Correcting Multiple Comparison Problems

Park, Si Hyung

Division of Life Sciences, Korea Univ.

December, 2017

Multiple Comparison Problem

- Multiple Comparison Problem
- Multiple Comparison correction

Family-wise Error Rate

- Bonferroni
- Holm, Hochberg

False Discovery Rate

- Benjamini and Hochberg
- Benjamini and Yekutieli
- Local FDR

Simulation & Comparison

Multiple Comparison Problem

Also known as "multiple hypothesis testing".

Occurs when simultaneously making inference about a set of hypothesis

e.g. Efficacy of drug

$$H_{0_1} : \text{not effective to patient 1}$$
$$H_{0_2} : \text{not effective to patient 2}$$
$$\dots$$
$$H_{0_m} : \text{not effective to patient m}$$

Family-level hypothesis is: "Is drug effective to the patients?"
If level of individual hypothesis is 0.05, then family-wise error rate is $1 - (1 - 0.05)^m > 0.05$

Multiple Comparison Correction

Controlling error rate of a family-level hypothesis.

Two different major types of error rate in MCP correction:

- FWER: probability of at least one type I error occurs in each individual hypothesis testing
- FDR: expected proportion of false positive to positives

Multiple Comparison Correction

Initiated by J. Tukey (1949) and H. Scheffé (1953), on pair-wise comparison of multiple sample means.

Early works utilized t or multivariate normal statistics.

After Bonferroni procedure was developed, p-values were used.

Notations - testing hypotheses

	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: notation of number of hypothesis in corresponding to each cell

R is an observable random variable.

U, V, S, T are unobservable random variables.

Family-wise Error Rate (FWER)

Probability that at least one type I error occurs in individual hypothesis testing.

- $\leftrightarrow$ experiment-wise error rate
- $P(V \geq 1) = 1 - P(V = 0)$
- If $FWER \leq \alpha$, then probability of one or more type I error of the family is controlled at level α

Tukey, Scheffé, Dunn, ... contributed to FWER controlling method

Bonferroni procedure

Tukey (1949) and Scheffé (1953) invented pairwise comparison method for sample means, both utilize t-statistics.

However It's cumbersome to use these methods in MCP based on other test statistics.

Bonferroni procedure was suggested by O. J. Dunn (1961), named after Bonferroni's inequality (generalized Boole's inequality).

Bonferroni procedure uses per-hypothesis p-values to control FWER, rather than a specific test statistic.

Bonferroni procedure

Suppose there is a set of m hypotheses. We want to test the family-wise hypothesis with level α.

By testing per-hypothesis on level $\frac{\alpha}{m}$, we can control $FWER \leq \alpha$.

Proof

By Boole's inequality: $P(\cup_{i=1}^{m} E_i) \leq \sum_{i=1}^{m} P(E_i)$,

$$FWER = P(\cup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m})) \leq \sum_{i=1}^{m_0} P(p_i \leq \frac{\alpha}{m}) = m_0 \frac{\alpha}{m} \leq \alpha$$

Bonferroni procedure

Suppose p_i is a p-value of individual hypothesis H_{0_i} in hypothesis family. Then, "Bonferronied" or Adjusted p-value

$$p_i^{adj} = p_i \times m.$$

Note that interpretation of adj. p-value is way different than that of original p-value.

- $p_i^{adj} \leq 0.05$: can reject H_{0_i} while controlling for $FWER \leq 0.05$
- $p_i \leq 0.05$: probability of type I error is less than 0.05.

Bonferroni procedure

Pros

- Do not need assumption of dependence
- Remarkably simple and computationally efficient

Cons

- **Excessive amount of loss in testing power** (too conservative)

Holm procedure

- One of the first sequential (step-wise) algorithms (1979).
- Holm procedure is especially, a step-down algorithm
- Uniformly more powerful than Bonferroni procedure, while controlling for FWER in strong sense.

Holm procedure

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 $p_{(j)} > \frac{\alpha}{m - j + 1}$.
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., (j - 1)$.

"Stepping down" the indices until we find the smallest j.

Adjusted p-value: $p_{(i)}^{adj} = (m - i + 1) \times p_{(i)}$

Holm procedure

Proof

Let I be the set of indices of the true H_0's. then length of I is m_0.

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

$$m-j+1 \geq m_0, \ \frac{1}{m-j+1} \leq \frac{1}{m_0}$$

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

By Boole's inequeality,

$$\therefore P(\cup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m_0})) \leq \sum_{i=1}^{m_0} P(p_i \leq \frac{\alpha}{m_0}) = \alpha$$

Thus, FWER is controlled at level α.

Holm procedure

- Do not need assumption of dependence.
- Holm procedure compares individual p-values with $\frac{\alpha}{m}, \frac{\alpha}{m-1}, ..., \frac{\alpha}{1}$, which leads to power increase.
- Holm procedure is uniformly more powerful than Bonferroni procedure.

Hochberg procedure

Sharpened from Holm or Sime's method

Step-up algorithm (1988)

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 **largest** j, in which $p_{(j)} \leq \frac{\alpha}{m-j+1}$.
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., j$.

