# 第一章 簡介

## 1.1 研究背景：

近代遺傳學的發展始自孟德爾的研究，在著名的豌豆實驗中他所選取觀察的性狀（例如顏色），恰巧是由單基因所控制，遺傳的規則性也因此被發現。顯性遺傳特徵的顯現只需父母中任一方具有顯性基因即可，隱性遺傳特徵則需要父母雙方都帶有此基因，才會有外顯的特徵。在醫學研究中所感興趣的重要性狀是疾病的罹患，早期對於遺傳疾病的看法深受孟德爾遺傳律的影響，研究集中於受單基因控制的疾病。有些遺傳疾病只需要來自父母任一方帶因就會得病， 例如：血友病，裘馨型肌肉萎縮症，亨廷頓舞蹈症等。有的遺傳疾病則是由隱性基因引起，因此一對外表正常但攜帶有隱性致病基因的父母，孩子有 25%的機率傳到兩個隱性致病基因。此時需要更多代親族的資料，才能夠觀察到孟德爾分離率的進行。

近代醫學的研究方向開始對複雜的遺傳疾病（complex diseases）產生興趣，這些疾病更具普遍性包含了大腸癌、乳癌、阿滋海默症、糖尿病、精神分裂症等。家族群聚的現象反映了遺傳或是共享環境因子的作用，然而這些疾病背後的致病機制卻十分複雜，非古典遺傳定律所能解釋。生物科技的發展使得與疾病發病有關的變異基因陸續被找到或是鎖定範圍，科學家發現帶有變異基因的人有更高的機率會得病，但是患病與否卻並非絕對。這類基因能使攜帶他的人“容易罹患”某些疾病，被稱為“susceptible genes”。以乳癌為例，研究已知第 17 號染色體上的 BRCA1 基因突變和第 13 號染色體上的 BRCA2 基因突變會增加乳癌和卵巢癌發生的罹患率，而且帶有突變基因的患者發病時間的分佈比不帶有突變基因的患者似有提前的傾向。有關乳癌研究的數據可以參考文獻 Iversen et al. (2000)、Ford et al. (1998)、Easton et al.(1997)、或是 Jeffery et al. (1997)。

## 1.2 研究目標

本論文提出分析家族存活資料的統計方法，探討上述複雜疾病在發病(disease occurrence)與發病時間(age onset)的家族群聚性，即家族成員在這些變數之關聯性。我們希望藉此對這些疾病背後的遺傳機制有所瞭解。

雖然基因科技已有突破發展，然而目前因其成本昂貴，且基於對個人隱私的自主權，要取得 population-based 的基因資訊並不容易。此外對於許多複雜疾病（如精神分裂），搜尋基因的直接證據的過程也並非如乳癌一樣順利。因此即使在生物科技有革命性發展的今日，分析家族史資料以間接方式探討遺傳機制依然具有重要的應用價值與意義。

## 1.3 研究方法與文獻背景簡介

我們的研究方法採用將 bivariate survival analysis 和 cure model 結合，以探討遺傳現象展現於發病(disease occurrence)和發病年齡間(age of onset)的關聯性。

早期存活分析以單變量的探討為主，例如 Kaplan-Meier 提出的無母數估計量，Cox 提出的等比例風險模式等。到了 80 年代研究的興趣拓展到探討多個存活變數之關連（multivariate survival analysis）。文獻出現許多二維函數的無母數存活函數估計量(例如 Dabrowska, 1988; Prentice and Cai, 1992; Lin and Ying, 1993; Wang and Wells, 1997)，但是這些估計量都有一些不合理處，例如違反單調遞減的性質。另一個研究的方向採半母數的分析方法，以較具有彈性的模式描述雙變量的存活變數，所提出的推論方法亦較具穩健性。我們將採此觀點以 copula model 做為模式的假設，並以 pseudo-likelihood 的方法估計代表關聯性的參數。

傳統存活分析在近年來發展的另一個重要的變革是由假設每個人都會發生感興趣的事件，到容許“不罹病”（immune or non-susceptible）的可能性，採用這個角度的模式被稱為 cure model 或是 survival analysis with non-susceptibility。文獻中多數針對 cure model 的分析均以混合模式(mixture model)來處理免疫者存在的問

題。混合模式的配置方式，把母體分為兩群：一群是會罹病的人（susceptible），另一群是不會罹病的人(immune)。所以整個母體可以視為是這兩類人的混合（mixture）。

治癒模式的文獻多集中在單變量的討論。直到 2001 年，Chatterjee & Shih 結合雙變量存活分析和混合模式，並應用在分析具有家族群聚性的乳癌疾病資料。這篇文章討論二個部分的關聯性—發病與否和發病時間。我們將在後面章節對此文章做更詳細的回顧。

我們的研究方法也是建構二維治癒模式（bivariate cure model），有別於Chatterjee & Shih(2001)的架構是我們的研究中多考慮了“死亡”這個競爭風險。我們認為死亡對發病是無法避免的競爭風險。另一動機是在討論所謂“susceptibility”時，我們希望釐清死亡對所觀察到疾病發生率所產的混淆（confounding）現象。我們以一個例子說明：假設我們感興趣的某個疾病其發病時間集中在老年的中末期，醫學的進步使其他容易致死的疾病得到有效的醫治，人類壽命的分佈因此得以延長，此時我們或許觀察到原本感興趣疾病的發生率也增加了，可是推究背後的原因，並非是這個疾病的危險因子增加，而是壽命普遍延長之故。考慮了死亡的治癒模式可以有系統的釐清來自死亡所造成競爭風險對死亡的影響，對探於討科學現象的本質應該有所幫助。

在第二章我們做了更深入的文獻回顧，第三第四章分別為單維度與雙維度的模式建構與推論。第五章為模擬實驗，用以檢驗所提出方法在不同情形下之表現。第六章為結論。

# 第二章　文獻回顧

## 2.1　單維度混合模式

傳統存活分析隱含的假設是個體一定會發生感興趣的事件，如果在分析資料時尚有個體未發生此事件，則此個體被視為設限(censored)。若觀測的時間得以延長，被設限的觀測值比例會越來小，終至個體觀測到事件發生的機率為 1。令 $X$ 為到感興趣的事件發生的時間，傳統存活分析隱含的假設為 $\Pr(X<\infty)=1$。當感興趣事件為“死亡”時，這個假設是合理的，因為人皆有死。然而當存活分析的方法拓展到死亡以外的事件時(例如感興趣的事件是“發病與否”)，這個假設就不盡合理，除非這是人人皆會罹患的疾病。

治癒模式(cure model)假設存在有部份的觀測值為免疫(immune)，永不可能發生此事件。在混合模式的架構下，令免疫與否的指標函數為 $\tilde{D}$，當 $\tilde{D}=1$ 時代表個體會發生事件(susceptible);當 $\tilde{D}=0$ 時代表個體為免疫(immune)。令 $X$ 為發生事件的時間，若是 $\tilde{D}=1$ 則 $X<\infty$;若是 $\tilde{D}=0$ 則可以給定 $X=\infty$。因此母體為兩個群體的混合，$X$ 的分配可以表示為

$$\begin{aligned}\Pr(X>t) &= \Pr(X>t\mid\tilde{D}=1)\Pr(\tilde{D}=1)+\Pr(X>t\mid\tilde{D}=0)\Pr(\tilde{D}=0)\\ &= \Pr(X>t\mid\tilde{D}=1)\Pr(\tilde{D}=1)+\Pr(\infty>t\mid\tilde{D}=0)\Pr(\tilde{D}=0)\\ &= \Pr(X>t\mid\tilde{D}=1)\Pr(\tilde{D}=1)+\Pr(\tilde{D}=0)\ 。\end{aligned}\qquad(2.1)$$

當 $\Pr(\tilde{D}=0)=0$，則此模式回歸到傳統的存活分析，即不存在免疫者的情形。文獻的興趣通常在估計 $\Pr(\tilde{D}=0)$ 與 $\Pr(X>t\mid\tilde{D}=1)$。

