行政院國家科學委員會專題研究計畫　成果報告

邊際列聯表的適用性探討

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-004-007-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：江振東

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 11 月 8 日

行政院國家科學委員會專題研究計畫成果報告

邊際列聯表的適用性探討

On Collapsibility of Contingency Tables

計畫編號：NSC 92-2118-M-004-007

執行期限：92 年 7 月 1 日至 93 年 8 月 31 日

主持人：江振東 國立政治大學統計系

一、計畫中文摘要

高維列聯表通常較不易分析，因此在不損及我們想要探討的因子間的相關性的前提下，如何利用邊際列聯表來取代原始列聯表作分析，在實務上是個相當有趣的課題。在此計畫中我們希望釐清文獻中幾種不同合併性(collapsibility)定義間的關係，並藉由對數線性模型中參數限制式的不同對充要條件所可能造成的影響來探討這些定義的適用性及提出適當的修正定義。

關鍵詞：邊際列聯表、可合併性

Abstract

A lower-dimensional contingency table is usually easier to understand than a higher-dimensional one. Although collapsing a larger table over its secondary factors and using the reduced marginal table to do the analysis is usually easier to carry out and easier to explain, the process should be exercised with care. In this study, we would try to sort out the differences among various definitions on collapsibility and their related necessary and sufficient conditions. We would also want to find out if the conditions are independent of the constraints imposed on the model parameters, and, if necessary, propose different definitions..

Keywords: collapsibility, marginal tables.

二、計畫緣由與目的

一個列聯表(contingency table)的維數如果太大，在作資料分析時往往會造成許多的不便，同時也會增加模型解釋上的困難度。因此一種直觀的思維方式便是設法藉由忽略其中某些次要因子(secondary factors)來降低維數，並且利用這些維數較小的邊際列聯表(marginal table)來作分析。這個方法如果可行的話，問題當然可以簡化許多。然而 Simpson's paradox 告訴我們，這樣的作法未必能夠完全適用。就某些結構的列聯表而言，主要因子(factors)在原始列聯表的關係在邊際列聯表中也能維持不變，這是我們最希望見到的情形。然而就絕大多數的列聯表而言，主要因子在原始列聯表中所呈現的關係在邊際列聯表有可能完全變了調；在這種情形下藉由這類簡化了的邊際列聯表來作資料分析，將會導致不正確的結論。因此在考慮降低列聯表維數，使用邊際列聯表來作分析的同時，我們必須先考慮合併列聯表的可行性，這也是這個研究計畫的主要目的。

針對合併列聯表的可行性，Agresti(2002)引述 Bishop et. al(1975)所提出的條件如下：

Suppose that a model for a multiway table partitions variables into three exclusive

subsets, A,B, C, such that B separates A and C. After collapsing the table over the variables in C, parameters relating variables in A and parameters relating variables in A to variables in B are unchanged.

這裡合併的可行性是就與集合 A 直接相關的參數以及連結集合 A、B 間的參數的不變性來作定義，此外這裡所提出的只是合併性的充分條件。由於 Agresti(1996,2002)的兩本書幾乎已經成為討論及學習類別資料分析的經典教科書或工具書，上述定義也彷彿就是可合併性的標準定義和條件。然而在文獻中關於可合併性則有不同的定義，依照條件要求的深淺程度以及適用情況的不同，由簡而繁可以區分為簡易合併性(collapsibility)，嚴格合併性(strict collapsibility)及強固合併性(strong collapsibility)(參見 Whittemore(1978), Shapiro(1982), Ducharme and Lepage(1986))。而相對於這些不同的定義所需要滿足的充分必要條件，在文獻中也有相關探討。這些條件有些適用在任意維數的列聯表，有些則僅適用在三維 IxJxK 列聯表甚至僅侷限於 2x2xK 形式的列聯表。然而在我們就這些文獻做過研讀後，意外發現這些定理的相關證明基本上都是架構在模型參數「和為 0」(sum-to -zero)的這個假設前提下來作探討。舉例來說，考慮模型

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$$

這是一個給定因子 X 的情形下，因子 Y 與 Z 為獨立的條件獨立模型。為了求得參數估計的唯一解，我們通常需要對這些參數設限。所謂的「和為 0」的限制式指的是

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z$$
$$= \sum_i \lambda_{ij}^{XX} = \sum_j \lambda_{ij}^{XY} = \sum_i \lambda_{ik}^{XZ} = \sum_k \lambda_{ik}^{XZ} = 0$$

。

然而我們也知道這種限制方式只是其中一種可能方式而已，其他最常被使用的方式還有「令最後一項為 0」(last=0)的方式，亦即 $\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0$ ，$\lambda_{Ij}^{XY} = \lambda_{iJ}^{XX} = \lambda_{Ik}^{XZ} = \lambda_{iK}^{XZ} = 0$ ，$\forall i, j, k$，這也是 SAS GENMOD 的內定形式。在此情形下，除了所謂的 estimable contrasts 外，限制式採用方式的不同會導致參數估計值的不一。由於前述幾類的合併性，多數建立在參數估計值在原始列聯表和邊際列聯表中維持不變的要求，這不免讓我們懷疑限制式採用方式的不一致是否會導致合併性的充要條件也會因而變更。實際上也確實如此，我們發現(見賴芬秀(1998))就簡易合併性而言，兩種不同限制式所對應的充要條件確實不一。至於嚴格合併性及強固合併性，我們也已發現就三維列聯表而言，可合併性的充要條件並不會因為採用「和為 0」或「令最後一項為 0」的限制條件而有所不同。然而由於我們只有針對這兩種限制方式來作討論，而限制條件並非只有這兩類，因此我們所不知道的是這個充要條件是否確實與限制式無關，也就是說無論我們給定任何的限制式，其充要條件是否依然不變？

