教育部（民83），**師資培育法**。

教育部（民84a），**高級中等以下學校及幼稚園教師資格檢定及教育實習辦法**。

教育部（民84b），**大學校院教育學程師資及設立標準**。

臺灣師範大學實習輔導處、彰化師範大學實習輔導處、高雄師範大學實習輔導處合編（民86），**教育實習手冊**。

Dawson, A. J.(1995). Reframing the clinical professor role: The faculty associate at Simon Fraser University. In M. F. Wideen, & P. P. Grimmett(eds.) ***Changing times in teacher education***. pp. 174-188. London: The Falmer Press.

Dittmer, A., & Fischetti, J.(1995). Foxfire and teacher preparation: practising what we teach. In M. F. Wideen, & P. P.Grimmett(eds.) ***Changing times in teacher education***. pp. 163-173. London: The Falmer Press.

Fullan, M., & Sheehan, N.(1995). Teacher education in Canada: A case study of British Columbia and Ontario. In M. F. Wideen, & P. P. Grimmett(eds.). ***Changing times in teacher education***. pp. 89-102. London: The Falmer Press.

Hauge, T. E.(1995). Teacher education in Norway: images of a new situation. In M. F. Wideen, & P. P.Grimmett(eds.). ***Changing times in teacher education***. pp. 67-78. London: The Falmer Press.

Hopkins, S.(1995). Using the Past; guiding the future. In G. A. Slick (ed.) ***Emerging trends in teacher preparation***. pp. 1-9. Thousand Oaks, CA: Corwin Press, Inc.

Kellett, C.(1994). Towards more school-based initial teacher education. In B. Field, & T. Field(eds.) ***Teachers as mentors: A practical guide***. pp. 96-118. London: The Falmer Press.

Selinger, M., & Pimm, D. (1995). The commodification of teaching: teacher education in England. In M. F. Wideen, & P. P. Grimmett(eds.). ***Changing times in teacher education*** . pp. 47-66. London: The Falmer Press.

Slick, G. A.,(1995). Connecting purposes-administrators' views of field experiences. In G. A. Slick(ed.). ***Making the difference for teachers***. pp. 103-129. Thousand Oaks, CA: Corwin Press, Inc.

Slick, G. A., & Burrett, K.(1995). Bits and Pieces-everything else you wanted to know about field experiences of the future. In G. A. Slick (ed.) ***Emerging trends in teacher preparation***. pp. 108-127. Thousand Oaks, CA: Corwin Press, Inc.

Tuinman, J.(1995). Rescuing teacher education: a view from the hut with the bananas. In M. F. Wideen, & P. P.Grimmett(eds.) ***Changing times in teacher education***. pp. 105-116. London: The Falmer Press.

Tyson, H.(1994). ***Who will teach the children***? San Francisco, CA: Jossey-Bass Inc.

Weiser, S.(1995). Rewarding the practicing professional. In G. A. Slick(ed.). ***Making the difference for teachers***. pp. 93-102. Thousand Oaks, CA: Corwin Press, Inc.

Wideen, M. F.(1995). Teacher education at the crossroads. In M. F. Wideen, & P. P. Grimmett(eds.) ***Changing times in teacher education***. pp. 1-16. London: The Falmer Press.

黃淑苓，國立中興大學教育學程中心副教授

教育研究資訊，1998.5，6(3)，頁 103-115

統計顯著性考驗的再省思

李茂能

本文旨在重新檢討統計顯著性考驗之意義、功能、特質與其在社會科學研究應用上之基本限制。統計顯著性考驗涉及對立假設與虛無假設兩種考驗；其中又以虛無假設爲統計顯著性考驗之直接目標。文中論及虛無假設之迷思與爭議、應用時機與注意事項。文末歸納出顯著性考驗之八點改進措施，並建議教育與心理研究者採用質化的統計顯著性考驗步驟，將最低效果值或差異值與統計考驗力亦納入虛無假設考驗之目標，期使顯著性考驗同時考慮到實質上之意義性與應用性。

關鍵詞：統計顯著性考驗、對立假設、虛無假設

Keywords: Statistical significance testing、alternative hypothesis、null hypothesis

