虛無假設顯著性考驗的演進、議題與迷思

李茂能

嘉義大學教育學系

摘要

本文旨在釐清虛無假設顯著性考驗過程中相關的議題與迷思，並凸顯 p 值與效果值的同等重要性。在相關議題方面，文中論及顯著性考驗之兩大派別與脈絡、點與區間虛無假設考驗、α 值與 p 值的定義、為何需事先訂出 α 值、為何不要用「*」表示顯著水準、效果值分析與顯著性考驗是否同等重要、研究樣本到底要多大、及研究結果可複製性等議題。在迷思方面，主在探究α、p 值的迷思、p 值與實得統計考驗力之迷思、第一與第二類型錯誤之迷思。為避免顯著性考驗之誤用，文末提議採用兩套關卡進行統計的顯著性考驗，其中第一道關卡是 p 值的分析，第二道關卡是效果值的分析。量化研究的品質透過此兩套關卡才能確保，其研究結論方能更具運用價值。

關鍵字：虛無假設、顯著性考驗、效果值分析、統計考驗力、樣本規劃

Null Hypothesis Significance Testing：Its Evolution, Current Use, Misuse, and Misconceptions

Mao-neng Fred Li

Department of Education,
National Chai-yi University

Abstract

This paper discusses some uses, misuses, and misconceptions of null hypothesis significance testing (NHST) in current social science research. First, the historical development of the Fisher test of significance and the Neyman-Pearson hypothesis testing are briefly described. Second, several critical issues related to NHST are discussed (e.g., α and *p*-values、point and range estimation). Third, several frequently asked questions about NHST are answered: Why should α be declared before data are collected? Why not use「*」to indicate a significant result? Are effect size analysis and statistical significance testing equally important? What is a sufficient sample size? Fourth, several misinterpretations of NHST are also discussed (e.g., α vs. *p*-value、*p*-value vs. statistical power, Type-I vs. Type-II error rates). Finally, to avoid misuse of NHST and ensure more practical or clinical significance, a two-step procedure (*p*-value analysis and effect size analysis) is proposed for evaluating a hypothesis and quality control.

Keywords：Null hypothesis、significance testing、effect size analysis、power、sample size

壹、前言

虛無假設顯著性考驗（null hypothesis significance testing，以下簡稱 NHST）是應用統計學的核心，本應是一般應用統計學教科書探討之主要重點，但這個核心內容之爭議性與迷思，卻很少被論及（Gliner, Leech, & Morgan, 2002）。歸納起來，NHST 主要的疑義有：為何要事先訂出 α 值、單側考驗的使用時機為何、效果值（effect size，請參見第參節中之定義）的分析為何與顯著性考驗同等重要、α 值與 p 值有何不同、為何勿用「*」以顯示顯著水準、在研究中統計考驗力分析與樣本規劃如何落實、第一類型錯誤比第二類型錯誤嚴重嗎？

過去統計學的教科書或學校相關課程，通常只教導Neyman-Pearson式「*p<.05」的顯著性考驗模式，因此學習者誤認為這一個考驗模式是唯一的。其實，NHST的地位至今仍處於質疑的狀態，它是經歷Fisher與Neyman-Pearson兩派的嚴厲論戰後之磨合產物，在1940~1960年代間，經由其它統計學者之整合，始漸穩固下來而成為今天量化統計考驗的主流；但也從1940年代開始出現了不少批判NHST的聲音（Kline, 2004；Thompson, 2001），文中他引述了402篇批判NHST的論文，值得大家細讀以充分了解NHST的分歧性與爭論性，並防範可能導致的迷思與誤用。為使讀者能全盤掌握NHST在學術研究上應用之演變，以下先從虛無假設（null hypothesis）與顯著性考驗（significance testing）蛻變之脈絡談起，再逐步探討NHST過程中所衍生的相關疑議與迷思，最後再提出解決之道。

貳、NHST之兩大模式演變與考驗步驟

欲談 NHST 的歷史演進，無法不談兩位當代統計學大師 Pearson 與 Fisher 的因緣際會。Pearson 曾長期擔任生物統計與 Biometrika 期刊主編，由於他的主觀性強烈，尤其與他看法異議的文章，即會加以刪改或退稿，當然 Fisher 的稿件亦不例外，導致與本世紀最有才華的統計學家 Fisher 結下不解之仇。Fisher 與 Pearson 原本均任職於 Galton 實驗室擔任統計分析工作，但因兩人意見長期相左，迫使 Fisher 於 1919 年拒絕在 Pearson 手下繼續工作，跑到 Rothamsted 農業實驗場去做研究，並戲稱自己是在「肥料堆理」工作的人。他負責的主要工作是植物播植實驗的設計及整理該實驗場 90 年來累積的實驗資料（翁秉仁，2000；Salsburg, 2001）。Fisher 在這裏發展出變異數分析方法及研究假設考驗之模式。在具有恩怨情仇的氣氛之下，兩位大師對於統計顯著性考驗之步驟自然交集不大。正統的 Neyman-Pearson 的假設考驗（hypothesis testing, 筆者姑且稱之為臨界值法）與 Fisher 式的顯著性考驗（significance testing, 筆者姑且稱之為 p 值法）在哲學觀與方法論上，事實上是具有差異性的（Denis, 2001; Gigerenzer, 2004）。以下先將 Fisher 早期所提出的虛無假設考驗步驟與應用時機摘述如下（Gigerenzer, 2004）：

一、只提出平均數沒有差異（或相關為 0）的虛無假設 H_0，但不提明確之對立假設或研究假設。

二、使用.05 之慣例顯著水準以拒絕虛無假設，假如達到顯著水準，即接納研究假設；以 $p<.05$、$p<.01$ 或 $p<.001$ 報告研究結果（視最接近實得 p 值的大小而定）。
三、不管任何情境，均使用本法。

本法早期為許多研究者所沿用，也被許多教科書所推崇；但因受到諸多的質疑，尤其是 Neyman-Pearson 學派的激烈反對，後來 Fisher 乃將原先所提出的虛無假設考驗步驟與時機修正如下：

一、提出統計上之虛無假設 H_0，但不一定是差異或相關為 0 的虛無假設。
二、根據實際資料，計算所得效果大小全由機遇所造成之機率：p 值。因為研究結果旨在提供資訊而非 Yes-No 之決策，只報告精確之實得 p 值（如 $p=.031$），而不使用.05 當作固定式的顯著水準，亦不論及虛無假設之接納或拒絕的問題。
三、假如您對研究問題所知有限時，才使用本法。

由此觀之，Fisher 認為研究上精確的 p 值報告比拒絕或接納的二分決策重要。他把 p 值視為衡量資料與虛無假設間差距的指標，p 值愈小差距愈大（Sterne, 2002），而且 p 值的顯著與否是事後主觀決定的。Fisher 主張「只提 H_0 但不事先設定 α，且只報告精確之 p 值而不作拒絕或接納的判斷，採取統計者與決策者分離之策略」。其次，為讓您比對兩派學說之差異性，將 Neyman-Pearson 顯著性考驗（簡稱 N-P 法）的考驗步驟摘要整理如下，再作兩派學說之比較。

一、同時提出 H_1 與 H_0。
二、決定適當之統計量與相關之抽樣分配。
三、實驗前先設定可容忍之 α、β 與樣本大小。
四、根據α值及自由度確定臨界值，區分拒絕區與接納區。
五、計算樣本統計量。
六、進行裁決，不是接納 H_0 就是 H_1（二分法）。

前述 Fisher 對顯著與否之主觀認定作法，受到二分法學派 Neyman-Pearson 的嚴詞批判，他們認為 N-P 法是比較客觀的理論決策模式。雖然對立假設 H_1 是研究者真正關切之所在，但對立假設有很多個且永遠無法被證實為真。因此，只能提出 H_0 間接加以求證，這是 Fisher 顯著性考驗與 Neyman-Pearson 假設考驗最大差異所在：同時提出 H_1 與 H_0。