此外這些充要條件在四維以上的列聯表是否仍舊適用，是否也不會受「和為 0」及「令最後一項為 0」這兩種限制方式的影響，是否可以更進一步的推論確實與限制式無關，則全然未知。由於推導方式並非單純只是處理三維列聯表的過程的直接衍生，而文獻中也並沒有任何討論，這些就是此研究計畫的主要著眼點。

三、計畫結果與討論

藉由這個計畫的執行，我們得到的主要結論可以歸納如下：

(一)簡易合併關切的重點在於某些特定的參

數值(比方集合 A 裡的因子)在合併其他次要因子前後，是否維持不變。就一個 hierarchical model 而言，我們知道除了最高階項的參數外，其餘低階參數可能因為限制式採用的不同，而導致不同的估計值。由於最高階項在降階後的列聯表已經不復存在，也就是說簡易合併所要探討的基本上都屬於低階項的參數，我們因此也就能夠說明何以簡易合併性的條件會隨著限制式的不同而有所改變。

(二)假定我們有興趣探討的不單單只是忽略集合 C 的所有因子前後，集合 A 裡的因子所對應的參數是否改變外，我們也希望知道連結集合 A、B 間因子的參數是否也不受影響，這就是嚴格合併性所關切的主軸。就三維列聯表而言，由於可合併性的充要條件之一是 $\lambda_{ijk}^{XYZ}=0,\ \forall i,j,k$，也就是我們所探討的模型不能是一個飽和模型(saturated model)，因此任意的二階項(亦即 $\lambda_{ij}^{XY},\lambda_{ik}^{XZ},\lambda_{jk}^{YZ}$) 就是最高階項。如前所述，由於最高階項並不會因為限制式的不同而有不同的估計值，因此在處理三維列聯表時，即使採用不同的限制式，嚴格可合併性的條件完全相同。但是就四維或四為以上的列聯表來說，這種現象不復存在，因此嚴格可合併性的條件依然取決於限制式的選用。此外，我們也分別得到「和為 0」及「令最後一項為 0」的限制條件下，嚴格可合併性個自所對應的充要條件。

(三)前述兩種合併性，主要著眼於如何藉由忽略若干次要因子來降低列聯表的維數，不過強固合併性則有不同的考量。其主要目的在於了解忽略某些因子的若干類別選項後，是否會對其他因子間的原始關係造成影響。針對這種類型的可合併性，我們也得到與嚴格合併性相同的結論。就三維列聯表而言，合併性的條件與限制式的選用無關。不過就四維以上的列聯表，我們可以找到反例來說明合併性的條件取決於限制式。

(四)如果我們關切的焦點，可以表示為 estimable contrasts 的一個函數的話，前述三種可合併性便與限制式的選用，完全無關。也就是說，可合併性的條件並不會因為限制式的不同而不同。由於 estimable contrasts 的估計值並不會受到限制式的影響，因此可合併性的條件也不受限制式的影響是可預期的。此外，我們也已經能夠經由數學推導來證實這個結果。

(五)就三維列聯表而言，無論是嚴格合併或是強固合併，如果我們想要忽略的部份與因子 Z 有關，其充要條件為因子 X 與 Y 中至少有一個需要與 Z 為條件獨立。除此之外，我們也發現這項結論的適用範圍還可以進一步作衍伸。我們可以理論證明，這個條件不僅僅只有在 X、Y、Z 都是類別變數的情況可以適用，當 X 與 Y 為連續型變數時，這個結論依然成立。

四、計畫成果自評

雖然這個計畫所要探討的課題，在幾年前我們曾經作過初步探討，不過由於之前對相關文獻的了解並不透徹，因此所得到的結果比較片面。藉由此計畫的執行，我們有機會再次對文獻作更仔細的研讀，並得以推導出有數學理論依據的結果。因此我想至少我們應該已經達到這個計畫最初設定的原始目標。不過在擬定計畫時，我們也曾期許能針對 Agresti (1996,2002)兩本書中提到可合併性時，所依據的主要工具 — Association Graph 來作探討。雖然我們一直覺得這一部

份應該還有相當多值得探討的空間，然而事與願違，截至目前為止，我們並無法得到具體成果，希望未來還能有機會針對這一部份繼續作了解。

五、參考文獻

1.Agresti, A (1996). *An Introduction to Categorical Data Analysis.* New York: Wiley.

2.Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

3.Bishop, Y. V. V., S. E. Fienberg, and P.W. Holland. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

4.Davis, L. J. (1986). Relationship between strictly collapsible and perfect tables. *Statistics and Probability Letter*. 4, 119-122.

5.Ducharme, G.R,. and Y. Lepage. (1986). Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society* B. 48, 197-205.

6.Long, J. S. (1984). Estimable Functions in Log-linear Models. *Sociological Method and Research*. 12,399-432.

7.Shapiro, S. H. (1982). Collapsing contingency tables: a geometric approach. *American Statistician*. 36, 43-46.

8.Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* B. 13, 238-241.

9.Whittemore, A. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society* B. 40, 328-340.

10.Wickens, T. D. (1989). *Multidimensional Contingency tables analysis for the Sciences.* Hillsdale, NJ: Lawrence Erleaum Associates.

11.賴芬秀 (1998). 「降低列聯表維數之可行性探討」，政治大學統計系碩士論文。