免疫與否的指標函數（$\tilde{D}$）是屬二元類別式變數(binary varible)，當資料含有解釋變數時許多文獻提出以邏輯式迴歸模式(logistic regression model)描述 $\tilde{D}$ 的分配：

$$\Pr(\tilde{D}=1)=\frac{\exp(\beta' Z)}{1+\exp(\beta' Z)}\ ,$$

其中 $Z: p\times 1$ 為解釋變數。這個方向的文獻有 Farewell（1982），Kuk and Chen（1992），Larson and Dinse(1985)，Taylor(1995)…等。雖然文獻對於描述 $\Pr(\tilde{D}=1)$ 的模式頗為一致，但是對於易感染者的發病時間的分配（稱為 latency distribution），就提出了許多模式，包含完全給定母數分配的模式，或是較彈性的半母數 Cox 等比例風險模式。在 Farewell（1977，1982）的論文中就假設發病時間分佈是 exponential 分配與 Weibull 分配，而 Larson and Dinse (1985)的論文則是選擇了 Cox 模式下基底函數具有 piecewise-exponential 分配，其他利用半母數 Cox model（即不給定基底函數分配）描述發病時間的有 Kuk and Chen (1992)，Sy and Taylor (2000)與 Peng and Pear (2000)。這些文章的差異性在於對基底風險函數(baseline hazard)的處理方法不同。另外，Taylor (1995)的論文中假設發病時間與解釋變數無關並利用無母數 Kaplan-Meier estimator 來當作 $\Pr(X>t\mid\tilde{D}=1)$ 的估計值。

多數討論治癒模式的文獻都未直接提供額外資訊以區辨“暫時設限”(susceptible but temporarily censored)與“免疫者”（immune）。Farewell（1986）曾提及難以分辨“發生率低但是發生時間短”和“發生率高但是發生時間長”這兩種不同的情形，因為兩者反映出來的設限資料是一樣的。因此無母數分析往往會有無法辨識問題，然而若是多做了模式的假設(透過假設分配型態或是迴歸模式)，則這些額外資訊有助於解決無法辨識的問題，可以參考 Li et al.(2001)的文章。

## 2.2 雙維度混合模式

Chatterjee & Shih (2001)將前述混合模式的概念擴展到雙變量的問題上，並應用在分析具有家族群聚性的乳癌疾病資料，因為實證研究顯示乳癌具有明顯的家族群聚的現象。這篇文章透過混合模式的表示法考慮家族成員在（1）發病與否（incidence)；（2）發病時間（age onset)的關聯性。親屬間發病與否的關聯性以

勝算比（odds ratio）來描述；對於發病時間的聯合分配，文章採取半母數 Copula 模式的架構。我們將在以下小節先介紹 copula 模式的概念，再回顧 Chatterjee & Shih 如何將此模式運用在雙變量的發病模式中，在 2.4 節中我們回顧這篇文章的資料分析。值得一提的是在此使用“incidence”的涵意是只有可能發病，並不是一定真正觀察到疾病發生。

### 2.2.1 二維 Copula 模式簡介 – 不存在免疫者的情形

一般多變量分析常用的多元常態的機率模式並不適合處理分佈呈偏斜存活變數。Copula 模式常被用於處理高維度存活變數，以二維(bivariate)的情形而言，令 $(X_1, X_2)$ 為一對彼此相關的存活變數，定義在 $[0,\infty)^2$ 之上，其聯合分配函數為

$$F(x_1, x_2) = \Pr(X_1 \le x_1, X_2 \le x_2) \text{ ，}$$

並令其邊際分配函數為 $F_j(x_j) = \Pr(X_j \le x_j)(j = 1,2)$。令 $F_j(X_j) = U_j \quad (j = 1,2)$，我們可得

$$\begin{aligned} F(x_1, x_2) &= \Pr(X_1 \le x_1, X_2 \le x_2) \\ &= \Pr\left(F_1(X_1) \le F_1(x_1), F_2(X_2) \le F_2(x_2)\right) \\ &= \Pr\left(U_1 \le F_1(x_1), U_2 \le F_2(x_2)\right) \\ &= C\{F_1(x_1), F_2(x_2)\} \text{ ，} \end{aligned}$$

其中 $C(.,.): [0,1]^2 \to [0,1]$，可視為 $(U_1, U_2)$ 的聯合分配函數。當 $X_j (j = 1,2)$ 為連續型隨機變數時，則 $U_j = F_j(X_j)$ 具有 $U(0,1)$ 分配。此時可得

$$\begin{aligned} C(u_1, u_2) &= \Pr(U_1 \le u_1, U_2 \le u_2) \\ &= \Pr\left(X_1 \le F_1^{-1}(u_1), X_2 \le F_2^{-1}(u_2)\right) \\ &= F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right) \text{ 。} \end{aligned}$$

Sklar (1959)稱 $C(u_1, u_2)$ 為 copula model，可以視為 $(U_1, U_2)$ 的聯合分配模式。這個模式亦可以拓展到更高的 $p$ 維度如下：

$$C(u_1,\ldots,u_p)=F\left(F_1^{-1}(u_1),\ldots,F_p^{-1}(u_p)\right)。$$

值得一提的是有時候模式定義在存活函數上（如 Chatterjee & Shih 與本論文），但是基本的推導架構是一樣的。Copula models 具有廣泛的應用性，尤其在描述存活分析上。主要原因之一是可以把邊際分配的影響與關聯性（用 copula 描述）分開討論。通常研究者會將$C(u_1,u_2)$予以參數化，也就是假設$(U_1,U_2)$服從某雙維度的母數分配，此時可表示其聯合分配函數為$C_\alpha(u_1,u_2)$，參數$\alpha$衡量相關的強度，與 Kendall's tau$(\tau)$的關係為$\tau=4E\{C_\alpha(U_1,U_2)\}-1$。Kendall's tau 的範圍介於-1 與 1 之間，其值只與資料的排序有關，因此比 Pearson correlation 更為穩健。因其且不易受到尾端分佈的影響，常被用來描述兩個存活變數間的關聯性。

以下是文獻中常見的 copula 模式:

A. Clayton model (Clayton, 1978):這個模式或許是最常見的 copula 模式，可表示為:

$$C_\alpha(u,v)=\begin{cases}(u^{1-\alpha}+v^{1-\alpha}-1)^{\frac{1}{1-\alpha}} & if\ \alpha>1\\ uv & if\ \alpha=1\end{cases}。$$

原始的 Clayton 模式設定$\alpha>1$，只能容許正相關（遺傳的例子多為正相關的應用）。當$\alpha=1$ 時代表相關性為零；而$\alpha$越大代表其關聯性越強。後續文獻的修正可容許負相關$(\alpha<1)$的可能。

B. Frank's model (Frank, 1979)

$$C_\alpha(u,v)=\begin{cases}\log\left\{1-\frac{(1-\alpha^u)(1-\alpha^v)}{1-\alpha}\right\}/\log\alpha & if\ 0<\alpha<1\\ uv & if\ \alpha=1\end{cases}。$$

C. Positive stable model (Hougaard, 1986)

$$C_\alpha(u,v)=\begin{cases}\exp\left[-\left\{(-\log u)^{\frac{1}{\alpha}}+(-\log v)^{\frac{1}{\alpha}}\right\}^{\alpha}\right] & if\ 0<\alpha<1\\ uv & if\ \alpha=1\end{cases}。$$

有關 copula model 的推論，研究重點集中在當$F_j(\cdot)$的邊際分配不給定時參數$\alpha$的估計，這個問題可視為半母數的推論。主要文獻所採用的方法是依據"擬概似函

數估計方法”（pseudo-likelihood estimation）：視 $\{(F_1(X_{1i}), F_2(X_{2i}))(i=1,2,...,n)\}$ 之無母數估計量為 $\{(U_{1i}, U_{2i})\ i=1,2,...,n\}$ 的“擬觀察值”（pseudo observations）；根據資料型態建構 $(U_1, U_2)$ 的概似函數；將先前建構的觀察值代入概似函數中，即為擬概似函數。可以參考 Genest, Ghoudi & Rivest (1995)針對完整資料，Shih & Louis (1995)針對二維度右設限資料，Wang & Ding (2000)針對 bivariate current status 資料所提出的方法，均是基於同一推論概念。

2.2.2 二維混合模式 (Chatterjee & Shih，2000)

為了與之後我們提出的方法作比較，我們修正了這篇論文的符號，以利統整。令 $(\tilde{D}_1, \tilde{D}_2)$ 為姊妹免疫與否的指標函數，$(X_1, X_2)$ 代表兩者的發病時間；對免疫者 $(\tilde{D}_j = 0)$，發病時間定義為∞。延續先前在 2.1 節的討論，令 $p_j = \Pr(\tilde{D}_j = 1)$ 代表個別發病與否的機率，令 $S_j(t) = \Pr(X_j \geq t \mid \tilde{D}_j = 1)$ 為“非免疫者”之發病時間的存活函數。這篇文章討論以下兩種關聯性：

