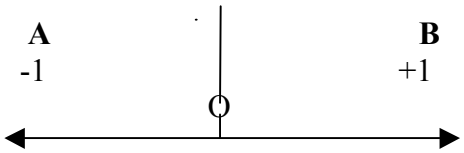
First, we examine GEOMETRIC representations for contrasts, most easily viewed as PLANNED, for comparing groups (a la' ANOVA); note general principles that start from geometric considerations & then move to algebra. NB: contrast coefficients sum to zero

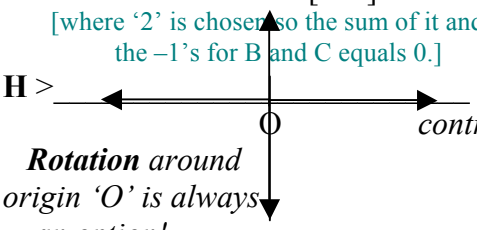
For ***two*** groups, only one comparison is possible: shown as a contrast between two treatments, here labeled as A,B; that is, we have a line w/ labels from -1 through to +1. *I'm defining a 'space' of dimension 1 here, where the* ***contrast coefficients*** *are -1 & +1. (Move now to panel on the right, for 3 gps.)*



Suppose there are ***three*** groups. Then we may *draw axes* (here ***mutually orthogonal***) that correspond to the ***plane*** on which these three points (groups) lie; the ***origin*** (O) corresponds to intersection (c.f. V & H)

+2 A [0] <- zero weight

[where '2' is chosen so the sum of it and the –1's for B and C equals 0.]



Rotation *around origin 'O' is always an option!*

-1 B -1 vs. C +1

contrast matrix

	V	H
A	2	0
B	-1	1
C	-1	-1

For ***four*** groups, we need a *three-space*; think of a *pyramid w/ regular sides*, also called a tetrahedron. So this figure *cannot* be seen on a plane; but we do know its *vertices* and they are the groups we wish to compare: A,B,C & D.

C *arrows make it too busy but the idea is same.*

D [← Above the plane ^]

A **B**

Now, imagine a ***plane*** parallel to this page (and to < A, B, C >); one can contrast D w/ the 'lower' three groups, and proceed then to make further contrasts among A,B & C, not unlike the method illustrated in upper right box here; note that the arrow points to a set of contrast coefficients where this idea is used, except C is compared w/ A,B at the second stage, and then A w/ B ignoring the others. Orthogonality for contrasts is NOT necessary, only helpful.

More generally, for k groups, there can only be at most k –1 independent (could be orthogonal) contrasts among the k groups. Geometrically, this becomes hard to see, but there are many examples that can be illustrated using CONTRAST COEFFICIENTS matrices.

e.g. Two possible X-sets w/ FOUR groups:

A	-1	-1	-1		-3/4	1/2	-1/4
B	1	-1	-1		-1/4	-1/2	3/4
C	0	2	-1		1/4	-1/2	-3/4
D	0	0	3		3/4	1/2	-3/4
	Helmert^				**Lin**	**Quad**	**Cubc**

And another for FIVE groups, following the same pattern as that above on the left.

