[研究新領域報導]

多重假設檢定之晚近發展

臺灣大學數學系 陳 宏

一、統計假設檢定之源起及架構

『假設檢定』爲一廣泛使用的科學方法，其核心思想乃將科學視爲由眾多理論所架構而成，而這些理論需與觀察到的數據相吻合。在這個精神之下，那些無法解釋觀察數據的理論，需被重新檢視及修正；至於那些能解釋現有觀察數據的理論，則需被未來的觀察數據持續的檢驗。

如何使用觀察數據來系統化進行『假設檢定』，這套方法在統計科學之領域，稱之爲『統計假設檢定』。此套架構的肇始，要從 1908 年在愛爾蘭都柏林 Guiness 啤酒廠工作的 William Sealy Gosset 所研究的問題說起，因其研究啤酒品質之需，而探討如何來檢驗該酒廠每天的產品是否與過往的產品一致。Gosset 使用高斯分配來描述過往的生產產品，令μ代表該分配的期望值，再依據每天生產產品中的 n 個樣本，計算其樣本平均 $\bar{x}$ 及標準差 s，計算 t 統計量 $t=\sqrt{n}\,(\bar{x}-\mu)/s$。然後使用 Monte Carlo 法求出當每天的產品是與此產品過往的生產一致時，t 統計量應該有的表現行爲，由此來判定每天的產品是否與此產品過往的生產一致，這個架構就爲現今每天在各個角落持續在進行之『統計假設檢定』，如品質管制、新藥物之檢測等。

在上述『統計假設檢定』的架構下，這包含了『虛無假設（null hypothesis）』H_0、『對立假設（alternative hypothesis）』H_1、檢定統計量、及決策四個要素。這可藉由下表來瞭解統計假設檢定之進行：

	虛無假設爲眞	虛無假設不眞
無足夠證據拒絕虛無假設	決策正確	第二型錯誤（Type II error）
拒絕虛無假設	第一型錯誤（Type I error）	決策正確

當對一假設（理論）進行驗證時，就可依據上述之過程驗證數據與理論彼此是否相互牴觸。在醫學診斷之領域，第一型錯誤稱爲僞陽性，而第二型誤差稱爲僞陰性。

但如何合宜的控制第一型錯誤及第二型錯誤，要等到 1930 年代的 Neyman-Pearson lemma 方才說清楚，當時急欲驗證的科學問題是『多施肥是確實會增加穀物產量』這個應該是對的理論。但因不可能同時降低犯第一型錯誤及第二型錯誤的機率，而也已有相當的旁證說明『多施肥是會增加穀物產量』這個應該是對的理論，所以奈曼（J. Neyman）及皮爾生（E.S. Pearson）提出將誤判後果較嚴重者定爲虛無假設，所以先控制第一型錯誤的機率不要超過某一事先設定的值，然後使第二型錯誤的機率愈小愈好。進而提出 Neyman-Pearson lemma，提供了方法如何在上述的架構下找出最好的檢定方法。

二、統計假設檢定之誤用及新的挑戰

在極多的科學研究中，需同時進行多重假設檢定，這在文獻中早有探討，但其所考慮的問題是僅當同時檢視的假設個數只有一、二十個[1]。但是由於不理解多重假設檢定問題之本質而產生繆誤之科學結果，則可以 1984 年 12 月份的 *Lancet* 期刊，Graham Martin 發表的一封公開信 "Munchausen Statistical Grid, that makes all trials look significant"，可看出其嚴重性。在信中指責統計學家及在醫藥學界使用統計工具者，蓄意的增加統計檢定的數目，以利於找到『發現』得以發表研究成果。因此在目前的科學期刊，會要求進行第一型式錯誤發生機率之修正。

如果對於多重假設檢定，採取控制會發生第一型式錯誤（文獻上稱爲 familywise error rate）的機率，最常用的修正是"The Bonferroni 修正"；簡單的說就是當進行 20 個統計假設檢定時，如欲將總體第一型式錯誤發生的機率控制在 5%之

下，此時對每一個假設檢定，第一型式錯誤發生的機率需控制在=0.25% (5/20)之下。

近十年來多重假設檢定所需考慮的檢定數目產生極大的變化，如使用微陣列（microarray）基因晶片，藉由生物體的檢體來發現，在不同的實驗情況之下哪些基因會有不同的表現行爲。通常在每一微陣列上有數千個基因[2]，這時要考慮的檢定數目將達到數千，使用「Bonferroni 修正」並不可行。

又如使用正子斷層攝影(PET)於腦部研究或診斷時，此時的觀察數據爲 radiotracer 影像的分佈，所關心的問題爲是否因病理、功能刺激之不同而造成 radiotracer 影像分佈的改變，腦部不同區域的活動力是否異於正常，如果有改變，改變發生在何處？此時可將腦部區分成多個區域，再使用統計檢定來檢測那些區域影像分佈產生改變。