（一）發病與否（incidence）：對遺傳性的疾病，$(\tilde{D}_1, \tilde{D}_2)$ 具有相關性，可採用勝算比（odds ratio）來描述相關性的強弱:

$$\gamma = \frac{P_{11}P_{00}}{P_{10}P_{01}}\ ， \qquad (2.2)$$

其中 $P_{ij} = \Pr(\tilde{D}_1 = i, \tilde{D}_2 = j), i = 0,1,\ j = 0,1$。

（二）發病時間（age onset）：假設當姊妹都有可能染病時，其發病時間的聯合存活函數服從 copula 模式：

$$\Pr(X_1 > t_1, X_2 > t_2 \mid \tilde{D}_1 = 1, \tilde{D}_2 = 1) = C_\alpha\{S_1(t_1), S_2(t_2)\}\ ， \qquad (2.3)$$

其中 $\alpha$ 衡量姊妹均為可能罹病時之發病時間的關連性強度，與 Kendall’s tau（$\tau$）有關。

在不給定$S(t)$母數分配的假設下，文章中提出參數$P_{ij}$與$\alpha$的半母數估計方法。值得注意的是在推導的過程中做了一個隱含的假設是:當成員中的某一個人有染病，其邊際發病時間的存活函數不受另一成員染病與否所影響，數學的表示法為：

$$\Pr\left(X_j \geq t \mid \tilde{D}_j = 1, \tilde{D}_i, i \neq j\right) = \Pr\left(X_j \geq t \mid \tilde{D}_j = 1\right)。 \qquad (2.4)$$

## 2.3 Chatterjee & Shih 提出之推論方法

### 2.3.1 設限下之資料型態

在 Chatterjee 和 Shih 論文之架構下，“免疫者”是混雜在設限觀察值中，令$\delta$代表觀察到發病與否的指標函數。當免疫者存在時，一定無法觀察到發病，因此$\delta = 0$，然而這些設限資料中卻同時包含$\tilde{D} = 1$與$\tilde{D} = 0$的個體。文章定義了對所有群體均適用的發病時間$X^*$，當$\tilde{D} = 1$時$X^* = X$；當$\tilde{D} = 0$時$X^* = \infty$。令$\tilde{C}$表示設限時間，在右設限架構之下，所觀察到的變數為$\tilde{T} = X^* \wedge \tilde{C}$與$\delta = I(X^* \leq \tilde{C})$。因為觀察時間有限，所以假設$\Pr(C \leq \infty) = 1$。當存在免疫者時，可發現

$$\lim_{t \to \infty} \Pr(X^* \leq t) = \Pr(\tilde{D} = 1) < 1，$$

因此$X^*$為非合適（improper）的隨機變數。文章主要討論雙維度的情形針對家族的兩個成員（如姊妹），資料可表示為$\left\{\left(\tilde{T}_{ij}, \delta_{ij}\right), i = 1,2,...,n;\ j = 1,2\right\}$，其中$\tilde{T}_{ij} = X_{ij}^* \wedge \tilde{C}_{ij}$和$\delta_{ij} = I(X_{ij}^* \leq \tilde{C}_{ij})$。值得一提的是當$\delta_{ij} = 1$，可推得$\tilde{D}_{ij} = 1$且$\tilde{T}_{ij} = X_{ij}$；當$\delta_{ij} = 0$，可推得$\tilde{T}_{ij} = \tilde{C}_{ij}$，但無法區辨$\tilde{D}_{ij}$的值。推論的目標包含估計$P_{ij}$ $(i, j = 0,1)$與$\alpha$。為了簡化估計問題，此篇文章假設：$p_1 = p_2 = p$和$S_1(t) = S_2(t) = S(t)$，即親屬間具有可交換性，目的是簡化未知參數的個數

文章提出兩個方法，整理在以下兩節。

### 2.3.2 二階段母數估計法

假設邊際分配已知，並以 $\Pr(X_j > t \mid \tilde{D}_j = 1) = S_j(t;\lambda) = S(t;\lambda)$ 表示之。

第一階段:以母數方法估計邊際分配的參數

參考（2.1），發病時間之存活函數可表示為

$$\Pr(X_j^* > t) = \Pr(X_j^* > t \mid \tilde{D}_j = 1)\Pr(\tilde{D}_j = 1) + \Pr(X_j^* > t \mid \tilde{D}_j = 0)\Pr(\tilde{D}_j = 0)$$

$$= S(t;\lambda)p + (1-p) \text{ ，} \qquad (2.5)$$

其中 $\Pr(\tilde{D}_j = 1) = p$。在忽略家族成員間的關聯性下，（$\lambda, p$）的概似函數（$L_1$）可寫成:

$$L_1(\lambda, p) = \prod_{i=1}^{n}\prod_{j=1}^{2}\left[p\left\{S\left(\tilde{t}_{ij}-;\lambda\right) - S\left(\tilde{t}_{ij};\lambda\right)\right\}\right]^{\delta_{ij}}\left\{pS\left(\tilde{t}_{ij};\lambda\right)+1-p\right\}^{1-\delta_{ij}} \text{ ，} \qquad (2.6)$$

對 $L_1(\lambda, p)$ 求極值，可以求得 $\lambda$ 和 $p$ 的最大概似估計量。雖然姊妹間的關聯性未被考慮，但因所估計之參數只關係到邊際分配，因此估計量仍會具有一致性。

第二階段:估計關聯性的參數

先前假設姊妹都會染病的存活函數具有以下 copula 模式：

$$\Pr(X_1 \geq t_1, X_2 \geq t_2 \mid \tilde{D}_1 = 1, \tilde{D}_2 = 1) = C_\alpha(u,v) \text{ ，} \qquad (2.7)$$

其中，$u = S_1(t_1), v = S_2(t_2)$。所以令

$$\frac{\partial^2 C_\alpha(u,v)}{\partial u \partial v} = C_\alpha^{11}(u,v) \text{ ，} \frac{\partial C_\alpha(u,v)}{\partial u} = C_\alpha^{10}(u,v) \text{ ，} \frac{\partial C_\alpha(u,v)}{\partial v} = C_\alpha^{01}(u,v) \text{ 。}$$

利用文獻隱含的假設，也就是

$$\Pr\left(X_1 \geq t_1 \mid \tilde{D}_1 = 1, \tilde{D}_2 = 0\right) = \Pr\left(X_1 \geq t_1 \mid \tilde{D}_1 = 1\right) = S_1(t_1) \text{ ，} \qquad (2.8)$$

$$\Pr\left(X_2 \geq t_2 \mid \tilde{D}_1 = 0, \tilde{D}_2 = 1\right) = \Pr\left(X_2 \geq t_2 \mid \tilde{D}_2 = 1\right) = S_2(t_2) \text{ ，} \qquad (2.9)$$

可得下列的關係式：

$$\left(\delta_{i1}, \delta_{i2}\right) = (1,1) \Rightarrow H_1(u_i, v_i) = C_\alpha^{11}(u_i, v_i)P_{11} \text{ ；}$$

$$(\delta_{i1},\delta_{i2})=(1,0)\Rightarrow H_2(u_i,v_i)=\Delta u_i P_{10}+C_\alpha^{10}(u_i,v_i)P_{11}\ ;$$

$$(\delta_{i1},\delta_{i2})=(0,1)\Rightarrow H_3(u_i,v_i)=\Delta v_i P_{01}+C_\alpha^{01}(u_i,v_i)P_{11}\ ;$$

$$(\delta_{i1},\delta_{i2})=(0,0)\Rightarrow H_4(u_i,v_i)=C_\alpha(u_i,v_i)P_{11}+v_i P_{01}+u_i P_{10}+P_{00}\ ,$$

其中，$u_i=S_1(\tilde{t}_{i1}),v_i=S_2(\tilde{t}_{i2}),$

$\Delta u_i=S_1(\tilde{t}_{i1}-)-S_1(\tilde{t}_{i1}),\Delta v_i=S_2(\tilde{t}_{i2}-)-S_2(\tilde{t}_{i2})$。

