描述性統計之正確應用

⊙林業經濟組・陳麗琴

前言

有不少研究人員辛辛苦苦好不容易作出試驗結果資料，然後請人把資料建立在EXCEL上，並且慣性的請人跑出資料的平均值及標準差或標準誤差，看了一看，就馬上要求作t-test，χ^2-test或是迴歸分析來檢驗試驗處理間是否有差異或相關，這樣的作法是很危險的。描述資料是非常重要的工作，它除了可以了解資料的長像，協助你直覺作專業的判斷，並因此選擇正確統計分析方法協助驗証判斷的準確性，而且可供讀者了解整個分析的背景。一個研究人員如果連資料本身的特質都不清楚，例如資料的測量方式(尺度)是名目的、順序的、等距的或等比的，資料的分布是單峰、雙峰或沒有型態(pattern)、左傾、右傾或是有極端值，冒然使用統計分析，並寫出研究報告供人參考應用，是不負責任的。就是因為資料有這些測量尺度及分布的不同狀況，才會有不同統計方法的產生，因此研究人員在進行何種差異分析前，必須先看清楚自己資料的本質，以下提供兩種簡單、常用且重要概念的檢視建議。

資料檢視方法

一、確認資料測量尺度(measurement scale)

在統計上資料類別分有四種，其尺度的高低等級從最低到最高依次是：名目尺度、順序尺度、等距尺度、等比尺度(表1)，依照資料不同尺度的特性及分布狀態，就會有無母數統計分析與有母數統計分析之區別。如果資料是屬於計量的、連續的，如等距尺度、等比尺度，且有一定型態分布，我們可以用有母數統計分析法；另一類資料是非計量的，僅僅是符號、等級或計數的，如名目尺度、順序尺度，也沒有特定分布，就要用無母數統計分法。

愈高等級資料可以用愈精密的統計分析方法，例如變異分析、迴歸分析等等，當然愈高等級的資料也可以反向使用較低等級的統計方法，不過所提供檢測資訊會變少，精準度也較差。相反的，有時候為了要應用較高等級的統計方法，我們會採取不那麼嚴格的方式，將順序尺度資料作為等距尺度方式來使用，但前提是樣本要夠大。舉例最常

表1 統計學上的資料類別分有四種，依其尺度的等級不同所適用的統計方式也有所不同。

類　別	定　義	特　性	例　子
名目尺度 (nominal scale)	符號，沒有數量的意義	分類	性別、黨派、職業
順序尺度 (ordinal scale)	數字、符號	分類、順序大小、強弱、高低	滿意度
等距尺度 (interval scale)	數字、等距、沒有絕對的0	順序、等距、不可倍數比較	海拔高、攝氏溫度
等比尺度 (ratio scale)	數字、等距、有絕對的0	順序、等距、可倍數比較	年齡、收入、生長量

使用的滿意度，如果以非常滿意、滿意、普通、不滿意與非常不滿意來分級，它是屬順序尺度，理論上是不能用等距方式統計方法來處理，但標準訂低一點，還是勉強可以使用，所以許多研究報告還是用了變異分析及迴歸分析等，但是如果去掉了中間值(如普通、沒意見)，我們就不認為是等距，僅僅是順序尺度而已，連平均數與標準差均無法使用，這時我們耳熟能詳的χ^2-test、K-S test或其他無母數檢定方法就派上用場(而非t-test)，用來比較不同等級下處理結果是否有差異。當然名目資料更是如此。

但如果這些具有符號、等級或計數特性的資料，具有某種特定分布型態(圖1)，如常態分布(normal distribution)、二項分布(binomial distribution)、卜瓦松分布(Poisson distribution)，當然仍是可用有母數統計分析法；換句話說，資料只要有某種特定型態的分布，均可使用該分布之有母數統計分析法；如果看不出分布型態，即便是等距、等比資料，也只能用無母數統計方式了，或者是嘗試將這些資料轉換成順序資料或名目資料後，再檢驗其分布，是否具有特定分布型態，如果是，則仍可採用有母數統計方式。

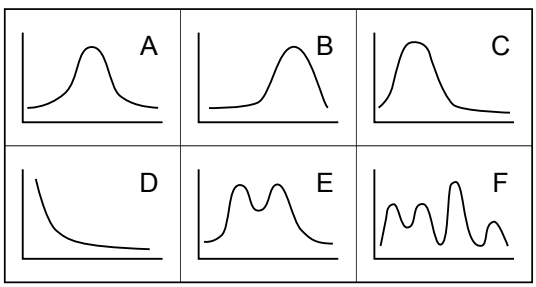

圖1 不同類型資料的分佈型態 (A：常態分布；B：左偏分布；C：右偏分布；D：指數分布；E：雙峰分布；F：無固定型態分布)

所以研究人員應先確認究竟所要分析的資料是屬於那一個測量尺度，並且請將資料的頻度或機率分布圖(例如line plot、histogram、pie chart)製作出來，作為協助判斷適合用那一種有母數或無母數統計方法分析。圖1所示是6種不同型態的資料分布，可供讀者參考檢查資料的屬性。

