流行病學暨生物統計專欄(卌)

資料分析的步驟

林明瀅

台北榮民總醫院感染管制委員會

前　言

大家都知道要對一堆未經整理的原始資料加以解釋是很困難的，甚至是不可能的，以往的各期向各位介紹了許多統計方法的運用，但它們只是資料分析的一部份，分析工作就是要了解目前資料的特徵，給予說明並加以推演，而運用統計方法較能客觀的分析出這些特徵；『分析』是指對所收集的資料加以整理、分類及摘要，分析的目的就是要簡化資料，使資料變成易於理解與解釋的形式，不過我們要知道資料分析的本身，並不能對所要探討的問題提供直接的答案，因此我們必須對分析的結果加以解釋，才能使人了解其中的含意。『解釋』意指了解分析的結果，對所研究的變項關係作種種推論，進而從這些關係中導出結論，並進一步與既有的通則、定律或理論相連繫。

表一是 12 份院內感染基本資料，想要由這些資料知道 A 病房有多少人感染或有幾人得院內泌尿道感染時，至少要動一動我們的指頭算一下，如果將它整理分類成表二，就很容易回答上述的問題，讓人一眼就看出 A 病房有 7 人感染，泌尿道感染有 3 人。

資料處理

資料分析的工具可分為二種，一種是手工計算，此方法對我們院內感染的工作伙伴一定不陌生，一大堆繁鎖的加減乘除，使人覺得很乏味，若是幾次的計算結果不一致時，更是令人哭笑不得，而整理後的資料大致上是一些描述性的結果，如感染率的多寡、不同病房感染率之分佈、各種院內感染部位的分佈、院內感染菌種之分佈 .. 等，最多再加上卡方檢定的結果，就可讓我們人仰馬翻了！另一種方法是將資料輸入電腦，再借助不同電腦軟體的功能，達成資料之分類及許多複雜統計方法的計算甚至印出報表，但是電腦是一台機器，要它為我們工作就必須給它程式命令，因此在資料電腦化的初期需要規劃很多事物並且投注相當多的心力，首先要決定那些院內感染的資料必須輸入電腦，一般以月報表的製作而言，至少要將病人的基本資料輸入：(1) 病人的病歷號；(2) 資料卡的編號；(3) 年齡；(4) 性別；(5) 病房；(6) 科別；(7) 感染部位；(8) 感染菌株；(9) 感染日期等九項資料欄，及各科別、年齡、病房的住（出）院人數及住院人日數；除此之外，為符合羣突發發生時，可迅速整理出可用的資料，需考慮某些資料是否亦要存入電腦，例如病人的住院日期、住院診斷、手術步驟、是否死亡、不同的潛在性疾病、侵入性的醫療措施、

病人的實驗室檢查結果、臨床症狀、檢體種類、感染菌株的抗生素感受性試驗等。

第二步驟就要考慮資料分析的方式、那些資料必須借助譯碼的技術（coding）、及譯碼格式如何設定？以表一的基本資料而言，如果我們想了解有插導尿管的病人，一般是幾天後得到感染時，表一中『是否使用導尿管』的資料就無法得知，因為收集資料時，並未填寫，若改填入『感染日期之前導尿管使用天數』，或是『插導尿管的日期』，再由電腦扣去感染日期，就可計算結果，因此資料分

表一　院內感染個案基本資料表

病房	病歷號	年齡	科別	性別	感染日期	感染部位	使用導尿管	感染菌株	臨床症狀
B-017	214779	69	一般外科	M	19930726	泌尿道	是	K. pneumoniae	fever
B-005	201660	42	一般外科	M	19930722	外科傷口	是	S. aureus	
A-002	196911	57	腸胃科	M	19930712	血流	否	S. aureus	fever, PMN>10000
A-004	132027	64	神經內科	F	19930710	呼吸道	否	E. cloacae	cough
A-011	131023	56	神經內科	F	19930712	泌尿道	是	E. coli	
A-003	237193	62	過敏科	M	19930714	泌尿道	是	Candida spp	
A-040	251730	38	感染科	M	19930709	血流	是	P. aeruginosa	shock
C-040	151730	41	婦科	M	19930717	外科傷口	是	E. coli	
A-022	141523	65	神經內科	F	19930710	呼吸道	否	E. cloacae	
B-008	101660	46	泌尿外科	M	19930722	外科傷口	是	S. aureus	
B-003	136611	77	心臟外科	M	19930712	血流	否	P. aeruginosa	fever
A-012	172822	64	腎臟科	F	19930710	呼吸道	是	E. cloacae	

