實証醫學中的統計原理

Statistics Principle Used in Evidence Based Medicine

邵文逸

台大醫學院臨床醫學研究所

實證 (evidence) 的基本精神在於採用符合科學準則的方法所蒐集的資料來做決策，而資料需要經過有系統的整理變成資訊之後，再經過合理的推論程序才能做出正確的決策。統計及研究方法在上述過程中，從資料收集、分析整理、到推論都扮演重要的角色。

統計的主要目的大致可以分兩種，描述與推論。統計學 statistics 原本的意思是收集與國家有關的資料，例如人口數、生育數、農業生產、商業活動等，這種調查的資料用來做爲國家施政與資源配置的依據，基本作法就是將各地區收集的資料彙整、分類、加總統計，以描述各地區的狀況，例如，平均家庭月收入北區 70000 元、南區 65000 元，這些結果就會被用來作爲規劃稅收的依據。但是如果調查結果不客觀，例如真正的收入其實只有 45000 元，那就會造成稅收短缺，影響預算使用。因此以描述爲主要目的統計調查必須確定所收集資料的來源是客觀的，也就是說選取的樣本需要具有代表性，所做的測量需要公正準確，得到的結果才不會有偏差(bias)，這是完成一個好的調查研究的首要條件。

怎樣才能收集到有代表性的樣本呢？一個基本原則就是調查範圍內的所有對象都必須有被收集到的機會。例如根據戶籍資料調查平均收入時，那些無業無居所的街友將不會或是不容易被調查到，以至於高估平均收入；又例如一般人怕稅徵機關課重稅，而傾向低報所得，結果將造成平均值的低估。像這樣，前者稱爲選樣偏差(selection bias)，後者稱爲測量偏差(measurement bias)，是研究首先必須積極避免的。最理想的選樣是調查範圍內的所有對象都有相同被抽到的機會，這樣的抽樣稱爲隨機樣本(random sample)。

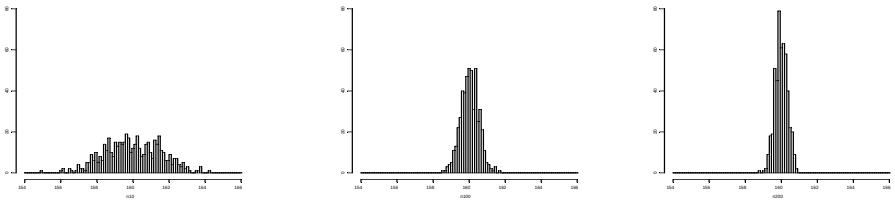
但是即使排除了偏差，一組好的隨機樣本還是會因爲資料收集過程的不確定性，使研究結果出現隨機誤差(random error)，隨機誤差主要受到兩個因素影響，第一是樣本數；第二是受調查樣本之間變異大小。樣本數越大誤差越小；受調查樣本間的變異越少，研究誤差也會變小。

下面的例子說明隨機樣本所產生的誤差。某研究利用隨機抽樣調查社區民眾的平均身高，研究者從一個身高平均值爲 160cm 的社區中隨機選取 10 名樣本，計算其平均身高，這個平均值應該在 160cm 附近，但是剛好等於 160cm 的可能性應該不高。現在有許多位研究者都做相同的隨機抽樣調查，然後個別計算他們自己研究樣本的平均身高，這些研究結果將呈現如下的情況：大部份接近 160cm，但有些較高有些較低。雖然個別研究的結果不一樣，但是將所有人的研究結果再作平均則會得到 160cm，這個 160cm 就是這些研究的預期值。當預期值與所要調查的社區民眾身高相同時，這些研究就是優良無偏差的研究。雖然一群研究可以達到無偏差的理想，但是各別研究與所要調查的真正目標卻可能有差異，這種差異的大小就是抽樣誤差。統計透過計算所有研究結果的標準差來描述研究之間變異的大小。在這裡樣本數爲 10 的研究結果，其抽樣誤差大小爲 1.54 ㎝。現在讓我們試試看增加樣本數爲 100 以及 200 分別進行抽樣並計算各組平均值，$\overline{X}_{100}$, $\overline{X}_{200}$,，並將相同抽樣研究重複 500 次，將這 500 個平均值作成直方圖，觀察其分布情況，並計算這些平均值的(再)平均以及標準差，結果如下(表一)：

這些樣本平均值都集中在 160cm 附近，再平

均的結果顯示三種樣本數的抽樣結果都是無偏差的，但是隨著樣本數增加，結果越來越集中，分散程度變小，標準誤差變小。而且這三組抽樣結果的分佈型態都是對稱近似鐘型的常態分布。這個現象就是統計上一個非常重要的「中央極限定理」，同時，這個定理也告訴我們抽樣誤差的大小就是，樣本族群的標準差除以樣本數開根號。