第一階段所得之估計量可得$\hat{u}_i=S_1(\tilde{t}_{i1};\hat{\lambda})$與$\hat{v}_i=S_2(\tilde{t}_{i2};\hat{\lambda})$。再利用所求得的

$H_j(\hat{u}_i,\hat{v}_i)\ (j=1,2)$來建構$(\alpha,\gamma)$的擬概似函數如下:

$$L_2(\alpha,\gamma)=\prod_{i=1}^{m}\left\{H_1(\hat{u}_i,\hat{v}_i)^{(\delta_{i1}=1,\delta_{i2}=1)}H_2(\hat{u}_i,\hat{v}_i)^{(\delta_{i1}=1,\delta_{i2}=0)}H_3(\hat{u}_i,\hat{v}_i)^{(\delta_{i1}=1,\delta_{i2}=0)}H_4(\hat{u}_i,\hat{v}_i)^{(\delta_{i1}=0,\delta_{i2}=0)}\right\},\tag{2.10}$$

其中$\gamma=\dfrac{P_{11}P_{00}}{P_{10}P_{01}}$，前面已假設$P_{10}=P_{01}$，因$P_{10}+2P_{01}+P_{00}=1$，且$P_{11}+P_{01}=p_1$已在前階段被估計出來，有關$P_{ij}$的參數，只剩一個參數未知，可將其表示為$\gamma$的函數。對$L_2(\alpha,\gamma)$求極值，可以得出$\alpha$和$\gamma$的最大概似估計量。

### 2.3.3 二階段半母數估計

無母數的估計方法有別於有母數是在於對$S(t)$的分配型態不做假設。由於資料$(\tilde{T}_{ij},\delta_{ij})$呈現了典型右設限的變數型態，其中$\tilde{T}_{ij}=X_{ij}\wedge\tilde{C}_{ij}$和$\delta_{ij}=I(X_{ij}\le\tilde{C}_{ij})$，所以加上獨立設限的假設，可用以下 Kaplan-Meier method 的方式估計$\Pr(X_j^*>t)$:

$$\hat{\Pr}(X_j^*>t)=\prod_{u\le t}\left\{1-\frac{\sum_{i=1}^{n}I(\tilde{T}_{ij}=u,\delta_{ij}=1)}{\sum_{i=1}^{n}I(T_{ij}\ge u)}\right\}。\tag{2.11}$$

令$t_{\max}<\infty$為最大觀測到的發病值，$\hat{\Pr}(X^*>t_{\max})$可以做為$1-p$的估計量。經由(2.1)

移項後，可得 $\Pr(X > t \mid \tilde{D} = 1)$ 的估計量為 $\frac{\hat{\Pr}(X^* > t) - (1 - \hat{p})}{\hat{p}}$。$\Pr(X > t \mid \tilde{D} = 1)$ 的估計量代入（2.11）並對 $L_2(\alpha, \gamma)$ 求極值，可以得出 $\alpha$ 和 $\gamma$ 的最大概似估計量。

## 2.4 乳癌資料的分析

Chatterjee & Shih 在文章中分析了 Washington Ashkenazi Study (WAS) 的資料，對象是住在美國首府 Washington DC 附近約 5000 人的猶太裔 Ashkenazi 一族。根據我們在網路蒐集的資料，美國國家衛生院 National Institutes of Health (NIH)希望藉這個研究以了解乳癌的遺傳機制，資料的蒐集到 1996 年截止。這篇文章分析了 WAS 資料中的家族成員，親屬關係只含一等親(first-degree relatives) ，目標之一是想了解親屬關係究竟影響致病與否、或是發病時間、或是兩者之結合。

因為文章假設親屬間的可交換性，用來估計邊際存活函數的個體包含有來自 4,856 個不同的家庭共 13,223 位的成員。在估計雙維度的關聯性時，共用了來自 3210 個家庭中的 6769 pairs，一個家庭可能提供超過一個以上的 pairs。研究發現當未知是否罹病時，在 100 歲前發病的 cumulative incidence probability 〔即 $\Pr(X^* \le 100)$〕大約是 0.23。無母數的估計量和假設 Weibull 分配的母數估計量相當吻合，代表以 Weibull 分配描述乳癌發病時間是合理的。

在這組乳癌資料中，最晚發病的乳癌患者發病年齡是 91 歲，最老的觀測值年齡是 103 歲，在這兩個年紀之間有 178 個觀測值都未被觀察到發病，作者認為這是充份觀測時間的證據，Kaplan-Meier 曲線應該不會再有下降的可能。若此看法為真，則 91 歲可視為接近乳癌發病年齡的上限，以先前的符號 $\tau_S \approx 91$。由此可以推論 $p$ 的估計量是 0.23，換言之有 0.77 的人對乳癌是免疫的。在估計關聯性時，作者在邊際分配上選擇了母數與無母數兩種方法，在 copula 模式選擇了 Clayton、Frank、stable frailty 三種模式，共跑了六種組合。分析發現代表罹病與否的 odds ratio ($\gamma$) 的估計量相差不大，大約是介於 2.67 到 2.94 間，這是頗強的相關性。然而發病時間的關聯性，當把模式參數的估計量轉換為 Kendall's tau 的

值時，只呈現微弱的關聯性，在不同的模式下有所出入，但都介於 0.1 到 0.2 之間。

此外模擬實驗中，作者以 Clayton 模式來描述發病時間的聯合分配，並以此生成樣本數為 5000 的成對資料。以如此大數目的個數做模擬和一般文獻的做法不同。我們的猜測是若罹病的比例不高，則會被觀察到發病的觀測值會更少，當資料有極高的設限比例時，需要極大的樣本數才能對免疫率和發病時間的分配做出滿意的估計。我們之後的分析也印證了這個看法。

# 第三章 單維度模式建構與推論

## 3.1 單維度模式之建構

我們所考慮的分析架構與 Chatterjee and Shih（2001）的模式最大的差別在於我們直接考慮了來自死亡的競爭風險，而非將其視為設限的原因之一。令 $\tilde{D}=1$ 表示帶有致病基因，$\tilde{D}=0$ 則是不帶有致病基因。以下敘述模式建構的基本假設:

假設一: 一旦帶有致病基因($\tilde{D}=1$)，就存在潛在的發病時間(latent onset time) ，令此時間為 $X$ 。

假設二: 因為存在"死亡"為"發病"的競爭風險，所以帶有致病基因卻不一定會發病。令 $Y$ 代表來自死亡競爭風險的發生時間。

假設三: 不帶有致病基因的人（$\tilde{D}=0$）一定不會發病。

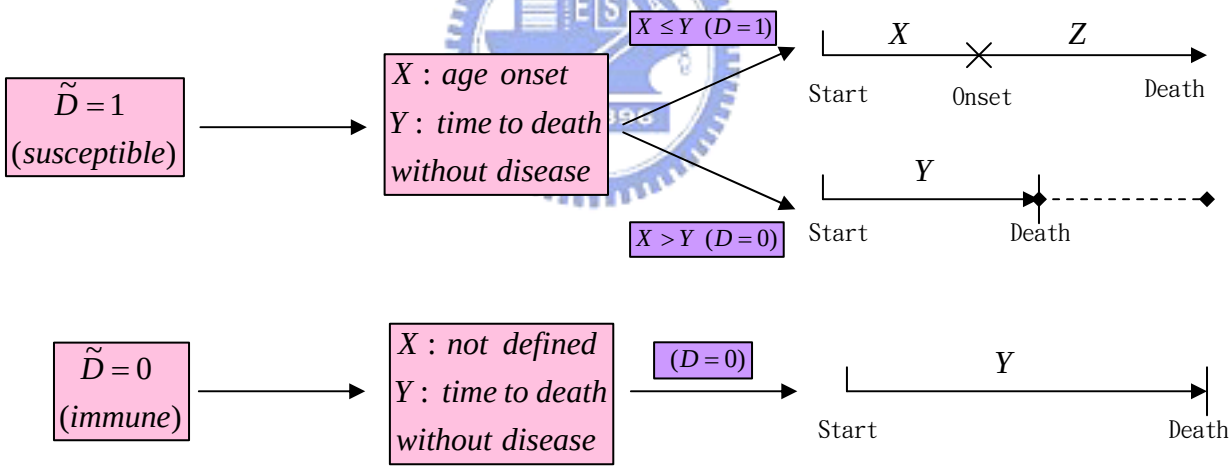


**圖 3.1:考慮死亡為競爭風險之發病模式**

在前述的架構下，一個人若是在生前發病，則 $\tilde{D}=1$ 且 $X<Y$ 。令 $D=I(\tilde{D}=1, X<Y)$ 代表在生前經歷發病的指標函數；若是 $D=0$ 代表生前未經歷發病，此時病人可能免疫($\tilde{D}=0$)或者是帶因但是發病卻被死亡設限($\tilde{D}=1, X>Y$)。在圖一中，若是一個帶因的人在生前發病($D=1$)，則發病會影響他的後續存活長