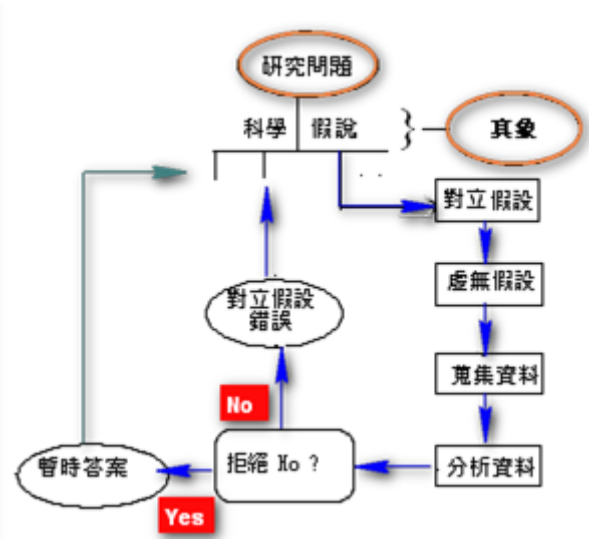
N-P 法最適用於研究者能針對 α、β 作出損益平衡（cost-benefit trade-off）分析與對研究問題有較明確之方向時。Neyman 與 Pearson 同時考慮第一類型錯誤與第二類型錯誤，以盡量降低考驗誤差。因而 N-P 臨界值法須事先設定α、β 及提出精確對立假設，最後進行拒絕或接納之決策。N-P 法行之有年，最常用在工業品管的決策上（Gigerenzer, 2004）。但如遇 p 值等於.049 與 p 值等於.051 時，前者達到統計上.05 之顯著水準，後者卻被判為未達到統計上.05 之顯著水準。雖然兩者之 p 值差異甚小，卻有天壤之別的決策。其實，研究者如能將決策二分法中，再增加一個未定論的選項則當更合乎情理，可參用 Steiger (2004) 之區間虛無假設考驗法（a range hypothesis 法，詳見後敘）；或者只看 p 值的實際大小，再視其他因素作出決策，較不會產生上述之困擾。Fisher 氏的顯著性考驗即是採取此策略，把顯著性考驗之重心從虛無假設是否達到顯著與否轉移到虛無假設為假的機率大小，因此其所謂的顯著水準（level of significance，α）

即等於 p 值，而 Neyman、Pearson 的α與 p 值則是兩回事，這可能是導致日後α與 p 值混淆的主因。

由於 Neyman-Pearson 與 Fisher 對於統計考驗的看法迥異，爭辯言辭亦甚激烈與刻薄，後來他們弟子為了撮合兩派之學說，但又怕得罪自己的老師而以匿名方式提出以「$p<\alpha$」進行統計決策之整合模式（Gill, 2007），稱之為虛無假設顯著性考驗（NHST），這可能是集體睿智的傳奇表現實例。這一個整合模式事實上 Neyman、Pearson 與 Fisher 統計大師仍不滿意，Hubbard 與 Bayarri (2003)、Gigerenzer (2004) 等人亦認為這是一個結合兩派理論的怪胎，主張揚棄傳統之虛無假設顯著性考驗。

比較上述兩派之統計考驗步驟可知，除了在資料收集前，須先固定α之顯著水準及第二類型錯誤β之外，Neyman-Pearson 力倡同時提出 H_1 與 H_0，但 Fisher 主張在研究問題所知有限時，只提 H_0，而反對提 H_1，因為對立假設有千千萬萬個，且我們無從知道對立假設是否正確（Denis, 2001）。例如，過去托勒密的天動說（太陽繞地球運轉），此種學說存在了 1400 年，之後才被哥白尼的地動說（地球繞太陽運轉）所取代，兩種假說接經過科學化的檢驗，卻出現迥然不同的學說，誰是誰非端視哪一學說較能提出合理說明。因此，Denis (2001) 提出統計上的對立假設與概念上（科學上）的對立假設（conceptual alternative）的論點，他認為研究者須先針對研究問題提出合理之假說，再將之轉換成統計上的對立假設。依照 Neyman-Pearson 的看法，當虛無假設被拒絕時，對立假設即被推論成立，但這僅是一個科學的假設獲得支持而已，這似乎是 N-P 法的弱點。概念上的對立假設是說明對立假設成立的原因假說，這些原因假說可能有無數個，因此他認為這兩種假說不可混為一談。例如，過去對於發高燒與沼澤之關係的研究，曾提出瘧疾是隨機因素所致當作虛無假設，並提出瘧疾是吸進有毒瘴氣作為對立假設，並推論對立假設成立。但後來卻發現瘧疾是生長在沼澤並帶有寄生蟲的蚊子所傳染的疾病，而非瘴氣所致。

由此觀之，統計上對立假設的成立，並不能視為概念上的對立假設一定成立。科學研究上的假說要變成學說，常需要千錘百鍊之後才能確立，亦反映出研究者不能任意提出對立假設，在拒絕虛無假設之後，亦須深思有無其他拒絕虛無假設的原因，在無法拒絕虛無假設之後，亦須再提出更合理之假設以利尋求真象。筆者乃以圖一說明前述科學探究的持續過程，因為當您找到了暫時的答案，但無法證實它，當您接納虛無假設，亦須繼續尋求其他合理的科學假說，再度去驗證它為真之可能性。因此，不管您的研究假設是否得到資料之支持，統計假設考驗的結果都只是暫時性的答案而已，真象的追求乃是永無止境的。



圖一 真象探究的循環流程

參、NHST 考驗過程中的重要議題

由於 N-P 法仍是 NHST 應用的主流，以下之論述雖主採 N-P 的論點，但仍顧及 Fisher 的主張。NHST 考驗所涉及之主要內涵有：對立假設與虛無假設之設定原則、α 值與 p 值的分辨、p 值與效果值的陳述方式、第二類型錯誤與統計考驗力的關係、點虛無假設考驗與區間虛無假設考驗、統計考驗力的主要影響因素、樣本大小之規劃；以下將針對這些問題逐一加以說明與論述，以利研究者之全方位的理解與運用。

一、對立假設與虛無假設之設定原則

Linacre (2010) 認為虛無假設的選擇需視我們的研究焦點而定。例如，在測驗資料的向度（dimensionality）分析時，當我們需要更強而有力的證據來拒絕測驗資料的多向性時，我們可以將這個測驗資料具有多向性的設定為虛無假設；假如我們需要更強而有力的證據來拒絕測驗資料的單向度時，也可將這個測驗資料具有單向性的建構設定為虛無假設。以上這兩種假設都是有效的。筆者則認為研究者可根據研究問題或探究之焦點，提出相對應之研究假設，其 H_1 與 H_0 之寫法可依下列任一原則加以陳述：

（一）將想要驗證的研究問題（例如：男生的 IQ 優於女生）定為對立假設，且對立假設之設定須有理論基礎或經驗依據，想要否定的研究假設定為虛無假設，

（二）將別人的主張(欲探究焦點)設為虛無假設（例如：某專家宣稱修高統可以治癒 80%以上的失眠患者），或

（三）將後果較嚴重「即錯誤拒絕比錯誤接受某一假設的後果」的假設定為虛無假設（例如：無罪者而被判刑），以控制 α 風險於可忍受的範圍內。

二、p 值與α值的定義與分辨

p 值（p-value）又稱統計考驗的實得顯著水準（observed significance level）。p 值是虛無假設 H_0 可以被拒絕的最低顯著水準（the smallest value of α for which H_0 can be rejected）。p 值可以說是根據實得統計量去拒絕虛無假設時，實際犯第一類型錯誤的機率，這與研究者事先所設定的犯第一類型錯誤的顯著水準α不同，一般研究者或讀者極易將兩者混為一談。簡言之，α值是研究者作為拒絕虛無假設的關鍵性 p 值（the critical p value）；而 p 值是：假定虛無假設為真的話，計算實得統計量與虛無假設中參數之差距，接著計算等於或大於此差距量的尾部機率（tail probability），此條件機率（$P(X \geq D \mid H_0)$）甚小時，代表虛無假設為真的話欲獲得此等統計量的機遇甚微（李茂能，1998）。p 值的計算得視假設考驗是單側或雙側考驗而定；例如，以單一母數的 z 考驗為例：

