鑑取遺傳資料之統計分析方法

臺灣大學公共衛生學院　戴 政

截切資料(truncated data)是統計學中不完整資料的一種，它的發生來自於取樣過程中，採用了非隨機的取樣方式(nonrandom sampling)，導致群體中一部分個體必然不會被取到所致。截切資料(或稱截切樣本)因爲必然不包含群體中的一部分個體，因此不能反映群體的全貌，故若依一般統計方法去直覺地(naively)估計群體參數(parameters)，所得估計值(estimates)必然是有偏誤的(biased)。遺傳研究的一個重要方向在尋找疾病的遺傳機制，但因爲遺傳疾病盛行率(prevalence)低，故於實務上收集家庭資料時，基於成本考量，常需以非隨機取樣方式進行，增加染病個體有效樣本數，但隨之而來引發的估計偏誤統計問題，需要發展正確的校正方法來解決。以下將就遺傳學研究採用非隨機方式取樣的背景、所產生的截切資料型式以及我們所發展校正截切資料估計偏誤的統計方法，作一簡單報告。

一、截切遺傳資料產生的背景及資料型式

研究遺傳疾病的一個困難處在於疾病盛行率一般都不高，故若想由群體中按隨機取樣方式取得足夠樣本數目的染病個體必須花費大量成本，對一般個人型研究於實務上並不易行。一種更爲節省成本的取樣方法爲採取鑑取方式(ascertainment scheme)收集資料，大致可分爲三類[1]：

1. 完全鑑取(complete ascertainment)

鑑取是指透過特殊的取樣方法先鑑定出家庭成員中某一個或某幾個染病個體，然後再經由這些個體進入該家庭取樣得到其他成員。若鑑取的方式相當周延，則有可能將目標群體中的所有染病個體都鑑取到，亦即每一個被取樣到的家庭中的染病個體，在第一階段鑑定染病個體過程中，都已被鑑定出，並無須透過另一個染病的家庭成員，進入該家庭，再取得之，這樣周延的鑑取方式稱爲完全鑑取。第一階段所鑑定出的染病家庭成員稱爲首被鑑病者(probands)，在完全鑑取方法，所有群體中染病者皆爲首被鑑病者。

2. 多重鑑取(multiple ascertainment)

多重鑑取是指鑑取方法會形成一個家庭可以因染病成員會由不同管道被鑑取到，而會被重覆鑑取到一次以上，但不是周延到完全鑑取程度，亦即，被鑑取到的家庭中，會出現一位或多位首被鑑病者。

3. 單一鑑取(single ascertainment)

當鑑取過程只有一個管道可以取到首被鑑病者則每個家庭中成員因年齡、個人因素差異僅可能只有一個成員在第一階段取樣過程中被鑑定出，稱爲單一鑑取。

鑑取遺傳資料依個體區分爲染病與正常二分法收集時，所形成的資料結構爲多項分布形式。最早思索如何正確分析這種截切遺傳資料的是 Weinberg[2]，其後 Li and Mantel[3]，Davie[4]，Li[5]分別針對三種鑑取資料提出校正偏誤的統計方法，以下將說明我們對這方面問題提出的想法與作法。

二、鑑取遺傳資料之自足分組估計方法

考慮一隱性遺傳基因座，其上有二對偶基因 D(顯性)和 d(隱性)。若收集資料爲(完全)鑑取染病孩子，再進入家庭收集正常之父母(即，父母基因型均爲 Dd)，則研究目的爲分析此種家庭產生染病孩子比率是否爲 0.25。若爲 0.25 則代表該疾病有可能受到單一隱性基因座控制，故正確估計比子代染病孩子所佔比例(遺傳學上稱爲分離率)，爲必須解決之統計問題。以三個孩子家庭爲例，三個孩子家庭中染病與正常孩子在群體中的分布與觀察樣本的資料結構如表 1 所示。假設以完全鑑取方式取樣時，則三個孩子正常的家庭不會被鑑取到，亦即表中樣本數 $n_1 = 0$。估計此種完全鑑取四項分布截切資料的正規作法爲

表一　父母均正常之三個孩子鑑取家庭資料的自足分組

同胞疾病狀態(正常、染病)	群體				樣本			
	基因頻率	染病狀態		分組和	觀察數	染病狀態		分組和
		正常	染病			正常	染病	
(3,0)	p^3	p^3	0		n_1	$3n_1$	0	
(2,1)	$3p^2q$	$2p^2q$	p^2q		n_2	$2n_2$	n_2	
(1,2)	$3pq^2$	pq^2	$2pq^2$	$p^2(C_1)$	n_3	n_3	$2n_3$	$g_1=3n_1+n_2$
(0,3)	q^3	0	q^3	$2pq(C_2)$	n_4	0	$3n_4$	$g_2=2n_2+2n_3$
				$q^2(C_3)$				$g_3=n_3+3n_4$
和	1	p	q	1				

