實用實證醫學，如何判讀統計檢定 —p 值的迷思
Statistical Hypothesis Test – Riddle of p Value

邵文逸

財團法人醫藥品查驗中心

前　　言

統計檢定在當前的臨床研究中被應用的非常普遍，檢定的結果也經常被視爲該研究是否有價值的一個決定性因素。多數的讀者們習慣上以檢定 p 值小於 0.05 來選擇該研究是否有參考價值，0.05 好像一個魔術數字一樣，掌握了一個研究的生死。雖然很多臨床研究跟統計學的專家不斷的呼籲大家不要只專注在 p 值的大小，但是因爲閱讀它很單純，因此在溝通研究結果的時候顯得非常方便 — 小於 0.05 顯著，大於 0.05 不顯著。不可否認的，在呈現研究結果時，p 值還是受到相當高程度的重視。因此如果能夠對 p 值所代表的意義有更深入的了解，以及知道判讀 p 值的時候需要注意些什麼，相信可以有效的提升大家掌握研究價值的能力。這裡我們盡量避免使用統計的數理原理，對於臨床研究設計跟統計原理介紹請大家參考之前的兩篇簡介[1,2]。

首先請思考下面兩個問題，

- p 值越小表示該研究的結果越好嗎？
- 樣本數越大，研究的 p 值一定越小嗎？

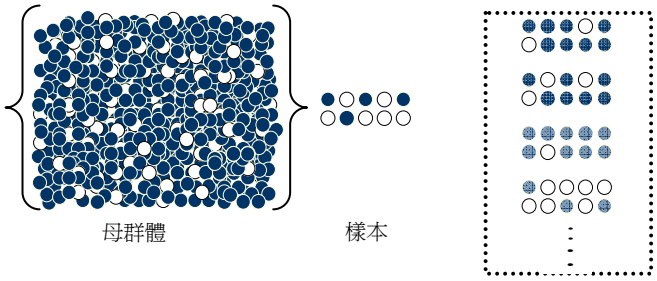
以上兩個問題的答案都需視情況而定，兩個問題都不是單純的“對”或“不對”。

統計檢定和 p 值

要了解 p 值，首先需知道研究跟統計分析的目的不僅僅在於分析收集到的那些有限的病人資料，而是利用手上有限的資料來了解自然界到底發生了什麼。

舉例來說，一個新的藥品 A 被發現可能對染上疾病 X 的病人有治療的效果。理想上所謂 A 對 X 病人的療效，就是把世界上每個角落符合疾病 X 診斷的每一位病人通通找出來，全部給予 A 治療，看看全部的病人裡面經過治療之後有多少病人復原了。但是這樣的作法並不切實際，實務上只能從研究收集得到的病人做起。因此一個重要的假設是，研究收集病人的過程是客觀的，研究所收集的病人跟那些沒有納入研究的病人，對於治療 A 有相同的表現。基於這樣的前提，研究所收集到的資料可以代表我們想要研究的“全部病人”。這個全部病人統計上稱爲母群體 (population)，而研究所收集到的病人就稱作樣本 (sample)。因爲每位病人都有些自己獨特的條件，例如遺傳、生活方式、之前的治療、對治療的偏好、順從性等等，以致每位病人被治療後的表現也多少會不同。雖然目前科技還沒有辦法透過掌握每位病人的所有特質來完美的預測每位病人治療後的結果，我們可以了解母群體中每位病人接受治療後的都會有自己的結果，因此母群體中這個治療的“真正效果”是存在的。我們用下面的簡圖來表示：○ 代表治療後有效的病人，● 代

表治療沒有效的病人。當前的研究就是從左邊的母群體中，抽選出 10 位病人，給予治療後觀察發現有 6 位病人有效。然後根據這樣的結果推論，這個治療 A 對疾病 X 的病人，治療效果是 60%。



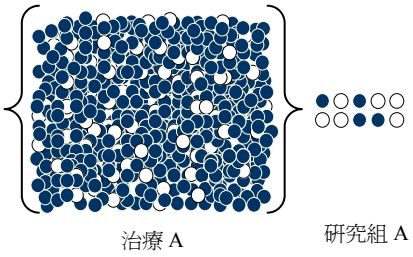
當我們做出“60%治療有效”的結論時，其實是對母群體病人接受治療的“真正效果”作出推論 (infer)。雖然推論的主要依據是研究所觀察的那 10 位病人，從上面的簡圖我們可以了解，這樣的研究結果 (10 位接受研究的病人當中 6 位治療有效)，並不是唯一可能的結果。其他還可能出現右邊虛線方塊裡的各種情況。因此即使是同樣的研究題目、類似的研究設計、也在相同的研究的母群體中進行選樣研究，結果出現稍有差異是很正常的。當接受研究的十位病人中只有一位治療有效時，我們就會推論該治療對這樣的病人群的治療效果爲 10%。上述這些研究結果的不同，稱爲抽樣誤差 (sampling error)，抽樣誤差純粹是因爲抽樣的偶然性造成，關於抽樣誤差的說明請參考[2]。