Uniformly more powerful than Holm procedure.

However, needs **non-negative dependence** assumption.
(as $p_{(i)}$ increases, probability of being part of true null increases)

Benjamini and Hochberg introduced a seminal concept, "False Discovery Rate" at 1995.

In many cases, we do not have to control FWER, especially when we want to explore as many potential effects as possible.

FDR controls for "the ratio of erroneous rejection to all rejected hypothesis", instead of probability of at least one type I error.

False Discovery Rate (FDR)

FDR Q_e: Expectation of proportion of 'falsely rejected nulls (V)' to 'rejected nulls (R)'.

- $FDR\ Q_e = E(Q) = E(V/R) = E(V/(V+S))$
- If $m_0 = m$, then $FDR = FWER$
  $\because s = 0,\ v = r.\ P(V \geq 1) = E(Q)$
- If $m_0 < m$, then $FDR \leq FWER$
  $\because v > 0,\ Q \leq I_{V \geq 1},\ E(Q) \leq E(I_{V \geq 1}) = P(V \geq 1)$

Instead of "p", we use "q" to denote the level of FDR.

Benjamini and Hochberg founded the concept of FDR (1995).

False Discovery Rate (FDR)

When to use FDR?

- Family-wise conclusion is not sensitive to individual failure.
- Exploratory analysis.
- Screening for potential discoveries.

B-H Procedure

Benjamini and Hochberg (1995), step-up algorithm

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 largest j, in which $p_{(j)} \leq \frac{j}{m} q^*$, where q^* is a level to control FDR for.
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., j$.

Compares $p_{(i)}$ to linear form of constant ($\propto i$) instead of hyperbolic form ($\propto \frac{1}{i}$), which leads to power increase.

Needs non-negative dependence assumption.

B-Y Procedure

Benjamini and Yekutieli (2001), for other forms of dependency.

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 largest j, in which $p_{(j)} \leq \frac{j}{m \cdot c(m)} q^*$.
   - q^* is a level to control FDR for.
   - $c(m) = \begin{cases} 1, & \text{under non-negative dependence} \\ \sum_{i=1}^{m} \frac{1}{i}, & \text{otherwise} \end{cases}$
3. Reject $H_{0_{(k)}}$'s, where $k = 1, 2, ..., j$.

By adding constant term $c(m)$, we can control FDR at level q^*.

B-Y procedure - Proof

Let $C_{v,s}^{(i)}$ be the event in which if p_i is rejected then $(v-1)$ true nulls and s false nulls are rejected alongside with it. Then,

$$C_k^{(i)} = \cup\{C_{v,s}^{(i)}:\ v+s=k\},\ E(Q) = \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}P(p_i \le q_k \cap C_k^{(i)}).$$

Denote $p_{ijk} = P(\{p_i \in [\frac{j-1}{m}q^*, \frac{j}{m}q^*]\} \cap C_k^{(i)})$. Then,

$$\sum_{k=1}^{m} p_{ijk} = P(\{p_i \in [\frac{j-1}{m}q^*, \frac{j}{m}q^*]\} \cap (\cup_{k=1}^{m} C_k^{(i)})) = \frac{q^*}{m},$$

$$\begin{aligned} E(Q) &= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\sum_{j=1}^{k} p_{ijk} = \sum_{i=1}^{m_0}\sum_{j=1}^{m}\sum_{k=j}^{m}\frac{1}{k}p_{ijk} \\ &\le \sum_{i=1}^{m_0}\sum_{j=1}^{m}\sum_{k=j}^{m}\frac{1}{j}p_{ijk} \le \sum_{i=1}^{m_0}\sum_{j=1}^{m}\frac{1}{j}\sum_{k=1}^{m} p_{ijk} \\ &\therefore = m_0\sum_{j=1}^{m}\frac{1}{j}\frac{q^*}{m} \end{aligned}$$

Under arbitrary dependence, FDR is increased by no more than $\sum_{j=1}^{m}\frac{1}{j}$.

Local FDR

Efron et al. (2001) used **empirical bayes approach** to FDR.

Originally developed to test microarray data.

Empirical bayes method

1. Estimate prior distribution using the whole data
2. Use 1. as a prior for each individual estimate

Local FDR

Local FDR is defined as: $fdr(Z) = \frac{p_0 f_0(Z)}{f(Z)}$.

$p_0 = 1 - p_1 =$ probability that a null is true.

$f_0(z) =$ the density of true nulls.

$f_1(z) =$ the density of false nulls.

$f(z) = p_0 f_0(z) + p_1 f_1(z)$

$fdr(Z)$ is the posterior probability that a null is true.

Local FDR gives answer to the question:
"How many decisions are wrong in the local area of this hypothesis of interest?"

Local FDR

e.g. microarray data

We have microarray data of 6,810 gene scores (Z's).

- $\underline{74}$ of 6,810 genes have scores in the interval $Z \in [1.9, 2.1]$.
- In Twenty permutated null score data sets, 676 fell into $[1.9, 2.1]$. Thus average of $\frac{676}{20} = \underline{33.8}$ fell into the interval.
- p_0 was estimated as maximum $\underline{0.811}$

If we declare all genes with score in $[1.9, 2.1]$ significant,

$\therefore fdr = \frac{0.811 \times 33.8}{74} = 0.37$ are expected to be erroneous.