（一）當 $H_1: \mu > \mu_0$ 右尾考驗時，

$p =$（$z >$ 考驗統計量）之機率；

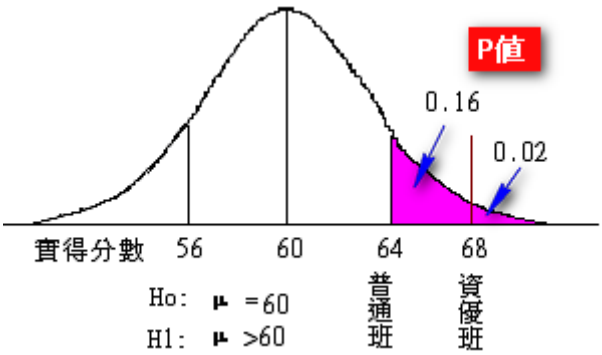
（二）當 $H_1: \mu < \mu_0$ 左尾考驗時，

$p =$（$z <$ 考驗統計量）之機率；

（三）當 $H_1: \mu \neq \mu_0$ 雙尾考驗時，

$p =$（$z >$ 考驗統計量）之機率＋（$z <$ 考驗統計量）之機率。

其實，假如您並無 100%的信心去預期所得結果之方向，那您還是進行雙側考驗吧，犧牲一點統計考驗力，進行雙側的統計顯著性考驗是保守但客觀的做法。



圖二 p 值的真義

以圖二為例，當對立假設為右側單側考驗（$H_1: \mu > 60$）時，其 p 值等於「$z >$ 考驗統計量」之機率；如為左側單側考驗（$H_1: \mu < 60$），其 p 值等於「$z <$ 考驗統計量」之機率；如為雙側考驗（$H_1: \mu \neq 60$），其 p 值等於「$z <$ 負的考驗統計量」之機率＋「$z >$ 正的考驗統計量」之機率。例如，假設本校學生高統考試過去的平均數為 60 分是事實，接著調查 36 位學生後發現：本年度普通班學生高統考試的平均數為 64 分，而其標準差為 24 分，進行 z 統計考驗後知，光靠機遇就會發生此種事件的機率為 0.1587，亦即普通班學生高統考試之真正的平均數如為 60 分及格的話，該班學生獲得高統的平均數為 64 分的可能性約有 0.16。其考驗統計量（observed z score）計算式為：

$$z=\frac{64-60}{\frac{24}{\sqrt{36}}}=1.0$$

p 值$=p(z>1.0)=.1587$

p 值其實是一種條件機率：$p(\overline{X}>64\,|\,H_0:\mu=60)\cong.16$。簡言之，研究者旨在探究，假設$H_0:\mu=60$為真的話，樣本平均數$\overline{x}>64$的機率有多少？

由此例之解說可知，p 值並不是虛無假設為真的機率，而是當 H_0 為真時，會發生所觀察事件之條件機率。當 p 值甚小時，我們拒絕虛無假設的原因乃是假如虛無假設為真的話，不太可能發生這樣的事件，因此可以考慮拒絕所提出的虛無假設。

三、p 值與α值的應用與陳述方式

由於 Neyman-Pearson 與 Fisher 對於統計考驗的論點迥異，導致目前對於 NHST 在論文上的實際作法分歧，有些研究者採用 N-P 臨界值法：查表決定臨界值,再利用*說明是否達到某一顯著水準,有些研究者則採用 Fisher 的 p 值法。筆者認為 p 值法具有以下兩個優點:

（一）一般常用之 SPSS/SAS 統計軟體均會主動報告 p-value，研究者可以根據α值評估是否拒絕虛無假設，不必查臨界值表。

（二）根據 p 值，研究者可以針對任何顯著水準進行統計考驗結果之評估。

另外，p 值與α值在論文中的陳述方式不一而足，但筆者認為以下兩種呈現方式較為完整且傳遞之訊息明確，可作為參考。以$H_0:\mu_1=\mu_2=\mu_3=\mu_4$考驗為例，說明如下：

1.精確 p 值報告法

（1）($p=0.01<\alpha=0.05$)，

（2）($F(3,76)=4.27, p=0.008, \alpha=0.05$)，

（3）($F=4.27, df=3/36, p=0.008, \alpha=0.05$)。

如為單側檢定：($F=4.27, df=3/36, p/2=0.008, \alpha=0.05$)。研究者應用時，取其一加以報告即可。研究者如欲加註效果值，亦可將之含入上列括弧之式子中。

2.粗略 p 值報告法

（1）($p<0.05$)，

（2）($F(3,76)=4.27, p<0.05$)，

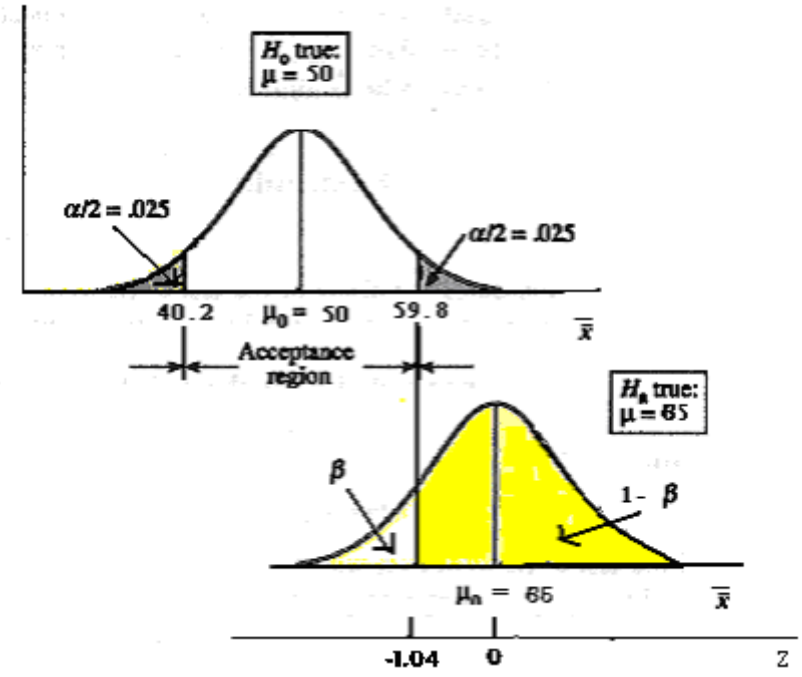
（3）($F=4.27, df=3/36, p<0.05$)。

如為單側檢定：($F=4.27, df=3/36, p/2<0.05$)。研究者應用時，取其一加以報告即可。研究者如欲加註效果值，亦可將之含入上列括弧之式子中。

四、第二類型錯誤與統計考驗力

可能由於 Fisher 的統計考驗模式偏重第一類型錯誤，第二類型錯誤經常被研究者所漠視。第二類型錯誤乃是當對立假設為真時，接納虛無假設的機率；而統計考驗力（power）乃是當對立假設為真時，正確拒絕虛無假設的機率（參

見圖四或圖五），亦即一個統計量可以偵測到最低效果值實際存在之能力。換言之，第二類型錯誤乃是對立假設為真時，研究者卻接納虛無假設的機率（β），而統計考驗力即是當對立假設為真時，研究者能正確拒絕虛無假設的機率（1-β）。套裝統計軟體如 SPSS 或 SAS，其統計考驗力的計算通常可借助於非中心性 t、χ^2 或 F 分配。統計考驗力之分析事實上是在控制第二類型錯誤與決定最低樣本數而又能發現具有臨床上或理論上意義的效果值大小。因此，統計考驗力有如化學試劑之敏感度，過低或過高均會造成不報與誤報之錯誤決定。一般研究者皆希望α與β均很低而統計考驗力可以很高，不過這樣的期望無法透過α與β的控制達成，因為α與β的關係成反比而β與統計考驗力的關係亦成反比，可謂魚與熊掌不可得兼。唯一可行的方法是透過樣本大小的控制及使用信、效度較佳的測量工具，來獲得所預期之統計考驗力，以能發現臨床上或理論上具有意義的效果值。為便利於讀者之理解，以下將以圖三之右尾考驗實例，說明如何逐步計算出統計考驗力。



