Multiple Comparisons

BIOS 6611

CU Anschutz

Week 10

Multiple Comparisons

The Problem

We may find ourselves in different cases where we wish to conduct more than one statistical test. For example,

- there is an overall/global hypothesis, but we then want to do post-hoc testing between groups (e.g., a *multiple comparisons* problem)
- in genomics we may wish to evaluate 1000s of SNPs in one study (e.g., a *multiple testing* problem)
- for a trial, we may define multiple primary outcomes of interest. (e.g., potentially both a multiple testing and comparisons problem)

Why is this problematic?

When we perform multiple statistical tests, the true *overall* type I error rate is larger than the type I error rate for the *individual* tests. These are known as **family-wise** (overall) and **marginal** type I error rates, respectively.

The Type I Error Trade-off

If we set $\alpha = Pr(\text{Reject } H_0 | H_0 \text{ is true}) = 0.05$ for a single test, then $1 - \alpha = Pr(\text{Fail to reject } H_0 | H_0 \text{ is true}) = 0.95$. For k tests, the *family-wise* type I error rate is then $1 - (0.95)^k$.

The following table displays the probability of rejecting at least one of the pairwise comparisons using a significance level of 0.05:

k	**FWER**
1	0.050
3	0.143
10	0.401
50	0.923

We see that as the number of tests increases we are more likely to have a true overall type I error rate that is far greater than the individual test type I error rate.

The Challenge

Although post-hoc multiple comparison procedures control the overall type I error rates, they inflate type II errors (the probability of failing to reject the null hypothesis when the alternative is true). With appropriate software, multiple comparisons can be incorporated into sample size and power analyses at the design phase.

There is generally not a single "correct" approach. If comparisons of interest are planned in advance (it also helps to limit comparisons to those with the most scientific relevance), certain procedures may be more appropriate (e.g., the LSD procedure). If a study is exploratory or hypothesis generating you may not be as worried about multiple testing.

Some of the Post-Hoc Comparison Methods

Many different post-hoc comparisons have been developed to better control the overall type I error rate (from less to more conservative):

- **Least Significant Difference (LSD)**: A sequential pairwise test of group means after ANOVA.
- **Duncan's Multiple Range Test**: Alternative range test that uses the harmonic mean of the sample size when the sample sizes are unequal.
- **False Discovery Rate (FDR)**: An approach to limit the proportion of false positive results to a reasonable level.
- **Dunnett's Test**: Used to compare several groups to a single control group; often used in clinical trials.
- **Tukey's Honestly Significant Difference (HSD) Test**: Uses the studentized range distribution to make all pairwise comparisons.
- **Bonferroni Adjustment**: Can be used for any C independent comparisons. Essentially you conclude that the p-value is significant if it is less than $\frac{\alpha}{C}$ instead of α. This is conservative, especially if the tests are not independent.
- **Scheffé's Test**: Can be used for any contrast of interest (not just pairwise comparisons), but can be very conservative.

False Discovery Rate

FDR Motivation

In some study settings several distinct hypothesis tests may be performed (e.g. genomic and proteomic experiments). The Bonferroni correction is sometimes used in these settings but can be overly conservative when 100s or 1000s of tests are performed.

One commonly employed, but not necessarily optimal, solution is to apply the False Discovery Rate (FDR) correction (latter part of section 12.4 in Rosner). The idea is to limit the number of falsely positive results to a reasonable level (e.g. 5% or 10%).

The FDR is still conservative when the tests are not independent. Permutations and bootstrap sampling of the original data to find adjusted p-values are good improvements over the FDR method since they work with the inherent dependence of tests in the original data.

An FDR Algorithm

A few different algorithms exist for false discovery rate calculations. The popular Benjamini-Hochberg algorithm for k tests is

1. Calculate the p-values for all comparisons/tests.
2. Rank the comparisons by p-values from smallest to largest.
3. Calculate $q = kp/rank$ for each test.
4. The FDR value for each test is the minimum of the q values for that test and all tests ranked higher than it.
5. The null hypothesis is rejected for all FDR values that are less than the pre-specified acceptable level (e.g. 5%, 10%).

FDR Example

We have conducted a study to investigate the relationship of 14 genes between cases for a given disease and controls without the disease. After conducting 14 tests we have the following results:

Test	Gene	p-value
1	A3	0.4883
2	A4	0.3169
3	HOXA5	0.4156
4	A7	0.2971
5	HOXA9	0.6393
6	A10	0.5606
7	B3	0.5842
8	B6	0.9442
9	B9	0.4741
10	MEIS1	0.7937
11	MEIS2	0.0451
12	PBX2	0.7554
13	PBX3	0.9901
14	ABC	0.0001

FDR Example

Based on our Benjamini-Hochberg algorithm:

Test	p-value	Gene	Rank	q (=kp/rank)	FDR: MIN(q for rank or higher)
14	0.0001	ABC	1	0.0014	0.0014
11	0.0451	MEIS2	2	0.3157	0.3157
4	0.2971	A7	3	1.3865	0.8950
2	0.3169	A4	4	1.1092	0.8950
3	0.4156	HOXA5	5	1.1637	0.8950
9	0.4741	B9	6	1.1062	0.8950
1	0.4883	A3	7	0.9766	0.8950
6	0.5606	A10	8	0.9811	0.8950
7	0.5842	B3	9	0.9088	0.8950
5	0.6393	HOXA9	10	0.8950	0.8950
12	0.7554	PBX2	11	0.9614	0.9260
10	0.7937	MEIS1	12	0.9260	0.9260
8	0.9442	B6	13	1.0168	0.9901
13	0.9901	PBX3	14	0.9901	0.9901

ABC is the only significant gene remaining after FDR, with MEIS2 no longer significant.

Multiple Corrections in R and SAS

R and SAS have functions we can use to implement multiple testing corrections to avoid having to implement corrections by hand each time we wish to account for multiple corrections.

In SAS you can use the PROC MULTTEST to implement both family-wise-controlling (e.g., Bonferroni, permutation, bootstrap) or FDR-controlling (e.g., FDR, FDR with permutation, FDR with bootstrap) corrections.

In R, you can use `p.adjust` to correct using methods like the Bonferroni or FDR. The output represents *adjusted* p-values for multiple comparisons, so even the Bonferroni will present an adjusted p-value (instead of just comparing the raw p-value to α/C).

FDR Example in R

```
pvec <- c('0.0001','0.0451','0.2971','0.3169','0.4156','0.4741','0.4883',
          '0.5606','0.5842','0.6393','0.7554','0.7937','0.9442','0.9901')
round(cbind('fdr'=p.adjust(pvec, method='fdr'),
            'bon'=p.adjust(pvec, method='bonferroni')),4)
```

```
##           fdr    bon
##  [1,] 0.0014 0.0014
##  [2,] 0.3157 0.6314
##  [3,] 0.8950 1.0000
##  [4,] 0.8950 1.0000
##  [5,] 0.8950 1.0000
##  [6,] 0.8950 1.0000
##  [7,] 0.8950 1.0000
##  [8,] 0.8950 1.0000
##  [9,] 0.8950 1.0000
## [10,] 0.8950 1.0000
## [11,] 0.9260 1.0000
## [12,] 0.9260 1.0000
## [13,] 0.9901 1.0000
## [14,] 0.9901 1.0000
```