有關分析有序多元間隔時間資料之統計方法的發展近況

台灣大學公共衛生學院流行病學研究所　張淑惠

在長期追蹤研究中經常會觀察到研究個體發生多次相同或不同形式的事件，其所收集形成之資料型態稱爲多元事件資料。而當研究個體之多次事件發生是有時序性時，此類資料稱爲有序多元事件資料。此種資料常見於各種不同科學領域包括生物醫學、公共衛生、工業等研究。例如於研究愛滋病疾病進展過程收集有關病人從發生 HIV 感染時間、到愛滋病發病時間、以及病發後到最後死亡時間，這三個不同事件發生有一定時間先後順序，此種有序多元事件資料其一系列事件的類型、數目、與發生先後順序必須事先決定。又例如於臨床醫學研究中比較不同藥物對於治療病人膀胱腫瘤復發情況之效果，於流行病學長期調查研究中探討青少年騎乘機車重複發生摔車的影響因素，或者是探討精神分裂症患者因病情好壞而一再地反覆出入醫院的模式與相關預後因素等，這種由長期追蹤研究所收集到相同事件或一序列事件（例如：入院與出院）的重複發生時間之資料，每一個體事件發生次數是由觀察所得的結果變數並非事先決定。以上兩種類型之有序多元事件資料其共同特性爲一連串事件的發生是有時間順序，故分析此類資料的重點可以放在兩相繼事件之間隔時間或是事件發生時間的模式，然而如何選取適當模式與統計分析方法，則取決於有序多元事件的資料類型、取樣方式、以及研究目的。

一、分析有序多元事件資料的研究發展背景

對於發展分析有序多元事件資料的統計方法，不論是考慮間隔時間或是事件發生時間模式，都必須面臨處理同一個體其發生多元事件之間相關性的問題，早期如 Prentice, Williams and Peterson [1]與 Andersen and Gill [2]等人分析有序多元事件資料主要是以不同時間發生之事件彼此獨立的馬可夫過程(Markov process)與多元間隔時間彼此獨立的半馬可夫過程(semi-Markov process)等隨機過程爲基礎下，發展出以事件歷史過程爲預測因子的各種對比條件風險(conditional hazard)回歸模式，而後續研究例如 Chang and Wang [3]將間隔時間的對比條件風險回歸模式推廣並應用於病患重複發生入院與出院這二種有序事件的復發資料，而 Strawderman [4]則考慮加速間隔時間條件回歸模式於相同事件復發資料。此以預測爲主的條件回歸模式必須確知個體過去觀察到的事件歷史影響後續事件發生率之參數形式。然而，在實務應用上，同一個體內多元事件之間的相關形式可能無法確切得知，因此近年來對於有序多元事件資料相關統計方法的主要研究發展是針對各種邊際模式，探討對多元事件間之相關性具穩健性(robustness)的統計分析方法。例如 Pepe and Cai [5]，Lawless and Nadeau [6]，Lin et al. [7]，與 Wang, Qin and Chiang [8]等人針對多元有序事件(serial events)及重複事件(recurrent events)資料之事件發生率（即事件發生時間）的不同邊際模式，發展出一系列穩健的統計分析方法。然而，對於分析有序多元間隔時間模式的統計方法，與分析事件發生時間模式不同之處是在於即使當設限時間與一序列間隔時間彼此獨立時，除了第一段間隔時間之外，分析第二段或以後之間隔時間會遭遇相依設限的問題[9,10]，因此對於有序多元間隔時間需要發展不同於事件發生時間的統計分析方法。以下第二、三、四節就有限有序事件、重複事件、以及截切有限有序事件等三種不同資料型態分述其多元間隔時間的穩健統計分析方法的發展。

二、有限有序間隔時間邊際回歸模式的統計分析方法

首先以起始事件→中間事件→終止事件之三階段過程爲例，說明二元有序間隔時間的資料結構以及其所面臨誘導訊息設限的現象。以 T_1 與 T_2 分別代表從起始事件到中間事件之第一段