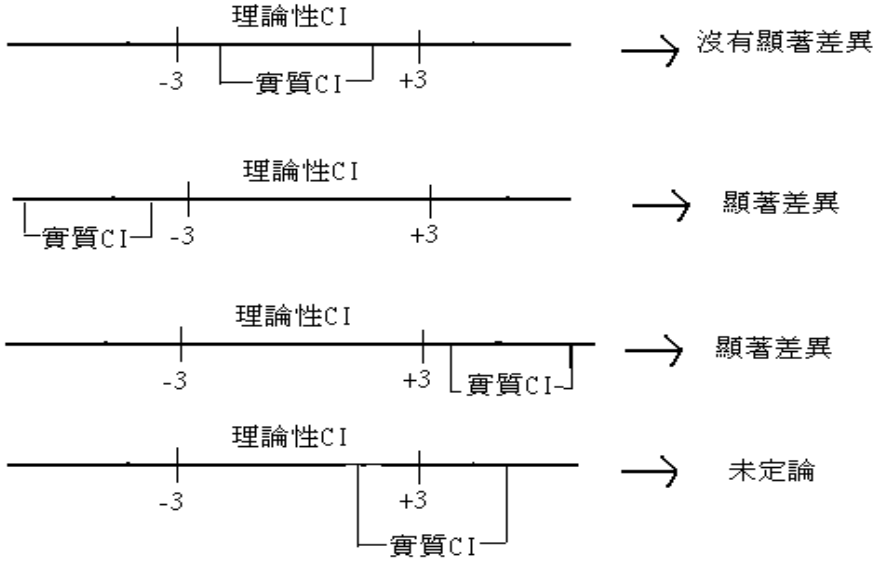
圖三 β與統計考驗力的計算：$H_0:\mu=50, H_a:\mu=65$

研究者如欲手算 z 分配下之統計考驗力，以 $H_0:\mu=50$，$\sigma_{\bar{x}}=5$ 為例，首先需查出在特定顯著水準之虛無假設下的臨界值（如 $\alpha/2=.025$ 時，上尾之臨界值為 1.96），接著計算 1.96 所對應之原始分數，如圖三中之原始分數 59.8(1.96*5+50)。其次，假定對立假設 $H_a:\mu=65$，$\sigma_{\bar{x}}=5$，在此分配下計算 59.8 遠離 65 的 z 分數((59.8-65)/5=-1.04)。由此 z 分數即可算出其遠離 $z=0$ 所對應之機率為.3508，而犯第二類型錯誤之機率β值等於.1492(=.5-.3508)。因此，其統計考驗力為：1-.1492=.8508。此項統計考驗力亦可以下式求得：.50＋.3508=.8508。

五、點虛無假設考驗與區間虛無假設考驗

Steiger (2004) 認為點虛無假設（point null hypothesis）非常不可能為真，認為區間虛無假設考驗法（loose/range null hypothesis）更貼近事實。有別於傳統的間接求證的歷程（reject-support hypothesis testing），有些研究者主張直接求證的"accept-support"考驗方法（Steiger, 2004），前者研究者以間接否定 H_0 來支持自己希望為真的 H_1，後者研究者則直接以 H_0 是否成立與否，來支持自己希望為真的 H_0。後者的統計考驗方式最常見於生物統計上，因為生物統計學者進行統計考驗的目的，經常在於檢驗兩種藥物藥效是否相同（a test of bioequivalence），而非兩者是否有差異，因此研究者通常希望 H_0 為真。由此觀之，第一類型與第二類型錯誤的角色恰好互換。為更客觀公正，他主張研究者最好同時提出兩個單側之假設考驗。例如，假如研究者要檢驗兩種減肥法效果是否相同，並認定兩種減肥法的差異小於 3 公斤，即視為沒有實質上意義之組間差異 trivial difference），因此這兩種減肥法效果視為相同。其理論上的虛無假設為 $H_0: -3 \le \mu_1 - \mu_2 \le +3$（此即理論上之信賴區間，乃是研究者希望為真的假設），而其對立假設則為 $\begin{matrix} H_1: \mu_1 - \mu_2 < -3 \\ H_1: \mu_1 - \mu_2 > 3 \end{matrix}$，此種雙重假設考驗等同於下述之兩個單側假設考驗: $\begin{matrix} H_0: \mu_1 - \mu_2 \le -3 \quad H_1: \mu_1 - \mu_2 > 3 \\ H_0: \mu_1 - \mu_2 \ge 3 \quad H_1: \mu_1 - \mu_2 < 3 \end{matrix}$。

當上述這兩個虛無假設都被拒絕時，表示這兩種減肥法間的差異($\mu_1 - \mu_2$)未達顯著差異，意即實得信賴區間[ex. 以α=.05 為例，$(\mu_1 - \mu_2) \pm 1.96 * (\sigma / \sqrt{n})$]完全落入理論信賴區間之內，實得信賴區間係指組間平均數差異的信賴區間；否則即表示這兩種減肥法間的差異具有顯著差異。上述考驗的結果亦可以下列之理論信賴區間與實得信賴區間是否重疊加以詮釋（參見圖四）。當實得信賴區間完全落入理論信賴區間之內時，即表示兩種減肥法的差異未達顯著差異。當實得信賴區間完全落在理論信賴區間之外時，即表示兩種減肥法的差異已達顯著差異。當實得信賴區間與理論信賴區間只有部分相重疊時，即表示兩種減肥法的差異未有定論，有待後續研究；注意此種決策法有別於傳統的二分法，較適合於當研究者希望虛無假設為真時的情境中。



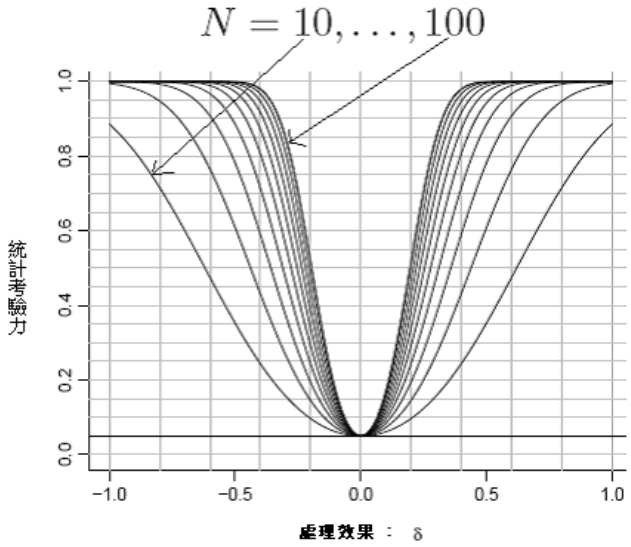
圖四 實得信賴區間與理論信賴區間在假設考驗上之應用

六 、統計考驗力的主要影響因素

統計考驗力（Power）是指正確拒絕虛無假設的機率，而影響統計考驗力的主要因素為：

（一）顯著水準，水準愈低統計考驗力愈低。
（二）樣本大小，樣本愈大統計考驗力愈高。
（三）母群變異量，變異量愈大統計考驗力愈低。
（四）效果值大小，差異量愈大統計考驗力愈高。
（五）測驗工具之信、效度，信效度愈好統計考驗力愈高。

這幾個關鍵因素與統計考驗力互為函數關係，研究者可善用它們進行樣本規劃。這些因素中，以樣本大小與處理效果值影響最大。在α與樣本大小保持恒定之下，處理效果愈大，統計考驗力則愈強，他們之間的關係可從圖五底部X軸之左右極端效果值看出來；在α與處理效果保持恒定之下，樣本愈大（例如，圖五內部曲線所代表的N由10增加到100時），統計考驗力也愈強。因此，研究者於解釋NHST達到顯著水準時，首先應分辨是樣本大小或是處理效果所致，假如是樣本過大所致，那麼該研究之結果就無應用價值可言，這是NHST應用上的最大盲點。因應之道，其一是研究者可將效果值分析融入虛無假設考驗中，亦即先決定臨床上或應用上之最低效果值，以此最低效果值作為H_0，考驗此H_0以決定所得之真正差異是否達到臨床或實際應用上之顯著水準。另一解決之道是利用此最低效果值，於研究之前先進行樣本規劃，以便能獲得可以偵測到最低效果值的統計考驗力，資料分析時並納入效果值大小之分析，以進行「顯著性考驗的品管」。



