How Not to be Too Wrong?

An Introduction to the Multiple Testing Problem

Zhang Ruiyang

University College London

ruiyang.zhang.20@ucl.ac.uk

29 Nov 2021

Abstract

This talk will be on the Multiple Testing Problem, which occurs when we are performing simultaneous hypothesis testing on the same set of data. The issue gets increasingly severe as the number of tests increases. The problem is hard to neglect, and is common in real life with examples including genetic research like GWAS and AB Testing for web designs. This talk will go over some basic remedies to the problem - the FWER controls like Bonferroni correction and the FDR controls like the BH procedure. It is self-contained and requires no prerequisite, but it would be helpful to have some knowledge of basic Statistics such as hypothesis testing and significance level, although they will be covered too.

Fisherian Test (1)

Setting: 8 cups - 4 milk then tea; 4 tea then milk

Experiment: Taste each cup, indicate whether it is milk then tea or tea then milk

Hypothesis: No such ability.

Significance Level: 5%

Prob. of 4 Correct Guesses at random: $1 \div \binom{8}{4} \approx 0.014 < 5\%$
Prob. of 3+ Correct Guesses at random: $(1 + \binom{4}{3}\binom{4}{1}) \div \binom{8}{4} \approx 0.24 > 5\%$

Fisherian Test (2)

Hypothesis: One (null) hypothesis in mind.

p-value: Probability p of having the given data assuming the null hypothesis is true.

Significance Level: A pre-determined, reasonable yet arbitrary threshold α to reject the null hypothesis.

Outcome: (1) $p < \alpha$, reject null (2) Otherwise, not reject null

Neyman-Pearson Test (1)

Hypothesis: Null Hypothesis H_0, the claim that we aim to reject, and Alternative Hypothesis H_1, some different claim.

p-value: Probability p of having the given data assuming the null hypothesis is true.

Significance Level: Predetermined. Probability of rejecting the null hypothesis when the null hypothesis is actually true.

Outcome: (1) $p < \alpha$, reject null (2) Otherwise, not reject null

Neyman-Pearson Test (2)

	H_0 is true	H_0 is false
Reject H_0	Type I Error False Positive α	Power True Positive $1 - \beta$
Not Reject H_0	True Negative $1 - \alpha$	Type II Error False Negative β

Multiple Testing Problem

Under true null, the random variable **p-value** follows Uniform(0, 1).

If $\alpha = 0.05$, we will make a false positive on average for every 20 true null tests.

20 true nulls, 1 false positive on average. 1000 true nulls, 50 false positives on average.

FWER

Given m null hypotheses denoted by $H_1, H_2, \cdots, H_m$:

	True Null Hypothesis	False Null Hypothesis	Total
Reject	V	S	R
Not Reject	U	T	$m - R$
Total	m_0	$m - m_0$	m

V, S, U, T are random variables, R, m, m_0 are known.

Family-wise Error Rate: $\text{FWER} = \mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0)$.

Bonferroni Correction

Boole's Inequality: For a countable set of events $A_1, A_2, A_3, \cdots$, $\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i)$, due to sub-additivity property of probability measure.

With this inequality, since we are rejecting each null hypothesis H_i when its p-value $p_i \leq \frac{\alpha}{m}$, we would have

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}\Big(\bigcup_{i=1}^{m_0} (p_i \leq \frac{\alpha}{m})\Big) && m_0 \text{ true null hypothesis} \\
&\leq \sum_{i=1}^{m_0} \mathbb{P}(p_i \leq \frac{\alpha}{m}) && \text{by Boole's inequality} \\
&= \sum_{i=1}^{m_0} \frac{\alpha}{m} && \text{p-value follows Uniform(0,1) under true null} \\
&= \frac{m_0}{m}\alpha \leq \alpha,
\end{aligned}
$$

which indicates that the FWER is controlled under level α.

Holm-Bonferroni Method (1)

(Holm, 1979)
For the m null hypotheses $H_1, H_2, \cdots, H_m$, we compute their respective p-values $P_1, P_2, \cdots, P_m$ and we rank them such that $P_{(k)}$ denotes the k-th smallest p-value. So, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, and we denote the corresponding null hypotheses as $H_{(1)}, H_{(2)}, \cdots, H_{(m)}$. We would want to control the FWER at α.

- Is $P_{(1)} \leq \frac{\alpha}{m}$? If so, reject $H_{(1)}$ and continue. Otherwise, EXIT.
- Is $P_{(2)} \leq \frac{\alpha}{m-1}$? If so, reject $H_{(2)}$ and continue. Otherwise, EXIT.