概似函數方法，但估計過程複雜，且不易推廣，Li[5]提出方法爲將所觀察到之多項分布資料(如，表 1 中三個孩子之(正常、染病)四種情形(3, 0)、(2, 1)、(1, 2)、(0, 3))依照正常與染病比例爲 p:q 重新分配爲 C_1、C_2 和 C_3 三組。依照這樣分組，當(3, 0)組不可能被鑑取到時，不只是對應(3, 0)組的 p^3 機率會被捨棄，與 p^3 同爲一組之(2, 1)組的部分 p^2q 機率也會被捨棄，只保留 C_2、C_3 組進行分離率 q 的估計，這種分組捨棄(subdivision discarding)方法可以得到與概似函數估計校正方法同樣結果，但計算更容易，且推廣至複雜情形時，概似函數難以進行，但這種方法仍可行。C_1、C_2 和 C_3 之組內正常者與染病者因爲都維持了 p 和 q 的比例，故這種分組捨棄方法 Li[5] 稱之爲自足分組法(self-contained subsets method)，簡稱 SCSM。捨棄 C_1 組後，所餘統計難度爲如何利用合併 C_2 和 C_3 組估計基因頻率，並討論估計量的特性(有偏否？變異數爲何？)。針對自足分組資料，Li[5]提出的 SCSM 研究是以最簡單的三項分布爲例子(此時只有一個自足分組可用於估計)說明截切資料估計方法，我的研究[6]將資料結構推展到一般化的多元體資料(如，表 1，此時有二個以上之自足分組)，說明當有多個自足分組可用於估計時，如何利用合併(combination)和插補(imputation)方法完成估計過程；更重要的是，Li[5]研究並未對 SCSM 作出詳盡的統計討論，我的研究則推導出 SCSM 估計量的分布，並利用這個分布導出自足分組之 SCSM 分離率估計量和變異數，證明這個估計量是有偏的(Li[5]以爲是無偏的)，但偏誤量可以估計出來。

根據 C_2 和 C_3 組內正常和染病的比例都維持 p:q，合併二組所得分離率估計量爲

$$\hat{q}=\frac{2n_3+3n_4}{(2n_2+2n_3)+(n_3+3n_4)}=\frac{(n_2+2n_3+3n_4)-n_2}{3(n_2+n_3+n_4)-n_2}$$

上式中 $\frac{(n_2+2n_3+3n_4)}{3(n_2+n_3+n_4)}$ 之分子爲樣本中所有染病孩子數，分母爲樣本中所有孩子數，此爲分離率 q 的直覺估計式(naive estimator)，爲截切資料未經調整的估計式，分子和分母同減 n_2(含單一染病孩子家庭數)後爲前述經調整截切資料偏誤的統計量。

三、鑑取遺傳資料之隨機化估計方法

在表 1 中自足分組法將三個孩子爲(正常，染病)的分布機率(p^3, $3p^2q$, $3pq^2$, q^3)重新分組爲 C_1、C_2 和 C_3 組，由於 C_1 組在鑑取過程因三個正常孩子不可能被鑑取到，而被捨棄，故可用之自足分組爲 C_2 和 C_3 組。根據 C_2 和 C_3 組內正常與染病比例爲 p:q 特性，我們[7]將這個問題推廣至一般化情形。考慮一個家庭有 s 個孩子，其中有 i 個染病孩子的發生機率爲 $\begin{bmatrix}s\\i\end{bmatrix}p^{s-i}q^i/(1-p^s)$，這個機率量可以拆解成：

$$\alpha_i=\begin{bmatrix}s-1\\i-1\end{bmatrix}p^{s-i}q^i/(1-p^s)$$

$$\beta_i=\begin{bmatrix}s-1\\i\end{bmatrix}p^{s-i}q^i/(1-p^s)$$

這裏，$\alpha_i/\beta_i=i/(s-i)$ 且 $\beta_i/\alpha_{i+1}=p/q$，$(1-p^s)$ 爲截切資料調整量。由於 $\beta_i/\alpha_{i+1}=p/q$，故 β_i