壹、序言

國內近幾十年來研究風氣日愈鼎盛，各種教育與心理研究報告或論文廣見於各類之期刊與學報中。雖然有些論文祇提出待答問題或研究目的，但其中不少量化的論文都沿用 Fisher 的「假設—驗證」之顯著性考驗做法，在論文當中都會先提出研究（對立）假設與虛無假設，再蒐集資料與予驗證；有些論文甚至祇提出虛無假設。到底哪一種寫作方式比較妥當，到目前爲止並未形成共識。不過不少的研究者似乎都視量化研究爲主流，而量化研究又以「顯著性」之有無當作研究價值之指標。事實上，國外有不少學者質疑與批判 Fisher 對於虛無假設之運用的作法 (Lykken, 1968; Morrison & Henkel, 1970; Meehl, 1978; Cohen, 1990; Serlin, 1987; Sohn, 1993; Falk & Greenbaum, 1995; Kirk, 1996)。Lakatos(1978) 更嚴厲指出此類社會科學研究結果只不過是增加一些垃圾知識而已 (nothing but an increase in pseudo-intellectual garbage)。Murphy(1990) 與 Kirk(1996) 認爲虛無假設既然被認爲不可能永遠爲眞，顯著性之考驗只是曠廢時日或多餘之舉措 (a trivial exercise) 而已。因而，Harcum(1990b)、Shaver(1992) 與 Kirk(1996) 咸認研究者應花更多時間與精力於研究結果的再複製性或一致性與應用性上；因此他們極力呼籲各教育與心理學雜誌的主編者嚴格管制統計顯著性考驗的品質，也規勸研究者儘量少用或謹愼使用統計顯著性之考驗。Orey, Garrison, & Burton(1989) 亦從哲學觀點評析各家對於虛無假設之論點，並指出其存廢至今仍爭議不休。國內教育學者賈馥茗（民73）曾呼籲教育研究者不要一味「玩弄工具」；黃政傑（民76）亦認爲教育研究應擺脫量化的支

配，不必「事事量化，處處量化」，不要弄到「科學必量化」、「不量化非科學」的地步；陳伯璋（民76）於分析我國近四十年來教育研究的品質後，認爲我們的教育研究「仍然脫離不了實證性、實用性、移植性、與加工性的性格」。儘管上述這些學者對於量化研究之嚴厲批評，統計顯著性之考驗仍在中外教育與心理學界大行其道；期刊上、學報上與學位論文上依舊把它做爲解釋量化資料之科學工具(Carver, 1978; Greenwald, et al., 1996)，仍舊把它視爲研究品質之保證。我們是否已墜入「科學的八股」的迷魂陣呢？爲了教育與心理研究之效率與品質，研究的生產者與消費者重新深入思考與檢討「統計顯著性考驗」的意義、特質、哲學基礎、統計理由、效能與缺失、應用時機與方法等問題，似乎已是刻不容緩。

貳、統計顯著性考驗之意義

人類下決策時可能訴諸感性也可能訴諸理性，而訴諸理性的途徑不外乎歸納與演繹。統計顯著性考驗就涉及歸納與演繹兩種推理歷程(Kirk, 1982)。研究假設或統計假設的提出涉及演繹的歷程：從過去相關之理論或文獻、與經驗推衍變項間的預期關係或差異，而支持與推翻研究假設就涉及歸納之歷程；從統計考驗後的結果歸納出研究假設的眞僞。不過人們因時間、資源等種種因素所限，常需要進行推論性的工作：從有限的代表性樣本的結果推論到母群體的特質。因此，統計顯著性假設考驗是人類追求理性推論之經濟有效途徑，其主要功能在於控制犯第一類型錯誤機率、導引推論方向與指導資料蒐集重點。統計顯著性考驗包含兩種假設考驗：對立假設(H1)與虛無假設(Ho)考驗。H1即是研究假設，爲研究者根據待答問題所擬的操作性定義。而Ho（發音爲Hsub-zero、H-on、或H-nought）考驗乃是一種間接求證的考驗歷程，研究者首先提出與對立假設相反的假設，並假定虛無假設爲眞；然後，研究者再蒐集資料反斥此種假設之繆誤與不合理。研究者透過此種間接求證的考驗歷程，可以有效控制將無效之研究結果誤判爲有效之機率（犯第一類型錯誤）；這也是爲什麼研究者常將誤判後果較嚴重的假設當作虛無假設以控制其風險。判斷此種假設之繆誤與不合理的準據有二：一爲虛無假設之抽樣分配(null hypothesis distribution)，二爲研究者事先所設定之α水準。α爲研究者犯第一類型錯誤時所能容忍之機率。α的設定多少帶有主觀之認定，不過我們都希望犯第一類型錯誤時所造成之代價愈少愈好。虛無假設之抽樣分配乃是當虛無假設爲眞時，連續抽樣試驗時所形成之次數分配。因此當虛無假設之抽樣分配之性質不明或不存在時，統計顯著性之考驗是無法進行的。

參、顯著性考驗之基本特質

研究者於敘寫對立假設與虛無假設時，應注意與遵行下列特性：

一、互斥性：當其一爲眞，另外一個必爲假。例如，Ho: $\mu \leq 100$ 與H1: $\mu > 100$ 無法同時爲眞。

二、互補性：對立假設與虛無假設應涵蓋考驗母數之所有值。例如，Ho: $\mu \leq 100$ 與H1: $\mu > 100$ 就合乎此一要求。

三、不確定性：除非考驗之資料來自於母群，否則不管拒絕他們或接納他們，我們永遠無法證明他們爲眞；而且，不管我們接納或拒絕他們，我們都可能會犯錯。

四、虛無假設考驗本質上爲一簡單假設考驗，祇涉及單一精確值(an exact value)的考驗，因爲只有簡單假設考驗我們才能直接加以考驗。譬如，複合式虛無假設Ho: $\mu \leq 100$ 所陳述的假設雖涵蓋一段範圍而非單一精確值，但該假設眞正考驗的對象僅是$\mu = 100$。假如$\mu = 100$的虛無假設可以拒絕，那麼虛無假設$\mu < 100$就自然可加以拒絕。