圖五 樣本大小、統計考驗力與處理效果間之關係（取自 Carstensen, 2002）

七、效果值之定義

為便利研究者深入了解效果值之定義及如何由 SPSS 之報表中獲得這些資訊，以下先摘要效果值之類別與定義，再進一步說明如使用 SPSS 計算效果值。效果值大小之估計，主要可分為三類：（一）標準化的差異指標（如Δ_1,Δ_2）、（二）相關指標（如 r,ϕ）、（三）變異解釋量的指標（如 R^2 與η^2）。茲依序定義如下：

效果值類別

❶ 平均數差異標準化

$$\Delta_1 = \frac{|\mu - \mu_0|}{\sigma}$$ （單一樣本）

$$\Delta_2 = \frac{|\mu_1 - \mu_2|}{\sigma}$$ （雙樣本）

❷相關係數: r, ϕ

❸決定係數 ： R^2, η^2, ω^2

決定係數會因不同統計方法而有不同之定義，進一步界定如下，以利讀者之驗證：

$$\omega^2 = \frac{SS_{effect} - (k-1)MS_W}{SS_T + MS_W}$$（固定效果 母群估計值）

$$\eta^2 = \frac{SS_{effect}}{SS_T}$$ …（η是曲線相關，樣本估計值）

$$\eta_p^2 = \frac{SS_{effect}}{SS_W + SS_{effect}}$$ …（partial η^2, SPSS提供）

$$\varpi^2 = \frac{MS_{effect} - MS_W}{MS_{effect} + (n-1)MS_W}$$ （隨機模式）

$$R^2 = \frac{SS_{reg}}{SS_T}$$

上式中η^2與另一效果值指標f^2具有密切互換關係，f^2的計算公式請參閱李茂能（2002）的論文。至於各種效果值大小的判斷，因為效果值的最低實用下限值常因各專業或研究性質而不同，多大是大，多小是小，甚難客觀加以論斷。為解決此難題，Cohen（1988, 1992）提出下列的經驗法則，供研究者之初步參考：

1. 大效果：差異值≥.80，相關係數≥.50，效果值：$R^2 \geq .26$，$\eta^2 \geq .14$，$f \geq .40$
2. 中效果：差異值≥.50，相關係數≥.30，效果值：$R^2 \geq .13$，$\eta^2 \geq .06$，$f \geq .25$
3. 小效果：差異值≥.20，相關係數≥.10，效果值：$R^2 \geq .02$，$\eta^2 \geq .01$，$f \geq .10$

接者以表 1 與表 2 之資料，示範如何利用 SPSS 計算 *p* 值、效果值與統計考驗力，以利研究者之實際操作。

表 1 **描述統計摘要表**

依變數: 成績

練習方式	平均數	標準差	個數
集中練習	7.83	2.69	12
分散練習	6.75	2.42	12
總和	7.29	2.56	24

表 2 **變異數分析摘要表**

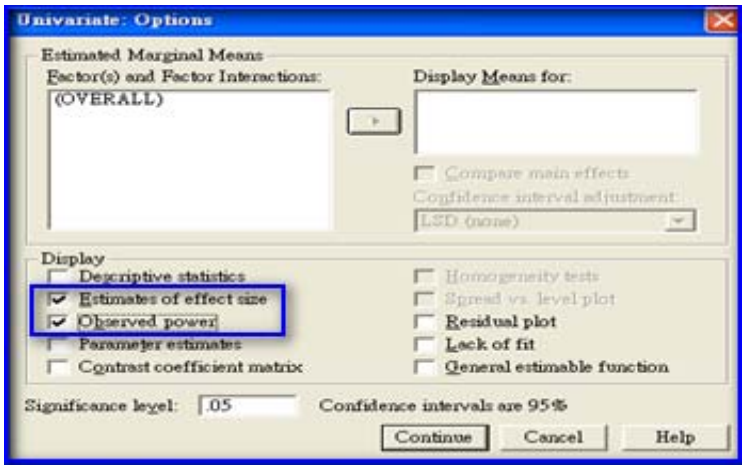
來源	自由度	平均平方和	F 檢定	顯著性	Eta 平方	Noncent. 參數	觀察的檢定能力[a]
校正後的模式	1	7.042	1.076	.311	.047	1.076	.168
Intercept	1	1276.042	195.064	.000	.899	195.064	1.000
A	1	7.042	1.076	.311	.047	1.076	.168
誤差	22	6.542					
總和	24						
校正後的總數	23						

a. 使用 alpha = .05 計算　依變數: 成績

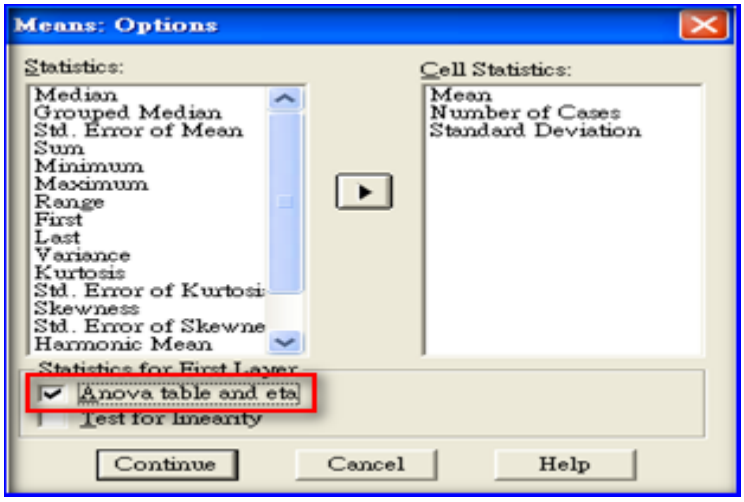
根據表2之資料，逐一演示效果值、統計考驗力、與*p*值之計算及結果，以利讀者具體理解以驗證三者間之關係。計算η^2、NCP、power與*p*值之公式條列如下：

1. NCP=SSA/MSE=7.042/6.542=1.076=F*df_b=1.076*1=1.076，
2. $\eta^2=\frac{SS_{effect}}{SS_{effect}+SS_w}=\frac{F*df_b}{F*df_b+df_w}=\frac{1.076}{1.076+22}$=.047，接著利用SPSS的內建函數計算Power與*p*值。
3. power=1-NCDF.F(q,df1,df2,NCP)=1-NCDF.F(4.3015,1,22,1.076)=.1682，式中q為顯著臨界值，
4. p=1-CDF.F(NCP,1,22)=.311。

因為*t*考驗係ANOVA之特例，計算*t*考驗之*p*值、效果值與統計考驗力亦可利用ANOVA間接求得。實際操作可利用SPSS的副程式「General linear model」之「Univariate」或「Multivariate」下的視窗中，點選「Options」（如圖六所示），在此視窗中點選「Estimates of effect size」及「Observed power」，即可得到η^2與統計考驗力；或使用「Means」下的視窗中，點選「Options」，再出現的視窗中點選「Anova table and eta」（如圖七所示），即可得到η^2。