度(令其為$Z$)，此時他實際的壽命為$X+Z$，而不再是$Y$了，然而我們目前的問題對於$Z$的推估並不感興趣。值得強調的是我們這裡所提來自死亡的競爭風險，是指發病前因非關疾病的原因導致的死亡。

對於"可被觀察到發病"($\tilde{D}=1, D=1$)的群體，其發病時間存活函數為 $\Pr(X>t \mid \tilde{D}=1, X \le Y)$，這個函數會受到死亡時間分佈的影響，因此並不是真正的潛在發病時間的存活函數($\Pr(X>t \mid \tilde{D}=1)$)。試想若是醫學進步使壽命得以延長，$\Pr(X>t \mid \tilde{D}=1, X \le Y)$ 亦會因此改變，因為 Y 的分配改變了。我們的架構得以釐清 $\Pr(X>t \mid \tilde{D}=1)$ 與 $\Pr(X>t \mid \tilde{D}=1, X \le Y)$ 的差別。估計前者的難度較大，因為

$$\Pr(X>t \mid \tilde{D}=1) = \Pr(X>t \mid \tilde{D}=1, X>Y)\Pr(X>Y \mid \tilde{D}=1) +$$

$$\Pr(X>t \mid \tilde{D}=1, X \le Y)\Pr(X \le Y \mid \tilde{D}=1)，$$

其中 $\Pr(X>t \mid \tilde{D}=1, X>Y)$ 無法直接由資料推估，所以需要做額外假設才能估計 $\Pr(X>t \mid \tilde{D}=1)$。

## 3.2 單維度混合模式

在我們架構下的帶因者（$\tilde{D}=1$）仍有可能因為來自死亡的競爭風險使得無法觀察到潛在的發病時間 X，這個現象也使得估計 $\Pr(X>t \mid \tilde{D}=1)$ 具有挑戰性。我們定義變數 $X^*$ 如下：

$$X^* = X \quad if \quad \tilde{D}=1,$$

$$X^* = \infty \quad if \quad \tilde{D}=0。$$

可導出 $X^*$ 的分佈具有以下的混合型態：

$$\Pr(X^* > t) = \Pr(X^* > t \mid \tilde{D} = 1)\Pr(\tilde{D} = 1) + \Pr(X^* > t \mid \tilde{D} = 0)\Pr(\tilde{D} = 0)$$

$$= \Pr(X^* > t \mid \tilde{D} = 1)\Pr(\tilde{D} = 1) + \Pr(\infty > t \mid \tilde{D} = 0)\Pr(\tilde{D} = 0)$$

$$= \Pr(X > t \mid \tilde{D} = 1)(p) + (1 - p)\ ,\qquad (3.1)$$

其中令 $\Pr(\tilde{D} = 1) = p$，代表母體裏可致病的比例；$\Pr(\tilde{D} = 0) = 1 - p$，代表母體免疫的比例。值得一提的是 $\Pr(X^* > t \mid \tilde{D} = 1) = \Pr(X > t \mid \tilde{D} = 1)$，而且 $\lim_{t \to \infty} \Pr(X^* > t) = 1 - p$。以上的分析顯示若我們可以估出 $\Pr(X^* > t)$，則可以利用其尾端機率估計 $1 - p$，亦可以估計 $\Pr(X > t \mid \tilde{D} = 1)$。

## 3.3 資料型態與無母數估計

當下的目標為估計 $\Pr(X^* > t)$，難度在於並非對每個個體均可觀測到其 $X^*$ 值，即無法取得 $X^*$ 的隨機樣本。在沒有人為設限的情形下，可觀察到的變數為 $T = X^* \wedge Y$，與 $D = I(X^* \le Y) = I(\tilde{D} = 1, X \le Y)$，可發現 $(T, D)$ 呈現了典型右設限的變數型態。我們將於 3.6 節討論加入“人為設限”的情況下（即病人失聯或是研究時間結束造成觀察不到事件發生）。令 $\{(X_i^*, Y_i, D_i)(i = 1, ..., n)\}$ 為 $(X^*, Y, D)$ 的隨機樣本，並且令 $T_i = X_i^* \wedge Y_i$，$D_i = I(\tilde{D}_i = 1, X_i \le Y_i)$，可觀察到的樣本可表示為 $\{(T_i, D_i)\ \ (i = 1, 2, ..., n)\}$。以下我們整理了指標函數 $D_i$ 的對應情形：

$$D_i = 1 \ \Leftrightarrow \left\langle \tilde{D}_i = 1 \quad and \quad T_i = X_i^* = X_i < Y_i \right\rangle ;$$

$$D_i = 0 \ \Leftrightarrow \left\langle \tilde{D}_i = 1 \quad and \quad T_i = Y_i < X_i \right\rangle \text{ 或是}$$

$$\left\langle \tilde{D}_i = 0 \quad and \quad T_i = Y_i, X_i^* = \infty \right\rangle 。$$

我們將先假設 $X^*$ 與 $Y$ 獨立，之後再檢討這個假設的合理性。在獨立右設限資料型

態下，我們可用以下 Kaplan-Meier 的方法估計 $\Pr(X^* > t)$:

$$\hat{\Pr}(X^* > t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I(T_i = u, D_i = 1)}{\sum_{i=1}^{n} I(T_i \ge u)} \right\} 。 \qquad (3.2)$$

令 $t_{\max} < \infty$ 為最大觀測到的發病值，Maller 與 Zhou (1992)討論 $\hat{\Pr}(X^* > t_{\max})$ 做為 $1-p$ 的估計量的可行性。經由(3.1)的移項後，可得 $\Pr(X^* > t \mid \tilde{D} = 1)$ 的估計量為

$$\hat{\Pr}(X^* > t \mid \tilde{D} = 1) = \frac{\hat{\Pr}(X^* > t) - (1 - \hat{p})}{\hat{p}} \ , \qquad (3.3)$$

其中 $\hat{p} = 1 - \hat{\Pr}(X^* > t_{\max})$ 。

儘管理論上(3.3)的無母數估計量可做為帶因者之潛在發病時間分配的估計量，但是此方法依賴一個很強的假設，稱之為"充分追蹤時間"（sufficient follow-up）。以下我們討論所謂充分的追蹤的具體涵義。首先定義以下的邊界值（boundary points）:

$\tau_S = \inf\left\{ t \ge 0 : \Pr(X \le t \mid \tilde{D} = 1) = 1 \right\}$：代表最大的潛在發病時間；

$\tau_F = \inf\left\{ t \ge 0 : \Pr(Y \le t) = 1 \right\}$：代表最大的競爭風險發生的時間。

如果 $\tau_S < \tau_F$，則追蹤時間是充分的，表示"帶因者"可能發病的最晚時間小於最長壽人瑞之壽命。在此條件成立時，利用 $t_{\max}$（最大觀測到的發病值）估計 $\tau_S$，進而以 $\hat{\Pr}(X^* > t_{\max})$ 做為 $1-p$ 的估計值理論上是正確的。反之如果 $\tau_F < \tau_S$，表示即使壽命最長的人瑞，仍有可能觀察不到發病時間。此時 $\hat{\Pr}(X^* > t_{\max})$ 應為 $\Pr(X^* > \tau_F)$ 的估計值，而非 $1-p$ 的估計值。充分追蹤的條件雖然是合理的假設，但是現實很難成立，因為要納入許多高壽的人在樣本中並不容易。無母數估計的另一個缺點是 K-M 曲線在尾端的估計值具較大變異性，若再依據（3.3）的公式，將 $\hat{p}$ 置於分母以估計 $\Pr(X^* > t \mid \tilde{D} = 1)$，則後者的變異性亦會膨脹。也因此我們亦考慮利用母數分析做為推論方法。

