Comparisons among means (or, the analysis of factor effects)

In carrying out our usual test that $\mu_1 = \cdots = \mu_r$, we might be content to just reject this "omnibus hypothesis" but typically more is required:

where are the differences coming from?

We will discuss comparisons among means as a way of investigating where the differences *are* and not merely that some differences exists

We make some distinctions at the outset:

Two classes of comparisons:

a) Planned (or a priori) comparisons: you have specific ideas of where differences in the means should be. These are to be used in place of doing the overall F-test

b) Incidental (or Post-hoc) comparisons: after you find a significant result from the omnibus F-test, you want to find out what differences in the means caused the rejection

We start with planned comparisons used in place of the omnibus F-test

Example:

Suppose we are testing the effect of a number of drugs on attention; four different drugs have been suggested as possibly helpful:

Drug 1	Drug 2	Drug 3	Drug 4	Placebo
μ_1	μ_2	μ_3	μ_4	μ_5

The subjects have been randomly assigned to the five conditions

I'm not particularly interested in the omnibus hypothesis but in specific questions about the differences in the means

For example:

a) the effect of each of the drugs versus the placebo: e.g., $\mu_1 - \mu_5$

b) the average effect of drugs 1 and 2 versus the average effect of drugs 3 and 4: e.g., $\frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4)$

c) differences between pairs of drugs: e.g., $\mu_1 - \mu_2$

d) the average of all drugs versus the placebo: e.g., $\frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4) - \mu_5$

A common aspect of all these questions is that they can be phrased as linear combinations of the means

Definition: a population comparison (or contrast) among population means is a linear combination of the population means:

$L = c_1\mu_1 + \cdots + c_r\mu_r = \sum_{i=1}^{r} c_i\mu_i,$

where $\sum_{i=1}^{r} c_i = 0$;

this latter condition will be discussed below

A sample comparison (or contrast) among sample means is a linear combination of the sample means:

$\widehat{L} = c_1\widehat{\mu}_1 + \cdots + c_r\widehat{\mu}_r = \sum_{i=1}^{r} c_i\widehat{\mu}_i,$

where $\sum_{i=1}^{r} c_i = 0$

Why the condition $\sum_{i=1}^{r} c_i = 0$?

$\sum_{i=1}^{r} c_i \mu_i = \sum_{i=1}^{r} c_i(\mu_\cdot + \tau_i) =$

$\sum_{i=1}^{r} c_i \mu_\cdot + \sum_{i=1}^{r} c_i \tau_i =$

$\mu_\cdot \sum_{i=1}^{r} c_i + \sum_{i=1}^{r} c_i \tau_i = \sum_{i=1}^{r} c_i \tau_i$

So, a comparison is unaffected by the grand mean, however defined when $\sum_{i=1}^{r} c_i = 0$

The sampling distribution of $\widehat{L}$:

$\widehat{L} = \sum_{i=1}^{r} c_i \widehat{\mu}_i$,

where $\widehat{\mu}_i \sim N(\mu_i, \frac{\sigma^2}{n_i})$, leads to

$\widehat{L} \sim N(L, \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i})$

$E(\widehat{L}) = E(\sum_{i=1}^{r} c_i \widehat{\mu}_i) = \sum_{i=1}^{r} c_i E(\widehat{\mu}_i) =$

$\sum_{i=1}^{r} c_i \mu_i = L$

$Var(\widehat{L}) = Var(\sum_{i=1}^{r} c_i \widehat{\mu}_i) = \sum_{i=1}^{r} c_i^2 Var(\widehat{\mu}_i) =$

$\sum_{i=1}^{r} c_i^2 (\frac{\sigma^2}{n_i}) = \sigma^2 \sum_{i=1}^{r} (\frac{c_i^2}{n_i})$

So, $\frac{\hat{L}-L}{\sqrt{\sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}}} \sim N(0,1)$

Or, $\frac{\hat{L}-L}{\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}}} \sim t_{n_T - r}$

So, we can test $H_o : L = 0$ or put a confidence interval on L

Suppose

$L = c_1\mu_1 + \cdots + c_r\mu_r$ and

$L' = ac_1\mu_1 + \cdots + ac_r\mu_r$

Then

$$\frac{\hat{L}' - L'}{\sqrt{MSE \sum_{i=1}^{r} \frac{(ac_i)^2}{n_i}}} =$$

$$\frac{a(\hat{L} - L)}{a\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}}} =$$

$$\frac{(\hat{L} - L)}{\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}}}$$

This gets rid of fractions in comparisons so the weights can be whole numbers

Arbitrary linear combinations:

$\widehat{L} \sim N(L, \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i})$ is appropriate for any arbitrary linear combination

Thus, when $c_1 = 1, c_2 = \cdots = c_r = 0$

$\frac{\widehat{\mu}_1 - \mu_1}{\sqrt{MSE/n_1}} \sim t_{n_T - r}$

But note that MSE is from the whole ANOVA layout (i.e., there are more degrees of freedom than usual)

The Bonferroni Discussion:

Suppose I have specified a collection of planned comparisons $(L_1, \ldots, L_K)$ that I would like to study in lieu of doing the omnibus test.