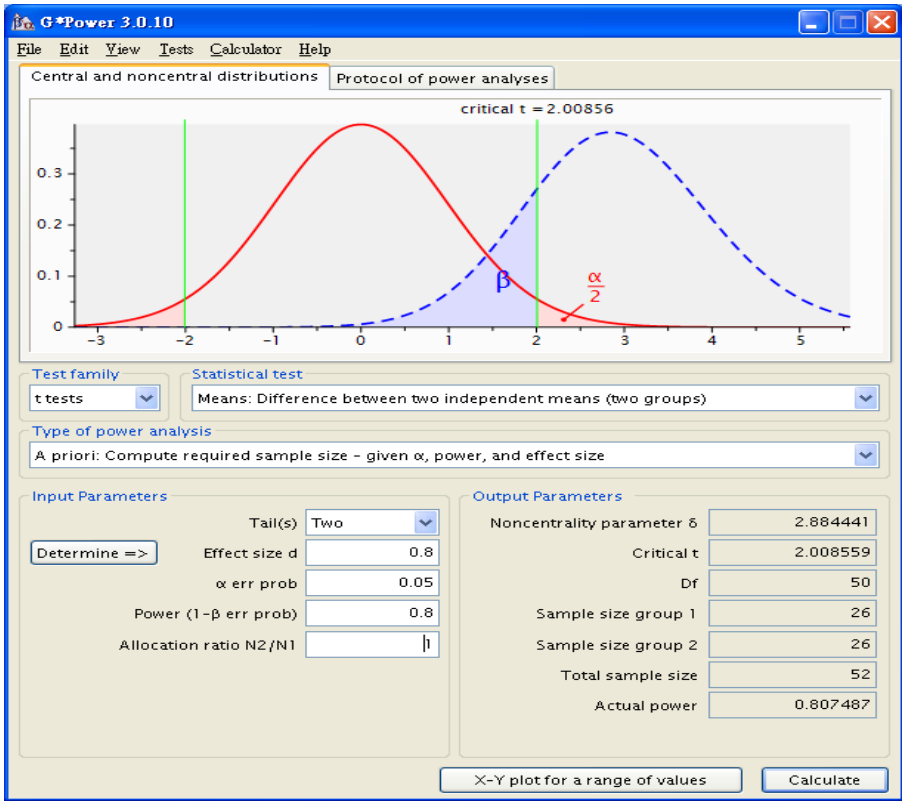
圖六 GLM單變項統計的效果值輸出之設定



圖七 Means分析中效果值之輸出設定

八、樣本大小之規劃

研究樣本之規劃是任何研究中的一個重要環節，它不只涉及成本的問題，也涉及統計考驗力及顯著性的問題，最好不要依據過去之經驗法則作決定，以免失之偏頗。研究者於研究計畫之初，即應慎重規劃樣本大小，所選定之樣本不要過大或過小。過去為規劃適當之樣本數，研究者常須查閱 Cohen（1988）書中為數眾多的統計考驗力圖表，甚為費時費力。為解決此困境，目前已有不少樣本規劃之軟體可供查用（李茂能，2002）。筆者在此推介 G*POWER 3.0 軟體（Faul, Erdfelder, Lang, & Buchner, 2008），該軟體不僅操作容易且是免費軟體共享。使用前須先決定α、統計考驗力、統計方法、單側或雙側、理論上或應用上的最低效果值之後，使用者接著在圖八左下角的「Input Parameters」的小視窗中填入相關參數，例如：雙側、效果值、α、power、與 N1/N2 之比值。最後，按下「Calculate」規劃之樣本大小即會在右下角的「Output Parameters」的小視窗中出現。使用者介面上半部則會出現虛無假設為真與對立假設為真的抽樣分配，可以讓研究者檢視臨界值、第一類性錯誤（α）的大小與第二類性錯誤（β）的大小。G*POWER 3.0 的下載網址：
http://www.psycho.uni-duesseldorf.de/aap/projects/gpower。



圖八 GPower 3.0 使用者介面

圖八係 GPower 3.0 使用者介面，其底部按鈕「X-Y plot for a range of values」，利用此按鈕研究者可以獲得 Power 值與樣本大小的函數曲線，以利研究者快速查看某特定統計考驗力所需之樣本人數。另外，不同的統計方法所需之樣本人數亦不同，但在同一研究中有可能同時用到多種統計方法，為便利取樣研究者可以所需樣本人數最多者，進行樣本規劃。由於顧及回收率及回收資料之可用性，研究者可依預計之回收率酌予調升樣本人數。具體之研究樣本規劃的實例，研究者可以觀摩柳敦仁（2006）、朱國聖（2004）等人之碩士論文。

肆、NHST 考驗過程中常見的迷思

歸納起來，NHST 考驗過程中最常見的迷思類別有三：

一、α與 p 值的迷思

α與 p 值的誤解與誤用甚為普遍，在論文中最常見之現象有：

（一）研究者在看到資料分析結果之後，再決定其α之顯著水準

因為未在資料分析前決定其α之顯著水準，因此易發生在同一研究問題不在同依變項上，出現不同的第一類型錯誤的設定，通常研究者是無法自圓其說的。例如，*p<.05，**p<.01，***p<.001 出現在同樣一個統計分析摘要表中，此即「roving α」的現象（Hubbard & Armstrong, 2005, 2006），α不是事前設定而是變動的第一類型錯誤之機率。此種報告方式亦可能產生 p 與α的混淆；例如，我們無法得知 p<.01 中，.01 是設定之α值或實得之顯著水準。如改寫如 $p = 0.01 < \alpha = 0.05$，即可避免前述之混淆。

（二）發現 p 值介於.05 與.10 之間時，改用單側考驗以達到.05 顯著水準，這亦是 roving α的另一現象。

（三）p 值常被解釋為虛無假設為真之機率

考驗統計量達顯著性時，即認為虛無假設不太可能為真，其實 p 值是 $p(x \geq D \mid H_0)$ 之條件機率，其虛無假設為真之機率 $p(H_0)$ 通常是未知的。

（四）過度簡化成接納或拒絕之模式（只看星號之有無），忽視了第二類型錯誤、power 分析與效果值等問題，這亦是筆者不建議使用*號的理由。

（五）誤將 p 值的大小作為研究價值或重要性的指標

過去有些研究者將*p<.05 解釋為研究效果達到顯著（significant），**p<.01 則解釋為研究效果達到高度顯著（highly significant），***p<.001 乃解釋為研究效果達到極顯著（extremely significant）。這種以*號之多寡詮釋 p 值的方法亦是過度依賴 p 值的結果，而忽略了效果值的大小。McLean (2001) 與 Gliner et al. (2002) 指出「*」不是研究者追求的唯一目標。假如研究者使用了很大的樣本，即使很小的效果值亦會達到統計上的既定顯著水準，而很大的效果值亦可能無法達到統計上的既定顯著水準。另外，研究者異常誤將統計上之顯著性等同於應用上（或臨床上）亦具有顯著性（Gliner et al., 2002）。為了避免思慮不周而誤判之現象，研究者最好是先進行樣本規劃與同時報告 p 值及效果值，以利於研究結果之正確解釋與應用。

基於前述幾個理由，筆者建議研究者盡量勿使用*於統計量上方標註顯著水準，但須同時報告α值與 p 值以利作決策，以避免誤用及誤解。研究者於研究前應先利用統計考驗力分析進行樣本規劃、慎選研究假設，研究結果中則應同時報告α、p 值，信賴區間或效果值大小。基於前述虛無假設顯著性考驗本質上之缺陷，有些學者（如 Steiger, 2004）因而建議採用「最低效果值假設」或「信賴區間估計法」進行顯著性考驗，而非沒有差異之「點虛無假設」假設考驗。

二、p 值與實得統計考驗力之迷思

實得統計考驗力係當效果值已經知道以後，所計算出來能夠拒絕H_0的機率，又稱為事後統計考驗力（post hoc power），它與p值具有簡單的函數關係（Carstensen, 2002）。以單一樣本之z考驗為例，當真正處理效果（true treatment effect）$\delta = 0$ 時，z與p值之計算公式為：