### 3.4 獨立右設限假設合理性之討論

我們以乳癌研究為例，討論假設 $X^*$ 與 $Y$ 彼此獨立的合理性。因為所牽涉到的背景知識遠超乎我們的專業，在此我們不期望所提出的看法或是解釋是完全正確的，只希望將複雜的問題作有系統的討論，若是其中有謬誤之處，也可以藉此在基本架構之下做進一步的修正。

根據實證資料顯示，具有 BRCA1 和 BRCA2 基因突變的人，有相當高的比例會產生乳癌或是卵巢癌。其中帶有 BRCA1 基因突變的女性在 50 歲以前罹患乳癌的比例是 49%，到了 70 歲以前罹患乳癌的比例則提高到 71%；帶有 BRCA2 基因突變的女性在 50 歲以前罹患乳癌的比例是 28%，到了 70 歲以前罹患乳癌的比例則向上攀升到 84%。但是也有得病者並不帶有這兩個基因，研究也顯示這類已得病但不帶變異基因的女性在 50 歲以前罹患乳癌的比例只有 4.5%，然而有 56% 到了 70 歲前會罹病，顯示此類人的發病高峰較晚。然而具有突變基因的人也未必患病，相關資料可見附錄一。研究發現具有乳癌和卵巢癌疾病的家庭中，有 81% 的疾病家庭是帶有 BRCA1 突變,而有 14%的疾病家庭是帶有 BRCA2 突變，可見此二個變異基因能解釋了大部分的家族群聚性。以上數據整理來源：Easton et al. (1997) ，Ford et al. (1998) ，Jeffery et al. (1997）。

乳癌的實證分析使我們對影響“個體帶因與否”和“發病時間長度”的生物機制感興趣，並希望提出統計分析方法對這個問題的探討有所助益。對於得病者並不帶有 BRCA1 或是 BRCA2 這兩個已被發現的突變基因的情形，可以猜測尚有未被找到的原因，其致病的機制似乎和 BRCA1 與 BRCA2 突變基因不同。其間差異最明顯的反映在發病的年紀，帶有 BRCA1 與 BRCA2 基因突變的人傾向早發病。然而有些帶有突變基因的人卻終生未罹病，或許可以解釋成發病被死亡設限了。

我們所提出的基本假設之一是只有帶因的人($\tilde{D}=1$)才具有潛藏的發病時間 $X$，這個假設應該是合理的。在此 $\tilde{D}=1$ 的人包含帶有 BRCA1 和 BRCA2 突變的

人及尚未找到原因的帶因者。比較值得爭議的或許是以下二個假設：（1）$X^*$與$Y$獨立；（2）對$\tilde{D}=1$與$\tilde{D}=0$的兩個群體，來自死亡的競爭風險$Y$均具有相同分配，即 $\Pr(Y>t\mid\tilde{D}=1)=\Pr(Y>t\mid\tilde{D}=0)\quad\forall t>0$。我們現在檢討這兩個假設的合理性。

A. 支持的論述:即使影響致病的基因不只一個，但是其個數應該仍是相當小的。因此這些控制$\tilde{D}$的值的致病基因在 DNA 序列上所處的位置與控制死亡競爭風險(決定$Y$的值)的基因位居在非鄰近的位置的可能性很大。基於猜測兩類基因的位置相鄰遠，而假設$\tilde{D}$與$Y$互為獨立應為合理。這個觀點也支持另一假設:$Y\mid\tilde{D}=1\sim^d Y\mid\tilde{D}=0$。

B. 反對的論述:以基因的功能討論，若是部份控制$\tilde{D}=1$的基因其功能是間接和外生因素（如環境）產生交互作用，也就是要有不良的環境因此才會啟動罹病的機制。同樣的外生因素也可能同時啟動影響$Y$值的基因。換言之可能存在共有的環境因子同時影響$\tilde{D}$與$Y$的分配。此時獨立的假設就不盡合理。若令$\xi$代表共同影響的因子，或許假設"條件獨立"（$\tilde{D}\perp Y\mid\xi$）比較合理。以此做為修正的方向則需要對$\xi$如何影響$\tilde{D}$與$Y$做額外假設，而且需要知道$\xi$的分佈。透過解釋變數或許可以擷取部份$\xi$的資訊（例如不良的飲食生活習慣同時影響多種疾病），但是不可能把所有影響的可能都考慮在內，此時可能會需要以 random effect 的方法處理。因為這個修正的方向牽涉的技巧複雜，且不影響我們提出的分析架構（只影響推論方法），我們在論文中依舊假設$X^*$與$Y$獨立且$Y\mid\tilde{D}=1\sim^d Y\mid\tilde{D}=0$，至於條件獨立的方向則做為後續的研究主題。

由前述乳癌的實證研究我們發現即使是$\tilde{D}=1$的群體(susceptible population)，仍具有相當的異質性，因為致病基因不只一個，對發病的影響機制亦不見得相同。假

設有 $J$ 種致病的基因，可以將帶有第 $j$ 個基因的人以指標函數 $I(G=j)$ 表示，此時如果 $\sum_{j=1}^{J} I(G=j) \neq 0$ 則 $\tilde{D}=1$，若是 $\sum_{j=1}^{J} I(G=j)=0$，則 $\tilde{D}=0$。乳癌研究顯示 $X \mid G=i$（例如具 BRCA1 突變者）與 $X \mid G=j\,(i \neq j)$（例如具 BRCA2 突變者）的分佈不見得相同。未來如果可以取得直接基因證據或許可以估計 $\Pr(X>t \mid G=j)\ (j=1,...,J)$，也可以檢定是否 $J$ 已包含所有致病的可能性。

## 3.5 單維度有母數分析

基於無母數的估計有其限制，因此我們亦以母數模式的假設用來描述帶因者發病時間的分配。在不失一般性的前提下，我們假設 $\Pr(X>t \mid \tilde{D}=1) = \exp(-\lambda t)$，其他更合理的母數模式的分析亦可以延用類似的概念。根據資料 $\{(D_i, T_i)\ (i=1,2,...,n)\}$ 可得概似函數

$$L(p,\lambda)=\prod_{i=1}^{n}\{f_\lambda(t_i)\cdot p\}^{I(D_i=1)}\{(1-p)+S_\lambda(t_i)\cdot p\}^{I(D_i=0)}$$

其中 $S_\lambda(t_i)=\exp(-\lambda t_i)$；$f_\lambda(t_i)=-\dfrac{\partial S_\lambda(t_i)}{\partial t_i}=\lambda\exp(-\lambda t_i)$。為了簡化計算，我們利用 E-M 演算法的想法，先建構完整資料 $\{(\tilde{D}_i, D_i, T_i)\ (i=1,2,...,n)\}$ 的概似函數 $L(p,\lambda)$，其中牽涉部分未知的 $\tilde{D}_i$，我們對此概似函數先求期望值(E-step)，再對期望概似函數取極大值(M-step)。以下是利用完整資料所建構 $(p,\lambda)$ 的概似函數：

$$\begin{aligned}
&L_C(p,\lambda)\\
&=\prod_{i=1}^{n}\Big\{\Big[\Pr(\tilde{D}_i=1)\Pr(X_i=t_i\mid\tilde{D}_i=1)\Big]^{I(D_i=1,\tilde{D}_i=1)}\Big[\Pr(\tilde{D}_i=0)\Big]^{I(D_i=0,\tilde{D}_i=0)}\times\\
&\qquad\Big[\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)\Big]^{I(D_i=0,\tilde{D}_i=1)}\Big\}\\
&=\prod_{i=1}^{n}\Big\{\big[\lambda p\exp(-\lambda t_i)\big]^{I(D_i=1,\tilde{D}_i=1)}\big[1-p\big]^{I(D_i=0,\tilde{D}_i=0)}\big[p\exp(-\lambda t_i)\big]^{I(D_i=0,\tilde{D}_i=1)}\Big\}。
\end{aligned}$$

(3.4)

對 $L_C(p,\lambda)$ 取 log 函數，可得

$$l(p,\lambda)=\log L_C(p,\lambda)$$

$$=\sum_{i=1}^{n}\left\{ I(D_i=1,\tilde{D}_i=1)\left[\log p+\log\lambda-\lambda t_i\right]+I(D_i=0,\tilde{D}_i=0)\log\left[1-p\right]+\right.$$

$$\left. I(D_i=0,\tilde{D}_i=1)\left[\log p-\lambda t_i\right]\right\}。 \qquad (3.5)$$