五、虛無假設考驗的對象爲母群而非樣本，因此撰寫對立假設與虛無假設時應使用希臘字母（如μ、σ、ρ、π），此爲慣例；而樣本的統計量數則以英文字母表示之（如X、SD、r、p）。

六、可考驗性(testable)：無法透過感官經驗或運用實徵或客觀測量方法驗證其眞僞的形而上之假設是不允許的；而且所提出之假設最好能於短期內（如數月內）加以檢驗。

肆、虛無假設考驗之功能與必要性

從哲學的觀點來看，量化研究乃是實證主義與邏輯實證論之具體實踐，他們追求的目標都是客觀、系統與實徵。統計顯著性考驗又是量化研究之主要工具，而統計顯著性考驗中之虛無假設考驗的反向思考概念卻與英國哲學家Karl Popper(1959)的否證論(falsificationism)具有較密切之關係。Popper非常重視邏輯的問題，他認爲科學上一個理論或假設的眞，往往涉及普遍性原則；然而普遍性原則所需之例證是無限多的。人類所能找到的實例經常是有限的，要驗證一項假設永遠爲眞異常困難（詹志禹，民82）。因而科學家乃採取反向思考的模式：先提出一個虛無假設，再不斷尋找證據去否證它。所謂「好的理論或假設」乃是經得起無數次的考驗都無法否證它，或找到一個反例。Popper認爲這樣的思考模式比較具有邏輯效力，因爲不管我們蒐集到多少證據都無法證明一個理論或假設永遠爲眞，我們唯一能做的就

是提出反證。而統計學者 Fisher(1966) 主張我們無法證明一件事情永遠爲眞，但可以證明它爲假(We can never prove something to be true, but we can prove something to be false)。譬如，一位研究者蒐集資料調查200個人，發現「每一個人都祇有一個頭」，並無法證實每一個人都祇有一個頭；但如果發現一個人擁有兩個頭就能反斥「每一個人都祇有一個頭的假設」。無疑的Fisher的統計考驗觀點與Popper的否證論前後呼應。爲了使讀者明白爲何使用虛無假設較客觀，請看以下法官辦案歷程之比喻。法官辦案爲了保持公正立場：不隨便羅織罪名，似應先假定受審之嫌犯無罪(Ho)，再去不斷蒐集證據證明他有罪；而不事先假定受審之嫌犯有罪(H1)，再去蒐集資料證明他無罪，以免產生自由心證的冤獄。

從統計的基礎來看，Fisher與Neyman、Pearson於1930年代前後完成了假設考驗所須之必備要件：點估計、一致性、充份性、隨機性、最大可能性估計法、第一類型錯誤、與第二類型錯誤之概念(Kirk, 1996)。根據Fisher的做法，虛無假設在統計考驗上的主要功能爲設立一個特定條件，以便獲得某一個統計量數的抽樣分配(The function of the null hypothesis in a statistical test is to establish a condition under which the sampling distribution of a statistic may be obtained)，因爲當虛無假設之抽樣分配的特質不明時，根本無法進行統計顯著性之考驗(Kiess, 1996)；而且虛無假設中所陳述的特定值乃是統計考驗之起始點。此外，虛無假設亦可以有效控制第一類型錯誤（無效實驗誤判爲有效）、第二類型錯誤（有效實驗誤判爲無效）、第三類型錯誤（誤判實驗效果方向）等之機率(Huck & Cormier, 1996)。直接考驗研究假設或對立假設是無法有效加以控制這三種類型錯誤。

伍、虛無假設考驗的迷思

根據前述統計顯著性之意義、特質與功能，知其間接推理之過程很容易導致誤解與誤用。Shaver(1992)與Kirk(1996)分析過去對於虛無假設之評論，歸納出下列顯著性考驗之能與不能。第一、統計顯著性之考驗並無法提供研究者虛無假設爲眞或爲假時之機率：P(Ho|D)，它祇能告訴研究者假如虛無假設爲眞時，重複試驗獲得某一研究結果之可能性：P(D|Ho)。Carver(1978)提供一有趣例子：假設P(Ho|D)代表一位已死犯人被絞刑的機率(D表犯人死亡，Ho犯人被絞刑)，而P(D|Ho)代表一位罪犯被絞刑死亡的機率。前者之機率甚低，如0.01，而後者可能高達0.97，這兩種機率是不能互換的。此種演繹推理之錯誤被Falk與Greenbaum(1995)譏爲：機率證據之錯覺(Illusion of probabilistic proof by contradiction)。第二、統計顯著性之考驗並無法提供研究者對立假設爲眞或假之機率，因爲抽樣分配是根據虛無假設而來。我們並無證據確知當對立假設爲眞時，複製該一研究結果的機率：P(D|H1)。因此，統計之顯著性考驗祇能告訴研究者當虛無假設爲眞時，重複隨機抽樣與分派