表一：樣本數與抽樣分布

樣本數	10	100	200
$\bar{X}$ 平均	160.0	160.0	160.0
$\bar{X}$ 標準差(誤)	1.54	0.50	0.34
抽樣分布			

用樣本資料推估族群的資訊

用上面的例子來說明甚麼是抽樣誤差，我們可以知道從一個已知的群體中選樣時，各種抽樣結果的變異情況。但是現實中實際進行研究時我們是由選出的有限樣本來推論族群的性質。例如，當我們看到研究結果樣本平均身高爲 159.2cm，此時到底社區整體民眾的身高平均爲多少？其實後者才是我們真正關心的答案，抽樣研究只是幫我們找出真正答案的線索，就像法庭上法官根據證據進行審判，所有證據背後的真相才是重點。統計理論讓我們有效的透過適當的研究方法與選樣來推論群體的性質。下面讓我們用一個新藥開發的例子來看看實際上研究資料的分析如何影響決策。

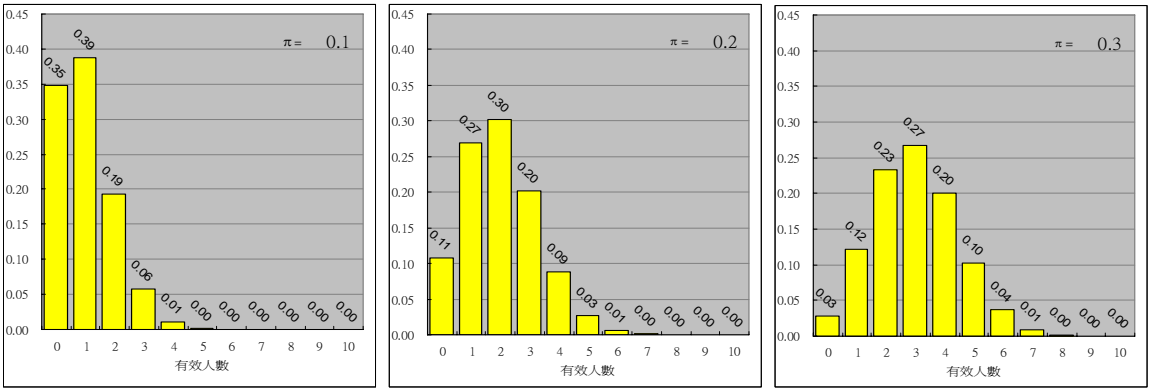
某實驗室的基礎研究結果發現，一項新的藥物 A 可能對於疾病 Z 的治療有效。經長期的努力後，現在研究終於準備進入臨床試驗階段，可以在病人身上試試看這個藥品的治療效果是否好到值得繼續投資開發。根據現有臨床資料，當前醫療對於疾病 Z 治療一年的存活率約爲 10%，所以如果 A 治療 Z 一年的存活率可以達到 20%，那就值得進一步開發，否則研究就中止。現在研究主持人決定先選 10 名病人進行小規模的試驗看看這個藥品的效果，如果結果發現 10 位病人當中有 2 位或超過 2 位病人存活超過一年，表示此藥品治療 Z 的一年存活率至少達到 20%，那就決定繼續投資開發。相反的如果只有一位或完全沒有病人存活超過一年，那這個藥品的開發就要中止。

就像之前社區抽樣會有抽樣的隨機誤差一樣，即使這個藥品治療效果一年存活率真的是 20%，在 10 名選出的研究樣本中，確實有可能連一位存活超過一年都沒有，根據統計理論這樣的結果出現的機會高達 0.11，如果加上只出現一位存活超過一年的結果出現的機會爲 0.27，這樣一個應該有效的藥品被放棄的機會將高達 0.38！超過 1/3 的機會研究者會損失這項有效的產品。

可是如果這個藥品治療一年真正存活率只有 10%，研究結果仍有機會出現超過兩位存活超過一年，也可能是 2、3、、、10 位，這些結果將導致藥品繼續開發，這樣的機會有多高呢？高達 0.26，也就是說有超過 1/4 的機會研究者將會浪費錢在一項沒有效的產品上。像以上這樣發生錯誤決策的機會其實高過一般人所想像的，該怎辦才能得到比較好的結果呢？要降低決策錯誤的機會可以透過增加樣本數來解決。如果研究的樣本數設計爲 100 名病人，當有效人數未達 15 位則決定不再開發，那麼真正效果爲 20%的藥品有 0.92 的機會可以繼續開發。相反的，真正效果爲 10%的藥品則有 0.93 的機會被終止開發，比前面只用 10 名病人的研究結果好多了。