二、正確使用平均值(mean)、中位數(median)、標準差(standard deviation，SD)與標準誤差(standard error，SE)

一般在寫研究報告的時候，一定先使用描述性統計來表達資料的特性，才進行下一步的統計分析，如t-test、χ^2-test、變異數分析(ANOVA)、迴歸分析(regression analysis)等等。不少研究人員在使用描述性統計時，究竟應該用SD或者用SE，並不清楚。因為$SE = SD/\sqrt{n}$，n是樣本數，因此SE會小於SD，感覺比較好，很多人就偏好用SE。

其實SD是用來表達各個樣本的離散或變異情形，而SE是測量各個樣本之平均值的準確度，是平均值的SD，是用來分析或檢測處理間差異用的，二者明顯用途不同，所以我們在描述資料時，自然要用mean±SD，如此讀者才能清楚的感覺到資料的分布狀況。SE常用在求取平均值的信賴區間(confidence interval, CI)的過程中，才會被用到，假設是95%CI以下，其公式為mean±1.96SE。這種CI常被拿來作為處理間比較分析之用。請大家注意，資料是常態分布　或接近常態分布時，SD與SE才具有意義。

其次，使用平均值或中位數，那一種較能代表資料的中央趨勢(central tendency)？如果資料是鐘型分配或兩邊對稱如常態分布，

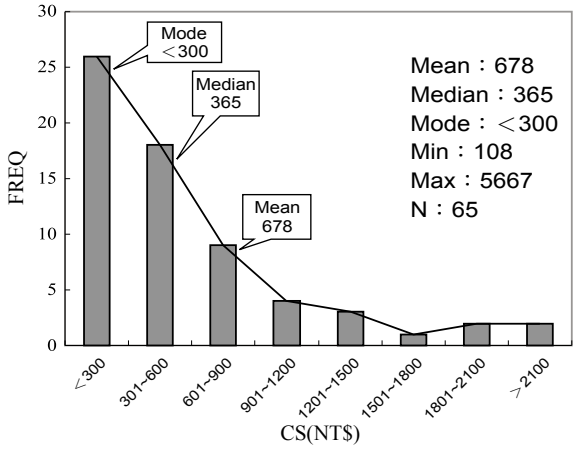

圖2 森林經濟價值的頻度分布圖

基本上使用平均值或中位數差異不會太大，但如果分布是屬於偏斜的、具遠離趨勢的極端值，或者並無特定型態的，中位數絕對是上選，而非平均值(這裡不討論眾數)，舉例說明如圖2所示，這是一個右偏的頻度分布圖，X軸代表民眾對某一森林資源經濟價值的看法，Y軸代表落在某一價值的樣本數，要如何選擇一個數值來代表這樣的森林價值？如果只看平均值，其值為678元，很顯然它無法代表真正的核心值，反倒是中位數，其值為365元比較具有資料的代表性，有一半的樣本小於等於365，平均值與中位數，二者數值幾乎差了一倍。

另外，如果樣本數夠大，根據中央極限定理，樣本平均值是可以形成常態分布，但請研究人員不要相信點估計mean的代表性，一定要看區間估計，即是信賴區間，因為信賴區間才是接近真實的。舉例說明，我們常看到一些研究報告很明白的寫出各處理的平均值及信賴區間，經檢定結果是處理間不顯著，但仍捨不得丟掉平均值，還說明它有差異；其實平均值也不是不能說明，只是它其實在統計分析上來講是個參考值。不同處理結果的信賴區間比較可如圖3所示，X軸代表不同森林資源(即處理)，Y軸代表不同森林資源經濟價值的信賴區間，從縱線與縱線之間重疊的狀況，就可一目了然究竟那個處理與那個處理結果的差異程度了。而若樣本數很小時，不適宜用百分比來代表資料的結構，以免誤導，只要說明樣本數有多少，其中那一個等級有多少數量就可以了，不要作太多的延伸解釋。如果資料中有極端值，必須小心謹慎，多半不宜視同其他樣本納入分析。

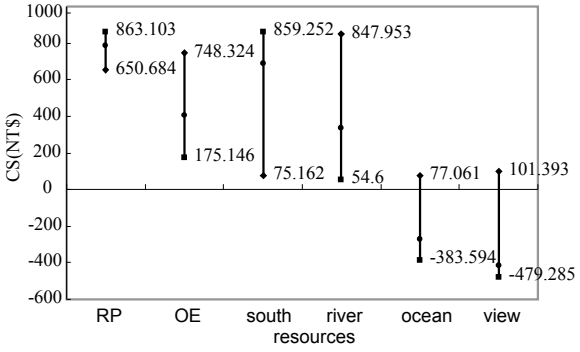

圖3 不同森林資源經濟價值的信賴區間

結語

一般研究者多可以毫無困難地將試驗資料作出描述性統計分析，但心中不免或多或少疑惑著，這些分布圖、平均值、中位數、標準差及標準誤該如何應用？應用是否適得其所？本文的目的即是企圖將這些統計值應用的背後基本概念作一番釐清，希望對讀者的信心有所幫助，輕鬆愉快地就能提升統計分析品質，有效表達正確的試驗資料特質。