資料整理與檢誤經驗談

以 SPSS 程式進行不合理值檢誤

蘇婉雯

數據資料檔的建立，必須包括資料檢誤的工作才能算完成。嚴謹的資料檢誤，可以增進資訊的正確性以及資料的可用性。一般來說，資料發生錯誤的來源可能來自於訪員、督導、過錄者或鍵入資料者。不過，有些時候只是程式語法錯誤，而造成資料錯誤的假象。基本的資料檢誤工作包括：不合理值檢誤及邏輯檢誤兩種。本期以 SPSS 統計軟體介紹不合理值的檢誤工作。

大部份變項都有其合理的值域或分佈，而超出這些值域、或落在合理分佈以外的觀察值，往往起因於建檔過程中發生的錯誤。一般而言，界外值(Outlier)亦屬於不合理值。茲將整理工作內容說明於下：

一、類別變項(Categorical Variables)

類別變項的合理數值應該是幾個固定的類別代碼，包含研究者設計的「跳答碼」、「遺漏值代碼」等。如果類別變項的資料中有任何數值非屬這些固定代碼，即為不合理值。研究者應該查明這些不合理值的來源，並做適當的修改或處理。

例 1：【原始問卷】如下：

3. 請問您父親的籍貫是
   □ 1.本省閩南人
   □ 2.本省客家人
   □ 3.大陸各省市
   □ 4.原住民
   □ 5.其他

表1-1【次數分配結果】如下：

3您父親的籍貫是哪裡？

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	86	4.4	4.4	4.4
	1	1424	72.7	72.7	77.0
	2	166	8.5	8.5	85.5
	3	236	12.0	12.0	97.6
	4	34	1.7	1.7	99.3
	5	2	.1	.1	99.4
	6	5	.3	.3	99.6
	7	7	.4	.4	100.0
	Total	1960	100.0	100.0	

從上面的次數分配結果表中，不容易發現是否有不合理值存在，必須核對問卷或過錄編碼簿之內容才能得知。如果能在資料檔或程式中補上選項數值說明(value label)，則能夠輕易的找出不合理值。

【自撰語法】- 選項數值說明

```
value label
a3   1 "本省閩南人"  2 "本省客家人"
     3 "大陸各省市"
     4 "原住民"  5 "其他"
     7 "不知道".
```

【點選視窗】- 選項數值說明

於SPSS 10.0版的Data Editor視窗中，選擇「Variable View」頁面，點選該變項之values欄，出現下面視窗並依序鍵入選項數值說明。

圖1. 點選視窗之選項數值說明

建立選項數值說明後，再執行次數分配分析，可得結果如表1-2：

表1-2

3 您父親的籍貫是哪裡？

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	86	4.4	4.4	4.4
	1 本省閩南人	1424	72.7	72.7	77.0
	2 本省客家人	166	8.5	8.5	85.5
	3 大陸各省市	236	12.0	12.0	97.6
	4 原住民	34	1.7	1.7	99.3
	5 其他	2	.1	.1	99.4
	6	5	.3	.3	99.6
	7 不知道	7	.4	.4	100.0
	Total	1960	100.0	100.0	

顯然「0」、「6」對於這個變項而言都是不合理的界外值。而「7」是原計畫設計的「不知道」編碼，單從問卷中無法得知此編碼。其中，選「6」的人很少，只有兩名，可能是過錄員或鍵入資料者所發生的錯誤；而「0」，則可能為「跳答+遺漏值」混合使用的代碼，該研究沒有明確定義「0」並區分跳答及遺漏值的結果。一般而言，資料檔中亦需要建立跳答、遺漏值等特殊編碼的選項數值說明。

類別變項的不合理檢誤，最方便的方法就是執行「次數分配」。只要將該變項做「次數分配」，則資料的值域、分佈就會一覽無遺。如果需要進一步檢誤可疑資料的編號(ID)，則需再配合其他程式。下面以 SPSS 語法說明如何挑出不合理值。

《步驟一》執行次數分配分析

【自撰語法】

可以利用「Frequencies Variable=varname1**」語法執行。**

【點選選單】

選取「Analyze→Descriptive Statistics→Frequencies**」再點選要分析的變項名稱**

《步驟二》挑出可疑值的 ID

【自撰語法】

可以透過下列語法，將可疑值及其 ID 列出。

```
Temporary.
Select If Any(Varname1, n1,n2,..).
List id Varname1.
```

【點選選單】

可以點選「Data→Select Cases**」再到對話視窗中設定條件，即可使工作中的資料檔僅剩下篩選過、符合條件的觀察值。此時，點選「**Analyze→Descriptive Statistics→Frequencies**」，再選定「**ID**」等編碼變項，即可得到可疑值的 ID 清單。**

