

A new graphic for one-way ANOVA

Bob Pruzek, SUNY Albany [July 2011]

This document describes and illustrates a new *elemental graphic* for one-way analysis of variance, i.e., ANOVA. The primary motivation for developing the central function was to facilitate a deeper understanding of the key features of analysis of variance by focusing on the central question of the method in the context of using modern graphics that can facilitate sound data analyses. It is also hoped that use of this function will facilitate development of modern data-analytic thinking and skills in ANOVA applications. Note that the `granova.1w` function in R that produces this graphic is found in package `granova`. (The `granova` package is currently in version 2.0; however, a parallel package called `GGgranova` is currently [July 13, 2011] being finalized, one based on the `ggplot2` package by Hadley Wickham. In addition, a comprehensive article (by R.M. Pruzek and J. Helmreich) that documents all functions in `granova` has been submitted for publication; it is entitled *Elemental Graphics for Analysis of Variance using the R Package granova.*)

The key (omnibus) F statistic at the heart of any inferential application of the simplest ANOVA model implies a particular way to compare means, one based on *data-based contrasts*. Indeed, there appears to be one distinctive approach to comparing several groups of quantitative data that is wholly consistent with the central question that drives (one-way) ANOVA, and that leads directly to what might be called an *elemental graphic* for this method. The method permits visualization of all data points for any number of groups where all group means necessarily lie on a straight line for any data system. (It is also straightforward to generalize the display to accommodate rows and columns in any two-way ANOVA as well; see page 7 below.) Additionally, this one-way ANOVA graphic can be used to visualize residuals, basic ‘effects,’ as well as all data points, remaining faithful to the central question that drives the method. A key feature of the graphic is that it facilitates visualization of the central variance estimates, i.e., the Mean Squares Between and Within, by displaying certain squares, the sides of which are based on standard deviation units. It follows that the conventional F statistic can be seen as a ratio of the areas of these squares.

It is anticipated that students, and users generally, using this function will better understand the basic principles of analysis of variance, and applied researchers will be able to better understand their data. In addition, as discussed briefly below, the function is easily used in to study the effects of various changes in one’s data, or to visualize the results of simulations, or the effects of repeated sampling (as in bootstrapping). Also, since any data system admits to alternative transformations, or re-expressions, the graphic can facilitate a better understanding of how choice of transformation effects not just means and variances and other summary statistics, but can help qualify or support inferences, or see the role of individual data points in their respective groups.

The initial illustration of the function uses *data* from a One Way ANOVA website:

<http://espse.ed.psu.edu/statistics/Chapters/Chapter11/Chap11.html#A%20Statlets%20Problem>

The function `granova.1w` is used below for the display and analyses of these data:

| | | |
|-------------|----------------------|---|
| <i>Data</i> | A: 26 25 29 21 20 18 | ← each row constitutes a treatment group, n = 6 each |
| | B: 24 17 16 13 21 19 | [NB: input of these data with the function can |
| | C: 32 34 29 19 27 28 | be based on input of yy , the <i>transpose</i> of this matrix; |
| | D: 23 29 26 20 24 25 | <i>i.e.</i> , input rows as columns of a matrix.] |

The data source (the website) states: “The data [were] collected by a research group investigating the whitening power of four new toothpaste formulas. The dependent variable ... (response) is a whiteness measure where the lower the number, the whiter the teeth. The independent variable Group codes the four different toothpaste formulas using letters” A - D.”

The function prints a number of standard numerical results and constructs the graphic. In its simplest forms, with all but one logical argument set to FALSE, the graphic shows only the essentials of ANOVA, each group of score values being centered on its mean. (Details about the arguments for the function are provided below.) Given that M_j represents the arithmetic mean of the j th group, and M , no subscript, represents the grand mean the contrast coefficients are constructed, each of form $c_j^* = M_j - M$, ordered from the smallest to the largest mean, say, from $j = 1 \dots J$. Each coefficient of the form $M_j - M$ is generally called an ‘effect’ in one-way ANOVA, and *the statistical (F) test for this method simply examines whether the effects cum contrasts are large enough in magnitude, relative to variation within groups, to conclude that their population counterparts are not all zero*. The graphic derives especially from the Sum of Squares Between in the numerator of the MS(Between) of the F-statistic, which has the form $SS(B) = \sum n_j (M_j - M)^2$; a little algebra shows that latter can be written as $\sum c_j n_j M_j$ where $c_j = n_j c_j^* = n_j (M_j - M)$, a data-based contrast or linear combination of group means defined by the c_j 's. The MS(W) is just the average of the respective subgroup variances if n 's are equal, a weighted average of group variances if the n_j 's differ across groups; one can think of MS(W) as a ‘scaling’ factor for MS(Between) where F is computed as MS(B)/MS(W).