註：假設資料

表二　院內感染部位與病房之分佈

	泌尿道	呼吸道	外科傷口	血流	合計
A病房	2	3	0	2	7
B病房	1	0	2	1	4
A病房	0	0	1	0	1
合　計	3	3	3	3	12

註：假設資料

表三　轉碼後之院內感染個案基本資料

病房	病歷號	年齡	科別	性別	感染日期	感染部位	感染菌株	臨床症狀
B-017	214779	69	GS	M	19930726	UTI	95	3
B-005	201660	42	GS	M	19930722	SWI	11	*
A-002	196911	57	GI	M	19930712	BSI	11	3,12
A-004	132027	64	NEUR	F	19930710	RTI	82	5
A-011	131023	56	NEUR	F	19930712	UTI	85	*
A-003	237193	62	AIR	M	19930714	UTI	553	*
A-040	251730	62	INF	M	19930709	BSI	184	20
C-040	151730	51	GYN	M	19930717	SWI	85	*
A-022	141523	65	NEUR	F	19930710	RTI	82	*
B-008	101660	42	GU	M	19930722	SWI	11	*
B-003	136611	77	CVS	M	19930712	BSI	184	3
A-012	172822	64	NEPH	F	19930710	RTI	82	*

註：假設資料

析的方式是要在資料收集前就要詳加考量。譯碼是屬於資料整理的程序，透過此一程序把資料加以類別化，並將原始資料轉化成文字或數字，以方便輸入電腦及易於統計及分析，這一程序的成功與否，對分析與解釋的工作將有莫大的影響。以院內感染而言有那些資料需要譯碼，例如科別、感染菌株、感染部位、臨床症狀..等，如表一所收集之院內感染資料，可將中文的科別以英文縮寫代碼表示，感染菌株以號碼代表如表三，以減少輸入錯誤的問題；不同臨床症狀，可一一舉列，填入『是』、『否』或是將不同臨床症狀給予分類編號後，再輸入電腦，這二種方式各有優缺點，須與各位的程式設計者討論才定案，而這些決定都要借助各位日常經驗及文獻上的建議，才可使往後的分析順利達成。表三中科別以英文縮寫代碼，感染菌株號碼，臨床症狀號碼，為了避免時間久了忘記這些代碼的意義，我們需要準備一本譯碼簿，說明每一變項上各種代碼所代表的意義，且說明要儘量詳細，編寫時應包括：變項編號、變項名稱、 欄位大小，及最大及最小值，譯碼代號，(見表四)。譯碼簿主要有四項功能可提供：(1) 資料輸入者的依據；(2) 電腦程式設計師編寫分析程式；(3) 統計分析完畢後，撰寫報告的參考；(4) 新舊人員交接的依循，以使資料的譯碼有其一致性。