透過統計理論的推算，我們可以知道各種可能

的研究結果出現的機會，下面(圖一)分別計算了從疾病的了解做比對，每項疾病診斷都有一群相對應



圖一：不同治療效果族群的結果分布

三種不同的真正治療效果族群中，每次抽選 10 名病人進行研究，各種有效人數出現的機率。例如從真正有效爲 0.3 的族群中選樣時(圖一最右圖)，結果出現只有 0 或 1 個病人有效的機率爲 0.03+0.12=0.15。因此既使效果高達 30%，研究結果還是可能導致停止繼續開發的錯誤的決策。

現在讓我們換一個角度來想想看以下的情況，在 10 名接受研究藥品 A 治療的病患中，結果有兩位存活超過一年，A 治療該疾病的一年存活率是多少呢？一般人直接的反應是 2/10=20%，但是有沒有可能真正的效果其實是 10%或是 30%，只是剛好因爲抽樣運氣的關係，碰巧出現兩名存活超過一年的病人？我們怎麼能夠確定到底真正的效果是 10%、20%、還是 30%呢？我們再一次用上面的結果分布圖來分析，讓我們看看，如果真正效果是 10%(圖一最左圖)，則結果出現剛好兩個有效的機率是 0.19；如果真正效果是 20%(圖一中央圖)，出現兩個有效的機率是 0.30；效果是 30%(圖一最右圖)時出現兩個有效的機率則是 0.23。從這些機率的比較我們可以發現，真正效果是 20%時，最可能出現我們目前的研究結果，因此根據目前觀察到的資料，真正效果爲 20%應該是最合理的選擇。以上的過程在統計分析上稱爲估計，估計就是根據所收集的資料推論群體真正的參數。這樣的過程跟臨床診斷其實是一樣的，我們幫病人問診、做檢查就像研究收集資料，將所收集的資料整理後與我們對的症狀表現，將這一群一群的表現與現在從該病人身上所收集到的資料做比對，看目前的資料比較符合哪一群表現，就將該病人歸納到該診斷，這就像我們之前選一個估計結果一樣。

臨床上經常會遇到資料不足，以致診斷模糊的情形。統計上也一樣會因爲資料不充足導致我們對該估計缺乏信心。讓我們再想想看前面的估計結果，如果比較三種估計的相對可能性，用目前研究資料出現的機率做爲指標，三種估計的可能性分別是 0.19、0.30、0.23，因此 20%比 10%更可能的程度只高了 1.6 倍，比 30%可能的程度也才多出 1.3 倍而已，如果我們考慮更多的可能，例如 21%、19%、我們將發現對原來估計的 20%越來越沒有把握了，因爲還有很多其他可能的答案。雖然如此，但是大多數人應該不會根據 10 位病人當中兩名有效來推論該藥品的真正效果可以高達 90%，除非有其他的原因讓我們強烈的認定該結果。從統計的理論來分析，真正效果是 90%時研究 10 位病人當中兩名有效的這種結果出現的機率只有 0.00000036 這麼低，相當於 274 萬次研究才會碰巧出現一次，實在是太難太不可能了，我們通常傾向將這種估計排除，但是我們要知道，雖然機率很低，理論上來說還是在合理可能出現的範圍裡，如果真的是這樣但是我們卻判斷爲不是，那在做這個就判斷時就犯了錯誤，所以理論上犯這種錯誤的機率就是 0.00000036，這就是統計上說的顯著 p 值，p 值代

表的意思就是根據目前這組研究資料，拒絕相信真正效果是 90%，這個決策可能犯的錯誤。這個決策分析的過程就是統計檢定。

兩組效果的差異是 0%，這個 0%就是統計檢定裡面所稱的無差異假設，英文名稱爲 Null hypothesis，也有人翻譯爲“虛無假設”。讓我們考

表二：兩個效果相同的族群其研究結果出現的分布機率

有效人數 A (0.2)	X (0.2) 出現機率	0	1	**2**	3	4	5	6	7	8	9	10
		0.107	0.268	0.302	0.201	0.088	0.026	0.006	0.001	0.000	0.000	0.000
0	0.107	0.012	0.029	0.032	0.022	0.009	0.003	0.001	0.000	0.000	0.000	0.000
1	0.268	0.029	0.072	0.081	0.054	0.024	0.007	0.001	0.000	0.000	0.000	0.000
2	0.302	0.032	0.081	**0.091**	0.061	0.027	0.008	0.002	0.000	0.000	0.000	0.000
3	0.201	0.022	0.054	0.061	0.041	0.018	0.005	0.001	0.000	0.000	0.000	0.000
4	0.088	0.009	0.024	0.027	0.018	0.008	0.002	0.000	0.000	0.000	0.000	0.000
5	0.026	0.003	0.007	0.008	0.005	0.002	0.001	0.000	0.000	0.000	0.000	0.000
6	0.006	0.001	0.001	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