To summarize the key idea of the graphic, the statistic F, which is central to one-way ANOVA, implies a particular way to compare (unstructured) group means, one based on a linear combination of group means for which the weights are data-based contrast coefficients. The denominator of F, the scaling factor, is just the average of the variances for the independent groups.

After organizing groups according to the sizes of the contrasts (from smallest to largest) and plotting all scores, *i.e.*, the $y(i,j)$ (vertically), and using a special symbol (a red triangle) to denote group means, a little reflection will show that the group means necessarily fall on a straight line. The two sets of numerical outputs provided by `granova.lw` are shown in Table 1, and the figure of principal interest is shown on page 2.

Table 1 Summary statistics for an illustrative application of one-way ANOVA

\$grandsum

| Grandmean | df.bet | df.with | MS.bet | MS.with | F.stat | F.prob | SS.bet/SS.tot |
|-----------|--------|---------|--------|---------|--------|--------|---------------|
| 23.54 | 3.00 | 20.00 | 99.15 | 17.12 | 5.79 | 0.01 | 0.46 |

\$stats

| | Size | Contrast Coef | Wt'd Mean | Mean | Trim'd Mean | Var. | St. Dev. |
|---|------|---------------|-----------|-------|-------------|-------|----------|
| 2 | 6 | -5.21 | 18.33 | 18.33 | 18.25 | 15.07 | 3.88 |
| 1 | 6 | -0.38 | 23.17 | 23.17 | 23.00 | 17.37 | 4.17 |
| 4 | 6 | 0.96 | 24.50 | 24.50 | 24.50 | 9.10 | 3.02 |
| 3 | 6 | 4.62 | 28.17 | 28.17 | 29.00 | 26.97 | 5.19 |

(NB: Examine numerical information first, what it alone provides, then on how the graphic presentation below complements & extends your understanding of these data.)

The graphic that follows illustrates how this works in the case of this particularly simple set of data, one with four groups, each of size six, for the toothpaste experiment. The function is called as: `granova.lw(teeth.df,main='ANOVA graphic for teeth whiteness data')`, where `teeth.df` is a 6 x 4 matrix (a `data.frame`) whose columns correspond to groups (see above; *i.e.*, the transpose of what is seen above). [On a Mac it can help to add `args kx=1.3 & px=1.3` to the call.]