與α_{i+1}結合構成一個自足分組。定義隨機變數 V_i 爲代表 s 個孩子家庭中有 i 個染病孩子發生的(條件)隨機變數，V_i : binominal(u_i, i/s)，其中 u_i 表示這類家庭樣本數。則(U_i–V_i)和 V_{i+1} 構成對應機率 β_i 與 α_{i+1} 之自足分組；據此，若定義 W_i=(U_i–V_i)+V_{i+1}，則可以定義出一個隨機化的分離率估計式

$$\hat{q} = \begin{cases} \sum_{i=2}^{s} V_i / \sum_{i=1}^{s} W_i \text{，若 } \sum_{i=1}^{s-1} W_i > 0 \\ c \qquad \text{，若 } \sum_{i=1}^{s-1} W_i = 0 \end{cases}$$

其中，c 爲介於(0,1)之一個給定數(如，設定爲 0.5)。我們的研究說明了這個隨機化估計式可以將以往分離率估計有偏誤的情形，幾乎消除，這在理論上徹底解決以往對鑑取資料估計偏誤校正問題的困惑。

四、鑑取遺傳資料之隱參數作用探討

根據上節說明可以了解當截切遺傳資料結構爲多項分布時(上述之基因頻率估計爲此種形式)，可以 SCSM 的方法迴避概似函數估計的複雜性，且達到校正偏誤的目的，但在處理其他遺傳問題時，概似函數制式化的作法仍是最被接受的方法。例如，處理複雜分離分析問題(complex segregation analysis)，所關心的問題是親代到子代基因的分離率(segregation ratio)，若分離率符合特定遺傳定律(如，顯性、隱性等)的數值，則可驗證遺傳基因的存在。Vieland and Hodge [8]針對以概似函數估計分離率提出概似函數估計偏誤之難以解決性(intractability)的問題，他們對問題的解釋相當的含混，說理並不夠清楚。我們的研究[9]根據遺傳資料特性，將構成概似函數的參數性質區分爲：目標(target)、設計(design)和干擾(nuisance)三種；並提出參數可依據於概似函數中可明顯表達(expression)及估計與否，再區分爲顯參數(exp- licit parameters)及隱參數(implicit parameters)的觀念(表二)。我們的研究理論上闡述了概似函數估計過程之難度及偏誤性來自於隱參數的難以表達，第一階段的鑑取方法或還可以定義鑑取機率做爲顯參數來表達，但第二階段家族內的取樣方式及家族結構就無法明白定義，成爲隱參數，非常困難處理。針對隱參數的處理，我們提出了一些初步解決的構想，例如，當家族結構爲隱參數無法將之於概似函數內參數化(parameterization)，此時可以採取 (i)以觀察到家族結構爲眞實結構策略，(ii)以樣本中最大觀察家族結構爲眞實結構策略，(iii)混合可能家族結構策略來試著解決。

表二　複雜家庭概似函數參數定義

參數分類	顯參數	隱參數
目標參數	θ =分離率	–
設計參數	D_1 =鑑取方法	D_2 =家族內取樣方法
干擾參數	δ =基因頻率，顯性度，表現力	τ =眞實家族結構

五、鑑取雙胞胎遺傳資料之一致性估計方法

雙胞胎資料是探討疾病是否受到遺傳控制的一個常用研究設計，分析雙胞對資料的統計方法是以估計雙胞對之間的一致性(concordance)來進行。定義一致性測度如下：

$$\theta = P(T_2 = D_1 | T_1 = D_1) = \frac{P[(T_1, T_2) = (D_1, D_1)]}{P(T_1 = D_1)}$$

其中，T_1 和 T_2 代表二位雙胞胎個體，D_1 代表染病(D_0 代表未染病)；分子部份是雙胞對都染病的機率，分母部分是單一個體染病機率；θ 是一個條件機率，稱爲個體間一致性(casewise concordance)。以往研究能夠處理疾病狀態爲二分法(染病與正常)的鑑取雙胞對資料的一致性估計，但對疾病狀態爲多分法(如，嚴重染病，輕微染病、正常三分法)的鑑取遺傳資料如何估計一致性，一直無法突破。瓶頸難度在於：(i)多分法的雙胞對一致性定義有許多可能情形，如何定義出一個理想的一致性測度，很困難，(ii)多分法的鑑取同胞對資料很難寫出完整概似函數，如何估計一致性便顯得更不清楚。

