is only constant in situations where there are constant binomial denominators.

The square root transformation

This is often used for counts, e.g., for Poisson variables with mean μ, when

$$\begin{aligned} X &\sim \text{Poisson}(\mu) \\ Y &= \sqrt{x} \\ E(Y) &\simeq \sqrt{\mu} \\ V(Y) &\simeq 1/4. \end{aligned}$$

The log transformation

When the variance of X is proportional to a known power α of $E(X)$, then the power transformation $Y = X^\beta$ will stabilize variance for $\beta = 1 - \alpha/2$. The angular and square root transformations are, of course, just special cases of this, but a singular case of interest is the constant coefficient of variation situation $V(X) \propto E(X)^2$ which justifies the log transform, as follows

$$\begin{aligned} E(X) &= \mu \\ V(X) &\propto \mu^2 \\ Y &= \log X \\ V(Y) &= k, \text{ a constant.} \end{aligned}$$

Overview of 1-way ANOVA

As this is the most frequently encountered situation and is the model for subsequent variants it will be discussed in some detail.

This procedure is used when you have groups (i.e. samples) of normally distributed measurements with the same variance and wish to test if all the population means are equal. With two groups it is equivalent to the two-sample unpaired t test, so it can be regarded as an extension of this test to cases with more than two groups. Suppose a random variable Y is measured for groups $i = 1, 2, \ldots, k$ and subjects $j = 1, 2, \ldots n_i$, and it is assumed that the appropriate general linear model for the $n = \sum_{i=1}^{k} n_i$ observations is

$$y_{ij} = \mu + \alpha_i + e_{ij}$$
$$\text{where } \sum_{i=1}^{k} \alpha_i = 0$$

and the errors e_{ij} are independently normally distributed with zero mean and common variance σ^2.

Then the 1-way ANOVA null hypothesis is

$$H_0 : \alpha_i = 0, \text{ for } i = 1, 2, \ldots, k,$$

that is, the means for all k groups are equal, and the basic equations are as follows.

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$$
$$\bar{y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}/n$$
$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$
$$\text{Total } SSQ = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2, \text{ with } DF = n - 1$$
$$\text{Residual } SSQ = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \text{ with } DF = n - k$$
$$\text{Group } SSQ = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2, \text{ with } DF = k - 1.$$

Here Total SSQ is the overall sum of squares, Group SSQ is the between groups (i.e. among groups) sum of squares, and Residual SSQ is the residual (i.e. within groups, or error) sum of squares. The mean sums of squares and F value can be calculated from these using

$$\text{Total } SSQ = \text{ Residual } SSQ + \text{ Group } SSQ$$
$$\text{Total } DF = \text{ Residual } DF + \text{ Group } DF$$
$$\text{Group } MS = \frac{\text{Group } SSQ}{\text{Group } DF}$$
$$\text{Residual MS } = \frac{\text{Residual } SSQ}{\text{Residual } DF}$$
$$F = \frac{\text{Group } MS}{\text{Residual } MS},$$

so that the degrees of freedom for the F variance ratio to test if the between groups MS is significantly larger than the residual MS are $k - 1$ and $n - k$. The SIMFIT 1-way ANOVA procedure allows you to include or exclude selected groups, i.e., data columns, and to employ variance stabilizing transformations if required, but it also provides a nonparametric test, and it allows you to explore which group or groups differ significantly in the event of the F value leading to a rejection of H_0.