表三：兩個效果不同的族群其研究結果出現的分布機率

有效人數 A (0.2)	X (0.2) 出現機率	0	**1**	**2**	3	4	5	6	7	8	9	10
		0.349	0.387	0.194	0.057	0.011	0.001	0.000	0.000	0.000	0.000	0.000
0	0.028	0.010	0.011	0.005	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.121	0.042	0.047	0.023	0.007	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	0.233	0.081	0.090	0.045	0.013	0.003	0.000	0.000	0.000	0.000	0.000	0.000
3	0.267	0.093	**0.103**	**0.052**	0.015	0.003	0.000	0.000	0.000	0.000	0.000	0.000
4	0.200	0.070	0.078	0.039	0.011	0.002	0.000	0.000	0.000	0.000	0.000	0.000
5	0.103	0.036	0.040	0.020	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000
6	0.037	0.013	0.014	0.007	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.009	0.003	0.003	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

統計檢定

一般常見的臨床試驗通常研究兩種治療方式的差異，例如使用新藥組，A，與使用安慰劑組，X，比較，在這樣設計下，通常統計檢定先設定兩種治療的結果是相同的，例如兩組都是 20%，所以

慮以下的例子，從兩個效果都是 20%的群體中各選 10 名樣本，結果會怎樣呢？可能兩組有效的人數一樣，也可能不一樣，我們如何根據這樣的研究結果下結論呢？如果結果 A 出現 3 個有效，而 X 出現 1 個有效，A 是否真的比 X 好呢？如果我們認定 A:X=3:1 是好的結果，哪麼所有 A>3 且 X<1 的結果也應該都是好的，在這樣的判定標準下，其實

A 與 X 本質上其實沒有差異，但是碰巧出現上述情況的機會有多高呢？這就是統計檢定考慮的問題。根據統計的理論，上述結果出現的可能性爲 0.12，差不多每 9 次研究就會因爲抽樣的隨機性而恰巧出現一次這樣的結果，讓我們造成誤判的結果。如果我們判斷 A 有效人數多過 X 兩個或以上就是好的結果的話，那麼類似 A=4, X=2 以及 A=5, X=3 的情況都會讓我們造成誤判，把這些可能性都加起來的話，誤判的機會將高達 0.2。而且也有可能出現 X 有效人數多於 A 的情況，我們將 X 誤判爲優於 A 的可能性也一樣高達 0.2，扣除這兩種誤判，正確作出判斷 A 與 X 沒有差別的機會剩下 0.6，比丟銅版隨便決定的結果並沒有好太多。表二是所有研究可能出現結果的機率表格。

灰色的部分就是作出 A 與 X 沒有差別決策的部分，表二左下方三角區域會讓我們得到 A 優於 X 的決策，而表二右上方三角部分會得到 X 優於 A。

讓我們再考慮另一個情況，表三是 A 真正有效的比率爲 30%而 X 只有 10%，所以 A 是真的優於 X。但是同樣因爲抽樣的隨機性，我們得到 A 至少比 X 多 2 個有效的機率有多大呢? 大概是 0.61，也不比丟銅板決定好太多，而這個得到正確決策結果的機率就是統計的檢力，power。此外，得到 A 與 X 差不多的機率大約爲 0.37，更嚴重的是有大概 0.02 的機率會得到 X 比 A 好，完全相反的決策。

但是如果我們每一組各選擇 100 名樣本做研究，在真正效果分別爲 30%對 10%的情況下，可以正確得到 A 優於 X 決策的機率就可以高達 0.975。爲什麼會有這麼大的差異？原因跟之前說明的道理一樣，當樣本變大時，研究的隨機誤差相對變小，結果趨向集中在族群真正效果的附近，因此大部分的結果將正確的反應族群的真正差異。

以上介紹的是運用統計原理在臨床研究上作決策的基本原理，其他因應不同的資料型態還有不同的計算方法，但是基本上都是遵循同樣的原理。正確使用統計方法讓我們在作決策時站在客觀的立場，全面的考慮完整的情況，以避免作出有侷限的決策，也因此幫我們有效的運用現有的證據，達到實證醫學的目的。

推薦讀物

1. Robert H., Fletcher, Suzanne W., Fletcher, Edward H., Wagner: extended reading: Clinical Epidemiology: the Essentials, 3rd ed, Willims & Wilkins, 1996.