$z = \dfrac{\overline{X} - 0}{\dfrac{\sigma}{\sqrt{n}}}$，其雙尾之下的$p$值=2×(1−CDF(|z|)), 而$z$值亦等於IDF(1−p/2)，反映出$z$值統計量為$p$值的一對一函數。至於其統計考驗力之計算（假如$\alpha = .05$及$\delta \neq 0$時），可由下式求得：

$$P\{|z| > 1.96\}$$

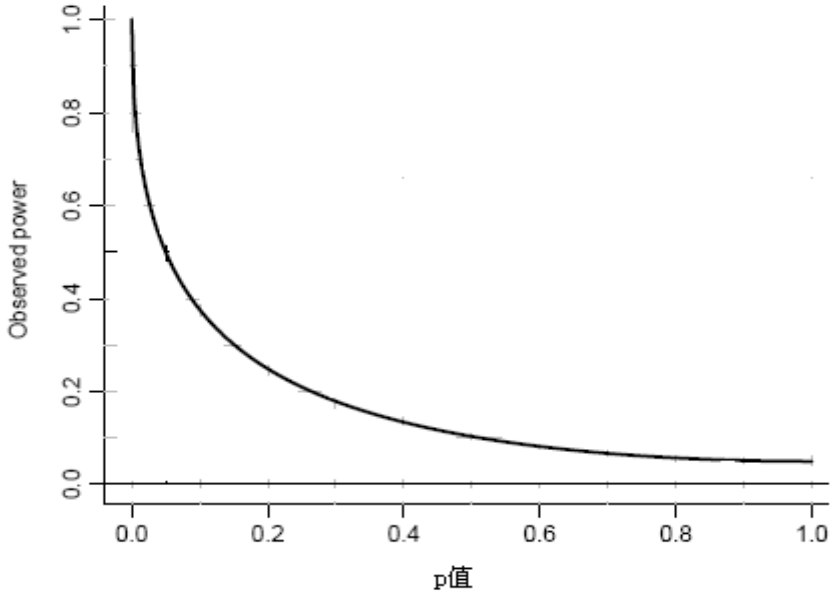
$$= 1 - P\{-1.96 < \frac{\overline{X} - 0}{\frac{\sigma}{\sqrt{n}}} < 1.96\}$$

$$= 1 - P\{-1.96 - \frac{\delta}{\frac{\sigma}{\sqrt{n}}} < \frac{\overline{X} - \delta}{\frac{\sigma}{\sqrt{n}}} < 1.96 - \frac{\delta}{\frac{\sigma}{\sqrt{n}}}\}$$

$$= 1+\text{CDF}(-1.96-\frac{\delta}{\frac{\sigma}{\sqrt{n}}})-\text{CDF}(1.96-\frac{\delta}{\frac{\sigma}{\sqrt{n}}})$$

式中 $\frac{\delta}{\frac{\sigma}{\sqrt{n}}}$ 相當於NCP（非中心性參數），當實得差異$\delta=\bar{x}$時，$\frac{\delta}{\frac{\sigma}{\sqrt{n}}}=z$。

由上式觀之，*p*值、*z*值與事後統計考驗力互為函數關係，事後統計考驗力（在SPSS中稱為observed power）亦僅是*p*值的簡單函數而已，而*z*值亦是*p*值的反密度函數。事後統計考驗力與*p*值之密切關係可由圖九值愈小統計考驗力愈大，反之則愈小。



圖九 事後統計考驗力與 *p* 值之關係

綜上所述知，事後統計考驗力乃由 *p* 值加以決定，而且 *p* 值與事後統計考驗力具有密切的一對一關係，未達顯著水準的 *p* 值必定伴隨出現低的事後統計考驗力。因此利用 *p* 值去計算事後統計考驗力，對於 *p* 值的解釋並無法提供更多的資訊；反而會產生一些誤用與誤解。例如，當 *p* 值未達顯著水準時，如發現事後統計考驗力亦甚低，就說是因統計考驗力不足所致。過去有些研究報告常宣稱「本研究不僅達到統計上之顯著水準，而且統計考驗力亦很高」宣稱「本研究未達到統計上之顯著水準，因統計考驗力較低所致」(Lenth, 2001)。其實，一個研究就是因統計考驗力高才會達到統計上之顯著水準，因為統計考驗力低才不能達到統計上之顯著水準（李茂能，2002)。爰此，進行事後統計考驗力的計算，似乎是沒有意義的。Carstensen (2002) 認為統計考驗力乃是對於一個未來研究能拒絕虛無假設的機率。因此，一當這個研究執行完畢, 其統計考驗力不是 1（假如 H_0 被拒絕的話），即是 0（假如 H_0 未被拒絕的話）。由此觀之，統計考驗力的分析主要用於研究設計之階段，如欲進行事後統計考驗力分析時，必須以理論上的效果值而非實得效果值去估計事後統計考驗力，才更具實質意義。倒是研究者可以利用事後統計考驗力與效果值大小的資訊，去解釋研究結果。例如，如果事後統計考驗力很大但效果值甚小，則可能樣本過大。又如事後統計考驗力很小但效果值並不小，則可能樣本過小，此項研究結果雖未

達顯著水準但仍可能具有應用價值。這大概是 SPSS 副程式 GLM 中提供事後統計考驗力的主要原因。

三、第一類型錯誤比第二類型錯誤重要

研究者進行任何統計決策都會犯第一類型錯誤或第二類型錯誤，一般都認為犯第一類型錯誤比較嚴重。其實，到底何者比較嚴重乃是主觀性認定之問題，端賴研究者如何去詮釋問題的後果。Yu (2006) 談及曾經有一巡防海域之戰艦，從雷達銀幕上發現一不明飛行物，而且無法辨識到底是敵機或友機。

艦長說:

（一）虛無假設是:發現的飛機沒有敵意。假如事實上它具有敵意而我沒有用飛彈攻擊它，這是犯了第二類型錯誤。犯第二類型錯誤的後果就是我艦會被攻擊，而導致我軍傷亡。

（二）對立假設是:發現的飛機具有敵意。但是假如事實上它未具有敵意而我用飛彈把它打下來，這是犯了第一類型錯誤。犯第一類型錯誤的後果就是我提前退役。

（三）由此觀之，犯第二類型錯誤的後果比較嚴重。因此，我不相信虛無假設，下令開火!

不過，副艦長卻辯說:

（一）假如虛無假設為假而我們沒有用飛彈去攻擊它，其後果就是我我軍會有所傷亡。

（二）假如對立假設為假，且該飛機商用民航機，犯了第一類型錯誤的後果就是使得數百個無辜的平民死亡及招致死亡更大的戰爭。

（三）由此觀之，犯第一類型錯誤的後果比較嚴重。因此，我不相信對立假設，不可開火！

Lipsey (1990) 則主張進行基礎研究時，應盡量降低犯第一類型錯誤，因為我們不希望輕易去改變已知的事或接納新的事實；進行應用性研究時，則應盡量降低犯第二類型錯誤，以避免實際應用價值的損失。前述這兩種主張也許可以改變過去大家都認為犯第一類型比較嚴重的說法，而漠視第二類型錯誤。Yu (2006) 引述 Curtis 的個人通信中，就認為犯第一類型錯誤與犯第二類型錯誤，有時其嚴重性是相當的。例如，美國 FDA 管制不良之新藥上市拯救了不少無辜生命，但 FDA 禁止有效之良藥上市亦使不少生命失去生機。足見個人的主觀判斷在平衡第一類型錯誤與犯第二類型錯誤上的關鍵地位（Hubbard & Bayarri, 2003），研究者似乎應針對研究之性質及目的作出理智判斷。

伍、研究結果的可複製性

研究結果的可複製性或推論性與抽出之樣本是否具有代表性具有密切關係，目前之統計技術已可利用原始資料進行內在複製分析（internal replicability），亦可利用外部資料進行外在複製性分析（external replicability）。研究結果的可複製性通常都利用重覆抽樣（resampling）的技術，進行研究結果的複核。假如研究者的樣本過小且非隨機抽樣時，可以利用 bootstrap 或 jackknife procedures 加以複製資料以解決資料常態性與變異同質性的問題，假如研究者的樣本足過大時，可以利用樣本分割技術，檢驗其複核效度（cross-validation），

以避免統計考驗力過大的問題。當然研究者如能重新取樣，另取一個具有代表性之樣本進行效度複核，其研究結果的外在可複製性，當然比前述之內在可複製性更佳。目前的統計分析軟體如 SAS 或 Amos 均能輕易進行 bootstrap 的複製技術。例如，SAS 從第八版起，已有內建的副程式（multtest for multiple testing）可以執行 bootstrapping 的樣本點的複製工作及統計考驗，參見以下之 SAS 程式：

```
multtest bootstrap nsample=200 seed=12345 pvals;
class Group;
test Mean（I1-I6）;
run;
```

式中「nsample」代表複製樣本之大小，「seed」值代表啟動初始值，省略時電腦系統時鐘值會取而代之，使用不同的初始值會使抽樣過程更具隨機性，「pvals」代表輸出原始 p 值與 bootstrapped p 值的比較分析。