	x_1	x_2	x_3	x_4
A	-1	-1	-1	-1
B	1	-1	-1	-1
C	0	2	-1	-1
D	0	0	3	-1
E	0	0	0	4

← *each* x_j *compares groups*

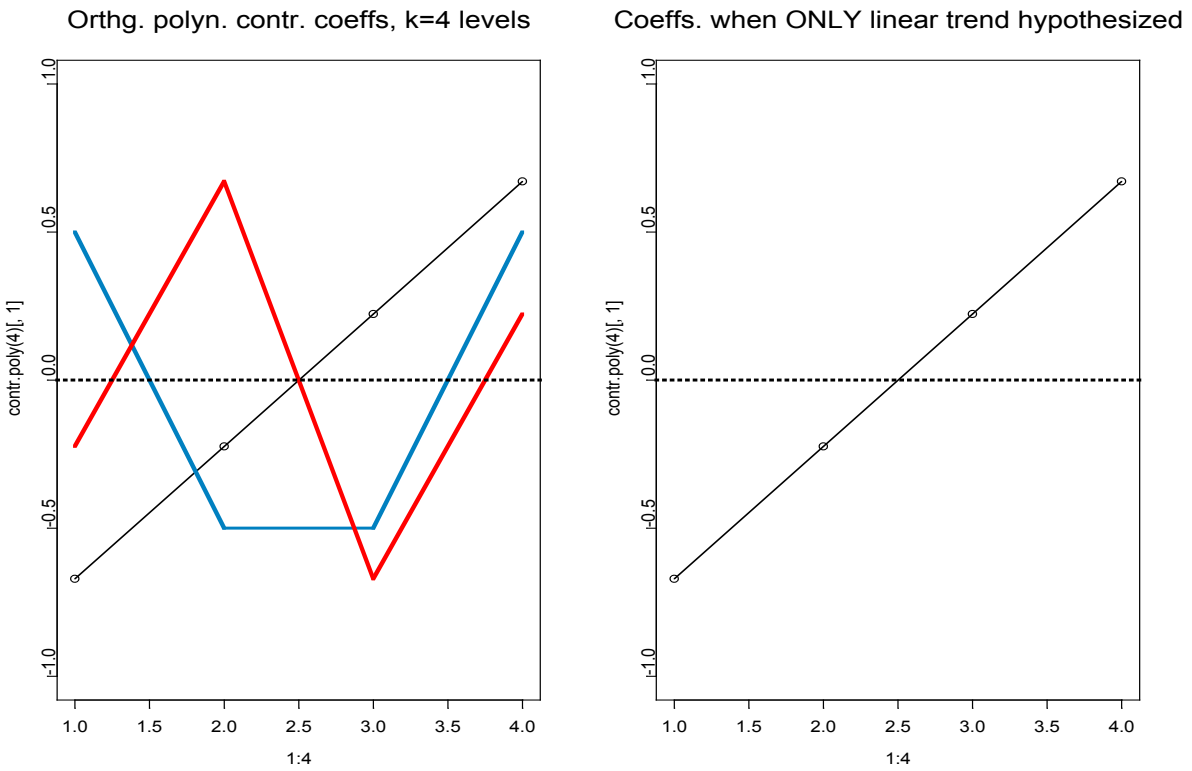
← This type of contrast matrix is called a *Helmert* form. The upper right variety involves orthogonal polynomials. (Relax)

Satisfy yourself that you can set one such matrix up for SIX groups, say, or more; there are always an INFINITE number of options for THREE or more groups (because coefficients could be decimal fractions).

Each system of ***contrast coefficients*** *can be seen as a* ***specification for a matrix*** *X for use in a regression analysis that involves comparing groups. That coefficients sum to zero in each column is central. If we used the FIVE group contrast matrix in the lower right portion of table (w/ 4 columns) to specify X, this would compare five groups of 'y' scores; we could set up the regression problem by merely* ***stacking*** *the y scores from these five groups on top of one another, to get* ***one vector*** *w/ n x 5 = N scores, and then* ***repeat each row of the matrix X n times*** *to accommodate individual y's in these groups. Regression w/ X's like this result in comparing MEANS of the groups, automatically!* **`granova.contr`** *follows from these facts. You need to try this function w/ real data for 3 or 4 groups asap. Choose equal sized groups for this.*

The following two figures depict the contrast coefficients for what are called **orthogonal polynomials.** Note that when you issue the command **`contrasts.poly(k)`** in R, that k – 1 mutually orthogonal columns of contrast coefficients are generated. Lines that depict the trends associated w/ these columns are shown on the left for 3 columns when k is 4. Try it, but note that I scaled the columns so the sums of absolute values equals 2 for each column – which I shall explain in class. The function I wrote follows.

Plots of contrast coefficients for orthogonal polynomial systems.