真正效果沒有差異，研究結果誤判顯著差別

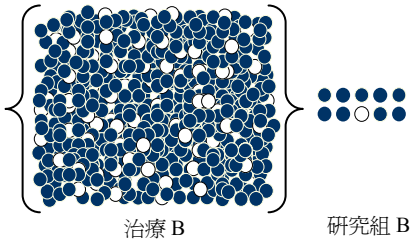
讓我們考慮下面的情況：治療 A 跟治療 B 的“真正治療效果”在母群體裡是相同的。但因爲抽樣的隨機性，研究結果出現 A 治療組 6 人有效，而 B 治療組只有 1 人有效，所以結果是 A 治療 60% 有效，比上 B 治療 10%有效。根據這樣的結果 — 60%比 10%，是否可以推論 A 治療優於 B 治療呢？當效果指的是一年存活率時，這樣的結果看起來頗具臨床意義。不過不要忘了，這樣的結果僅僅是因爲“隨機”的抽樣誤差偶然出現。像這樣來自原本應該沒有差異的群體，研究結果看起來卻有差別的機會就是統計檢定的 p 值。

使用分析這類資料最常用的費雪確切檢定 (Fisher's exact test) 對上述這組研究資料進行統計檢定，結果顯示 p 值爲 0.029，小於 0.05 達到統計顯著差別，根據這個檢定結果，判斷 A 治療跟 B 治療效果的差別達到統計顯著，因此大多數人就會認定“A 治療比 B 治療好”。

看到這裡，我們發現這個結論顯然違背事實 — A 治療跟 B 治療在母群體裡面的“真正治療效果”是一樣的。正確解釋費雪確切檢定 p 值爲 0.029 的意思是：從這樣兩個相同的群體裡，抽選上述樣本 (每組各十位病人)，結果出現一組只有 1 人有效而另一組卻多達 6 人有效，有效人數差別高達至少 5 人的機率 (加上更極端 0 比 7 結果的機率)。這樣的機會雖然不高，但確實有可能發生。



	A	B
有效	6	1
無效	4	9



p 值小，表示該樣結果出現的頻率不高。0.029 表示大約每 34 次相同的研究，平均會出現一次這種結果。這樣的頻率是根據母群體沒有差別的前提計算出來的，而這樣的前提就是一般所稱的“虛無假說”(null hypothesis)。當 p 值太小的時候，表示在這樣的前提下，看到這種結果的可能性不高，因此這個前提有被重新考慮的必要，這就是通常所說的“推翻虛無假說” — 研究者宣稱該前提不是真的，因此“研究結果達到統計顯著”。當研究者根

據其所研究的 p 值推翻虛無假說這個前提，但這個前提其實是真確的，那研究者就犯了判斷上的錯誤。從以上推論的過程我們可以知道，犯這種錯誤的機會其實就是這個研究的 p 值。因此如果研究結果 p 值爲 0.01，依據這個結果宣稱研究達到顯著差異，“推翻虛無假說”，那麼犯錯的機會就是 0.01。

現在大家習慣採用 0.05 當作一個臨界，當研究的 p 值小於這個臨界值的時候就宣稱研究結果達到統計顯著，也就是說大家普遍同意接受 5%犯錯的可能性。這個願意接受的犯錯可能性稱爲“顯著水準”(significant level) — 超過這個水準就宣稱研究達到統計顯著。其實顯著水準是可以調整的，如果研究者願意接受較高的犯錯可能，那顯著水準可以提高到 0.1；當然也可以採取比較嚴格的標準，例如 0.01，會降低犯錯的可能，但同時也降低了研究結果達到統計顯著的機會。但是不管採用怎樣的數值，重要的是這個水準必須在研究收集資料之前就決定好，不可以等到資料收集完畢，計算出 p 值之後才看情況決定。

因此當母群體真正效果沒有差別時，每進行 100 次相同的研究，大約有 5 個研究會因爲抽樣偶然的機會，出現 p 值小於 0.05 的結果而被誤判爲有統計顯著差異。這種誤判的機會適用在所有的統計檢定，包括不同的統計方法、各種研究設計、也適用在各種樣本數。也就是說，當母群體的真正治療效果沒有差別時，不管研究選用 10 名樣本、還是選用 100,000 名樣本，統計檢定結果因抽樣誤差偶然出現 p 值小於 0.05 的機會都維持在 0.05。研究將沒有真正差別的效果誤判爲統計顯著的機會並不因樣本數或是研究設計的不同而改變，這也就是爲什麼幾乎所有的研究均普遍採用 0.05 當作顯著水準。在母群體沒有真正差別的前提下，只要是運用正確的統計方法進行檢定，p 值不會因爲樣本數的增加、或研究設計的不同而變小或變大。但是如果母群體的真正治療效果有差別，情況就不一樣了。