ANOVA graphic for teeth whitening data, N=24

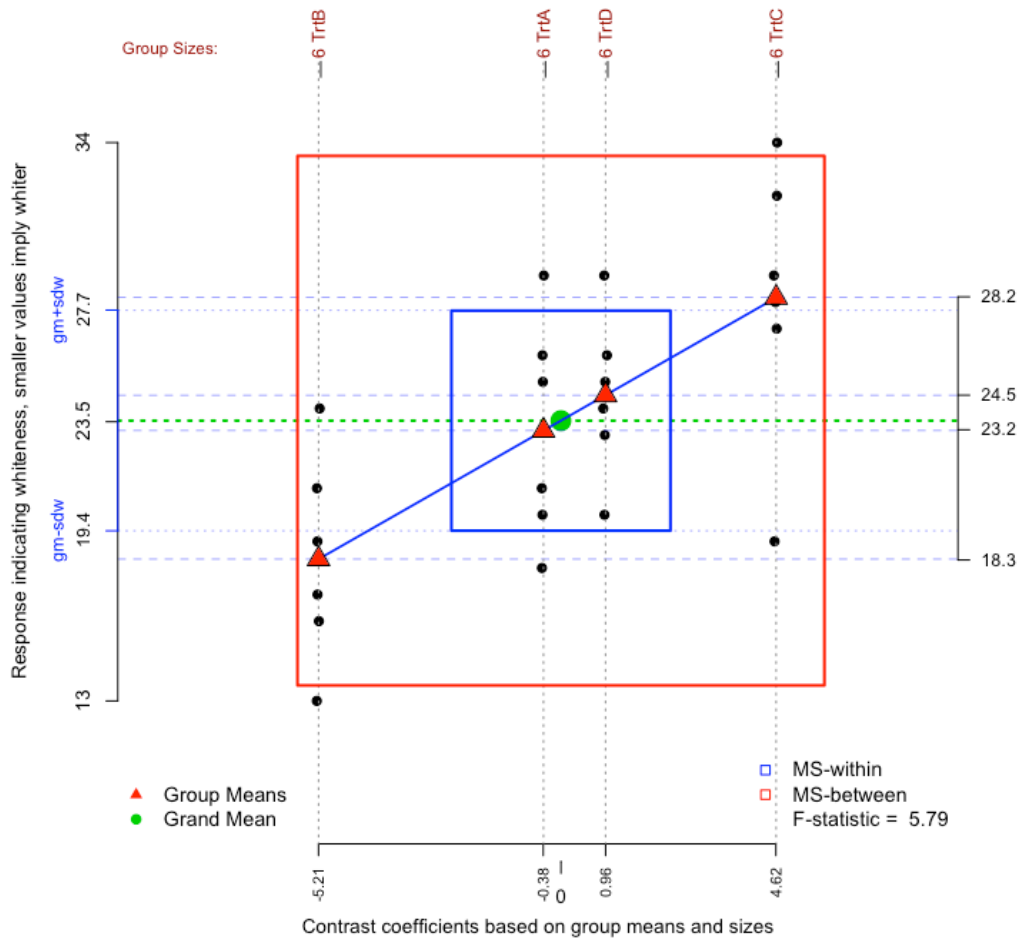


Figure 1 Graphic that illustrates use of function `granova.1w`

Figure 1 depicts a scatterplot of *contrast coefficients* (column two of `$stats` above) for the ordered groups, smallest to largest means, versus the sets of scores for the respective groups, for the horizontal and vertical axes respectively. Each contrast coefficient attaches to a single score, according to that score's group so there are n_j contrast coefficients for each group. The contrast coefficients have been jittered to help distinguish points (see below) and the means of groups are signified by (red) triangles within each of the score-sets; also means themselves are printed on the right margin. The plot of the means necessarily follows a straight line because the means are essentially being plotted against themselves, except that the grand mean is initially subtracted from the first set of means. The grand mean [here, 23.54] is coupled with the contrast coefficient of zero, visualized as the green point at the center of the plot.

The overlaying squares in the center of the plot correspond to the within group estimate of variance (blue) [MS(W)] & the between group (red) variance estimate [MS(B)]; each side (of each square) corresponds to two standard deviation units so that their areas correspond to variances. The left side of the figure shows numerical values of edges for the MS(W). The ratio of the areas of the two squares, the red area divided by the blue, is just the F-statistic. In this case, the area of the red square is 5.79 times larger than area of blue square, so it is seen that $F = 5.79$. Study of the two squares, preferably for several examples, is likely to help you see, perhaps for the first time, how variances (*i.e.*, mean squares) can be represented visually; and take note that when means are homogeneous enough, the red square will be smaller than the blue in which case the F statistic will be less than unity (one). The range of scores is given on left/vertical axis; group means, which correspond to the red triangles in the graphic, are printed at the right margin. Here, the four means

spread past two s.d.(within) units, implying ‘notable statistical effects,’ which follows (loosely) from seeing that the red square is substantially larger than the blue.

Figure 1a, below, shows the same basic graphic, but in this case a rug plot showing all residuals is shown at the right margin, centered on grand mean. In addition, green crosses show 20% trimmed means, these generally being effective robust replacements for the arithmetic means (that are clearly not robust estimates of location). Note that the trimmed means tend to be near their non-trimmed counterparts, except for the case of the third treatment, treatment C. In the latter case a low-outlier appears, which is the reason the 20% trimmed mean (see the green cross for the ‘G-3’, label at top) is discernibly larger than the conventional mean.