The three lines on the left, two jagged, correspond to *linear, quadratic and cubic trends*, those associated with the usual three columns of the orthogonal polynomial coefficients matrix. To use *all three* in a regression analysis is equivalent to stating that you are searching for regression coefficients for each term in a fully saturated model, one where all the between groups degrees of freedom are incorporated in the fitting and testing. Use of ONLY the linear coefficients in regression, i.e., -.67, -.22, .22, .67 (or, equivalently, -3, -1, 1, 3, or the downward trend values 3, 1, -1, -1), as in the right hand plot, entails prespecification of a *stronger, more specific, non-saturated* model. Simpler models of this form are not only easier to use, but they ***might*** yield more persuasive results. This depends on whether the particular ***linear prediction of the y-mean trend*** is supported by the data used in analysis. (Note that more degrees of freedom remain for the error term if only linear trend is used.) *More generally, when prior theory or previous experimental results support a particular prediction, then use of contrasts whose coefficients correspond to that* ***prediction*** *of y-values (usually group mean y values) can often be justified in a planned comparison approach to analysis.* From one rather purist point of view, this is THE way one should often approach design, given specific a priori, or advance, hypotheses. Modern practice, however, usually regards purist views as unrealistic. Much greater latitude is usually given to the analyst, to use exploratory as well as confirmatory strategies in analyses, where different models are compared, diagnostics are used, and less concern is given in practice to strict probabilistic interpretations of inferential results. Note that all this comes under the heading of planned comparisons, which must be distinguished from post-hoc comparisons in ANOVA settings, for which further discussion is provided below.

Comparing Multiple Groups

As you know, the method for comparing the means of two groups can be generalized to multiple groups and summarized by the formula. This is done by calculating an omnibus F statistic, written as F = MS(effect) / MS(error), where each MS is an estimate of a corresponding population variance. As in the above case, there is a null hypothesis where each group is seen to represent a population, for which all population means are equal (say H_0: $\mu_1=\mu_2=\mu_3$).

MS(effect), sometimes called MS (treatment), or MS (between), is usually thought of as variance of hypothesized populations means. MS(error) is usually taken to be a simple pooled within variance. (It is not uncommon, however, that the MS (error) term may estimate some combination of within sample variance and interaction.) The numerator MS for an *F*-statistic may be associated with one, two or several group comparisons, making the nature of F quite broad. In general, one needs to distinguish between ANOVA models with 'FIXED, MIXED, or RANDOM' factors, and this is strictly a matter of *a priori* judgment of the researcher.

The most common application is to think of ANOVA as *fixed* effects, where typically at least some groups are formed or created as a result of manipulation (*e.g.,* random assignment to treatment groups, where the treatments being compared are the only ones of interest, and these have been selected, not sampled from a larger set, by the investigator). The (pooled) variance within groups is taken to represent 'error variance' and so that dividing the variance between groups (*i.e.,* the mean squares between) by the pooled variance within groups (the mean square within) yields the *F*-statistic in such a fixed effects situation. This computed sample *F*- statistic can be compared to a theoretical distribution for F, by comparing it with tabled *F* to which it corresponds (at some alpha level, for a particular pair of numerator and denominator degrees of freedom). As is the case of the *t*-test, one commonly uses .05 or .01 'levels of significance' (alphas) as reference values in such comparisons. Should the observed *F*-statistic be larger than the tabled value, the NHST results in a finding of statistical significance. Such a result is taken to imply *SOME* difference in postulated population means that correspond to the sample groups being compared.

Omnibus F-statistics generally do not have direct counterparts in confidence intervals, since CIs are feasible ONLY in situations where TWO groups are under comparison. More on this below.

Planned Comparisons

Conventional ANOVA questions can often be approached more effectively through the use of specific planned comparison contrasts; these can obviate the complications of post-hoc comparisons (to be discussed later) in many situations, as well as improve the power and focus of one's analysis. Planned comparisons can be approached in several ways, but what is discussed below is largely grounded in John Tukey's work on this topic (see Benjamini & Braun, 2002). Note that we shall refer to effect sizes as well as mean comparisons; population ESs are generally of the form ES = $(\mu_A - \mu_B)$ /sigma where a difference between two population means is divided by an appropriate standard deviation, sigma; note that sample Effect Sizes are of the same form, but statistics replace the parameters for sample ESs.