真正效果有差異，研究正確判斷顯著差別

現在我們考慮另一種可能性：如果 A 治療在病人母群體中真正有效是 60%，而 B 治療真正有效是 10%，那麼研究結果出現 A 組 6 位病人有效，且 B 組出現 1 位病人有效的機會是不是大多了呢？確實是這樣，當研究病人來自這樣兩群真正治療效果不一樣的母群體時，上述研究結果出現兩組有效人數相差達到至少 5 人的機會將高達 0.62。這時根據統計檢定 p 值爲 0.029，判定差異達統計顯著，進而推論兩個治療的真正效果不同，這時研究者就做出了一個正確的推論了。

在母群體有差別的情況下，研究結果可以正確得到統計顯著的機會，稱爲統計的“檢力”(power)。檢力是正確推翻虛無假說的機率。研究設計的目的之一是希望在可行的範圍內，盡量提高檢力，具體的策略就是設計合理的樣本數來達成。

真正效果有差異，研究結果誤判沒有顯著差別

兩個真正有差異的母群體也可能因爲抽樣的機會，產生兩組有效人數差異不多的研究結果，例如下面表一中的三個研究結果。

表一

		有效人數 / 樣本數			
	真正效果	第一次研究	第二次研究	第三次研究	第四次研究
A 治療	60%	4 / 10	6 / 10	5 / 10	…
B 治療	10%	2 / 10	3 / 10	1 / 10	…
差異	50%	20%	30%	40%	
p 值		0.314	0.185	0.070	

這三個研究雖然都是由相同的母群體中選

出，但只因爲多了或是少了一兩位病人治療有效，造成研究結果治療有效病人數在兩個研究之間的差異稍微看起來比較少了，也因此統計檢定出現了不顯著的結果。

當研究的樣本數少的時候，一兩位病人結果的差異，很容易使統計結果出現幅度較大的差別。例如，A 治療真正效果是 60%，雖然 10 位病人接受治療最可能的結果是 6 位有效，但只要少了一位有效，結果就會變成 10 位當中 5 位有效，研究結果變成 50%，跟母群體的真正效果就差了 10%。但是，如果研究選取 100 位病人，雖然同樣會有因爲機會的關係多或少一兩位病人有效，造成的結果差別卻僅有 1%或 2%，幅度比樣本數只有 10 位的時候少多了。樣本數越大，研究結果變動的幅度隨之縮小，研究結果也越趨近母群體的真正效果。根據統計理論，使用每組 100 位病人來研究 A 治療 (60%有效) 跟 B 治療 (10%有效) 的差別時，幾乎百分之百會得到統計顯著差異的結果。

微小的差異

並不是所有被研究的效果都像上面的例子一樣大，讓我們看看以下的兩個治療：對於疾病 Y 的病人母群體，標準治療 C 真正的效果是一年存活 20%，新的治療方法 D 的真正存活率是 21%。即使是存活率，從 20%變成 21%可能大多數人都不會覺得是很重要的改變。但是如果每組各選 10,000 名病人進行比較研究，得到結果如下面的表二。

統計檢定 p 值爲 0.041 達到統計顯著，推論兩個治療效果不一樣。這個推論其實是正確的，因爲兩個治療在母群體中的效果是真的不一樣 (20%比上 21%)，只不過這個差別小到幾乎可以忽略，可能不會影響臨床上處理病人的決定。因此 p 值小，達到統計顯著差異，並不一定表示研究的結果好，也不一定等於被研究的治療比較優秀。

一個“好研究”應該是針對重要問題、設計適當且執行品質優良的研究。“好的研究結果”有助於回答問題，提供肯定的資訊，降低原來問題中不確定的部份。

肯定的資訊可以是治療效果有不同 (正向結果)，也可能是肯定各種治療的效果差別不大 (負向結果)。例如：原本大家對於新治療 D 是否比標準治療 C 好沒有確定的資訊可以參考，因爲系統性回顧整理當時所有的臨床試驗研究結果如表三的前四個研究，雖然新治療 D 的治療效果似乎比原治療 C 好一些，但所有研究的統計檢定都沒有達到顯著差異。也就是說在兩種治療沒有真正差別的前題下，只是因爲抽樣而隨機出現這些研究結果的可能性很高。因此大家對於是否該使用新治療 D 可能有較大的爭議，因爲根據所累積的資訊，D 跟 C 比較，效果從較差到較好都有，因此爭議較大。