As for an overall summary of what these data have to say, as seen in the graphic, the second treatment, *i.e.*, group B, appears to have had the ‘best’ effect (since smaller scores mean better results according to the authors); furthermore, treatments A & D are not particularly competitive with B, and treatment C is notably less effective than the others. (These statements are based on the tacit but fundamental assumption that the ‘material’ had been randomly assigned to treatment groups so that the entities were more or less comparable from group to group before the introduction of treatments.) As noted, there seems to be (only) one *anomalous* data point, this the lowest score for treatment C; the green \times for group C, the 20% trimmed mean, is seen not to be effected by the low outlier. Note that variances or standard deviations are not too dissimilar; except that treatment C yielded more variable scores. (We recall that *formal inferential* application of ANOVA entails

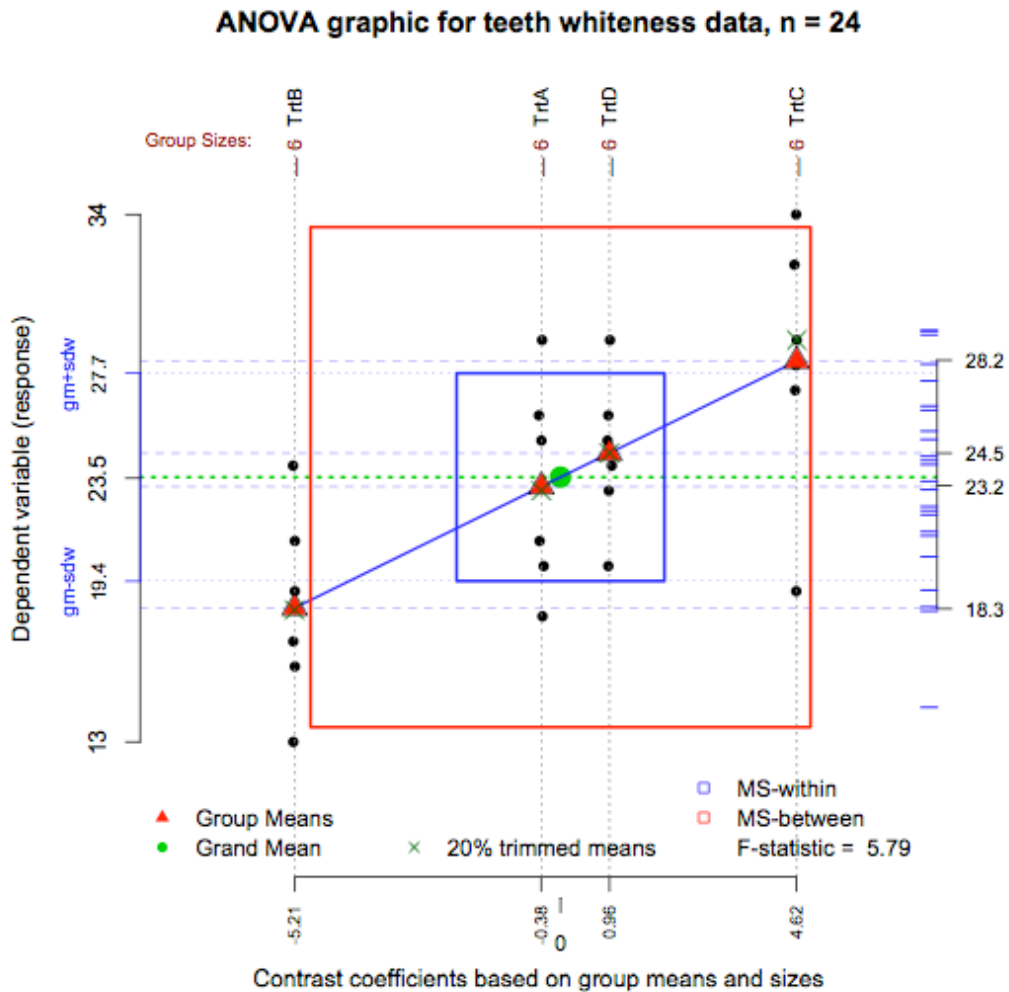


Figure 1a: A counterpart of Figure 1 that adds residuals and trimmed means to the graphic

an assumption of equal variances, as well as normality in so-called treatment populations.) Finally, given that the F- statistic is associated with a small p-value (see `$stats` above), we have causal evidence (assuming random assignments) that these treatments have different effects (in some putative system of populations), *i.e.*, the formal inferential test implies that the population means corresponding to these four samples are different from one another, and that the treatments were the likely causes of these differences. (Although there is *evidence* of causal effects, note that statistical results like these say nothing about what aspects of the treatments may have caused these effects. Generally, statisticians can provide tools to study effects of causes; but the causes of effects are quite another matter!)

The second graphic, Figure 1a, is more informative than the first in that residual scores are now shown in the so-called ‘rug’ plot on the right side (but where instead of centering on zero, centering is with respect to the grand mean), and trimmed means are shown. It seems clearer now that the lowest point in the right-most group (C) stands out more clearly as an outlier. While the trimmed mean for treatment C is larger than its untrimmed counterpart the other trimmed means are more or less the same as their untrimmed counterparts.

