Multiple Testing

Guduguntla Vamshi, Jianhong Chang , Dan Chen
NC State University

1. Overview & History

While performing a large number of statistical tests, some will have p-values less than the significance level purely by chance, even if all of the null hypothesis are really true. This is corrected by using controlling the Family wise error rate and the False Discovery Rates. We present two methods to address this problem - bonferroni adjustment and Benjamini-Hochberg procedure and demonstrate the ideas.

1.1 The problem with multiple comparisons:

Say you have a set of hypotheses that you wish to test simultaneously. The first idea that might come to mind is to test each hypothesis separately, using some level of significance (alpha). At first, this doesn't seem like a bad idea. However, consider a case where you have 40 hypotheses to test, and a significance level of 0.05. What's the probability of observing at least one significant result just due to chance?

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1-0.05)^{25} \\ &\sim 0.72 \end{aligned}$$

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1-0.05)^{40} \\ &\sim 0.88 \end{aligned}$$

Example: Garcia et al. Calorie intake, olive oil consumption and mammographic density among Spanish women.

So, with 25 tests being considered, we have an 72% chance of observing at least one significant result, even if all of the tests are actually not significant. There is no universally accepted approach for dealing with the problem of multiple comparisons; it is an area of active research.

Methods for dealing with multiple testing frequently call for adjusting alpha in some way, so that the probability of observing at least one significant result due to chance remains below your desired significance level.

1.2 Line of research:

- Bonferroni, C. E. "Teoria statistica delle classi e calcolo delle probabilità." Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3-62, 1936.
- Simes, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1995;57:289–300

2. Technical details:

2.1. FWER

Suppose interest lies in testing N contrasts (θ_i, i =1, 2, N) simultaneously. So the hypotheses being tested form a complete hypothesis: $H_0 : \{\theta_1 = \theta_1^0, ..., \theta_k = \theta_N{}^0\}$ vs. H_a : at least one $\theta_i \neq \theta_i^0, i \in \{1,2,...,\mathrm{N}\}$. (For two-sided tests, could also be one-sided tests). The family wise error rate (FWER) = P (reject one or more hypotheses | all hypotheses are null) =P (at least one type I error happened among the k tests). We hope to control the FWER so that overall, we have at least $(1 - fwer) \times 100\%$ chance to reject the complete null hypothesis and make the inference that not all the k contrasts are equal to their null.

2.1.1 Bonferroni

The Bonferroni inequality produces a simple and conservative method to control the FWER. For each of the test, the usual type I error rate α is replaced by $\alpha' = \alpha / \mathrm{N}$. As a result, the FWER could be adjusted to $\alpha^* = 1 - (1 - \alpha')^N = 1 - (1 - \alpha / \mathrm{N})^N \leq \alpha$.

2.1.2 Tukey-Kramer Correction:

If we want to compare the effects of levels of a factor, Tukey is the most useful correction for pairwise comparisons in balanced design (equal sample size for each level of the factor). The assumption is that each test statistic follows the studentized range distribution. Point estimator and the estimated variance are the same as those for a single pairwise comparison. The only difference between the

confidence limits for simultaneous comparisons and those for a single comparison is the multiple of the estimated standard deviation. So the simultaneous $(1-\alpha)\times 100\%$ confidence intervals are $\hat{\theta}_i \pm q(\alpha,t,t(n-1))\sqrt{\hat{\text{var}}\,\theta_i}$, for $i \in \{1,2,...,N\}$, where t is the number of levels of a factor, n is the sample size of each level.

2.1.3 Scheffe Correction:

For simultaneous $(1-\alpha)\times 100\%$ confidence intervals for any number of contrasts, use $\hat{\theta}_i \pm \sqrt{(t-1)(F(\alpha,t-1,t(n-1))}\sqrt{\hat{\text{var}}\,\theta_i}$, for $i \in \{1,2,...,N\}$, where t is the number of levels of a factor, n is the sample size of each level.

2.2 FDR

By controlling the FWER at α, we are able to control the chance of making any false positive conclusions (type I errors). However, we do not need to control FWER in many cases, and instead some false positives are allowed as long as the rate of false positives is low. With this consideration the use of false discovery rate (FDR) is motivated. The FDR is the proportion of rejected hypotheses that are in fact null. For N hypothesis tests, the result could be summarized in the table below:

	Accepted Null	Rejected Null	Total
True Null	N00	N01	N0
True Alternative	N10	N11	N1
Total	N-Nr	Nr	N

$$FDR = E\left(\frac{N_{01}}{\max(N_r,1)}\right) = E\left(\frac{N_{01}}{N_r} \mid N_r > 0\right) P(N_r > 0)$$

2.2.1 The Benjamini-Hochberg procedure:

BH method controls the FDR to a value no greater than a specified α, assuming p-values under the null are independent. Firstly, the p-values of the N tests are ordered: $p_{(1)} \le p_{(2)} \le ... \le p_{(N)}$. Then let

$\hat{k} = \arg\max_{1 \le k \le N} \left\{ k : p_{(k)} \le \frac{\alpha k}{N} \right\}$. If there is no such k, then reject nothing. If there is such a k, reject all nulls associated with $p_{(1)} \le \ldots \le p_{(\hat{k})}$. As a result, the FDR is controlled to $FDR \le \frac{\alpha N_0}{N}$.

3. Demonstration & Code

Comparison of Adjusted p-values

Bonferroni, BH (FDR) adjusted p-values and raw p-valus for a series of 8 ordered p-values is plotted. The 5th most significant outcome remains statistically significant at the alpha=0.05 level for all but the Bonferroni procedure. The Bonferroni method controls the family-wise error rate and attempt to limit the probability of even one false discovery, so is relatively strong (conservative). BH attempts to control the expected proportion of false discoveries and is relatively weak.

Demonstration of Benjamin-Hochberg procedure.

A series of ordered p-values are compared with their corresponding critical values. Find the test with the highest rank, for which the p value is less than or equal to its critical value, then reject all hypotheses associated with the p value smaller than it. In this example, test 1, 2, 3 and 4 are declared as significant, even though the p value of test 3 is greater than its critical value.

Demonstration of how family-wise error rate (FWER) control works

1000 simulations are run to generate p values for 10 tests, with 5 true nulls and 5 false nulls. We assume the true nulls have p-values distributed uniform (0, 1), and the false nulls have p-values distributed uniform with a user-determined maximum. Without any adjustment, the probability of rejecting at least one true null, i.e. type I error, is 23.9%. In contrast, Bonferroni controls FWR and the chance of rejecting one or more null is only 3.3%.

Demonstration of how false discovery rate (FDR) control woks

We run 10000 simulations of 100 tests with 90 true nulls and 10 alternatives. Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1. Under the alternative hypothesis p-values are skewed towards 0, so we use the beta distribution and control the strength of alternative by changing the shape parameter. BH procedure is used to calculate the number of rejected tests. Comparing with the true null to get the number of

false rejection, then we get realized FDR for each simulation. The mean realized FDR is an estimate of the true FDR, and always controlled to be smaller or equal to alpha * N0/N, which is 0.045 in this example.

Code

The link to code can be found here:

https://github.com/vamshing/Statistical-Techniques-for-Big-Data/blob/master/MultipleTesting/Presentation_example2.ipynb