資料整理與檢誤經驗談

王文心

一、 過錄的原則

就量化研究而言，過錄（Coding）簡單的說，是將問卷或其他來源的資料轉換成數字，以利使用電腦來進行統計分析。過錄包括四個主要的步驟：給予與每個問題（或變項）答案相對應的過錄碼（Code）、安排適當的欄位供電腦讀取過錄碼、製作過錄編碼簿、以及檢查過錄結果。

過錄碼的設計具以下幾項特性：

1. 包容性：問題的每一種答案都有一種可歸入的類別
2. 互斥性：每一種答案只能歸入一種類別
3. 適當性：分類後能夠取得研究所需要的資訊，或給予的方式符合一般常規。

除此之外，在進行過錄時應把握幾項要領，第一，每筆資料應留有識別變項，例如：研究計畫編號、樣本編號等，以利事後資料有錯誤時，用以查找原始問卷；第二，沿用封閉式問卷中的選項代碼，以免過錄員混淆；第三，封閉式問題答案的分類或加總，應由研究者事後依其研究目的與需要再進行，過錄時應保留最原始的資訊。

二、過錄編碼簿（Codebook）

在完成過錄碼的給予之後，應將所有過錄規則與內容彙整成一份文件存檔，這樣的文件便稱為過錄編碼簿（參考下面範例），過錄編碼簿包括的項目如下：

1. 題號：即該變項位於原始問卷中之題目的題號
2. 變項名稱：供電腦辨識及分析使用，例如用變項意義為代表的 birthyr（出生年）、birthmth（出生月份）、sex（性別）、edu（教育程度）等；或是以變項順序命名的 V1、V2、V3 等。在撰寫統計分析程式時，會需要使用變項名稱。
3. 卡數/欄位：記錄資料存放的位置。在以 ASCII 格式建立資料檔時，一般習慣將資料的欄位一列控制在 80 欄之內，若訪問的問卷太長，一列不夠過錄時，便登錄到第二列、第三列；像這樣多列代表同一樣本的資料檔，每個樣本的第一列資料我們稱為第一卡，第二列的資料稱為第二卡，依此類推。SAS、SPSS 讀取多卡資料時，須注意其特定語法。
4. 變項說明（variable label）：解釋變項名稱代表的意涵，譬如：birthyr 代表出生年、birthmth 代表出生月份、sex 代表性別等。在製作過錄編碼簿時，變項說明的完整度應一如問卷的題目，以避免日後與其他變項混淆不清。
5. 選項數值說明（value label）：解釋各數值的涵義，舉例來說，在性別（sex）變項所對應的選項數值說明中，1 代表的選項數值說明為男性，2 代表的選項數值說明為女性。
6. 備註：記錄提醒研究者注意的資訊。例如：使用的紀元方式，如西元年、民國年；反向題；答案分類判斷的標準；遺漏值的過錄方式等。

過錄編碼簿範例

題號	變項名稱	卡數/欄位	變項說明	選項數值說明	備註
	ID	1/1-3	樣本編號		
1	Sex	1/4	請問您的性別是	1 男性 2 女性	9:missing
2	Birthyr	1/5-6	請問您的出生年月？ 年	依受訪者回答填入	99:missing 民國前一年 91，以此類推，民國前六年以前出生者，皆鍵 96
	Birthmth	1/7-8	請問您的出生年月？ 月		99:missing

過錄編碼簿有其製作的必要性，當研究者將回收後的問卷交由他人編碼時，過錄編碼簿是編碼員的基本指引，可避免標準不一致或錯誤的發生，過錄編碼簿亦利於資訊的保存，方便將資料與他人共享，也能避免因時間久遠而對過錄代碼的意義不復記憶。

三、何謂資料檢核

在進入統計分析之前，首先應確保資料檔的正確性，就資料檔有無存在不應出現的數值或是受訪者回答不符合邏輯之處進行檢查，以免得到錯誤的分析結果。資料檢核可分為類別變項或連續變項的基本檢核，以及牽涉到兩個或兩個以上變項之間的邏輯檢核。

四、類別變項的基本檢核

在筆者任職的單位中，我們以不合理值（illegal value）來指稱資料檔的類別變項中所出現研究者事先定義以外的數值，有學者將這種檢查資料的方式稱為「可能性檢核」（possible code cleaning）（Barbie；1989），

不合理值（illegal value）也有人稱為超過範圍的數值（wild code 或 out-of-range value），若是某個變項的過錄碼只有三種時，資料檔中就不應該出現這三種之外的數值，例如：問到性別的題目時，研究者在過錄編碼簿中只採用 1 代表男性，2 代表女性，9 代表不知道或沒有填答，若吾人在檢查資料檔時，發現性別的資料欄位中出現 4 或 5 時，4 或 5 就稱為不合理值。

出現不合理值的可能原因是：過錄員過錄錯誤、資料鍵入時按錯鍵，或是還有一個可能性是，已有新增過錄碼，但在過錄編碼簿上未及記載，以致進行資料檢核的人員將之視為不合理值。對於不合理值的處理，一般會建議研究者利用其樣本編號，找出原來的問卷來加以更正資料。

五、連續變項的基本檢核

連續變項可以用是否出現偏離值（outlier）的方式來檢查資料的正確性，偏離值是指偏離常態分布的數值，若是數值落在三個標準差之外，即屬於偏離值。例如新生兒平均體重為 3250 公克，標準差為 450 公克，若出現 4500 公克以上則屬於偏離值，因為真的會出現 4500 公克以上的機會不大，需要進一步查證。可以用次數分配或作圖的方式，例如：直方圖、莖葉圖或盒狀圖等來找出偏離值。

六、邏輯檢誤

除單變項的基本檢核之外，還有牽涉到兩個或兩個以上變項之間關係正確性的檢核，我們再進一步區分為跳答檢核（Filter check）與邏輯檢核（logical check）。所謂的跳答檢核是指在問卷中常會見到的跳答題，受訪者若回答了某些特定答案後，則不須要再回答續問的問題，例如：當我們詢問受訪者家中有無桌上型個人電腦，若受訪者回答沒有時，就不會再續問有關家中個人電腦的等級。這類的跳答檢核，可依問卷中所設計的跳答方式進行。

另外在邏輯檢誤方面，還可針對問卷中並未註明跳答或不適合設計為跳答題，但是問題的答案有前後邏輯關係的變項進行檢核，亦即受訪者回答了某一類的答案後，便不可能會回答另一類的答案。譬如問卷前半部提到是否有購買彩卷的習慣，若受訪者答否時，在問卷後半問到受訪者每週的固定支出時，就不應在支出項目上出現樂透彩。

在上述的例子中，兩個變項分開看時，答案都在合理的範圍之內，但綜合一起看時便會出現不合邏輯的情形；再者，不僅是類別變項之間會有邏輯上的關係，類別變項與連續變項、或是連續變項與連續變項之間都可能會有邏輯存在。

出現邏輯錯誤時的處理方式與不合理值相似，都須找出樣本編號後，查原始問卷來確定受訪者回答的答案，若原始問卷中的填答記錄確實有邏輯上的問題時，則研究者可能須要利用受訪者留下的聯絡方式，再度與受訪者確認其答案。

本期我們先介紹過錄與資料檢核的基本概念，在後續幾期中，我們將從實務面介紹單選題、複選題、開放式問題等如何編碼，以及如何撰寫統計程式來檢查資料檔中有無不合理值、或是邏輯有不一致的問題等。