Two final points: the argument `dosqrs`, defaulted as T, can be set to FALSE in which case the entire figure becomes somewhat less busy; this also affords the option of teaching about what ANOVA shows without reference to inference, this being the sole purpose of the F statistic, or the overlying squares. Furthermore, it is recognized that comparing only two (independent) groups of scores is most often done using a two group *t*-test, in which case the red and blue squares have no special value.

Suggestions for use of this graphic function to gain experience learning about 1w ANOVA

Since R makes it so easy to simulate data, it is straightforward to use this function to visualize simulated data (and possibly to compare various re-expressed or sampled versions of the data using this graphic or others). For one-way ANOVA, a simulation might proceed initially by sampling so that approximate sample normality is a realistic expectation. For example, suppose we wish to generate data from a five normal population, so that all means are 10 with a common sd of 2; and further, suppose the simulated data points are to be randomly assigned to five groups of varying sizes. We could accomplish this using `>granova.lw(yy=rnorm(100,10,2), sample(1:5,100, repl=TRUE))`

To make the figure a bit more interesting, we have sampled from a t-distribution w/ df = 6 (null H true) where again there are five groups: `>granova.lw(yy=rt(100,6)*2+10, group=sample(1:5,100, repl=TRUE), re=T, tr=T, kx=1.2, px=1.2, size.line=-4, main='')`
`title('One-way ANOVA based on simulated data [null H true], varying group sizes, N = 100')`

Figure 2 below shows one particular result of such a specification.

