多重檢定問題，false discovery rate (FDR)與 q 值

王紋璋

在統計假設檢定分析過程中，根據收集的資料我們得到檢定統計量（test statistic）之樣本觀察值；然後，我們計算在虛無假設（null hypothesis）下檢定統計量出現與觀察值相同或更極端之值的機率，稱為 p 值。p 值越小，我們越傾向於否定虛無假設。當 p 值小於給定的門檻 α ，稱為顯著水準（significance level），我們拒絕虛無假設，並稱檢定結果顯著。當虛無假設為真時，仍有可能因所得 p 值小於給定的 α ，而獲致拒絕虛無假設的錯誤結論，這種錯誤稱為型一錯誤（type I error）。進行單一檢定時，犯型一錯誤的機率等於給定的顯著水準 α。

當同時進行多個檢定時，犯型一錯誤之機率會增加。以同時進行兩個檢定為例，當兩個虛無假設為真且 $\alpha = 0.05$ 時，個別檢定不犯型一錯誤的機率為 0.95。若兩個檢定獨立，則同時不犯型一錯誤之機率為 $0.95^2 = 0.9025$ 。因此，至少一個檢定犯型一錯誤的機率為 $1 - 0.9025 = 0.0975$ 。檢定個數越多，犯型一錯誤之機率就越高。這個現象被稱為多重檢定問題（multiple testing problem）。

傳統上，Bonferroni 校正是處理多重檢定問題最常使用的方法。但是當檢定數量大時，執行 Bonferroni 校正常面臨檢定力（power）不足的情況。本文主要將介紹另一類常用的方法：false discovery rate (FDR)與 q 值。

令研究中總檢定個數為 m，且令其中虛無假設為真之檢定個數為 m_0，則對立假設（alternative hypothesis）為真之檢定個數為 $m - m_0$。 m 是由研究者所決定的常數，m_0 是未知的常數。給定一個拒絕虛無假設的標準（如 p 值<0.05），令拒絕虛無假設之檢定個數為 S。在這 S 個拒絕虛無假設之檢定中，令虛無假設為真（即犯型一錯誤）之個數為F，對立假設為真之個數為 T，則 $F + T = S$。由於 S 、 F 與 T 的值會隨著樣本的不同而產生變動，因此是隨機變數。其中，S 是觀察得到的變數，而 F 與 T 是觀察不到的變數。我們可將 m 、 m_0 、 S 、 F 及 T 這些數量之關係整理如表一。下面，我們將以表一為基礎，介紹處理的多重檢定問題的方法。

表一

	拒絕虛無假設	不拒絕虛無假設	總數
虛無假設為真	F	$m_0 - F$	m_0
對立假設為真	T	$m - m_0 - T$	$m - m_0$
總數	S	$m - S$	m

Familywise error rate（FWER）

傳統上處理多重檢定問題最常使用的方法是控制 familywise error rate（FWER）。FWER 之定義為 $\Pr(F \geq 1)$，亦即發生一個以上型一錯誤的機率。控制 FWER 最簡單的方法是進行 Bonferroni 校正：欲確保 FWER ≤ α，需將個別檢定可容許之型一錯誤率定為 $\frac{\alpha}{m}$。固定 α，當檢定個數越多，$\frac{\alpha}{m}$ 之值越小，越不容易拒絕虛無假設，因此對大量檢定進行 Bonferroni 校正時，會大幅降低檢定力。關於進行 Bonferroni 校正可能會面臨的其它問題，讀者可參考林彥光老師在生統期刊第五期所撰寫的「淺談 Bonferroni 事後校正」一文。

False discovery rate（FDR）

另一種度量型一錯誤程度的方法，是考慮顯著結果中型一錯誤的比例：

$$\frac{\text{型一錯誤個數}}{\text{拒絕虛無假設個數}} = \frac{F}{S}。$$

$\frac{F}{S}$ 是一個隨機變數。Benjamini & Hochberg (1995)提出以 $\frac{F}{S}$ 之期望值，作為所犯型一錯誤程度之度量，稱為 false discovery rate（FDR）：

$$\text{FDR} = \text{E}\left[\frac{F}{S}\right]。$$

Benjamini & Hochberg (1995)提出一個根據所得 p 值，設定拒絕虛無假設的標準，以控制 FDR 的方法。令 $p_1, p_2, \ldots, p_m$ 為 m 個檢定所得之 p 值。給定一個門檻 q^*，