因為隨機變數 $\tilde{D}$ 無法觀測到屬缺失值（missing data），所以提出用 EM 演算法求 $l(p,\lambda)$ 的極值。

**E-step:**在給定已觀察到資料下，對 $l(p,\lambda)$ 取條件期望值，會牽涉到以下計算：

$$E\left[I(D_i=0,\tilde{D}_i=1)\mid D_i=0,T_i=t_i\right]$$

$$=I(D_i=0)\times\frac{\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}{\Pr(\tilde{D}_i=0)+\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}$$

$$=I(D_i=0)\times w_{1i}\ ， \qquad (3.6)$$

其中 $w_{1i}=\dfrac{\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}{\Pr(\tilde{D}_i=0)+\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}$；與

$$E\left[I(D_i=0,\tilde{D}_i=0)\mid D_i=0,T_i=t_i\right]$$

$$=I(D_i=0)\times\frac{\Pr(\tilde{D}_i=0)}{\Pr(\tilde{D}_i=0)+\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}$$

$$=I(D_i=0)\times w_{2i}\ ， \qquad (3.7)$$

其中 $w_{2i}=\dfrac{\Pr(\tilde{D}_i=0)}{\Pr(\tilde{D}_i=0)+\Pr(\tilde{D}_i=1)\Pr(X_i>t_i\mid\tilde{D}_i=1)}$ 。

**M-step:** 對 $E\left[l(p,\lambda)\mid Data\right]=\breve{l}(p,\lambda)$ 求極值。將 $l(p,\lambda)$ 個別對 $p$ 與 $\lambda$ 微分，可得以下兩項微分式為：

$$\frac{\partial\breve{l}(p,\lambda)}{\partial p}=\sum_{i=1}^{n}\left\{ I(D_i=1,\tilde{D}_i=1)\left[\frac{1}{p}\right]-I(D_i=0)w_{1i}\left[\frac{1}{1-p}\right]+I(D_i=0)w_{2i}\left[\frac{1}{p}\right]\right\}，$$

$$(3.8)$$

$$\frac{\partial \breve{l}(p,\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left\{ I(D_i = 1, \tilde{D}_i = 1)[\lambda - t_i] - I(D_i = 0) w_{1i} t_i \right\} 。 \quad (3.9)$$

令微分式子等於零(即 $\frac{\partial \breve{l}(p,\lambda)}{\partial p} = 0$ & $\frac{\partial \breve{l}(p,\lambda)}{\partial \lambda} = 0$ ),可以解出兩項式子的單根(此二單根即為 $p$ 和 $\lambda$ 的估計量),分別為:

$$\hat{p} = \frac{\sum_{i=1}^{n} \left[ I(D_i = 1, \tilde{D}_i = 1) + I(D_i = 0) \times w_{1i} \right]}{n} , \quad (3.10)$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} I(D_i = 1)}{\sum_{i=1}^{n} \left[ I(D_i = 1) + I(D_i = 0) \times w_{1i} \right] t_i} 。 \quad (3.11)$$

可發現 $\hat{p}$ 的分子為樣本中估計為帶因的個數。當母數模式非指數分配時,所得之解可能不會有 explicit form,此時可藉助如牛頓法以求得數值解。

## 3.6 當存在人為設限時之修正

### 3.6.1 單維度之資料型態

先前章節中我們僅考慮來自死亡競爭風險發生所造成的設限,即假設每個觀察對象均可追蹤到死亡或者是發病。為了符合一般實際情形,我們在此章節中考慮單維度下發生"人為設限"的情形。所謂"人為設限"是指病人失聯或是研究時間結束造成觀察不到死亡或發病事件發生。令 $C$ 為設限時間。延續前述之定義,令 $X^*$ 代表個體發病的時間, $Y$ 代表競爭風險發生的時間。因為 $C$ 是一種外生的因素,因此我們假設設限時間( $C$ )、發病時間( $X^*$ )和競爭風險( $Y$ )是完全獨立(mutually independent)。當外生設限情形存在時,令可觀測到的時間為 $\tilde{T}$ ,其中 $\tilde{T} = X^* \wedge Y \wedge C$ 。令 $\left\{ (X_i^*, Y_i, \delta_i^x, \delta_i^y)(i = 1,...,n) \right\}$ 為 $(X^*, Y, \delta^x, \delta^y)$ 的隨機樣本,其中 $\delta_i^x = I(X_i^* \le Y_i \wedge C_i)$ 和 $\delta_i^y = I(Y_i \le X_i^* \wedge C_i)$ 。 可觀察到的樣本可表示為 $\{(\tilde{T}_i, \delta_i^x, \delta_i^y)\ (i = 1,2,...,n)\}$ ,其中 $\tilde{T}_i = X_i^* \wedge Y_i \wedge C_i$ 。值得一提的是我們只記錄個體所

發生最早的事件的時間，即“發病”，“死亡”或“人為設限”其中的最短的時間，但是透過指數函數可以區辨事件的種類。因此依據可觀測到的指標函數$(\delta_i^x, \delta_i^y)$的值，我們可以整理出以下三種狀況:

$$(\delta_i^x, \delta_i^y) = (1,0) \Leftrightarrow \left\langle \tilde{D}_i = 1 \quad and \quad \tilde{T}_i = X_i^* = X_i < Y_i \wedge C_i \right\rangle ;$$

$$(\delta_i^x, \delta_i^y) = (0,1) \Leftrightarrow \left\langle \tilde{D}_i = 1 \quad and \quad \tilde{T}_i = Y_i < X_i \wedge C_i \right\rangle \text{ 或是}$$

$$\left\langle \tilde{D}_i = 0 \quad and \quad \tilde{T}_i = Y_i < C_i, X_i^* = \infty \right\rangle ;$$

$$(\delta_i^x, \delta_i^y) = (0,0) \Leftrightarrow \left\langle \tilde{D}_i = 1 \quad and \quad \tilde{T}_i = C_i < X_i \wedge Y_i \right\rangle \text{ 或是}$$

$$\left\langle \tilde{D}_i = 0 \quad and \quad \tilde{T}_i = C_i < Y_i, X_i^* = \infty \right\rangle 。$$

3.6.2 外生設限下之無母數估計

前面已提及根據（3.1）可推導得$X^*$的分佈具有以下的混合型態：

$$\Pr(X^* > t) = \Pr(X > t \mid \tilde{D} = 1)(p) + (1 - p) ,$$

其中令$\Pr(\tilde{D} = 1) = p$，代表母體裏可能致病的比例。推論主要的目標為估計$\Pr(X^* > t)$，當“人為設限”的情形存在時，$X^*$受到變數$Y \wedge C$的設限，獨立設限的假設仍成立，此時新的指標函數為$\delta^x = I(\tilde{D} = 1, X^* \le Y \wedge C)$。所觀察到的資料$\{(\tilde{T}_i, \delta_i^x, \delta_i^y)\ (i = 1,2,...,n)\}$亦呈現獨立右設限的資料型態，仍可利用 Kaplan-Meier 方法估計$\Pr(X^* > t)$，其估計量如下：

$$\hat{\Pr}(X^* > t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I\left(\tilde{T}_i = u, \delta_i^x = 1\right)}{\sum_{i=1}^{n} I\left(\tilde{T}_i \ge u\right)} \right\} 。$$

令$\tilde{t}_{\max} < \infty$為最大觀測到的發病值，令$1 - \hat{\Pr}(X^* > \tilde{t}_{\max})$做為$\tilde{p}$的估計量，進而可得

$\Pr(X > \tilde{t} \mid \tilde{D} = 1)$ 的估計量為

$$\hat{\Pr}(X > \tilde{t} \mid \tilde{D} = 1) = \frac{\hat{\Pr}(X^* > \tilde{t}) - (1 - \tilde{p})}{\tilde{p}} \text{ 。}$$

以上討論可發現在單維度的情形考慮了死亡的模式和 Chatterjee 和 Shih 其實是相同的分析。值得一提的是後者考慮的設限變數其實是 $Y \wedge C$，而非僅含外生的設限變數 $C$。我們的架構直接考慮了死亡為競爭風險的好處之一是可以推估除 $\Pr(X^* > t)$ 與 $p$ 之外其他的參數，如 $\Pr(X > t \mid \tilde{D} = 1, X \le Y)$，這是在生前會發病但到 $t$ 時間尚未得病的比例。可發現

$$\Pr(X > t \mid \tilde{D} = 1, X \le Y) = \frac{S(t) - E[S(t)]}{1 - E[S(t)]} \text{ 。} \qquad (3.12)$$