我們[10]的研究解決了上述二個難題。對定義多分法一致性，我們先定義出二個部分一致性測度：

$$\theta_1 = P(T_2 = D_1 | T_1 = D_1)$$

$$\theta_2 = P(T_2 = D_2 | T_1 = D_2)$$

其中，D_1(輕微染病)、D_2(嚴重染病)爲三分法下染病分類(D_0 爲正常)，將θ_1 和θ_2 依不同權數合

表三　三分類性狀，不完全鑑取資料之自足分組

雙胞對資料疾病狀態(樣本數)	資料分組				
	D_2^*	D_2	D_1^*	D_1	D_0
$D_2^*D_2^*$ (m_{22D})	$2m_{22D}$	0	0	0	0
$D_2^*D_2$ (m_{22S})	m_{22S}	m_{22S}	0	0	0
$D_1^*D_1^*$ (m_{11D})	0	0	$2m_{11D}$	0	0
$D_1^*D_1$ (m_{11S})	0	0	m_{11S}	m_{11S}	0
$D_2^*D_1^*$ (m_{21D})	m_{21D}	0	m_{21D}	0	0
$(D_2^*D_1, D_2D_1^*)$ (m_{21S})	βm_{21S}	$(1-\beta)m_{21S}$	$(1-\beta)m_{21S}$	βm_{21S}	0
$D_2^*D_0$ (m_{20})	m_{20}	0	0	0	m_{20}
$D_1^*D_0$ (m_{10})	0	0	m_{10}	0	m_{10}

註：$D_2^*D_2^*$代表雙胞對染病程度均爲 D_2，且二個體均被鑑取到(成爲首被鑑病者，用*表示)，$D_1^*D_1$ 代表雙胞對染病程度均爲 D_1，而其中之一爲首被鑑病者，依此類推；符號 m 表示觀察樣本次數，下標代表對應之疾病配對；β爲同胞對 D_2D_1 中，D_2 被鑑取到($D_2^*D_1$)之比例。

併，得到個體間綜合一致性測度

$$\theta_c = a\theta_2 + (1-a)\theta_1$$

其中，$a = Q_2\pi_2/(Q_2\pi_2 + Q_1\pi_1)$ 爲加權數，Q_1 和 Q_2 是 D_1 和 D_2 的盛行率，π_1 和π_2 是描述因所採取之鑑取方法使得 D_1 和 D_2 染病者被鑑取到之鑑取機率(顯參數)。θ_c 的優點爲將鑑取機率納入一致性定義，且當處理問題只有二分法染病分類時，它可以回復到θ 定義，並且可以證明它是最大概似估計量。在估計θ_c時，我們利用自足分組法克服包含設計參數(鑑取機率π_1、π_2)在內之截切資料於概似函數分析時之難處理性(概似函數不易完整設定、一致性估計不易進行)。例如，當雙胞對資料爲非完全鑑取時，雙胞對鑑取資料之自足分組如表三。

表三中，D_2^*和 D_1^*符合一致性測度定義考慮之條件狀態(scenario)，即 T_1 爲染病者，且被鑑取到，故被選爲二自足分組。由此二自足分組估計出之一致性估計爲：

$$\hat{\theta}_c = \frac{2(m_{22D}+m_{22S}+m_{11D}+m_{11S})-(m_{22S}+m_{11S})}{(m+m_{22D}+m_{22S}+m_{11D}+m_{11S}+m_{21})-(m_{22S}+m_{11S}+m_{21S})}$$

式子中，m 爲總樣本數；分子、分母中正值部分爲未經調整之估計式，所減去部分爲對鑑取資料之調整量。

六、結　語

我們開始研究鑑取(截切)遺傳資料始自 Li[5]之後，至今有 20 年。這些研究對鑑取遺傳資料作了理論性探討[11,9]，也發展出實用的統計方法[6]，並將這些方法推展到原始問題之外[10]。這些研究成果主要來自國科會多年來長期性計畫支持，非常感謝。

參考文獻：

[1] 戴政，遺傳流行病學—基因定位之遺傳設計與分析方法，台北(藝軒)(2002)。

[2] W. Weinberg, *Arch. Rass.-u. GesBiol*, **9**, 165 (1912).

[3] C. C. Li and N. Mantel, *Am. J. Hum Genet*, **20**, 61 (1968).

[4] A. M. Davie, *Ann. Hum. Genet.*, **42**, 507 (1979).

[5] C. C. Li, *Ann. Hum. Genet.*, **50**, 259 (1986).

[6] J. J. Tai, *Genet. Epidmiol.*, **14**, 465 (1997).

[7] Y. C. Yao and J. J. Tai, *Biometrics.*, **56**, 795 (2000).

[8] V. J. Vieland and S. E. Hodge, *Am. J. Hum. Genet.*, **56**, 33 (1995).

[9] J. J. Tai and C. K. Hsiao, *Hum. Hered.*, **51**, 192 (2001).

[10] J. K. Huang and J. J. Tai, *Statistics in Medicine* (in press) (2006).

[11] Y. C. Yao and J. J. Tai, Biometrics., 56, 795 (2000).