We know how to test each $H_o : L_k = 0$ and how to put a confidence interval on L_k

Should I just "blast away" and do each test or confidence interval at say, $\alpha = .05$? Or should I try to be more safe and control the overall error rate (and engage in "safe statistics")

To be more formal about the problem of multiple testing, suppose there are K hypotheses to test, $H_1, \ldots, H_K$, and for each, we set the criterion for rejection at the fixed Type I error value of α_k, $k = 1, \ldots, K$.

If the event A_k is defined as the incorrect rejection of H_k (that is, rejection when it is true), the Bonferroni inequality gives

$$P(A_1 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} \alpha_k \ .$$

Noting that the event (A_1 or $\cdots$ or A_K) can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experiment-wise (or overall) error rate is bounded by the sum of the K α values set for each hypothesis.

Typically, we let $\alpha_1 = \cdots = \alpha_K = \alpha$, and the bound is then $K\alpha$.

Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual αs at some small value such as $.05/K$;

the overall error rate is then guaranteed to be no larger than .05.

Orthogonal Comparisons:

Consider two comparisons:

$L_1 = \sum_{r=1}^{r} c_{1i}\mu_i$, where $\sum_{i=1}^{r} c_{1i} = 0$

$L_2 = \sum_{r=1}^{r} c_{2i}\mu_i$, where $\sum_{i=1}^{r} c_{2i} = 0$

$\hat{L}_1$ and $\hat{L}_2$ are orthogonal (i.e., statistically independent) if

$\sum_{i=1}^{r} \frac{c_{1i}c_{2i}}{n_i} = 0$

If the n's are all equal, this reduces to

$\sum_{i=1}^{r} c_{1i}c_{2i} = 0$

For example, in our drug illustration (where the n's are all equal), consider the following four comparisons:

$$L_1 = \mu_1 - \mu_2$$

$$L_2 = \mu_3 - \mu_4$$

$$L_3 = \mu_1 + \mu_2 - \mu_3 - \mu_4$$

$$L_4 = \mu_1 + \mu_2 + \mu_3 + \mu_4 - 4\mu_5$$

For r groups, there are at most $r - 1$ mutually orthogonal comparisons

Consider two comparisons:

$L_1 = \sum_{r=1}^{r} c_{1i}\mu_i$, where $\sum_{i=1}^{r} c_{1i} = 0$

$L_2 = \sum_{r=1}^{r} \frac{n_i}{n_T}\mu_i$, the weighted grand mean

Then, $\sum_{i=1}^{r}((c_i\frac{n_i}{n_T})/n_i =$

$\frac{1}{n_T}\sum_{i=1}^{r} c_i = 0$

In other words, a comparison is independent of the weighted grand mean

If I define a Sum of Squares for a comparison as:

$SS(L) = \hat{L}^2 / \sum_{i=1}^{r} \frac{c_i^2}{n_i}$, and with a single degree of freedom

To test $H_o : L = 0$, $SS(L)/MSE \sim F_{1,n_t-r}$

If $L_1, \ldots, L_r$ are mutually orthogonal, then $SS(L_1) + \cdots + SS(L_{r-1}) = SSTR$

And a simultaneous test of $L_1, \ldots, L_r$ against zero is the same as our omnibus test based on $MSTR/MSE$

There is one point of view on dong planned comparisons that I will refer to as the "orthogonal contingent" (Hays would fall into this crowd)

If planned comparisons are orthogonal, then sample analogues are "independent". Thus, it makes sense to test each at, say, .05, and not worry about inflating the overall error rate.

Problems with this:

1) Comparisons may be independent, but the tests are not (they all use the same MSE)

2) Even ignoring dependent tests you still should control the overall error rate by using, say, a Bonferroni correction

3) With unequal n's in particular, orthogonality is very weird; and even with equal n's you may not be able to ask all the questions that one would like to

If you use Bonferroni, there is no need to limit the number or kind of planned comparisons carried out

Incidental or Post-hoc comparisons:

In planned comparisons, no overall omnibus test was performed

Now, suppose an overall test was done and we reject the null hypothesis of equal means

We now want to find out where the difference are

In other words, we want to test the significance of *any* comparison

So, choose any comparison L

To test $H_o : L = 0$, compare $SS(L)/MSE$ to $(r-1)F_{r-1,n_T-r}$

this is called "Scheffe's procedure"

note the multiplier of $(r-1)$ and the increase in the numerator degrees of freedom compared to testing one planned comparison against zero using F_{1,n_T-r}

So, post-hoc confidence intervals would have the form:

Scheffe Procedure:

$$\widehat{L} \pm (\sqrt{(r-1)F_{\alpha,r-1,n_T-r}})(\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}})$$

One planned comparison:

$$\widehat{L} \pm (\sqrt{F_{\alpha,1,n_T-r}})(\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}})$$

Bonferroni correction with K planned comparisons:

$$\widehat{L} \pm (\sqrt{F_{\frac{\alpha}{K},1,n_T-r}})(\sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}})$$

Scheffe's procedure has two important properties:

a) If the overall F test is significant at α, then at least one comparison will be significant at α or better

This may not be the one you can ever make sense of, however

b) The experiment-wise (overall) significance level (i.e., the probability of rejecting one or more comparisons against zero by chance) remains at α no matter how many you do.