陸、統計顯著性考驗的品管

統計顯著性考驗的品管乃是利用效果值分析，進行研究結果的應用性分析。在過去的顯著性考驗過程中，因研究者過度重視 p 值的分析，而漠視了理論效果值（hypothesized effect size）與實得效果值（observed effect size）的地位。理論效果值乃是研究者認為將來在臨床應用上的最低效果值，低於此值即無應用價值可言；實得效果值則是事後實徵資料所獲得的效果值。由於一個研究結果是否達到統計上的顯著性與否，主要是受到α、β、樣本大小、效果值與測量工具的信、效度所影響。因此，研究者在詮釋研究結果時，當然必須全方位去評估研究結果，否則易導致結論的偏差。為避免 NHST 之迷思與誤用，筆者主張採用兩套關卡進行統計顯著性考驗的解釋，其中第一道關卡是量的管制，第二道關卡是質的管制：

一、進行α與 p 值的比較，

二、進行理論效果值（簡稱 HE）與實得效果值（簡稱 OE）的比較。

表 3　**統計顯著性考驗的品管**

	$\alpha \le p$(Non-Sig)	$\alpha > p$(Sig)
OE > HE	可能具應用價值	重要結果
OE ≤ HE	H_0 為真	無應用價值

使用此套顯著性考驗，研究者須先設定甘冒犯第一類性錯誤（α）的大小、甘冒犯第二類性錯誤（β）的大小、及最低理論效果值。由表 3 可知，當$\alpha \le p$ 且 OE ≤ HE 時，H_0 為真的機率就很高了；當$\alpha > p$ 且 OE > HE 時，H_1 為真的機率就很高了，不但虛無假設被拒絕了，而且研究結果亦具應用價值；當$\alpha > p$ 但 OE ≤ HE 時，雖然達到統計上的顯著水準，但研究結果卻不具應用價值；當$\alpha \le p$ 但 OE > HE 時，假如 N 也甚小且測量工具的信、效度亦不錯時，研究結果則可能具應用價值。至於各種效果值大小的判斷，因為效果值的應用下限值常因各專業或研究性質而不同，多大是大，多小是小，可能甚難客觀加以論斷。為暫時解決此難題，可參考前述 Cohen (1988) 的經驗法則，將來理論成熟時再加以

修正。經過此兩套關卡，傳統 NHST 過度偏重 p 值的誤用與迷思情形自然可以避免，統計的顯著性考驗結果必然更能反映真相與具有應用價值。

此外，研究者在進行研究之前，應先決定α值與依據預估最低效果值進行「樣本之規劃」與前述之「顯著性考驗的品管」，經過此完整之配套措施之後，進行統計顯著性考驗才有應用上之價值。研究者且應從以下三種標準去詮釋研究結果：

一、統計顯著性考驗（statistical significance）

二、效果值大小或實用性（practical significance）

三、研究結果的可複製性（replicability），方能作出正確的資料分析及解釋（McLean & Ernest, 1998）。

柒、結語

不管是「Reject-Support」或「Accept-Support」的統計顯著性考驗途徑，Neyman-Pearson 的臨界值法與 Fisher 的 p 值法在 NHST 的運用上各有長短處，而在 NHST 運用上，學術界對於α與 p 的應用與解釋方式，似乎仍有歧見。研究者應體認到統計上達到顯著並不等於臨床上的應用價值（statistical significance ≠ practical importance）。臨床上具有應用價值的差異，亦有可能因統計考驗力較低而無法發現到重要之研究結果。同樣地，統計上達到顯著亦有可能係統計考驗力過強所致，在臨床上並無應用價值。因此，研究者更應事先規劃出適當的統計考驗力與關心實際效果值的估計，而不是考驗效果值是不是等於 0。因此，傳統虛無假設顯著性考驗並非適用於所有之研究情境，有時研究者可改用最低效果假設顯著性考驗、或最低效果值的信賴區間估計法，反而更恰當。二分法的統計決策有時亦不是最佳決策或結論，增加一個「尚未定論」決策可能與事實更貼切，以利更佳對立假設的考驗。

另外，研究者亦應同時報告精確的 p 值、信賴區間，或應用筆者所主張的兩套關卡進行統計的顯著性考驗，似乎更能提供更多的評估資訊，對於研究結果的解釋更具周延性與應用性。當研究者希望虛無假設為真時的「Accept-Support」情境時，生物學上常用的「區間虛無假設考驗」似乎比「點虛無假設考驗」來得貼切。

最後，筆者呼籲應用統計學的教科書或統計課程，似乎有必要將上述之 NHST 的演變、迷思、效果值分析、及樣本規劃納入基本教材中，以利研究者對於研究樣本、統計分析與研究結論作出更正確之應用與解釋，方能確保量化研究的品質。

參考文獻

中文部分

朱國聖（2004）。**國民小學教師實施資訊科技融入教學之評估與其應用障礙程度之研究**。國立嘉義大學國民教育研究所碩士論文。

李茂能（1998）。統計顯著性考驗的再省思。**教育研究資訊，6，3，**103–115。

李茂能（2002）。量化研究的品管：統計考驗力與效果值分析。**國民教育研究**

學報，8，1-24。

柳敦仁（2006）。**雲嘉南五縣市國小初任校長行政表現與學校效能關係之研究。**國立嘉義大學國民教育研究所碩士論文。

翁秉仁（2000）。Fisher, Ronald Aylmer。Retrieved June 1, 2006 from the World Wide Web: http://episte.math.ntu.edu.tw/people/p_fisher

英文部分

Carstensen, B. (2002). Post-hoc power: Don't do it. Retrieved May 1, 2006 from the World Wide Web:http://www.biostat.ku.dk/~bxc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155-159.

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

Denis, D. (2001). Inferring the Alternative Hypothesis: Risky Business. *Theory & Science, 2,* 1. Retrieved Oct 4, 2006 from the World Wide Web: http://theoryandscience.icaap.org/content/vol002.001/03denis.html

Faul, F., Erdfelder, E., Lang, A-G., & Buchner, A. (2008). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences(in press). Behavior Research Methods.

Gigerenzer, G. (2004). Mindless statistics, *The Journal of Socio-Economics, 33,* 587-606.

Gill, J. (2007). How do we do hypothesis testing? Retrieved July 4, 2007 from the World Wide Web: http://artsci.wustl.edu/~jgill/papers/hypos.pdf

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71(1),* 83-92.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence(p's) versus errors (alphas) in classical statistical testing. *American Statistician, 57,* 171-178.

Hubbard, R., & Armstrong, J. S. (2005). Why we don't really know what statistical significance means: *A major educational failure.* Retrieved Dec 12, 2009 from the World Wide Web: http://marketing.wharton.upenn.edu/ideas/pdf/Armstrong/StatisticalSignificance.pdf.

Hubbard, R., & Armstrong, J. S. (2006). Why we don't really know what statistical significance means: Implications for educators. *Journal of Marketing Education, 28, 2*, 114-120.

Kline, R. B. (2004). Beyond sign*ificance testing: Reforming data analysis methods in behavioral research*. Washington, DC：American Psychological Association.

Lenth, R. (2001). Some practical guidelines for effective sample-size determination, *The American Statistician, 55*, 187–193.

Linacre, J. M. (2010). Which hypothesis is the null hypothesis? *Rasch Measurement Transactions, 23(4)*, 242.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park: Sage.

McLean, A. (2001). On the nature and role of hypothesis tests. Retrieved July 1, 2006 from the World Wide Web: http://Alan.mclean@buseco.monash.edu.au.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5(2),* 15-22.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in*

the twentieth century. New York: Freeman & Company.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*, 163-182.

Sterne, J. A. C. (2002). Teaching hypothesis tests-time for significant change? *Statistics in Medicine, 21*, 985-994.

Thompson, B. (2001). 402 Citations questioning the indiscriminate use of null hypothesis significance tests in observational studies. Retrieved July 1, 2006 from the World Wide Web: http://www.warnercnr.colostate.edu/~anderson/thompson1.html

Yu, C. H. (2006). Don't believe in the null hypothesis? Retrieved Dec. 1, 2006 from the World Wide Web: http://seamonkey.ed.asu.edu/~alex/computer/sas/hypothesis.html