Suppose three groups (randomly formed) are to be compared, and are labeled A, B, & C. Two planned comparisons can be generated in this case, and they might take forms such as:

I: $\mu_A - \mu_B$ and

II: $(\mu_A - \mu_B)/2 - \mu_C$

to indicate two independent comparison of two groups for these three groups. We can in this situation define two population effect size parameters as:

ES_I: $ES_I = (\mu_A - \mu_B)/\sigma$

ES_{II}: $ES_{II} = [(\mu_A - \mu_B)/2 - \mu_C]/\sigma$

where σ represents the *common* population variance. These comparisons are mutually orthogonal (at least when sample sized are equal) which serves the purpose of ensuring that each comparison/effect can be interpreted independently of the other. The idea of planned comparisons extends to any number of groups in principle, at least for fixed effect designs, but there cannot be more than J-1 orthogonal (or mutually independent) effects of this kind (where J is the number of groups under comparison) and the comparisons are planned in advance (*i.e., a priori*).

Consider an example where one wishes to examine the effects of FOUR treatment conditions (medication (MED), behavioral (BI), combined (MB) and control (CTRL)) while working with children who exhibit attention deficit disorder behaviors. A typical strategy would be to randomly assign children to one of the four conditions; then following treatments, a dependent variable would be observed, perhaps using some kind of standardized assessment measure for ADHD behavioral disorders. Although standard ANOVA methods (based on use of an omnibus F-statistic, in conjunction with some kind of post hoc comparison) to compare group differences (on the dependent variable) could be used this approach will often be problematic. That is, an overall F-statistic may not reach level of significance due to insufficient power (more later) which case one is led erroneously to conclude that no notable differences exist among treatments even when the treatments really do have differential effects. In such situations, a planned-comparison approach may yield several advantages.

If the problem of assessing outcomes is addressed using planned comparisons (i.e., contrasts) then specific (combinations of) groups are compared to learn whether individual effects, *i.e*. specific planned comparisons, yield significant statistics, or not. The use of an *a priori* or planned comparison approach typically provides more support for causal inferences, but it also has the potential to be more powerful in the sense that real differences between specific treatment groups may be found that could have been missed by an omnibus F-test. In general, the specific contrasts to be made are informed by theory, prior knowledge, or previous research.

In the context of the ADHD intervention example, prior research may makes it reasonable to study whether any or all of the following specific comparisons are 'notable.'

(C1) The CTRL group mean differs from that of the other three (combined) groups.
(C2) The combined MB treatment mean differ from the average of the BI and MED means, ignoring the control group.
(C3) There is a difference between the BI and MED intervention treatments.
The three *contrast sets,* columns C1 – C3 in the following table, represent the above questions using (orthogonal*) planned comparisons. The set-up for these contrasts is seen to have a 'Helmert' form, where contrasts are now shown by columns:

	C1	C2	C3
CTRL	1	0	0
MB	-1/3	1	0
BI	-1/3	-1/2	1
MED	-1/3	-1/2	-1

where, again, C = Control, BI = Behavioral Intervention, MED = Medication and MB = Combined treatments.

*To check if contrasts are orthogonal, take the *cross products of any two column vectors*, see if the total is zero (*e.g*., for C1 & C2: 1*0 = 0, -1/3*1=-1/3, -1/3*-1/2=1/6, -1/3*-1/2=1/6; and 0 – 1/3 +1/6 +1/6 = 0).

The system of contrasts entails comparison of the control group with the three treatment conditions (C1), the combined treatment (MB) against two (pure) groups, (C2), and the direct comparison of the medication and behavioral intervention groups (C3). When contrasts are mutually orthogonal, the corresponding questions are mutually independent of one another, so that each comparison can be interpreted independent of others. This provides for the enhanced power as well as more focus, as noted earlier.

For each such comparison, one can compute either a sample *t*-statistic, or a sample effect size (ES) to aid interpretation. As noted above each *t*-statistic has the basic form $(\text{mean}_1 - \text{mean}_2)/$ MS(error) where the 's correspond in general to *combined groups* (*e.g.* all the groups w/ +'s, vs. all groups /w –'s). In the case of effect sizes, the only thing that differs is that the denominator becomes a pooled standard deviation, usually computed as the square root of the mean square within groups.