執行以下程序可使 FDR $\leq q^*$：

(1) 令 $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ 為由小到大排序之 p 值，並令 $H_{(i)}$ 代表對應於 $p_{(i)}$ 之虛無假設。

(2) 令 k 為最大的 i 值，使 $p_{(i)} \leq \frac{i}{m} q^*$ 成立， $i = 1, \cdots, m$ 。

(3) 拒絕虛無假設 $H_{(i)}$， $i = 1, \cdots, k$。

q 值

當進行多重檢定時，Storey & Tibshirani (2003)對每個檢定分別估計一個 q 值，作為其顯著性的一種度量。對任意一個檢定 i ，令 p_i 及 q_i 為對應之 p 值與 q 值。 q_i 的意義為，當稱檢定 i 及其他 p 值 $\leq p_i$ 之檢定結果為顯著時，預期的型一錯誤比例。因此，q 值是一種以 FDR 為基礎的度量。

令 $S(t)$ 、 $F(t)$ 及 FDR(t) 分別代表，以 t 為檢定之顯著水準時，表一中的 S 、 F 及對應的 FDR。欲估計每個檢定對應之 q 值，需先估計 FDR(t) ，$0 < t \leq 1$。當檢定個數 m 大時，

$$\text{FDR}(t) = E\left[\frac{F(t)}{S(t)}\right] \approx \frac{E[F(t)]}{E[S(t)]} \quad ,$$

故可分別估計 E$[S(t)]$ 與 E$[F(t)]$ 。由於 $S(t)$ 是可觀察到的，因此可用其觀察值 $\#\{p_i \leq t\}$ 估計 E$[S(t)]$。另一方面，雖然 $F(t)$ 是觀察不到的，但虛無假設為真時，檢定所得之 p 值會均勻分布於 0 與 1 之間，此時 p 值 $\leq t$ 之機率即為 t ，故

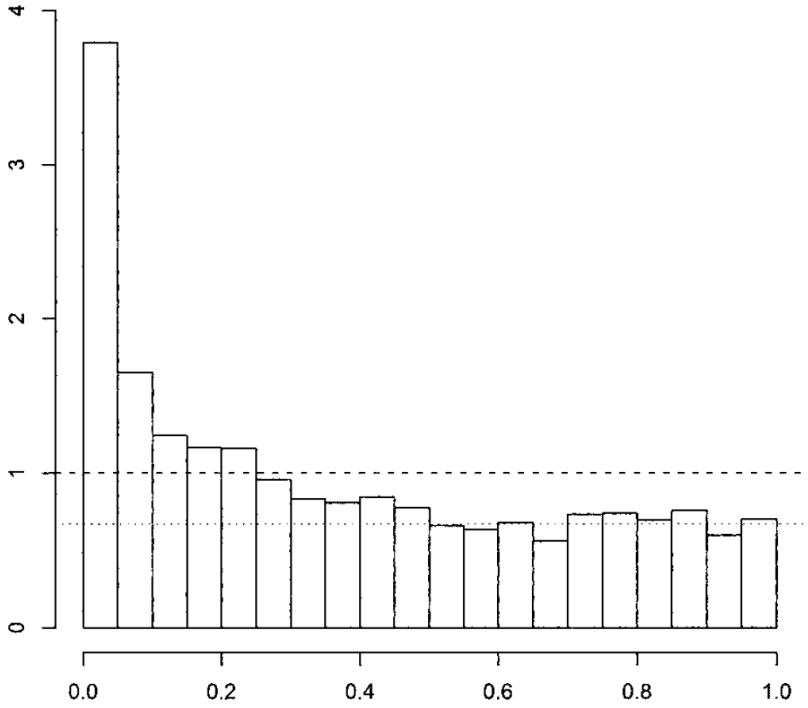
$$\text{E}[F(t)] = m_0 \cdot t \ 。$$

上式中， m_0 是一個未知常數，其值可以下式估計：

$$\widehat{m_0} = \frac{\#\{p_i > \lambda \ ; \ i = 1, \cdots, m\}}{1 - \lambda} \ 。$$