間隔時間與從中間事件到終止事件之第二段間隔時間，C 爲從起始事件到研究結束之右設限時間。由於有限研究時間會造成右設限取樣限制，例如當 $T_1 \le C$ 且 $T_1+T_2>C$ 時，則 T_1 可完整觀察到而 T_2 僅有訊息爲大於 $C-T_1$。故即使當 C 與(T_1, T_2)獨立時，由於 T_1 與 T_2 彼此相依，使得 T_2 與其設限時間 $C-T_1$ 並非獨立，而造成所謂誘導訊息設限(induced informative censoring)的問題，若忽略此問題而以傳統存活分析方法分析第二段間隔時間之資料，則會得到偏差的統計分析結果[10-12]。然而當分析第二次事件發生時間 T_1+T_2 時，T_1+T_2 仍與右設限時間 C 是獨立，故傳統存活分析方法仍然適用。當考慮加速間隔時間模式 $\log T_j = -\beta_j Z+\varepsilon_j$ (j=1,2)時，其二元隨機誤差(ε_1, ε_2)之聯合分布的形式未知且與解釋自變數 Z 無關，而基於此多元事件依序發生之特性，Chang [12]提出可將間隔時間 T_1 與 T_2 做適當轉換後再依序相加成爲一系列新的轉換事件發生時間：$T_1 e^{\beta_1 Z}$ 與 $T_1 e^{\beta_1 Z} + T_2 e^{\beta_2 Z}$，此時它們的聯合分布與 Z 無關，且仍與轉換之設限時間 $Ce^{\beta_1 Z}$ 與 $Ce^{\min\{\beta_1,\beta_2\}Z}$ 分別獨立。因此可利用轉換後之資料建立 β_j (j=1,2)的對數排序估計式，在 T_1 與 T_2 之相關形式未知下仍可得到 β_2 的正確估計。然而此種人爲轉換資料的方式可能會使得原來觀察到第二次事件的個案變成右設限的個案而損失原資料之部份訊息。Huang [13]則以相同的時間轉換方式提出使用資料中所有發生事件訊息的加權估計式。Huang [13]所提估計式需要 C 與(T_1,T_2)和 Z 皆獨立的假設。而 Chang [12]所提出的估計方法則可推廣至 C 與(T_1,T_2)並非獨立之相依設限的情形。

三、重複事件之間隔時間回歸模式的統計分析方法

對於重複相同事件的資料而言，事件發生次數也是結果變數之一，以 T_j 爲第 j-1 次與第 j 次事件之間隔時間，當個體在追蹤研究結束時其事件發生次數爲 m 時，即表示觀察到 m 個完整間隔時間$\{T_1, T_2, \ldots, T_m\}$與最後誘導訊息設限間隔時間 $T_{m+1}^{+} = C-(T_1+\ldots+T_m)\}$ 需滿足此取樣條件 $\sum_{j=1}^{m} T_j \le C < \sum_{j=1}^{m+1} T_j$。以臨床試驗爲例，其主要研究目的是探討以某藥物治療對延長疾病下一次復發時間的總體效果，但不同個體之復發次數可能有很大差異，故考慮所有間隔時間$\{T_j, j\ge 1\}$的加速時間的混合模式爲 $\log T_j = \alpha - \beta Z+\varepsilon_j$ (j≥1)，此時固定參數 β 爲解釋變數 Z 對所有間隔時間的總體邊際效果，潛在變數 α 則代表個體之未知特質且服從一未知分布，隨機誤差$\{\varepsilon_j, j\ge 1\}$彼此獨立且服從一未知相同分布，而$\{\varepsilon_j, j\ge 1\}$則與 α 彼此獨立。雖然在此混合模式下同一個體內的一序列間隔時間彼此獨立且具相同分布，但是在重複事件之取樣條件下，間隔時間之抽樣分布不能反映原群體之全貌。更具體的說，最後一個間隔時間 T_{m+1} 其邊際抽樣分布有偏長之趨勢，而所觀察到前 m 個完整間隔時間$\{T_1,T_2,\ldots,T_m\}$雖具相同邊際抽樣分布卻有偏短之趨勢。Chang [14]推廣 Wang and Chang [10]對間隔時間的邊際分布的估計方法，對同一個體的每段間隔時間依其觀察到的事件發生次數 m 給予不同權數 $w = \{I(m \ge 1) + I(m = 0)\} / \max\{m,1\}$（在此 $I(\cdot)$爲指標函數），以此加權方式可以得到不偏的風險集合並建立對總體效果 β 的加權對數排序估計式，在$\{T_j, j\ge 1\}$之相關性未知下仍可得到 β 的正確估計。當不同間隔時間的邊際分布不同時，由上節所提出對有序間隔時間的轉換方式與以事件發生順序作爲分層變數，則可以得到對 β 的分層對數排序估計式[14]。上述加權與分層之估計方法皆可推廣到 C 與$\{T_j, j\ge 1\}$不獨立時的相依設限情形。另外，Huang and Chen [15]針對每個間隔時間之比例邊際風險模式，採用相同加權方式來估計邊際風險的總體比值，他們的方法則無法直接推廣至當不同間隔時間的邊際分布不同的情況。

不同型態的事件也可能依序重複發生，以精神分裂症患者因疾病反覆發作而進出醫院之過程爲例，則住院天數 T_{1j} (j≥1)與病情控制後出院至下一次發作入院之間隔時間 T_{2j} (j≥1)爲第 j 段之有序二元間隔時間，可考慮 T_{1j} 與 T_{2j} 的二元加速時間之混合模式，由上節所提出對有序間隔時間的轉換方式與推廣上述估計方法，亦可得到對$\{T_{1j}, j\ge 1\}$與$\{T_{2j}, j\ge 1\}$之兩個不同總體效果的加權與分層等兩種不偏聯合估計式[14]。