在 data mining 方面，企業界希望能將其現有之資料，經由轉化成爲知識，這包括了趨勢、特徵及相關性，如以探討相關性爲重點的 basket analysis，是想去發掘哪些事物總是同時發生。舉例來說，買 A 商品的通常同時購買 C 商品。美國一個應用 data mining 做購物籃分析的著名實例是零售連鎖商 Walmart 發現的「星期四、尿布和啤酒」。也就是由購物籃分析發現在禮拜四晚上，消費者通常會同時購買尿布和啤酒。這樣的發現達成交叉銷售的方法。如以一千種產品爲例，當檢測那兩種產品的相關係數不爲 0，這需執行約五十萬個統計檢定，這在統計學理上「可行嗎」？

就只對這兩個應用來看，很顯然的與奈曼（J. Neyman）及皮爾生（E.S. Pearson）所欲處理的問題大不相同，在這些應用是否仍應該先控制第一型錯誤的機率不要超過某一事先設定的值，然後使第二型錯誤的機率愈小愈好？如果 Neyman-Pearson lemma 之架構並不適宜，此時提供了方法如何在上述的架構下找出最好的檢定方法。

三、研究目的及偽發現率

當進行假設檢定的個數爲 G，其中未知的眞實虛無假設個數爲 G_0，未知的錯誤虛無假設個數爲 G_1，當依據特定方法使用數據進行假設檢定時，被拒絕之虛無假設個數 R 爲已知，如下圖。

	未拒絕之虛無假設個數	拒絕之虛無假設個數	
眞實的虛無假設個數	U	第一型錯誤 V	G_0
錯誤的虛無假設個數	第二型錯誤 T	S	G_1
	$G-R$	R	G

如不考量假設檢定的個數，仍依一個假設檢定的 5%第一型錯誤機率，此時 $E(V)=0.05\ G_0$，當 G_0/G 接近 1 時，在所有被拒絕之虛無假設個數 R 中，大多爲不該拒絕的虛無假設。由此不難看出此種方式之不恰當性，此種處理所控制的是所謂的 per comparison error rate（PCER）。

當使用 PCER 來處理微陣列基因晶片之多重假設檢定問題不恰當時，又該採取那種錯誤率？因多數的生物學家相信，「不同表現的基因個數不會太多」，同時被認定爲不同表現基因的「後續確認研究」極爲耗時及昂貴，爲了避免因機率的緣故而產生過多的『科學發現』，較合宜的準則是 V/R 小，V/R 是被錯誤拒絕的虛無假設之比例。這就是 Benjamini and Hochberg [3]建議使用僞發現率(false discovery rates, FDR)，FDR= $E(V/R)$。且其證明 Simes [4]所提之調整第一型式錯誤機率的辦法，確能控制 FDR。

由此開始了一系列的文章，這包含了新的錯誤率（見 Storey [5,6]; Genoverse and Wasserman [7] ; Dudoit, Shaffer, and Boldrick [2] ）、optimality、empirical Bayes approach (Efron [8])、Dependency (Benjamini and Yekutieli [9] ; Meinshausen and Rice [10])。

四、探討

由上述之討論，不難看出多重假設檢定之研究方興未埃，新的應用及問題與日俱增，其重要性顯而易見，但直至目前爲止，尚缺乏如 Neyman-Pearson lemma 型式的結果，告訴使用者該如何系統化的來處理多重假設檢定。當然由於問題之複雜性，也極可能不會有一套完整之方法論。

不過與傳統之單一假設檢定相較，這其中仍有相當不一樣的訊息可被使用。如假設檢定統計量彼此獨立，則這些假設檢定統計量可用混合分配來描述，這與 empirical Bayes、無母數密度函數產生緊密之結合，已有相當的進展。但在實際問題中，假設檢定統計量常常並非獨立，如是又該如何處理？

參考文獻

[1] R. Miller, *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag (1981).

[2] S. Dudoit, J. P. Shaffer and J. C. Boldrick, *Statistical Science*, **18**, 71 (2003).

[3] Y. Benjamini and Y. Hochberg, *Journal of the Royal Statist. Soc., Ser. B*, **57**, 289 (1995).

[4] R. J. Simes, *Biometrika*, **73**, 751 (1986).

[5] J. D. Storey, *Journal of the Royal Statist. Soc., Ser. B*, **64**, 479 (2002).

[6] J. D. Storey, *Annals of Statistics*, **31**, 2013 (2003).

[7] C. Genoverse and L. Wasserman, *Annals of Statistics,* **32**, 1035 (2004).

[8] B. Efron, *Journal of the Amer. Statist. Assoc.*, **99**, 96 (2004).

[9] Y. Benjamini and D. Yekutieli, *Annals of Statistics*, **29**, 1165 (2001).

[10] N. Meinshausen and J. Rice, Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. To appear *Annals of Statistics* (2004).