Note that contrasts can be generalized to more complex studies. For example, further support of causal inferences might be obtained from using the planned comparison with paired sample data. Using a pre-post measure of behavioral disorders, the subjects could be matched on some variable(s) and then assigned to the above groups. This would require a series of block and interaction contrasts.

Post Hoc Comparisons

Although planned comparisons offer greater power and control than post hoc tests, a brief discussion of these latter approaches are offered here since they are commonly used with ANOVA. Exploratory data analysis may seem to compel one to engage in an omnibus *F*-test to learn if there is any significant difference among group means. Supposing that one rejects the null hypothesis that the treatment means are equivalent in the population, the researcher is often left to do further study to learn whether evidence exists to support an argument that particular subgroup means differ in the population, based on more specific differences observed in his or her sample. The latter step is called *post-hoc* comparisons of treatment group means, and it is generally more flexible but less powerful than a corresponding planned comparison.

So a typical ***exploratory*** approach is to first determine that a difference exists by using a formal null hypothesis (*i.e.*, H_0: $\mu_j=\mu_{j^*}$ for all *j* not equal to *j**) and then follow up with multiple comparison tests to determine which means are different. *In the context of this approach*, this is necessary as one may be working with several means, and the F only indicates that a difference exists and does not say where. To address this problem, a multiple comparison technique such as the Scheffé test is used to compare two specific means of interest. The Scheffé formula entails a version of calculating the difference between two means and dividing by it by MS (error).

To summarize, the issue at hand when considering whether to engage in contrasts or post hoc tests is based on whether one is asking specific questions in advance, as is required by planned comparisons, or to letting the data itself show if differences exist among groups by taking an exploratory approach. Depending on the context of the research, either approach might be justified, but *it is advisable to pursue planned comparisons when possible*, especially in contexts of experimental data analysis. Use Tukey's (**`hsd`**) & Scheffe's **`sheffe.test`** in R.

Assumptions of Normality & Homogeneity & Transforming Data

The use of ANOVA significance tests entails an assumption of normality in each hypothesized population distribution, and further, that these different populations have the same variance. The

normality assumption is routinely violated in applications of ANOVA, but this is not such a bad thing as there is good evidence to suggest that one can do so without invalidating the test result (Box, 1953; Howell, 2002). Violations of the homogeneity of variance assumption are to be viewed more seriously, as these can more often invalidate test results.[1]

The homogeneity of variance issue seems not to be addressed adequately in most applications of ANOVA, an issue that is exacerbated when dealing with groups of unequal sample sizes (Howell, 2002). Simply put, conventional ANOVA entails the assumption that population variances across groups being analyzed are equal for the dependent variable in question. All is not lost however when such is not the case, as one often can *transform scores* for the dependent variable to help insure that distribution variances are similar across groups, and then run ANOVA on the transformed values. Tukey prefers the term *re-expression*, instead of transformation. This is because by transforming a data value, one does not alter its inherent value, but re-expresses it so that it is easier to manage. For example, when reporting a GRE score with a mean of 500 and standard deviation of 100, one is working with data that might initially have been expressed using raw scores (but this is a linear transformation, so use of GRE scores vs. their raw scores would have no notable effect in the ANOVA context).

The logarithmic transformation is among the more common transformations, where a score is re-expressed using a power of say 10 (*i.e.*, $\log_{10}$). Such transformations may work well with data that are positively skewed; other options are square root and reciprocal transformations. Review of these and other transformations are available in many sources (see Howell), so suffice it to say here that this approach can be useful when dealing with skewed data or heterogeneous variances. It is often helpful to try different re-expressions when attempting to get data into a shape such that analyses may be conducted without having to be concerned about violating so many assumptions that any inference is unreasonable.