統計分析計劃

製作譯碼簿應同時決定要分析那些變項，以及採用那些計算公式及統計方法，以便電腦程式的設計者可據以編寫計算程式。擬定書面的統計分析計劃，可幫助我們思考研究目的及方向，其內容可分為三部份：變項的界定、變項的分組及分析指引。何謂變項的界定？例如表一中所有的變項都是『原始變項』，即是我們收集的原始資料，表二中的變項則是『組合變項』，它是由原始變項經由計算而得，另外若要看不同年齡層的感染率時， 便須將年齡予以分組，至於要以那一點為分組點， 亦要看我們的研究目的，假使我們想探討新生兒 (出生一個月內) 的感染率是否不同時，表五中的分組就不恰當。若是有涉及兩組資料的比較時，如表五中，45 -54 歲的每千人日感染率是否比 55-64 歲這組低時，更要寫下使用的統計方法及其計算公式作為日後指引，當然也可以列明報表的格式，以便將來的查閱與分析工作。例如要計算出 A 病房有幾例院內泌尿道感染，須由表一中『病房』及『感染部位』兩變項交集計算而得，表二的每一變項就是如此產生。如果要計算各病房每千人日的感染率時，就要多增加一項資料，即是每一病房的住院人日數 (見表五)。以上各項在統計分析計劃書中都要逐項明確的列出。

統計方法的選用

一、選擇適當的統計方法

表四　院內感染資料之譯碼簿

變項編號	變項名稱	欄位型式	欄位大小	最大值	最小值	說明
1	Ward	文字	7			病房
2	Hxno	數字	6			病歷號
3	Age	數字	3	110	1	年齡
4	Ser	文字	4			科別 GI 腸胃科 NEUR 神經內科 AIR 過敏科 NEPH 腎臟科 INF 感染科 GYN 婦科 GS 一般外科 GU 泌尿外科 CVS 心臟外科 ..
5	Sex	文字	1			性別 M 男性 F 女性
6	Infdate	日期	10			感染日期
7	Infsite	文字	5			感染部位 UTI 泌尿道感染 SWI 外科傷口感染 RTI 呼吸道感染血 BSI 流感染
8	Organism	數字	3			感染菌株 11 S. aureus 82 E. cloacae 85 E. coli 95 K. pneumoniae 184 P. aeruginosa 553 Candida spp ..
9	Symptom	數字	2			3 fever 5 cough 12 PMN > 10000 20 shock ..

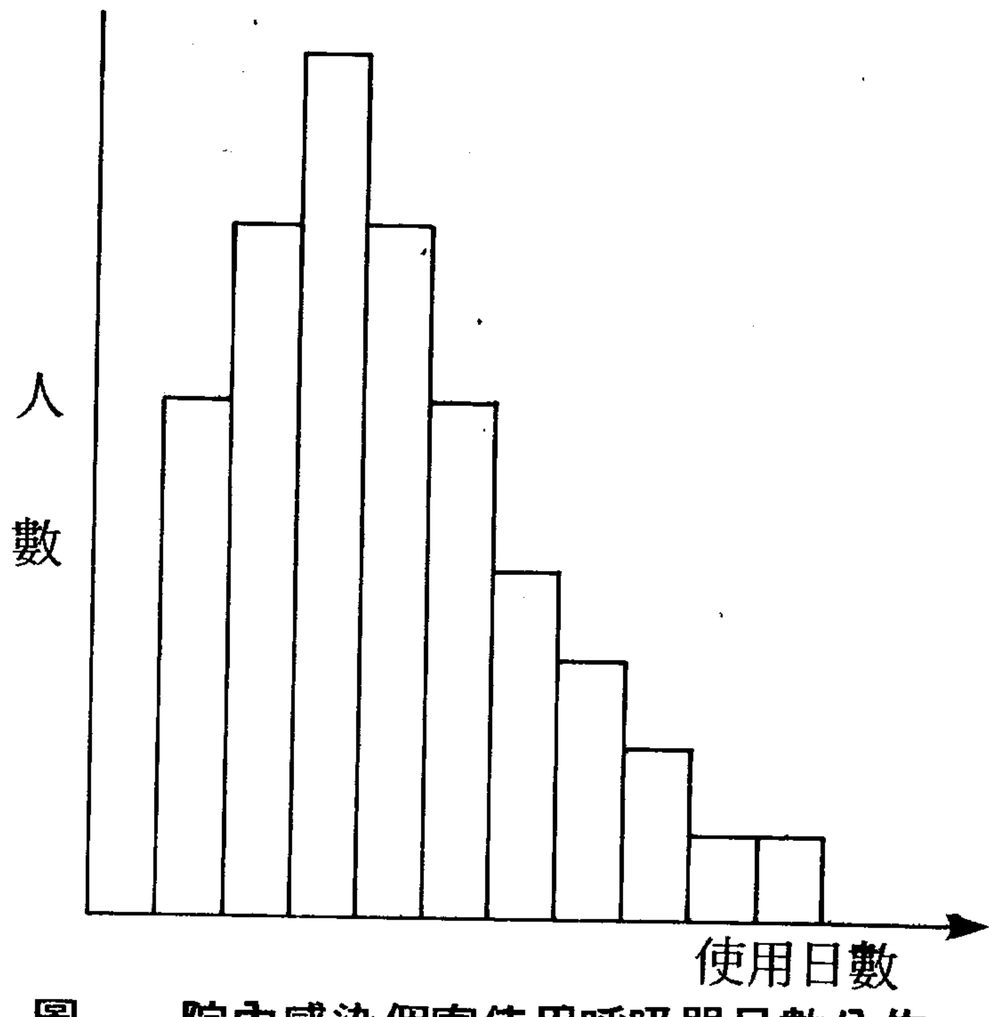
檢定統計的目的，是要決定兩個或兩個以上的樣本所收集的資料是否來自同一母羣體，如果不是，便要進一步決定這些樣本之間的任何差異或關聯，究竟有多大可能性可因機率所造成的。選擇適當統計方法時，下列兩個因素是值得我們加以考慮的 (1) 研究問題的性質：先確定所欲探討之問題是屬於描述性的問