進行試驗時，隨機性出現某一研究結果之機率，我們並無法確知該一研究結果隨機出現之機率。研究者切勿把1-P(D|Ho)解釋爲P(D|H1)，否則會導致以下之迷思：P(D|Ho)值愈小，顯著性愈高（例如，P<.01比P<.05更顯著(more significant)）。第三、統計顯著性之考驗並無法顯示研究結果之重要性。例如，Ho：$\rho_{xy}=0$遭到拒絕時祇表示在母群中X與Y變項間之關係達統計上之顯著水準，並未說明在母群中X與Y變項間之關係大小。而且，統計顯著性之考驗與樣本大小具有密切關係，只要樣本夠大，且研究工具之信度夠好，不管其實質理論之眞相如何，幾乎所有研究結果都會達統計上之顯著水準(Meehl, 1978)。同樣的，當樣本很小時幾乎很少研究結果會達統計上之顯著水準。因此，只知道研究結果之統計顯著性並無法確知研究成果之大小與重要性。Dar(1978)也指出過分依賴虛無假設考驗，容易衍生下列誤解、誤用之現象：

一、社會科學研究過於強調統計顯著性

研究者只關注是否達到統計假設之顯著水準（如.05），而不注意到邏輯的連貫性、理論的預測力、闡釋力等方面的問題。事實上，理論的建立(theory building)重於顯著性之考驗；如有優良的理論，統計的顯著性之考驗甚至可以不要。

二、統計顯著性常被自動視爲與實質意義性相等

理論、研究假設、與現實(reality)間常未具有直接關係。很不幸的，統計顯著性常被有意或無意地解釋爲實質上具有意義性、重要性。因此，統計圖表上星號(*)變成了社會科學研究的目標；而忽略了所考驗之理論是否有意義、是否反映現實眞相。

三、依賴電腦統計套裝軟體以獲得精確性與嚴謹性的科學特質

功能強大而複雜的電腦統計套裝軟體（如SAS、SPSS、BMDP）的出現與普及，使得研究者常願多跑幾次電腦以分析出最令自己滿意的結果，而不願再去檢視原始資料，思索與發展更有預測力的理論。上述扭曲顯著性考驗原意與以工具爲內容之偏頗與弊端，使得量化的社會科學研究常被指控爲「玩弄工具」、「玩弄統計」。

陸、虛無假設之改善辦法與途徑

由於我們事先既已知道虛無假設爲假，拒絕虛無假設所能提供的有用訊息事實上並不多。拒絕虛無假設祇能讓研究者確定差異之方向性，並不能告訴研究者差異或相關之大小；且其所帶來之誤解亦不少。因此虛無假設考驗的創始人Fisher早在1925年即已建議使用其他相關性指標（如相關比，η）補足統計考驗之不足。Kirk

(1996) 整理出過去學者所提出的 40 種指標，並將這些代表效果大小 (effect magnitude) 的指標分爲三類：(1)相對性關係強度指標：如 Fisher 的 η^2，Hays 的 ω^2；(2)效果值大小指標（例如，標準化的平均數差異值）：如 Cohen 氏之 d 指標與 f 指標；(3)其它很少被使用到的指標：如 Cohen 氏之 U_1，U_2，U_3 指標。他呼籲研究者在報告考驗之統計量數（如 t,F）時，也應同時報告自變項與依變項間之關係強度或效果大小。研究者可以根據 d 指標與 ω^2 指標的大小判斷與解釋處理效果之大小 (Cohen, 1988, 1992)。D 值等於 0.5 時、ω^2 值等於 0.06 時，屬於中等效果值 (medium effect)，研究者用肉眼即可輕易看出；D 值等於 0.2 時、ω^2 值等於 0.01 時，屬於小效果值 (small effect) 的最低限度；D 值等於或大於 0.8 時、ω^2 值等於 0.15 時，屬於大效果值 (large effect)。這個解釋準則適用於其它標準化的效果值指標。事實上，研究者只要根據考驗之統計量數（如 t, F）、自由度、樣本大小都可計算出效果值大小。因此，這些效果值大小最好與統計指標一併在報告中呈現，這對研究結果之詮釋具有佐證與澄清之效果。

總之，當虛無假設考驗結果瀕臨顯著水準之臨界值、或實質意義有疑慮時，可以用下列途徑澄清：

一、採用信賴區間考驗虛無假設

二、複製研究結果

三、計算效果值大小

四、計算變項間關聯指標

五、同時考慮應用上之損益 (costs and benefits)