式中 λ 為在 0 與 1 之間適當選取之數值，λ 之選取在 Storey & Tibshirani (2003)中有詳細的說明。上述 m_0 之估計乃基於：虛無假設為真時，檢定所得之 p 值會均勻分布於 0 與 1 之間；對立假設為真時，檢定所得之 p 值會有向 0 靠攏的趨勢。下面以一個實例說明。Hedenfalk et al. (2001) 探討基因表現量在兩種癌細胞間的差異性。圖一呈現研究中

3170 個基因表現量差異檢定所獲得之 p 值的分布。若所有基因表現量均無顯著差異，p 值應均勻分布於 0 與 1 之間，高度應接近圖中上方虛線。但從圖中可看到，0 附近之 p 值的比例較高，代表有些基因表現有顯著差異。另一方面，大於 0.5 之 p 值的分布則相當均勻，高度貼近下方虛線，顯示介於 0.5 與 1 之間的 p 值所對應的基因中絕大部分表現量無顯著差異。以此例子而言，0.5 是合適的 λ 值，而 $\frac{\text{介於 0.5 與 1 之間的 p 值個數}}{1-0.5}$ 則是表現量無顯著差異之基因數的合理估計。



圖一，3170 個基因表現量差異檢定之 p 值的分布（此圖取自 Storey & Tibshirani (2003)）。

根據上述計算，可得到 $\text{FDR}(t)$ 之估計為

$$\widehat{\text{FDR}}(t) = \frac{\widehat{m_0} \cdot t}{\#\{p_i \le t\}} \quad 。$$

以所得 $\widehat{\text{FDR}}(t)$ 為基礎，令第 i 個檢定之 p 值為 p_i ，則此檢定對應之 q 值估計為

$$\widehat{q_i}(p_i) = \min_{t \ge p_i} \widehat{\text{FDR}}(t) \quad 。$$

依據上式，給定多重檢定中的任兩個檢定 i 與 j ，若 $p_i \le p_j$ ，則 $\widehat{q_i} \le \widehat{q_j}$ 。

根據以上所定義的 q 值，在進行多重檢定分析時，當給定一個門檻 q^* ，並稱所有

q 值 $\leq q^*$ 之檢定結果為顯著時，所得結果之 FDR $\leq q^*$ 。

在統計假設檢定分析中，對於所犯型一錯誤之程度的控制越嚴格，所得顯著結果越可靠。處理多重檢定問題時，FWER 與 FDR 是兩種常用來衡量所犯型一錯誤之程度的方法。FWER 度量型一錯誤出現的機率， $\Pr(F \geq 1)$。檢定個數越多，結果中出現型一錯誤的機率就越高。因此，當對大量檢定進行 Bonferroni 校正以控制 FWER 時，個別檢定所要求的顯著水準 α 就會非常嚴格，造成檢定力大幅下降。另一方面，FDR 度量顯著結果中型一錯誤的比例， $\mathrm{E}\left[\frac{F}{S}\right]$ 。由於對立假設為真，檢定所得之 p 值會有向 0 靠攏的趨勢。因此，進行大量檢定時，對立假設為真之比例越高，越適合以 FDR 處理多重檢定問題。實務上，可藉由 Benjamini & Hochberg (1995)所提的檢定程序或是 Storey & Tibshirani (2003)所提出的 q 值，控制 FDR。

<u>參考文獻</u>

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Serier B-Statistical Methodology* **57** (1): 289-300.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP et al. 2001. Gene-expression profiles in hereditary breast cancer. *The New England journal of medicine* **344**(8): 539-548.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16): 9440-9445.

林彥光. 2015. 淺談 Bonferroni 事後校正. TMU 生統期刊第五期.