You pay for this, however, because any comparison is generally difficult to declare significant post-hoc – for example, no pairwise comparison may be significant post-hoc

Tukey's Procedure (Honestly Significant Difference – HSD)

Suppose I plan to do all $r(r-1)/2$ pairwise comparisons of the means

We could do a Bonferroni correction to control the overall error rate but a better procedure (in the sense of shorter confidence intervals) is Tukey's procedure

We begin by assuming $n_1 = \cdots = n_r = n$; consider the r sample means in an ANOVA each deviated from their means: $\bar{Y}_{1\cdot} - \mu_1, \ldots, \bar{Y}_{r\cdot} - \mu_r$

Each of these is $\sim N(0, \frac{\sigma^2}{n})$

Thus,

$$\frac{max(\bar{Y}_{i\cdot} - \mu_i) - min(\bar{Y}_{i\cdot} - \mu_i)}{\sqrt{MSE/n}} \sim q(r, n_T - r)$$

this is the "studentized range" statistic

Turning this around, we have

$$P(|(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) - (\mu_i - \mu_{i'})| \leq q_\alpha(r, n_T - r)) = 1 - \alpha$$

Thus,

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \pm \sqrt{(MSE/n)}\ q_\alpha(r, n_T - r)$$

holds for all i, i'

This can be rewritten as

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \pm \sqrt{(MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i})}\ (\frac{1}{\sqrt{n}})\ q_\alpha(r, n_T - r)$$

This is conservative if we use the above for unequal n's – this is called the Tukey-Kramer method

Tukey's procedure is an alternative to our overall F test

We look at the confidence interval for the most extreme pair – if zero is included, don't reject

If we use this as our overall test, then at least one pairwise comparison is significant post-hoc

In a sense, however, we just *plan* to do all pairwise comparisons

Some have called it a Type IV error to use the overall F test and then followup it up "post-hoc" with Tukey's procedure

Some caveats on multiple testing:

The problem of multiple testing and the failure to practice "safe statistics" appears in both blatant and more subtle forms.

For example, companies may suppress unfavorable studies until those to their liking occur.

A possibly apocryphal story exists about toothpaste companies promoting fluoride in their products in the 1950s and who repeated studies until large effects could be reported for their "look Ma, no cavities" television campaigns.

This may be somewhat innocent advertising hype for toothpaste, but when drug or tobacco companies engage in the practice, it is not so innocent and can have a serious impact on our collective health.

It is important to know how many things were tested to assess the importance of those reported.

For example, when given only those items from some inventory or survey that produced significant differences between groups, be very wary!

People sometimes engage in a number of odd behaviors when doing multiple testing. We list a few of these below in summary form:

(a) It is not legitimate to do a Bonferroni correction post hoc; that is, find a set of tests that lead to significance, and then evaluate just this subset with the correction;

(b) Scheffé's method (and relatives) are the only true post-hoc procedures to control the overall error rate. An unlimited number of comparisons can be made (no matter whether identified from the given data or not), and the overall error rate remains constant;

(c) You cannot look at your data and then decide which planned comparisons to do;

(d) Tukey's method is not post hoc because you actually plan to do all possible pairwise comparisons;

(e) Even though the comparisons you might wish to test are independent (such as those defined by orthogonal comparisons), the problem of inflating the overall error rate remains; similarly, in performing a multifactor analysis of variance (ANOVA) or testing multiple regression coefficients, all of the tests carried out should have some type of control imposed on the overall error rate;

(f) It makes little sense to perform a multivariate analysis of variance before you go on to evaluate each of the component variables. Typically, a multivariate analysis of variance (MANOVA) is completely noninformative as to what is really occurring, but people proceed in any case to evaluate the individual univariate

ANOVAs irrespective of what occurs at the MANOVA level; we may accept the null hypothesis at the overall MANOVA level but then illogically ask where the differences are at the level of the individual variables. Plan to do the individual comparisons beforehand, and avoid the uninterpretable overall MANOVA test completely.

Note the article on Random Field Theory on the class web site:

This chapter is an introduction to the multiple comparison problem in functional imaging, and the way it can be solved using Random field theory (RFT).

In a standard functional imaging analysis, we fit a statistical model to the data, to give us model parameters.

We then use the model parameters to look for an effect we are interested in, such as the difference between a task and baseline.

To do this, we usually calculate a statistic for each brain voxel that tests for the effect of interest in that voxel.

The result is a large volume of statistic values.

...