One-way ANOVA based on simulated data [null H true], varying group sizes, N=100

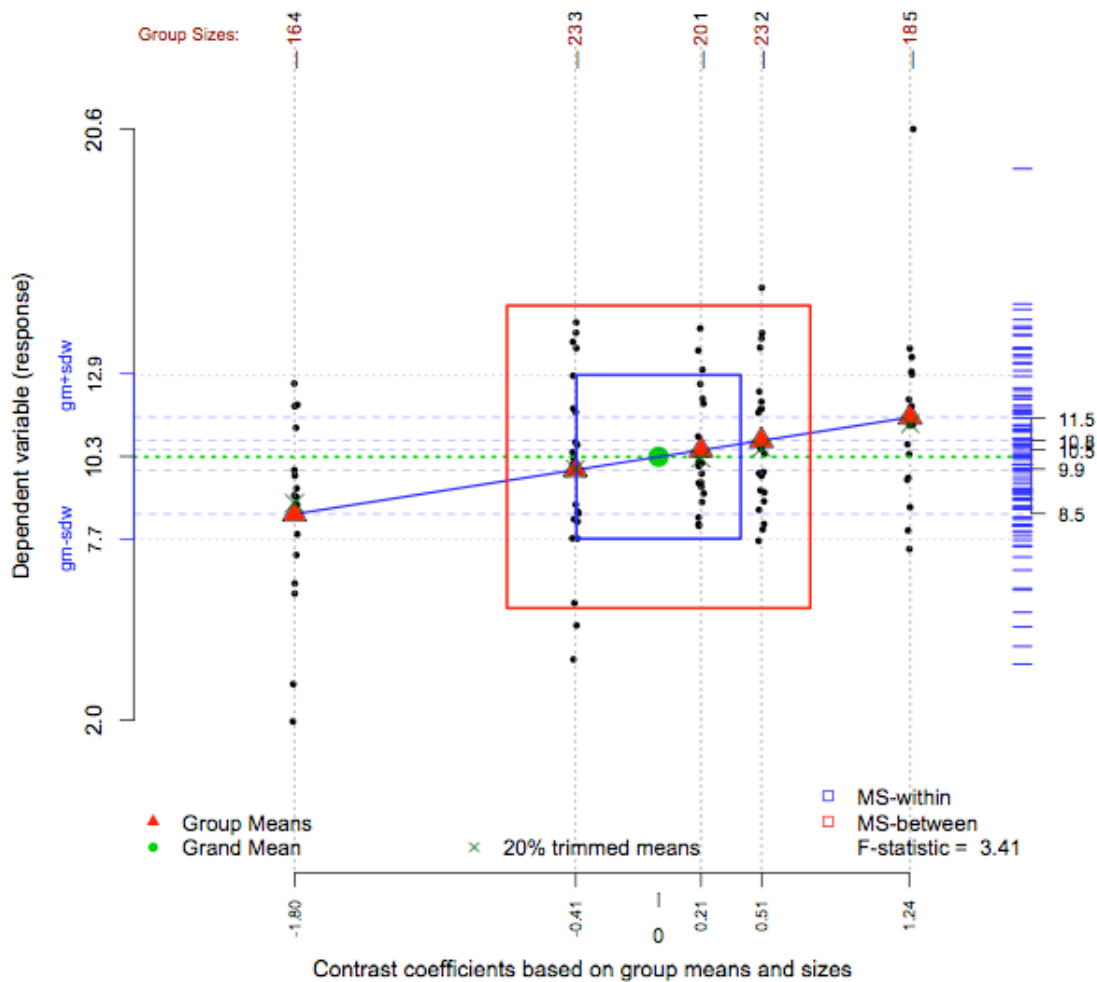


Figure 2 Randomly simulated data for ANOVA with five groups of varying sizes

The idea here is that the standard assumption used in the mathematical derivation of the ANOVA test function entails the assumption of population normality, whereas applications regularly lead to samples with longer-than-normal tails, so there may be good reasons in some contexts to study effects of various longer-than-normal tailed distributions on results when ANOVA is used. If samples are to be held to the same size, the second argument for `gp` could be constructed as `rep(1:#gps,ea=n)`, in obvious notation. A further variation on this theme might entail use of ranks in place of the initial scores in the `yy` vector; for example, we define `yy = rank(rt(100,3))` for the first of the two arguments in the basic function; this would provide what is known in the basic literature of statistics as the Friedman version of non-parametric one-way ANOVA. Visual comparisons of rank-transformed data with its unranked counterpart can be quite revealing.

Another method of interest entails bootstrapping. The function `samp.mat.boot`, given below (with only two lines of code) is easily employed when the input data `yy` take the form of a matrix (and `gp` is left unspecified). Data (e.g. `teeth.df` above) would be in the form of a matrix as in: `granova.lw(yy=samp.mat.boot(yy=teeth.df))` so that 'new' data to be analyzed and plotted

generally consists of bootstrap samples from respective columns (treatment groups) of the input matrix. How many groups to compare, of what sizes, possibly equal or varying, and what distributional specifications to make are among the choices that one must make, but applications based on S- language software are straightforward to make, as well as easily tailored to particular needs in defined applications in the context of R. In applied statistical practice, as distinguished from simulation studies, there is the possibility that covariate information will be available for individual entities. In such cases, it may be useful or revealing to employ a variety of colors, shapes, sizes, etc. to characterize individual scores, here seen as small black dots. One further option (not shown) provided by `granova.1w` is that of identifying individual data points, which can be accomplished with mouse clicks if argument `ident=T`. Once one decides to visualize data points, any information thought likely to be informative in the context of the application is of potential interest in such graphics. In observational studies, as contrasted with experiments, covariate information may be usefully summarized using propensity scores that in turn can be used in constructing graphics of this form.

There is special merit in the teachings of estimable statistician John Tukey, who argued compellingly that data are the proper focus of applied science, not statistics, not statistical theory, and not necessarily formal inference. This generally implies that investigators should aim to take account of all that is known about the data's source and context. Graphics in particular have potential to suggest new hypotheses, to help ensure sound qualifications, and sometimes to show there may be good reasons to modify or revise initial research questions, depending what one's data have to say and on the background information that may be available. The elemental ANOVA graphic seen here has the virtue that it connects directly with the central question that drives one-way ANOVA while providing all the data points (possibly with identifying information) as well as the standard summary statistics for this method. Moreover, when generalized to the case of two-way ANOVA (`granova.2w`), and related functions, a variety of other advantages can be seen in sharper relief. (Write to me at rmpruzek@yahoo.com to get illustrations of graphics provided by `granova.2w`, as well as other functions in the `granova` package. Or better yet, download the package, and use the examples in the function documentations to generate your own graphics. Or try these functions with your own data.)

* * * * *

AN AUXILLARY FUNCTION TO FACILITATE BOOTSTRAPPING; where input `x` is intended to be matrix `yy` in `granova.1w` (where equal-sized groups are stipulated).

```
samp.mat.boot <- function(x) { y <- x
for(i in 1:ncol(y)) y[,i] <- sample(x[,i], nrow(x), repl=T)
list(y=y) }
```