表五　院內感染部位與年齡群之分佈

年　齡	泌尿道	呼吸道	外科傷口	血流	合　計	住院人日數	每千人日感染率
< 1 歲	0	0	0	0	0	112	0.00
1-4 歲	0	0	0	0	0	43	0.00
5-14 歲	0	0	0	0	0	0	0.00
15-24 歲	0	0	0	0	0	98	0.00
25-34 歲	0	0	1	0	1	435	2.29
35-44 歲	0	0	1	1	2	322	6.21
45-54 歲	0	0	1	0	1	417	2.39
55-64 歲	2	2	0	1	5	783	6.38
> 64歲	1	1	0	1	3	603	4.93
合計	3	3	3	3	12	2813	4.26

註：假設資料

題或關係性的問題，例如表五的資料是屬於描述性，所得到的訊息是變項的分配情況，若是想了解插導尿管是否比較容易得到院內泌尿道感染則屬於關係性的問題。(2) 研究資料的性質，表一中『是否使用導尿管』屬於類別變項，若改成『插導尿管的天數』則屬於連續性變項，不同的變項，所選用的統計方法就不一樣，（見表六）。另外要注意數據分佈的型態，因為統計分析所用的公式，都是在一些有關分佈型態的假定之下所導衍出來的，因此在使用每一方法之前，我們都要先確定資料的分佈型態是否與公式背後的基本假設相符，例如在探討資料的集中趨勢時，當我們在分析院內呼吸道感染使用呼吸器之天數時，就不宜使用平均使用呼吸器天數，而以眾數表示較具代表性，因為使用呼吸器的天數，不呈常態分佈，而呈偏態分佈（見圖一）。因而其分佈型態不同，不同變項種類，所選的統計方法不同，若呈常態分佈選用『母數統計方法』，不呈常態分佈則選用『無母數統計方法』（見表七），常見院內感染資料的統計方法選用（見表八）。



圖一　院內感染個案使用呼吸器日數分佈

表六　變項類別不同之統計方法選擇

變項種類 \ 統計方法	二項類別變項	類別變項	等距及等比變項
二項類別變項	卡方檢定 比率差別 波以松檢定 費歇恰當檢定	卡方檢定	
類別變項	卡方檢定	卡方檢定	
等距及等比變項	Z 檢定 *t* 檢定	ANOVA	相關 迴歸

