Review of Multiple Comparison Correction methods

- From Bonferroni procedure to Efron's Local FDR -

2014140089 Park, Si Hyung

1 Introduction

Ever since works on multiple comparison problem (MCP) were started by Tukey (1949) and Scheffé (1953), who have both utilized $t-$statistic for pairwise sample mean comparisons, numerous procedures for controlling family-wise error rate (FWER) have been developed. Achieving higher testing power while controlling FWER for the same level has especially been a major concern in the field of MCP, rulling the field for almost 35 years before Benjamini and Hochberg suggested the concept of false discovery rate (FDR) in 1995.

In this brief report, I would like to review two major concepts in MCP - FWER, FDR - and some of the selected procedures which controls FWER or FDR in certain level. Historical meanings and important proofs were described to better highlight modifications and developments of a statistical concept over time. Simple simulation of multiple testing was also conducted to compare different methods of MCP correction.

1.1 Multiple comparison problem

Multiple comparison problem (MCP), also known as multiple hypothesis tesing problem or multiplicity problem, is a problem which occurs when a family-level hypothesis testing is made based on a set of individual hypotheses. Due to the fact that a family-level type I error rate (family-wise error rate; FWER) is always greater than experiment-level type I error ($\because$ $1-(1-\alpha)^m > \alpha$, if $m \geq 2$), an appropriate correction for the error control of family-level hypothesis testing is needed.

1.2 Adjusted p-values

Adjusted p-values are the values that are calculated from original p-values. We can compare adjusted p-values with the level which we control family-level error for. e.g., if an adjusted p-value of a hypothesis H_i is less than or equal to α_0, then we can reject H_i while controlling $FWER$ for α_0. Unlike p-values, adjusted p-values do not give us the information about type I error probability of that hypothesis.

1.3 Regression dependency of hypotheses

For some family-level error controlling procedure, assumption on dependence structure is needed. Let I_0 be a subset of indices of hypotheses in a set of tests. If there exists no regression dependence between non-rejected test statistics and true-null test statistics, then the hypotheses in the set are said to be independent. If there is a positive regression dependence between non-rejected test statistics and true-null test statistics, then the hypotheses in the set are said to have positive regression dependency on each one from a subset I_0 ($PRDS$ on I_0). Benjamini and Yekutieli (2001) defined $PRDS$ on I_0 as follows: "For any increasing set D, and for each $i \in I_0$, $P(X \in D|X_i = x)$ is non-decreasing in x". More intuitive explanation uses p-values. For each $i \in I_0$, If p-values increases, then probability of corresponding null hypothesis is true does not decreases.

1.4 Notations

	Declared as non-significant	Declared as significant	Total
True H_0	U	V	m_0
False H_0	T	S	$m_1 = m - m_0$
Total	$m - R$	R	m

Table 1: notation of the number of hypotheses in corresponding to each cell

Notations used to define family-level error concepts mathematically are as described on Table 1. This notation follows that of Benjamini and Hochberg (1995). Total m hypotheses in a set is being tested in this situation. $m_0 \leq m$ hypotheses are true null. Note that R, the number of rejected hypotheses, is an observable random variable while U, V, T, S are unobservable random variables.

2 Family-Wise Error Rate (FWER)

Family-wise error rate is a probability that at least one type I error occurs when testing each individual hypothesis in a set. Using previous notations, it can be defined as $P(V \geq 1) = 1 - P(V = 0) = E(I_{V \geq 1})$. If FWER is controlled at level α_0, then probability of one or more type I error occurs in the set is less than or equal to α_0.

In early works, FWER was controlled by using a specific type of test statistic - t, normal, multivariate t, or multivariate normal. After Bonferroni procedure was introduced, procedures using p-values from individual testings became dominant.

2.1 Bonferroni procedure

Bonferroni procedure, widely used nowadays, was suggested by Dunn (1961). Dunn originally suggested a method to construct confidence interval in MCP using Bonferroni's inequality, which the procedure was named after. By testing individual hypotheses on level α/m, we can control FWER at level α. We can prove this with Boole's inequality (1).

$$\begin{aligned} &\text{By Boole's inequality } P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i), \\ &FWER = P(\cup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m})) \leq \sum_{i=1}^{m_0} P(p_i \leq \frac{\alpha}{m}) = m_0 \frac{\alpha}{m} \leq \alpha \end{aligned} \tag{1}$$

Adjusted (or "Bonferronied") p-value of ith hypothesis is equal to $min(p_i \times m, 1)$.

Algorithm 1 Bonferroni Procedure

```
procedure BONFERRONI
    $pvals$ ← array of p-values
    $adj\_pvals$ ← empty array of length same as $pvals$
For $i$ in 1:$length(pvals)$:
    $adj\_pvals[i] \leftarrow pvals[i] * length(pvals)$
    if $adj\_pvals[i] \leq \alpha$ then reject $i$th hypothesis
```

Dunn simplified MCP correction in two major ways. (i) Dunn introduced a control procedure which is based on individual p-values. It is possible to conduct any kind of multiple testing, even it does not uses specific type of test statistic. (ii) Bonferroni procedure is remarkably computationally efficient. By setting the procedure as *algorithm 1*, Bonferroni correction only costs $O(n)$. In many big-data analysis where high complexity algorithms are not feasible to use, Bonferroni procedure can help. Also, Bonferroni procedure does not require the assumption of regression dependence. However, excessive loss in testing power in Bonferroni procedure led to the development of stepwise FWER controlling procedures.