此外，有關年齡、年度、月份、小孩數…等變項，皆因題目與受訪者的不同，而有不同的合理值域。

二、連續性變項

檢查連續性變項的基本方法，可以

由簡單的描述性統計值(平均值、標準差、極大值、極小值)、分佈圖等幾方面來看。

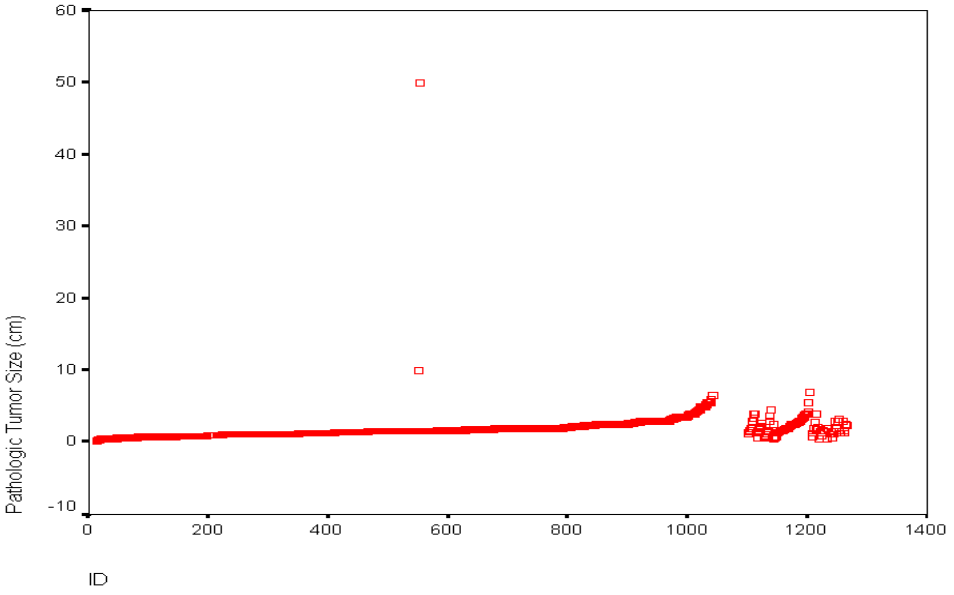
例 2-1：以下舉一般性的觀念為範例：

以平均值而言，某社區青年收縮壓平均值只有 70 mmHg，並不符合一般生理分佈。

以標準差而言，某社區青年收縮壓標準差高達 50 mmHg，亦不符合一般文獻的記載。

以極大值和極小值來看，學童身高落在 80 到 170 公分以外者，亦不符合學童正常生理狀態。

圖 2：本例以分佈圖來檢查界外值，資料為 SPSS 軟體所附的資料範例。



從圖 2 的分佈看出遠離集中分佈的觀察值(outlier)。上述為一個腫瘤直徑的散點分佈圖，我們可以看出有兩個點各為 10 公分、50 公分左右，遠遠離開一般腫瘤大小的分佈範圍。其資料的正確性值得查驗。

如果這是一個罕見的病例、嶄新的發現，也必須建立在資料正確性無庸置疑的基礎上！所以，資料檢誤是量化研究重要的程序之一。

【自撰語法】

可以使用「EXAMINE VARIABLES = varname/PLOT =BOXPLOT.」語法，即可同時得到詳盡的描述性統計值(包括：平均值、標準差、峰度、偏度、極大值、極小值、樣本數…等)，以及箱型圖(box plot)。另外，可以依照專業知識設定連續變項合理

的值域，做界外值檢誤，語法如下：

```
Temporary.
Select if (varname<100 or varname>500).
List id var.
```

如此可將varname這個變項超過500或低於100者全列到output檔中，並將ID一併列出，方便查明。

【點選選單】

可以點選「Analyze→Descriptive Statistics→Explore」，將要分析的變項選至「Dependent List」中，計算平均值、標準差、極大值、極小值、樣本數；再點選「Graphs →Boxplot」(設定 Summaries of separate variables)，可得到箱型圖。

在上述方法中所挑出的個案編號，應詳細檢查原始回卷的內容後，將錯誤的資料更正，整個不合理值檢查的工作才告完成。如果資料鍵入時發生錯誤，而錯誤的資訊仍在合理值範圍中，則相當不容易找出。這就是為什麼需要double key in後再核對兩資料檔是否一致。嚴謹的資料整理工作，可以早期發現資料的錯誤，並早期修改，以免影響分析結果。不合理值檢查工作更應該於邏輯檢查之前先執行。在下一期通訊中，我們將介紹如何以 SPSS 進行邏輯檢查的方法，敬請期待。