值得一提的是『再驗證性』在顯著性考驗中佔有輔助澄清之功能。尤其在教育與心理研究中隨機抽樣與隨機分派通常無法達成，研究者再找另外一個樣本，交互驗證研究結果至爲重要，否則推論的結果易招致以偏概全。請細思下面三個各說各話的經驗：

圖一　張三、李四、王五之住所分布圖

據聞張三、李四、王五分別住在地球的北迴歸線、赤道、與南迴歸線上。他們三個人分別在居住處所附近進行調查南北極之氣溫。住在北迴歸線的張三一再發現他愈往北走天氣愈冷，愈往南走天氣愈熱，因此下結論：南極最熱、北極最冷。住

在赤道的李四不斷的發現他愈往北走或往南走天氣均愈冷，因此下結論：南、北極最冷。住在南迴歸線的王五也不斷發現他愈往北走天氣愈熱，愈往南走天氣愈冷，因此下結論：南極最冷、北極最熱。

柒、顯著性考驗的應用時機與型態

研究者於研究計畫階段時可能會面臨要不要進行顯著性之考驗或虛無假設考驗的問題。Frick(1996) 認爲虛無假設考驗最適合於次序性之考驗（如條件之次序、效果之次序、相關之方向），而較不適合於效果值大小 (size of effect) 之預測。一般來說，當您想進行通則性之研究、理論之驗證時可提虛無假設，而普查時、研究者主要目的不在推論母群之特質而旨在描述樣本特質時、研究假設之似眞性未明時，尤其在敘述性或探索性研究上，就不必進行統計顯著性之考驗，研究者祇提待答問題即可。

檢視目前教育與心理的博碩士論文或期刊上之論文，文中呈現待答問題、研究假設（對立假設）、與虛無假設之敘寫格式大概可以分爲以下幾種型態：

一、祇提待答問題。

二、祇提出研究假設。

三、祇提出虛無假設。

四、同時提出待答問題、研究假設。

五、同時提出待答問題、研究假設與虛無假設。

質化的研究採第一種型態居多，其餘爲量化研究所採行之型態。筆者認爲在正式之論文上量化研究最好使用第四與第五種型態。第二與第三種型態讓人感覺有點唐突，尤其論文中祇提出虛無假設常令許多讀者無法理解。第四種寫作型態雖然未提虛無假設，但在資料分析時仍採取虛無假設考驗之步驟，進行資料分析與研究結果之推論。

捌、虛無假設考驗應注意事項

一、無法拒絕虛無假設時

邏輯上，無法推翻一個不可能爲眞的虛無假設是說不通的 (Cook, Gruder, Hennigan, & Flay, 1979; Harcum, 1990a; Murphy, 1990)。除非您的研究能符合以下幾項嚴格標準，接受虛無假設才能讓人信服：*1.* 產生實驗效果的各種條件均已明確而具體

地加以界定；2.與實驗效果相關的變項與無關的干擾變項影響力量都已控制或排除；3.研究設計佳，統計考驗力強；4.實驗操弄與工具均精確、正確有效；5.α夠大，β很小。

而且，未達統計上之顯著水準而接納虛無假設並不能被解釋爲：研究者所提出的對立假設未獲所蒐集之資料支持(Kirk, 1996)。假如所蒐集之資料顯示出支持對立假設的趨勢，接納虛無假設之考驗結果必須格外小心。例如，教育部爲了瞭解常態編班與能力分班是否能改善學校學生之EQ，隨機選擇了12個國中，隨機分派到實驗組（常態編班）與控制組（能力分班）接受爲期三年之實驗。假如三年後研究結果經過統計考驗發現：$\overline{X}$實驗組=103.0，$\overline{X}$控制組=90.0；α爲.01，實得t值爲1.61，p=.14。根據此項結果虛無假設：實驗組與控制組的學生其EQ沒有差異，我們無法加以拒絕。但此項結果並不意謂著：實驗組與控制組的學生其EQ沒有差異，其實此項結果只是告訴我們：兩組資料所顯現之差異係因機遇或抽樣誤差所造成。也許從實驗組與控制組學生的EQ平均數看來，兩組之間之差距仍達13之多。再進一步計算Cohen氏之d指標，得到d=.86；這個效果是蠻大的。更進一步查看原始資料，研究者可能會發現：大部分實驗組學生之EQ一致性地高於控制組的學生。這是不是仍暗示著：常態編班比能力分班好。此時研究者應繼續進行複製研究或計算其信賴區間方可解開這個謎：實驗組與控制組學生的EQ到底有沒有差異？此外，效果值大小與其對理論考驗之重要性不一定成正比，這端視研究之精密性質與目的而定。有時「小」的效果值亦可能具有「重大」意義的。因此，研究者於接納虛無假設時，發現「小」的效果值的理論亦可能具有「重大」意義的。研究者也可透過統計考驗力(power)之分析（可查表或使用相關公式計算之），瞭解研究結果是否因樣本過小而造成無法拒絕虛無假設(Cohen, 1988)。統計考驗力(1-β)最好能介於.80到.95之間(Keppel, 1991)，以控制犯第二類型錯誤之機率(β)。計算前述改善學生EQ之實驗的統計考驗力約爲0.40，足見其離最低標準0.80尚有一段距離，犯第二類型錯誤之機率過高。由前述之D值與1-β值綜合判斷，本實驗效果似應具有應用之價值，值得研究者再加以驗證。