  $\vdots$

- Is $P_{(k)} \leq \frac{\alpha}{m-k+1}$? If so, reject $H_{(k)}$ and continue. Otherwise, EXIT.

Holm-Bonferroni Method (2)

To show that this method does in fact keep the FWER at α, we let I_0 be the set of indices corresponding to the true null hypothesis. This set is unknown to us and has size m_0. Let us assume that we make the first false positive decision at $H_{(h)}$. Based on the procedures stated above, all the decisions for null hypothesis $H_{(1)}, H_{(2)}, \cdots, H_{(h-1)}$ are true positives. Also, we know for a fact that $h-1 \leq m-m_0$ due to the definition of $m-m_0$. This implies that $m-h+1 \geq m_0$, and then $\frac{1}{m-h+1} \leq \frac{1}{m_0}$. Now, since $H_{(h)}$ is rejected, we would have $P_{(h)} \leq \frac{\alpha}{m-h+1}$ by definition, so we will then have $P_{(h)} \leq \frac{\alpha}{m-h+1} \leq \frac{\alpha}{m_0}$. This means, if there is any false positive, we have at least one true null hypothesis with p-value less than $\frac{\alpha}{m_0}$.

Holm-Bonferroni Method (3)

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}\Big(\bigcup_{i \in I_0} (P_i \leq \frac{\alpha}{m_0})\Big) \\
&\leq \sum_{i \in I_0} \mathbb{P}(P_i \leq \frac{\alpha}{m_0}) \qquad \text{by Boole's inequality} \\
&= m_0 \frac{\alpha}{m_0} \qquad \text{p-value follows Uniform(0,1) under true null} \\
&= \alpha,
\end{aligned}
$$

which indicates that the FWER is controlled at level α.

Holm-Bonferroni Method is uniformly more powerful than Bonferroni.

FDR

	True Null Hypothesis	False Null Hypothesis	Total
Reject	V	S	R
Not Reject	U	T	$m - R$
Total	m_0	$m - m_0$	m

(Benjamini & Hochberg, 1995)
Define a new random variable $Q = V/(V + S)$ where $Q = 0$ when $V + S = 0$. This is the proportion of the false positives over all the rejected null hypotheses. This is unobservable since we do not know V, or S, or their realisations v or s. We will define the False Discovery Rate, or **FDR**, as the expectation of Q,

$$\mathrm{FDR} = \mathrm{E}[Q] = \mathrm{E}[V/(V + S)] = \mathrm{E}[V/R].$$

To avoid the division by zero issue, we would have the alternative formula for FDR as

$$\mathrm{FDR} = \mathrm{E}[V/R|R > 0]\mathbb{P}(R > 0).$$

FDR Properties

(1) If all the null hypotheses are true, FDR = FWER. When $s = 0$ and $v = r$, $Q = 0$ if $v = 0$ and $Q = 1$ if $v > 0$, which means $\mathbb{P}(V \geq 1) = \mathrm{E}[Q]$. This means a control of FDR is a control of FWER in the weak sense.

(2) When $m_0 < m$, FDR is no bigger than FWER. Given $m_0 < m$, $v > 0$ implies $v/(v+s) \leq 1$, which means $V \geq 1 \implies V \geq v/(v+s)$ and $\mathbb{P}(V \geq 1) \geq \mathrm{E}[Q]$. This means a control of FWER will control FDR.

Benjamini-Hochberg Procedure

For the m null hypotheses $H_1, H_2, \cdots, H_m$, we compute their respective p-values $P_1, P_2, \cdots, P_m$ and we rank them such that $P_{(k)}$ denotes the k-th smallest p-value. So, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, and we denote the corresponding null hypotheses as $H_{(1)}, H_{(2)}, \cdots, H_{(m)}$. We let the level of FDR that we would want to control at as α. The procedure works as the following:

- Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$.
- Reject all $H_{(i)}$ where $i = 1, 2, \cdots, k$.

Theorem

*For **independent** test statistics and for any configuration of false null hypotheses, the above precedure controls the FDR at α.*

Extensions

- Weaken Independence Criteria for BH Procedure
- FDR Control using Empirical Bayes methods
- Online Multiple Testing
- Knockoff for Variable Selection

Reference

- 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*, **57** No. 1, 289-300.
- Benjamini Y. (2010) Discovering the False Discovery Rate. *Journal of the Royal Statistical Society, Series B*, **72**, Part 4, 405-416.
- Holm S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*. **6**, 65-70.