2.2 Holm procedure

In addition to improve power of test while controlling for same level of FWER, Holm (1979) designed a step-up procedure. In fact, it is one of the first stepwise algorithm for MCP correction. Holm procedure can be conducted as follows.

1. Sort p-values in increasing order: $p_{(1)}, p_{(2)}, ..., p_{(m)}$.

   - Each of the p-value corresponds to $H_{0_{(1)}}, H_{0_{(2)}}, ..., H_{0_{(m)}}$.

2. Find the smallest j, in which $p_{(j)} > \dfrac{\alpha}{m-j+1}$.

3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., (j-1)$.

By rejecting hypotheses with Holm procedure, we can control FDR for level α_0 (2).

If the smallest index which satisfies $p_{(j)} > \dfrac{\alpha}{m-j+1}$ is j, then

$$m-j+1 \geq m_0,\ \frac{1}{m-j+1} \leq \frac{1}{m_0},\ p_{(j-1)} \leq \frac{\alpha}{m-j+2} \leq \frac{\alpha}{m_0}. \tag{2}$$

$\therefore$ By Boole's inequality, $FWER = P(\cup_{i=1}^{m_0}(p_i \leq \dfrac{\alpha}{m_0})) \leq \displaystyle\sum_{i=1}^{m_0} P(p_i \leq \dfrac{\alpha}{m_0}) = \alpha$

Adjusted p-value of ith hypothesis for Holm procedure is equal to $min(p_i \times (m-i+1), 1)$. As Holm procedure compares individual p-values to $\dfrac{\alpha}{m}, \dfrac{\alpha}{m-1}, ..., \dfrac{\alpha}{1}$ instead of uniform level $\dfrac{\alpha}{m}$ as in Bonferroni procedure, Holm procedure is uniformly more powerful than Bonferroni procedure, though power gain is little. Holm procedure also does not assume specific regression dependence of hypotheses.

2.3 Hochberg procedure

Hochberg (1988) sharpened Holm's method to be more powerful. In his two-paged paper, he modified Holm procedure and suggested step-up algorithm with similar rejection critetion.

1. Sort p-values in increasing order: $p_{(1)}, p_{(2)}, ..., p_{(m)}$.
2. Find the largest j, in which $p_{(j)} \leq \dfrac{\alpha}{m-j+1}$.
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., j$.

With a little modification, Hochberg procedure is uniformly more powerful than Holm procedure. However, this algorithm requires non-negative regression dependence assumption to properly control FWER at certain level. Although testing power increase in Hochberg procedure is almost negligible, it later inspired Hochberg to develop a step-up FDR controlling procedure.

3 False Discovery Rate (FDR)

While the field concentrated on improving testing power of FWER controlling method, some insisted to develop a new concept of family-level error, other than FWER. FWER has drawbacks: (i) Controlling for FWER has much less power than controlling for per comparison error. (ii) In many cases it is not needed to control FWER in strong manner made. Saville is the one who recommended using per comparison error rate (PCER) than FWER, because of these reasons.

Inspired by the fact that the ratio of the number of false rejection to the number of total rejection might be a useful error rate to control for, Benjamini and Hochberg (1955) introduced a seminal concept, false discovery rate (FDR). FDR is defined as "expectation of ratio of the number of erroneous rejection to the number of total rejection", which can be denoted as $E(Q) = E(V/R) = E(V/(S+V))$. By defining FDR as $E(Q)$, relationship with FWER can be derived.

1. if $m_0 = m$, then FDR = FWER.
   $\because s = 0,\ v = r.\ P(V \geq 1) = E(Q)$

2. if $m_0 < m$, then FDR $\leq$ FWER.
   $\because v > 0,\ Q \leq I_{V \geq 1},\ E(Q) \leq E(I_{V \geq 1}) = P(V \geq 1)$

Because of this relationship, controlling for FWER also controls FDR in strong manner. In contrary, controlling for FDR weakly controls FWER (does not assure that FWER is also being controlled).

3.1 Benjamini-Hochberg (B-H) procedure

Alongside the introduction of FDR, Benjamini and Hochberg developed a step-up procedure to easily control FDR at certain level q^*.

1. Sort p-values in increasing order: $p_{(1)}, p_{(2)}, ..., p_{(m)}$.
2. Find the largest j, in which $p_{(j)} \leq \dfrac{j}{m} q^*$.
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., j$.

Unlike Hochberg procedure for FWER control,

3.2 Benjamini-Yekutieli (B-Y) procedure

3.3 Brief introduction to local FDR

4 Discussions - Simulation and Comparisons

5 References

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure". Scandinavian Journal of Statistics, 6(2), 65-70

Abdi, H. (2010). ”Holm’s Sequential Bonferroni Procedure”. University of Texas at Dallas

Aickin, M., Gensler, H. (1996). ”Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods”. American Journal of Public Health, 86(5), 726-728

Benjamini, Y., Hochberg, Y. (1995). ”Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300

Benjamini, Y., Yekutieli, D. (2001). ”The control of the false discovery rate in multiple testing under dependency”. Ann. Statist. 29(4), 1165-1188

Robinson, D. (2015). ”Understanding empirical Bayes estimation (using baseball statistics)”. From Variance Explained, retrieved at Dec. 2017

Efron, B., Tibshirani, R., Storey, J., Tusher, V. (2001). ”Empirical Bayes Analysis of a Microarray Experiment”. Stanford University Technical Report. 216

Efron, B. (2005). ”Local False Discovery Rates”. Stanford University Technical Report. 234