此外，有些學認爲當我們所求得的P(D|Ho)值大於α時，接納(accept)虛無假設P(Ho|D)的解釋是不正確的，正確的說法應該是無法拒絕(fail to reject)虛無假設，因爲接納強烈暗示著該虛無假設爲眞(Huck & Cormier, 1996)。事實上，尚有其他之虛無假設亦可能爲眞，我們不能武斷地說當中之一的虛無假設爲眞。此種裁決虛無假設考驗結果的解釋用語，值得研究者參考應用。例如，研究者最好說：「研究結果暗示著(suggest)」而不是「研究結果顯示出(indicate)」。

二、拒絕虛無假設時

事實上，初次拒絕虛無假設的結論是相當脆弱的，且所獲之有用訊息並不多，我們尚需要其它證據加以驗證與佐證；它亦須再經千錘百鍊之後才能建立較可靠而有用之理論（芮涵芝，民85）。研究者於呈現研究結果時應將下列有關資訊一併報告出來：描述統計量數、推論統計量數、效果值或關聯性指標、樣本大小、顯著水準。更令人困擾的是：一個否證或反例就可推翻一個科學理論嗎？雖然邏輯推論是如此，但人心並非如此（詹志禹，民82）。數學的定理可以一成不變，但自然科學所建立的理論就可能不是永恆性的眞理，社會科學所發現的眞相更是如此。譬如：

牛頓提出的理論，確實比克卜勒的理論好得多，可是事實上它仍然有很多問題無法解決（例如：運動的速度大到光速時則非靠量子論來解釋不可），一些所謂的不正常現象或是反例，其實自始至終就存在的，但是這都無損於後來牛頓理論強勢的發展……爲什麼我們不會因爲那一兩個反例，在當時就把牛頓理論推翻掉（詹志禹，民82）？

又如：

您可能想窮一生之精力與時間去求證：所有天鵝都是白的。經年累月之證據顯示在美國所發現的1000隻天鵝都是白的。直到有一天您去澳洲訪問親友時，赫然發現了一隻黑色天鵝。您對於最初的所有天鵝都是白的假設要如何下結論呢？該假設被證明是錯誤，或該假設之論據被消弱而已(Orey, Garrison, & Burton, 1989)？

事實上，我們通常都是採取「夠好原則」(good enough principle) (Meehl, 1978)或經驗法則（芮涵芝，民85），去判斷一個人文或社會科學理論或假設之眞象。只要正向的證據遙遙超過於負向的證據，我們即會接受該假設或理論。同樣的，一個或少數幾個實證或例證就可推翻一個繆論（虛無假設）嗎？事實上，社會科學的因果預測力與穩定性均不如自然科學，所研究之現象、情境或行爲非常複雜且控制不易，硬要仿照自然科學之做法，尋找普遍性之原理原則或不變之定理，恐是難上加難。另外，當研究者使用大樣本而又拒絕虛無假設時，務必檢視效果值之大小是否具有實質上之意義。因爲只要樣本夠大，統計考驗力亦隨之昇高，任何一個統計量數都會達到統計上之顯著水準。

三、顯著水準α之設立與解釋

在教育與心理學界，通常顯著水準α（先驗機率）都定在.05。這個先驗機率必須在蒐集資料前設定，且不能隨著P值之大小而更動；例如α原本設定爲.01，但事後發現P值爲.03而將α更改爲.05以便使得研究假設達到.05之顯著水準。或

