Comparing several means

STA441: Winter/Spring 2016

One-way Analysis of variance

- Categorical explanatory variable
- Quantitative response variable
- *p* categories (groups)
- H_0: All population means equal
- Normal conditional distributions
- Equal variances

Analysis means to split up

- With no explanatory variable, best predictor is the overall mean.
- Variation to be explained is SSTO, sum of squared differences from the overall mean.
- With an explanatory variable, best predictor is the group mean.
- Variation still unexplained is SSW, sum of squared differences from the group means.

$$SSTO = SSB + SSW$$

$$
\begin{aligned}
SSB &= \sum_{j=1}^{p} n_j(\overline{Y}_j - \overline{Y})^2 \\
SSW &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{i,j} - \overline{Y}_j)^2 \\
SSTO &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{i,j} - \overline{Y})^2.
\end{aligned}
$$

ANOVA Summary Table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	$p-1$	SSB	$MSB = SSB/(k-1)$	MSB/MSW	p-value
Error	$n-p$	SSW	$MSW = SSW/(n-k)$		
Corrected Total	$n-1$	$SSTO$			

$$H_0 : \mu_1 = \ldots = \mu_p.$$

R^2 is the proportion of variation explained by the independent variable

$$R^2 = \frac{SSB}{SSTO}$$

Contrasts

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

$$\widehat{c} = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p$$

$$\text{where } a_1 + a_2 + \cdots + a_p = 0$$

Overall F-test is a test of p-1 contrasts

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

Sample Question

Give a table showing the contrasts you would use to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

There is one row for each contrast.

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

(This is a good format.)

Multiple Comparisons

- Most hypothesis tests are designed to be carried out in isolation
- But if you do a lot of tests and all the null hypotheses are true, the chance of rejecting at least one of them can be a lot more than α. This is **inflation of the Type I error probability**.
- Multiple comparisons (sometimes called follow-up tests, post hoc tests, probing) offer a solution.

Multiple comparisons

- Protect a *family* of tests against Type I error at some *joint* significance level α.
- If all the null hypotheses are true, the probability of rejecting at least one is no more than α.

Multiple comparisons of contrasts in a one-way design: Assume all means are equal in the population

- Bonferroni
- Tukey
- Scheffé

Bonferroni

- Based on Bonferroni's inequality:

$$P\{A_1 \text{ or } A_2 \text{ or } \ldots A_k\} \leq P\{A_1\} + P\{A_2\} + \ldots + P\{A_k\}$$

- Applies to *any* collection of k tests
- Assume all k null hypotheses are true
- Event A_j is that null hypothesis j is rejected.
- Do the tests as usual
- Reject each H_0 if $p < 0.05/k$
- Or, adjust the p-values. Multiply them by k, and reject if $pk < 0.05$

Bonferroni

- Advantage: Flexibility
- Advantage: Easy to do

- Disadvantage: Must know what all the tests are before seeing the data
- Disadvantage: A little conservative; the true joint significance level is *less* than 0.05.

Tukey (HSD)

- Based on the distribution of the largest mean minus the smallest.
- Applies only to pairwise comparisons of means
- If sample sizes are equal, it's the most powerful, period.
- If sample sizes are not equal, it's a bit conservative.

Scheffé

- Find the usual critical value for the initial test. Multiply by p-1. This is the Scheffé critical value.
- Family includes *all* contrasts: Infinitely many!
- You don’t need to specify them in advance

Scheffé

- Scheffé family includes *simultaneous* tests of *s* contrasts.
- So for example with 12 treatment conditions, one could test

$$\mu_7 = \mu_8 = \mu_9 = \mu_{10}$$ as a follow-up to

$$H_0: \mu_1 = \ldots = \mu_{12}$$

- Scheffé critical value is usual critical value for the initial test times *(p-1)/s*.

Scheffé

- Follow-up tests *cannot* be significant if the initial overall test is not. Not quite true of Bonferroni and Tukey.
- If the initial test (of p-1 contrasts) is significant, there *is* a single contrast that is significantly different from zero (not necessarily a pairwise comparison)
- Adjusted p-value is the tail area beyond the product $[F \times s/(p-1)]$.

Which method should you use?

- If the sample sizes are nearly equal and you are only interested in pairwise comparisons, use Tukey because it's most powerful

- If the sample sizes are not close to equal and you are only interested in pairwise comparisons, there is (amazingly) no harm in applying all three methods and picking the one that gives you the greatest number of significant results. (It's okay because this choice could be determined in advance based on number of treatments, α and the sample sizes.)

- For simultaneous tests of multiple contrasts, must use Bonferroni or Scheffé. Tukey is out.

- If you are interested in contrasts that go beyond pairwise comparisons and you can specify *all* of them before seeing the data, Bonferroni is almost always more powerful than Scheffé. (Tukey is out.)

- If you want lots of special contrasts but you don't know exactly what they all are, Scheffé is the only honest way to go, unless you have a separate replication data set.

How far should you take this?

- Protect all follow-ups to a given test?
- Protect all tests that use a given model?
- Protect all tests reported in a study?
- Protect all tests carried out in an investigator's lifetime?

We will be very modest. If we follow up a test for difference among several means, we will hold the joint significance level of the follow-up tests to 0.05 somehow.

Copyright Information