表三最下方的“最新研究”結果統計檢定發現兩種治療效果的差別達到統計顯著。因此對之前不確定的問題提供了肯定的資訊。但是這個例子的結果是兩種治療的結果差不多。在沒有這個研究結果之前，大家對於治療效果的認定差別可能比較大。而最新的研究結果，提供了一組比較肯定的資訊 — 兩個治療的效果分別是 20%跟 21%，就統計檢定的角度，這個研究結果的差別是顯著的；但是從臨床的角度，只差 1%的效果，可能不足以令大家從原本的治療 C 更換成新治療 D，因此這個結果沒有達到臨床上的顯著意義，但是肯定了兩種治療差異的程度很微小。

表二	真正效果	研究結果：有效人數 / 樣本數	
C 治療	20%	2,000 / 10,000	
D 治療	21%	2,100 / 10,000	p 值 0.041
差異	1%	1%	

表三	原治療 C			新治療 D			p 值
	有效	樣本數	%	有效	樣本數	%	
第一個研究	3	13	23%	4	15	27%	0.588
第二個研究	3	15	20%	5	28	18%	0.583
第三個研究	25	110	23%	27	115	23%	0.510
第四個研究	61	321	19%	77	310	25%	0.083
最新研究	2,000	10,000	20%	2,100	10,000	21%	0.041

所以研究的統計顯著跟研究結果必須分開兩個階段來看，統計檢定評估可能的不確定性，排除隨機的不確定性可能造成的影響後，再來看結果在臨床的使用上有什麼樣的意義。統計 p 值顯著，不必然表示該研究結果在臨床上對病人的處理有重要的影響，也不一定表示受到研究的治療效果一定是比較好。

回到最初的兩個問題

- p 值越小表示該研究的結果越好嗎？
  p 值越小，虛無假設爲真實的可能性越低，推翻虛無假設可能犯錯的機會越低。因此宣稱研究有統計顯著差異的肯定程度越高。

  但是p值跟研究所要探討的臨床效果是否重要之間不是絕對的關係。一個 p 值顯著的結果，臨床上的效果卻可能很微小。像前面所舉的例子，20%比 21%臨床上只是很微小的差別，但是只要研究樣本數夠大，統計檢定就會達到顯著。

  因此看到 p 值小，還需要分別從統計推論跟臨床意義兩個角度來判讀。

- 樣本數越大，研究的 p 值一定越小嗎？
  p 值是根據母體真正效果沒有差異的“虛無假說”來計算，所以：

  如果**母群體真正效果沒有差異**，研究結果的 p 值將隨機出現在 0 到 1 之間，跟樣本數大小沒有關係。

  可是只要**母群體的真正效果有差異**，研究結果的p值就會隨著研究樣本數增加而越來越小。

請看以下比較 A 治療跟 B 治療的例子：
我們做系統性回顧發現 100 個小型的研究，這些研究每組 10 只用了名病人；另外發現了 100 個大型的研究，這些研究每組收集了 300 名病人。這些研究的 p 值會呈現怎樣的結果呢？

如果 A 治療跟 B 治療的真正效果沒有差別：

不管是每組 10 人還是每組 300 人的研究，我們將發現，每 100 個研究當中大約都有 5 個研究的統計檢定 p 值會小於 0.05。這些研究的 p 值相當平均的分布在 1 跟 0 之間。

如果 A 治療跟 B 治療的真正效果有差別：

我們將發現，每 100 個研究中，統計檢定結果 p 值小於 0.05 的研究將多於 5 個；且每組 300 人的那些研究中，有更多的研究 p 值小於 0.05。所有研究結果的 p 值普遍分布在較小 (靠近 0) 的那端。

當幾個相同題目的類似研究，出現不一致的結果，且沒有特別的理由可以解釋彼此之間的差別時，那麼這些差異就有可能是抽樣誤差所造成。統合分析 (meta analysis) 透過整合這些研究研究降低抽樣誤差。一個研究有可能因爲偶然的機會，結果出現比較小的 p 值，因而造成我們對研究的誤判。可是當好幾個題目相同的類似研究都出現比較小的 p 值時，結果純粹只是因爲抽樣隨機出現的可能性就降低了，因此我們對結果判讀的肯定程度就隨著升高。這也就是爲什麼一系列結果一致的研究，所提供證據的強度較高的原因。

參考資料

[1] 邵文逸. 如何嚴謹地研判醫學文獻.台灣醫學 2003 年 7 卷 4 期 535-542.

[2] 邵文逸. 實証醫學中的統計原理.台灣醫學 2005 年 9 卷 4 期 531-535.