者，當您發現P值爲.009時，也不能將α從原本的.05更改爲.01以便使得您的研究假設達到更顯著的水準。這種做法不僅違反研究倫理，而且顯著與更顯著之區分在邏輯上亦是沒有意義的(Shine, 1980)，因爲1-P(D|Ho) ≠ P(D|H1)。依照慣例，P值小於.05則推翻虛無假設，大於.05則接受虛無假設。因此.05變成下決策之臨界點。其實，此種二分式決策方式頗有爭議的：過於武斷與主觀。人類對於一項研究結果之信心應是隨著P值（當虛無假設爲眞時之後天機率）之變小而逐漸增加，懸崖式的全有、全無的信心劃分法是不合理的(Harcum, 1990a; Kirk, 1996)。因此，當您的P值接近.05或其它之臨界值時，您對於虛無假設之拒絕與接納就必須格外謹愼。研究者必須尋求其它佐證（如效果值大小、樣本大小、複製研究結果）一併考慮，才不致誤判。至於研究者呈現統計考驗結果的方式，可採取Stallings(1985)的做法：P值與α值併陳，例如，P=0.31<.05；即先列出P值再將顯著水準α放在不等式後面，此種陳述方法簡潔明瞭、易懂易行，值得推廣。研究者於解釋時，亦應瞭解它爲條件機率，P(D|Ho)意指著所有的假設考驗皆爲無效的實驗而被誤判爲有效的機率，不應將他解釋爲(1)虛無假設爲眞（眞正無效之實驗）之機率，(2)全部統計考驗（含有效與無效之實驗）所犯第一類錯誤之機率。事實上，我們並無法估計虛無假設爲眞之機率，α爲犯錯之上限（這是最糟糕的狀況：所有的假設考驗皆爲無效的實驗）。所有我們能知道α=0.05涵義僅是：不管進行了多少無效的實驗，大約有百分之五的無效的實驗被宣判達統計上顯著水準(Cohen, 1996)。

玖、結論

根據前述對於顯著性考驗之剖析與批判，爲了質化、意義化量的研究，研究者似乎應有以下幾點體認：

一、描述統計應與推論統計併陳。

二、P值與α值應併陳報告。

三、使用統計方法不能保證研究結果之嚴謹性、精確性。

四、理論建構與發展重於統計考驗。

五、研究結果之複製性重於統計方法的深奧性。

六、實質性意義(practical significance)重於統計顯著性(statistical significance)。

七、效果大小重於統計假設考驗之顯著水準。

八、重視研究結果的整體性關係之脈絡與解釋。

爲落實上述各點呼籲，研究者似可根據Huck & Cormier(1996)之建議，採行下列步驟，進行統計顯著性之考驗：

一、根據待答問題擬出研究（對立）假設與擬出虛無假設。

二、決定適當之顯著水準。

三、界定效果值。

四、界定統計考驗力。

五、決定樣本大小。

六、蒐集與分析資料。

七、選擇適當的統計考驗值。

八、比較統計考驗值與臨界值，裁決是否拒絕虛無假設。

此外，研究者亦可兼採Serlin(1987)，Murphy(1990)，與Frick(1996)的建議措施：先設定一應用上或臨床上之最低效果值做爲虛無假設考驗之目標，或先設定一應用上或臨床上之最小差異值做爲虛無假設考驗之目標。如此一來研究者必將研究的關切重心從統計上之「顯著性」轉移到實質上的「意義性」與「應用性」了。而且，「顯著性」與「意義性」或「應用性」都兼顧到。

研究者遵行上述之共識與考驗步驟，當能避免統計顯著性考驗的誤用與誤解或濫用，亦能增強與確保研究結果之內、外在效度。Huff(1954)在其著作：How to Lie with Statistics一書中引述英國政治家Disraeli的看法「世間存在三種謊言：一般性謊言、詛咒謊言、與統計(Lies, damned lies, and statistics)」。統計不是高明之騙術，研究者似乎應以此爲誡，注意統計運用之妥當性，才不致像醉漢把燈柱當支撐用而非照明。因爲徒有科學工具或方法並不能保證研究結果的科學性，社會科學研究者似應更努力於研究假設所根據之理論內涵的探討與研究結果的連貫性與推論性，而非光利用自然科學之方法取得科學之表徵放在貧乏空洞的理論上而已。此外，研究假設考驗結果之應用與解釋更應著眼於情境脈絡性、實質意義性與應用可行性上。誠如陳伯璋（民76）所言：「教育研究不僅要從客觀的量化中來發現事實，更需要以質的方法來詮釋這些事實背後的意義」。也誠如黃政傑（民80）、邱天助（民78）、林彩岫（民82）所呼籲：質、量研究應相互爲用，加以統合。質化量的研究具體做法有二：使用質的方法發展研究假設或量化工具、使用質的方法闡釋量的研究發現。而質化與量化研究之整合方式可以下列四種模式進行：順序式、並行式、融合式、互動式（林彩岫，民82）。質、量研究是互補而不是互斥的，相互截長補短方能提高研究品質，並掃除量化研究「玩弄統計」、「有術無學」的科學八股。