令 $\tilde{S}(t) = \tilde{\Pr}(X > t \mid \tilde{D} = 1)$，則 $E[S(t)]$ 的估計量為 $\sum_{i=1}^{n} \tilde{S}(\tilde{t}_i)\hat{H}(\Delta\tilde{t}_i)$，其中 $\hat{H}(t)$ 為 $H(t) = \Pr(Y > t)$ 的 K-M 估計量：

$$\hat{H}(t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I\left(\tilde{T}_i = u, \delta_i^y = 1\right)}{\sum_{i=1}^{n} I\left(\tilde{T}_i \ge u\right)} \right\}, \qquad (3.13)$$

$\hat{H}(\Delta\tilde{t}_i)$ 為 $H(t)$ 在 $t = \tilde{t}_i$ 的估計機率值。

### 3.6.3 外生設限下之有母數估計

基於無母數的估計有其限制，我們在此考慮將帶因者的發病時間的分配，給予母數模式。令帶有致病基因者發病時間之存活函數定義為 $\Pr(X > t \mid \tilde{D} = 1) = S_{\lambda_1}(t)$，其中 $\lambda_1$ 為模式參數。另外，我們也令 $Y$ 和 $C$ 的分配函數各別為 $\Pr(Y > t) = H(t)$ 和 $\Pr(C > t) = G(t)$。為了簡化計算，我們利用 E-M 演算法的想法，先建構完整資料 $\{(\tilde{D}_i, \tilde{T}_i, \delta_i^x, \delta_i^y), (i = 1,2,...,n)\}$ 的概似函數 $\tilde{L}(p, \lambda_1)$，其中牽涉部分未知的 $\tilde{D}_i$，我們對此概似函數先求條件期望值(E-step)，再對期望概似函數取極大值(M-step)。以下是利用完整資料所建構 $(p, \lambda_1)$ 的概似函數：

$$\tilde{L}(p,\lambda_1)=\prod_{i=1}^{n}\Big\{\Big[S_{\lambda_1}(\Delta t_i)H(t_i)G(t_i)p\Big]^{I(\tilde{D}_i=1,\delta_i^x=1,\delta_i^y=1)}\times$$

$$\Big\{\Big[S_{\lambda_1}(t_i)H(\Delta t_i)G(t_i)p\Big]^{I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=1)}\times$$

$$\Big[H(\Delta t_i)G(t_i)(1-p)\Big]^{I(\tilde{D}_i=0,\delta_i^x=0,\delta_i^y=1)}\times$$

$$\Big[S_{\lambda_1}(t_i)H(t_i)G(\Delta t_i)p\Big]^{I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=0)}\times$$

$$\Big[H(t_i)G(\Delta t_i)(1-p)\Big]^{I(\tilde{D}_i=0,\delta_i^x=0,\delta_i^y=0)}\Big\}, \qquad (3.14)$$

其中 $S(\Delta t)=S(t-)-S(t)$ ， $H(\Delta t)=H(t-)-H(t)$ ， $G(\Delta t)=G(t-)-G(t)$ 。

對 $\tilde{L}(p,\lambda_1)$ 取 log 函數，可得

$$\tilde{l}(p,\lambda_1)=\sum\Big\{I(\tilde{D}_i=1,\delta_i^x=1,\delta_i^y=1)\Big[\log f_{\lambda_1}(t_i,\lambda_1)+\log(p)\Big]$$

$$I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=1)\Big[\log(S_{\lambda_1}(t_i))+\log(p)\Big]+$$

$$I(\tilde{D}_i=0,\delta_i^x=0,\delta_i^y=1)\Big[\log(1-p)\Big]+$$

$$I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=0)\Big[\log(S_{\lambda_1}(t_i))+\log(p)\Big]+$$

$$I(\tilde{D}_i=0,\delta_i^x=0,\delta_i^y=0)\Big[\log(1-p)\Big]\Big\}+\text{常數}, \qquad (3.15)$$

其中 $-\frac{\partial S_{\lambda_1}(t)}{\partial t}=f_{\lambda_1}(t)$ ， $-\frac{\partial H(t)}{\partial t}=h(t)$ ， $-\frac{\partial G(t)}{\partial t}=g(t)$ ，常數項與 $\lambda_1$ 和 $p$ 無關。因為隨機變數 $\tilde{D}$ 無法觀測到屬缺失值(missing data)，所以提出用 EM 演算法求 $\tilde{l}(p,\lambda_1)$ 的極值。

**E-step:** 需要求

$$E\Big[I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=1)\,|\,\delta_i^x=0,\delta_i^y=1,\tilde{T}_i=t_i\Big]$$

$$=I(\delta_i^x=0,\delta_i^y=1)\times\tilde{w}_{1i}\ ;$$

$$E\Big[I(\tilde{D}_i=1,\delta_i^x=0,\delta_i^y=0)\,|\,\delta_i^x=0,\delta_i^y=1,\tilde{T}_i=t_i\Big]$$

$= I(\delta_i^x = 0, \delta_i^y = 0) \times \tilde{w}_{1i}$ ，

$E\left[I(\tilde{D}_i = 0, \delta_i^x = 0, \delta_i^y = 1) \mid \delta_i^x = 0, \delta_i^y = 1, \tilde{T}_i = t_i\right]$

$= I(\delta_i^x = 0, \delta_i^y = 1) \times \tilde{w}_{2i}$ ；

$E\left[I(\tilde{D}_i = 0, \delta_i^x = 0, \delta_i^y = 0) \mid \delta_i^x = 0, \delta_i^y = 1, \tilde{T}_i = t_i\right]$

$= I(\delta_i^x = 0, \delta_i^y = 0) \times \tilde{w}_{2i}$

其中 $\tilde{w}_{1i} = \dfrac{S_{\lambda_1}(t_i)p}{S_{\lambda_1}(t_i)p + (1-p)}$ ， $\tilde{w}_{2i} = \dfrac{1-p}{S_{\lambda_1}(t_i)p + (1-p)}$ 。

**M-step:** 對 $E\left[\tilde{l}(p, \lambda_1) \mid Data\right] = \tilde{\tilde{l}}(p, \lambda_1)$ 求極值，可將 $\tilde{\tilde{l}}(p, \lambda_1)$ 個別對 $p$ 和 $\lambda_1$ 微分，可得以下兩項微分式為：

$$\frac{\partial \tilde{\tilde{l}}(p, \lambda_1)}{\partial p} = \sum_{i=1}^{n} \left\{ \left[ I(\tilde{D}_i = 1, \delta_i^x = 1, \delta_i^y = 1) + I(\delta_i^x = 0, \delta_i^y = 1)\tilde{w}_{1i} + I(\delta_i^x = 0, \delta_i^y = 0)\tilde{w}_{1i} \right](\frac{1}{p}) + \right.$$

$$\left. \left[ I(\delta_i^x = 0, \delta_i^y = 1)\tilde{w}_{2i} + I(\delta_i^x = 0, \delta_i^y = 0)\tilde{w}_{2i} \right](\frac{-1}{1-p}) \right\} ;$$

$$\frac{\partial \tilde{\tilde{l}}(p, \lambda_1)}{\partial \lambda_1}$$

$$= \frac{\partial}{\partial \lambda_1} \left\langle \sum_{i=1}^{n} \left\{ I(\tilde{D}_i = 1, \delta_i^x = 1, \delta_i^y = 1) \left[ \log(f_{\lambda_1}(t_i)) \right] + \right. \right.$$

$$\left. \left. \left[ I(\delta_i^x = 0, \delta_i^y = 1) + I(\delta_i^x = 0, \delta_i^y = 0) \right] \tilde{w}_{1i} \log(S_{\lambda_1}(t_i)) \right\} \right\rangle 。$$

對解 $\dfrac{\partial \tilde{\tilde{l}}(p, \lambda_1)}{\partial p} = 0$ 和 $\dfrac{\partial \tilde{\tilde{l}}(p, \lambda_1)}{\partial \lambda_1} = 0$ 即為 $p$ 和 $\lambda_1$ 估計量。