A Nonparametric Alternative Based on Bootstrapping

Again, one will not necessarily get into trouble when violating an assumption, since many test statistics are robust. But there may be many cases where it is better to make no assumptions about the parameters of an underlying population. One can free the analysis from some of the aforementioned assumptions via bootstrapping procedures (Efron & Tibshirani, 1993). *The approach entails taking a given sample and assuming that the population from which it was drawn is distributed in the same manner as the samples (Howell, 2002). Once this assumption is made, one can sample (with replacement, using a computer), and this can be done several thousands of times, starting from the extant sample*. Any number of statistics might be generated for each bootstrap sample. From these, one takes a nonparametric approach (in the sense that no assumptions need be made about populations) to computation of confidence intervals, tests of significance and so on using information from derived bootstrap distributions. More discussion is needed here of course.

A Brief Note on Interactions

When joint manipulation of two independent factors is conducted, one may analyze the data to learn whether there is evidence about whether, say, row and column treatments interact to influence outcomes. For example, it might be found that a positive outcome of some form of therapy is found only when it appears in conjunction with patients' expectation that the therapy will be successful. This is called an interaction. Interactions are probably more common than are recognized; often the best evidence for interactions derives from use of graphical methods. The prospect of interaction should nearly always be considered in practice.

A Big Caveat Regarding NHST

Although the use of tests using NHST is by itself not particularly problematic, several knowledgeable analysts (*e.g.*, Cohen, 1990, 1994) have concerns about the use of NHST in design unless steps are taken to be cognizant of what information significance tests do and do not offer. Arguably, the most fundamental problem raised by Cohen is that many researchers exhibit faulty logic when making NHST inferences. For example it is common to conclude that a rejected null hypothesis provides the basis for inferring that theory is established, the *p* value is often thought of as representing the probability that the null hypothesis is true, and the dichotomous decision associated with acceptance or rejection may promote disregard of data that could prove interesting or useful when seen in another light; effect sizes may offer one such benefit. Effect sizes can be especially helpful in planned comparison contexts.

As noted above, using ANOVA has many forms, and potentially many uses. It is also true, however, that one must remain cognizant of common fallacies in logic that may be associated with NHST. Certainly, these concerns will be mitigated somewhat when using the more informative assessments of confidence intervals effect size estimates. In general, confidence intervals are to be preferred to significance tests, and this also means that planned comparisons are to be preferred to omnibus tests, since the former lend themselves to computation of both CIs and ESs, whereas the latter do not. See http://forrest.psych.unc.edu/jones-tukey112399.html for an important, readable alternative logic to underpin NHST; this was written by Lyle Jones & John Tukey, and it is well worth reading.

References

Benjamini, Y. & Braun, H. (2002). John Tukey's approaches to multiple comparisons. Available on-line: http://www.ets.org/research/dload/RR-02-24.pdf

Box, G.E.P. (1953). Non-normality and tests on variance. *Biometrika, 40*, 318-335.

Cohen, J. (1994). The earth is round (p<.05). *The American Psychologist, 49,* 997-1003.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Efron, B. & Tibshirani, R. J. (1993) *An introduction to the bootstrap*. New York: Chapman and Hall.

Howell, D.C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Jones, L.V., & Tukey, J.W. (2000). A sensible formulation of the significance test. Available on-line: http://forrest.psych.unc.edu/jones-tukey112399.html

--------The function that standardizes contrast vectors, by columns follows:

```
std.contr <- function(cont, tol = sqrt(.Machine$double.eps)^0.6)
{ #generates 'standardized contrast vector'; positive & abs(negative) values
# each sum to 1; cont assumed to consist of contrast(s) vector/matrix w/ mean zero;
# otherwise stops
if(!is.matrix(cont)) cont <- as.matrix(cont)
if(abs(mean(cont)) > tol)
stop("Input vector/matrix must have mean zero (for each column)")
if(ncol(cont) == 1)
```

```
cont <- matrix(cont, ncol = 1)#Strong assumption that cont is matrix
dg <- apply(abs(cont), 2, sum)
if(length(dg) == 1)
dg <- as.matrix(dg)#print(paste("length(dg)", dg))
s.cont <- round(2 * cont %*% diag(1/dg),3)
s.cont       }
```