參考書目

邱天助（民 78 ），研究方法：質與量的辨證。**社會教育期刊，** ***18,*** 133-157 。

林彩岫（民 82 ），教育研究的兩個典範：質與量之討論。載於賈馥茗，楊深坑主編之**教育學方法論**(pp. 209-225) 。台北：五南。

芮涵芝（民 85 ），**科學的本質**。台北：桂冠。

高敬文（民 81 ），**未來教育的理想與實踐**。台北：心理。

高敬文（民 85 ），**質化研究方法論**。台北：師大書苑。

陳伯璋（民 76 ），**教育思想與教育研究**。台北：師大書苑。

黃政傑（民 76 ），教育研究亟須擺脫量化的支配。載於中國教育學會主編之**教育研究方法論**(pp. 131-140) 。台北：師大書苑。

黃政傑（民 80 ），質量研究的對立與統合。載於**教育理念革新**(pp. 107-120) 。台北：心理。

詹志禹（民 82 ），二十世紀科學哲學的發展。載於嘉師國教所主編之**國民教育學術演講集**（第一集， pp.29-47 ）。

賈馥茗（民 73 ），教育研究的反省。載於師大教育研究所主編之**教育學的展望**(pp. 3-15) 。台北：五南。

Carver, R. P.(1978). The case against statistical significance testing. ***Harvard educational review, 48(3),*** 378-399.

Cohen, B.(1996). Explaining psychological statistics. ***Pacific grove,*** CA: Brooks/Cole.

Cohen, J.(1988). ***Statistical power analysis for the behavioral sciences(2nd ed.).*** Hillsdale, N. J.: Lawrence Erlbaum.

Cohen, J.(1990). Things I have learned so far. ***American psychologist, 45,*** 1304-1312.

Cohen, J. (1992). A power primer. ***Psychological Bulletin, 112,*** 155-159.

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. ***Psychological Bulletin.*** 86, 662-679.

Dar, R.(1978). Another look at Meehl, Lakatos, and the scientific practices of psychologists. ***American psychologists, 42,*** 145-151.

Falk, R., & Greenbaum, C. W.(1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. ***Theory and psychology, 5,*** 75-98.

Fisher, R. A.(1925). ***Statistical methods for research workers.*** London: Oliver & Boyd.

Fisher, R. A.(1966). ***The design of experiments(8th ed.).*** New York: Hafner.

Frick, R. W.(1996). The appropriate use of null hypothesis testing. ***Psychological methods, 1,*** 379-390.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values : What should be reported and what should be replicated? ***Psychophysiology, 33,*** 175-183.

Harcum, E. R.(1990a). Guidance from the literature for accepting a null hypothesis when its truth is expected. ***Journal of general psychology, 117,*** 325-344.

Harcum, E. R.(1990b). Deficiency of education concerning the methodological issues in accepting null hypotheses. ***Contemporary educational psychology, 15,*** 199-211.

Huck, S. W., & Cormier, W. H.(1996). ***Reading statistics and research***(2nd ed.). New York: Harper Collins College Publishers.

Huff, D.(1954). ***How to lie with statistics.*** New York: Norton.

Keppel, G.(1991). ***Design and analysis: A researcher's handbook(3rd ed.).*** Englewood Cliffs, New Jersey: Prentice Hall.

Kiess, H. O.(1996). ***Statistical concepts for the behavioral sciences(2nd ed.).*** Boston:Allyn & Bacon.

Kirk, R. E.(1982). ***Experimental design: procedures for the behavioral sciecne(2nd ed.).*** Pacific Grove, CA: Prooks/Cole.

Kirk, R. E.(1996). Practical significance: A concept whose time has come. ***Educational and psychological measurement, 56,*** 746-759.

Lakatos, I.(1978). Falsification and the methodology of scientific research programs. In J. Worral & G. Currie(Eds.), ***The methodology of scientific research programs: Philosophical papers(Vol.1).*** New York: Cambridge University Press.

Lykken, D. T.(1968). Statistical significance in psychological research, ***Psychological Bulletin, 70,*** **151-159.**

Meehl, P. E.(1978). Theoretical risks and tabular asterisk: Sir Karl, Sir Ronald, and the slow progress of soft psychology. ***Journal of counseling and clinical psychology, 46,*** 806-834.

Morrison, D. E., & Henkel, R. E.(Eds.)(1970). ***The significance test controversy.*** Chicago: Aldine.

Murphy, K. R.(1990). If the null hypothesis is impossible, why test it? ***American psychologist, 45(3),*** 403-404.

Orey, M. A., Garrison, J. W., & Burton, J. K.(1989). A philosophical critique of null hypothesis testing. ***Journal of research and development in education, 22(3),*** 12-21.

Popper, K. R.(1959). ***The logic of scientific discovery.*** New York: Basic Books.

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. ***Journal of counseling psychology, 34,*** 365-371.

Shaver, J. P.(April, 1992). ***What statistical significance testing is, and what it is not.*** Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Shine, L. C.(1980). The fallacy of replacing an a priori significance level with an a posteriori significance level. ***Educational and psychological measurement, 40,*** 331-335.

Sohn, D.(1993). Psychology of the scientist: LXVI. The idiot savants have taken over the psychology labs! Or why in science using the rejection of the null hypothesis as the basis for affirming the research hypothesis is unwarranted. ***Psychological reports, 73,*** 1167-1175.

Stallings, W. M.(March , 1985). ***Mind your p's and alpha.*** Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Il.

李茂能，嘉義師院初教系副教授