四、截切有序事件資料之多元間隔時間聯合分布的無母數估計方法

目前所探討的有序多元事件資料皆是取自

某段時期發生起始事件世代的不偏隨機樣本，以三階段事件發生過程爲例，Lin, Sun and Ying [16]與 Wang and Wells [17]等人針對取樣自新發生起始事件之世代的長期追蹤資料提出對二元有序間隔時間的聯合分布的不同形式無母數估計函數，而他們估計方法均以設限時間其存活函數的倒數爲權數修正誘導訊息設限所造成的取樣偏差。而 van der Laan [18]與 Gürler [19]等人則針對取自截切樣本且無設限之資料，提出以截切時間分布的倒數來修正截切之取樣偏差，而得到二元事件時間的聯合分布的無母數估計函數。故取樣自已發生起始事件之盛行世代(prevalent cohort)的長期追蹤資料，則會同時產生左截切與右設限兩種取樣偏差，而估計二元有序間隔時間的聯合分布無法以調整風險集合的方式同時校正這兩種取樣偏差，可以左截切與右設限時間之聯合分布的倒數來修正這兩種取樣偏差。估計左截切與右設限時間之聯合分布的首要難題在於左截切時間必須小於右設限時間，而且對於三階段事件過程而言至少有兩種盛行世代的取樣方式：一是已發生起始事件且未發生中間事件之盛行世代，二是已發生起始事件且未發生終止事件之盛行世代，再加上長期追蹤會形成不同的左截切與右設限的取樣模式。因此 Chang and Tzeng [20]則針對取樣自盛行世代的二元有序間隔時間資料所可能面臨之各種左截切與右設限的取樣模式，對左截切與右設限時間之聯合分布提出不同的無母數估計方法，以不同加權方式得到二元有序間隔時間的聯合分布的一系列無母數估計函數。他們也將此加權估計方法推廣至出現其他相依且互斥競爭終止事件的情況。

五、結語

發展重複與多元有序事件資料的統計分析方法是近十年來存活分析重要研究領域之一。在實務上，不同取樣方式如雙向截切與區間設限等取樣限制會使得多元事件發生時間無法完整觀察到，或者在取樣過程中，由於世代效應或其他潛在因素使得取樣過程與多元事件發生過程可能相關而形成所謂相依取樣的情況，均會造成分析多元事件資料的難度。因此未來值得進一步探究與發展在各種取樣限制或相依取樣的情況下多元事件資料之統計分析方法。

參考文獻

[1] R. L. Prentice, B. J. Williams and A. V. Peterson, *Biometrika*, **68**, 373 (1981).

[2] P. K. Anderson and R. D. Gill, *The Annals of Statistics*, **10**, 1100 (1982).

[3] S.-H. Chang and M.-C. Wang, *Journal of the American Statistical Association*, **94**, 1221 (1999).

[4] R. L. Strawderman, *Biometrika*, **92**, 647 (2005).

[5] M. S. Pepe and J. Cai, *Journal of the American Statistical Association*, **88**, 811 (1993).

[6] J. Lawless and C. Nadeau, T*echnometrics*, **37**, 158 (1995).

[7] D. Y. Lin, L. J. Wei, I. Yang and Z. Ying, *Journal of the Royal Statistical Society*, *series B*, **62**, 711 (2000).

[8] M.-C. Wang, J. Qin and C.-T. Chiang, *Journal of the American Statistical Association*, **96**, 1057 (2001).

[9] R. D. Gelber, R. S. Gelman and A. Goldhirsch, *Biometrics*, **45**, 781 (1989).

[10] M.-C. Wang and S.-H. Chang, *Journal of the American Statistical Association*, **94**, 146 (1999).

[11] M.-C. Wang, *Statistica Sinica*, **9**, 999 (1999).

[12] S.-H. Chang, *Biometrics*, **56**, 183 (2000).

[13] Y. Huang, *Journal of the Royal Statistical Society*, series B, **64**(1), 17 (2002).

[14] S.-H. Chang, *Lifetime Data Analysis*, **10**, 175 (2004).

[15] Y. Huang and Y. Q. Chen, *Lifetime Data Analysis*, **9**, 293 (2003).

[16] D. Y. Lin, W. Sun and Z. Ying, *Biometrika*, **86**, 59 (1997).

[17] W.-J. Wang and M. T. Wells, *Biometrika*, **85**, 561 (1998).

[18] M. J. van der Laan, J*ournal of Multivariate Analysis*, **58**, 107 (1996).

[19] U. Gürler, *Journal of the American Statistical Association*, **91**, 1152 (1996).

[20] S.-H. Chang and S.-J. Tzeng, *Lifetime Data Analysis*, **12**, 53 (2006)