表七　分佈型態不同與變項類別不同之統計方法選擇

	樣本數大（符合常態分佈） 母數統計方法	樣本數小（不符合常態分佈） 無母數統計方法
類別變項		
兩組資料的比較	卡方檢定	Fisher's exact test
兩組資料配對比較	McNemar's test	Sign test
多組資料比較	卡方檢定	Cochran's Q test
連續性變項		
兩組資料的比較	*t* 檢定、Z 檢定	Mann-Whitney U test
兩組資料配對比較	配對 *t* 檢定	Wilcoxon Signed rank test
多組資料比較	ANOVA *t* 檢定 + 多重比較	Kruskal-Wallis test

表八　院內感染資料之統計方法選用

母數統計	無母數統計	院內感染資料
卡方檢定	Fisher's exact test	是否得院內泌尿道感染 vs 是否使用導尿管
t-檢定	Mann-Whitney U test	是否得院內泌尿道感染 vs 使用導尿管天數
配對 *t* 檢定	Wilcoxon signed rank test	使用藥物前後 vs 血壓降低毫米數
ANOVA	Kruskal-Wallis test	科別（內、外、兒） vs 使用靜脈滴注天數
線性回歸分析	—	住院天數 vs 使用呼吸器天數

二、避免統計檢定的濫用

常被濫用的統計檢定如下：

㈠ 由一堆雜亂無章的資料中企圖尋找顯著差異

在分析資料時要將研究的虛無假設常記在心，千萬不要盲目的接受＂顯著差異＂，要與臨床的現象互相配合，若使用多重比較可以減少誤差。

㈡ 統計上的顯著差異經常被研究者認為是實質差異 (substantive significance)

統計上 p 值的大小如："0.01"、"0.05"、"0.1" 並不能用來描述變項間相關強度。檢定時有顯著差異有兩種可能：(1) 變項間有實際相關 (2) 樣本數目超大。因此我們要隨時問自己，『這種相關是否是重要且有實質意義』，並了解檢定『是否有顯著差異』並非最終目的，而是最初步驟。

㈢ 過份強調發現差異

只強調統計上的差異，而未探討這些差異對臨床現象的改善程度

㈣任意的改變變項的形態

為了進行卡方檢定，故意將等距變項改為類別變項，而造成下列誤差：

1. 用了敏感度低的統計方法。
2. 任意分組，造成統計結果的不同
3. 將序位變項轉為類別變項

分析結果的解釋

我們以表五為例做解釋，『本月中有 2813 人日數，共有 12 位院內感染個案發生，每千人日院內感染率為 4.26 人次，泌尿道感染 3 例、呼吸道感染 3 例、外科傷口感染 3 例及血流感染 3 例，其中 55-64 歲這組每千人日院內感染率最高為 6.38 人次，而 25 歲以下無院內感染個案發生』；括號內即是一般的描述性資料的解釋。至於兩組間的比較，請參考本專欄以往各期的解說。

另外對於研究結果的解釋，有時會遇到以下兩個問題

1. 研究結果不確定或與假設不符

當得到結果與原先的假設背道而馳時，我們必須探討：為什麼會得到這樣的結果？此時需詳細審查研究程序的每一環節；其可能為一個或數個因素造成的 (1) 理論或假設不正確；(2) 方法或步驟不適當或不正確；(3) 測量方式或定義不明確；(4) 分析錯誤。如果經過仔細審查之後，發現方法、測量、分析都適當無誤時，則可斷定問題出在假設或理論的不正確，可以所得結果來修正假設或理論，反而對科學的進展有所貢獻。

2. 未曾預測到的發現

我們在處理這類發現時更須抱有懷疑的態度，在決定接受此發現之前，應單獨加以研究，以驗證其真實性，只有經過精心設計、控制必要因素、從事有系統分析考驗的結果我們方能採信。

合理的分析解釋是建立在適當的研究設計及統計方法，其目的不僅在描述個別事實的存在，而且還要找出事實間的關連，因此欲得到可靠的解釋，資料分析的每個步驟都要謹慎處理。