Simulation & Comparison

Simulation scheme

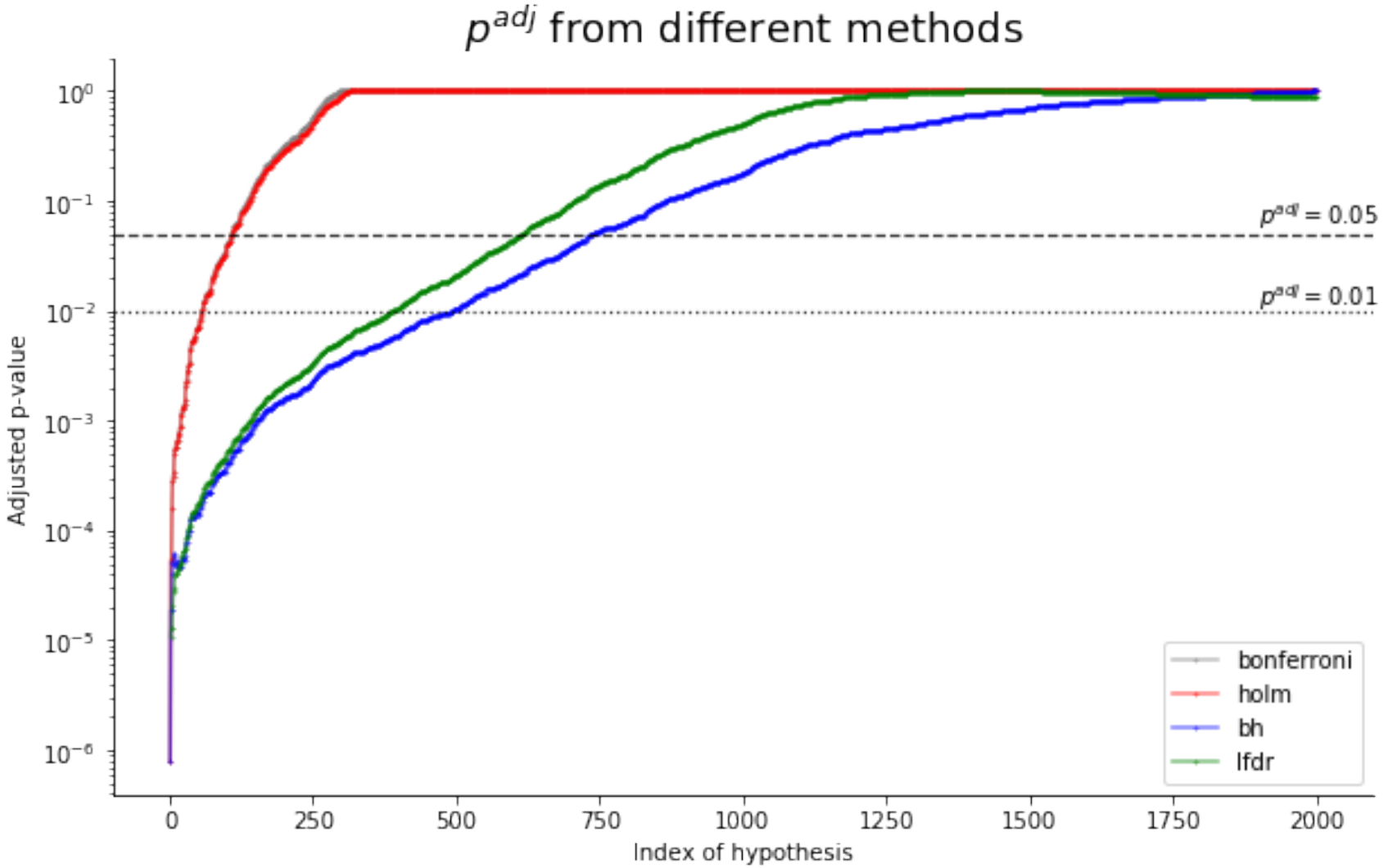
1,000 samples were generated from $N(0, 1)$,
and 1,000 other samples were generated from $N(3, 1)$.

p-values were calculated by Z-testing each sample.

Two FWER methods (Bonferroni, Holm) and Benjamini-Hochberg FDR method, local FDR were compared.

[implementation] [notebook]

Simulation & Comparison



Simulation & Comparison

	Bonferroni	Holm	B-H
True rejection	107	109	**721**
False rejection	1	1	18
Total rejection	108	110	739
Ratio of false rejection	.009	.009	**.014**
Ratio of true acception	.529	.529	.780

Table: simulation result at $FWER \leq 0.05$ or $FDR \leq 0.05$

References

- Savin, N. (1980). "The Bonferroni and the Scheffé Multiple Comparison Procedures". The Review of Economic Studies, 47(1), 255-273
- Weisstein, E. "Bonferroni Inequalities". From MathWorld - A Wolfram Web Resource, retrieved at Dec. 2017
- 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". Technical Report. 216