循時間演變的生命資料

陳珍信

生命有始有終, 卜卦算命論前世今生, 所關心的是個體的存續。統計學用於生命資料的研究, 則爲了探知族群的總體動勢。

生死之間, 衍生了與病、老機轉有關的資料。這些研究的素材, 推動著生物統計學的發展。

一、人口普查與生命統計表

直到十九世紀中葉前,「統計」都泛指國家 (State) 的政治資訊。由此我們可明瞭英文字 Statistics 的來源。羅馬帝國曾施行人口普查, 以評訂人民服兵役及繳稅之義務。我國漢朝時代也辦過人口普查來計算各地方之歲入與兵力。這自古以來的國力調查, 到了十七世紀變成所謂的「政治算術」(Political Arithmetic), 才加上現代科學的影響。當時因疫病的流行, 興起死亡率研究, 是人口學的濫觴。

十六世紀歐洲的教區登錄, 開始記載教友受洗、喪葬、結婚的日期等資料, 非正式但較常規的形成了人口統計的登錄系統。人口學始祖 John Graunt 於 1662 年出版了 Natural and Political Observations Upon the Bills of Mortality, 即利用倫敦地區約五十萬人在教堂的喪葬與受洗資料, 以死因將資料分層爲同質的子群, 再比較各子群間的死亡率差異。當然, 將受洗資料視同出生資料, 忽視了未受洗或受洗前死亡的新生兒, 是當時教區登錄裡隱藏的資料不完整性。但 Graunt 使用「分層」與「比較」的方法, 將近代統計學的原理灌注於人口資料的研究, 是重大的貢獻。

Edmund Halley 於 1693 年在英格蘭發展出「生命統計表」(Life Table), 將死亡率研究創下了嚴謹的計算方法, 幾乎沿用至今。若我們想建構的生命統計表其年齡的分隔爲 $\{a_i\}$, 就可利用 (1) 某普查年之年中 (即 6 月 30 日) 在年齡 $[a_i, a_{i+1})$ 的人口總人數, 與 (2) 當年死亡通報所得在同一年齡層的死亡總數, 計算得 $[a_i, a_{i+1})$ 年齡別死亡率爲 (1)÷(2)。要注意的是這個年齡別死亡率是個條件機率, 即爲已知活到 a_i 歲, 而死於 $[a_i, a_{i+1})$ 的機率。在生命統計表的計算中, 我們通用的是從零歲新生兒設有十萬人開始, 以條件機率乘法律來計算「從出生存活至 a_i 歲的機率」。設 T 表示存活歲數之隨機變項, 可表爲

$$\Pr(T > a_i) \;=\; \prod_{j=0}^{i-1} \{1 - \Pr(T \in [a_j, a_{j+1}) |$$

$$T > a_j)\}\,,$$

其中 $a_0 = 0$ 歲。這計算過程須假設我們有個恆定族群 (Stationary Population), 即普查年之後各年間, 每一年齡層的存歿人數皆與普查年之人數相同。其實, 上述所提的條件死亡率之計算, 在實用上還須調整出生年月日與出年年之間的差異, 就不再贅述, 因爲在此只介紹基本概念。在生命統計表中, 也計算平均餘命 (Average Remaining Lifetime)。因此, 生命統計表成爲保險精算學對保險費率及理賠額計算不可或缺的工具。

當我們再回到時光隧道, 發現英國的教區登錄直到十八世紀仍是其人口統計的資料來源。於1836年才建立政府人口登錄系統, 1874年才又將有違公定登錄的行爲之罰則立法。然而, 最古老的政府人口統計登錄系統是1628年首創於芬蘭。該世紀間, 丹麥、挪威與瑞典也紛紛建置其政府登錄系統。北歐國家至今在人口學、精算學及統計學之研究, 於國際間有其不可忽視的地位, 其來有自。

英國在北美殖民地的大城波士頓也早在同個十七世紀就陸續建立政府人口統計登錄系統與違法罰則的立法。在人口學研究逐漸融入更新、更廣義化的統計學研究的變遷時代, 據說由於關切波士頓城所蒐集人口統計資料的精確及含義, 也促成了1839年 American Statistical Association 的成立。美國統計學會對全世界統計科學的研究、教學與推廣, 有其重大的貢獻。

現代國家的政府不僅有定期的人口普查, 其他尚有工商、衛生等調查工作。愈自由民主的國家, 愈能將普查資料與學術相結合, 作爲施政的彈性規劃及調整之參考。我國衛生署對死因診斷資料的電腦化行之十多年了, 對於生命統計表的標準化已漸有績效。

二、癌症研究與設限資料

十九世紀 Adolphe Quetelet 等人開始以數學模式來解釋人口數目的變化, 譬如說用 logistic curves。一直到二十世紀初葉, 這種機制式的或其他生物決定論式的模型, 都無法顯示其說服力。但這人口學的歷史方向, 卻導致生物統計的另一傳統發展。從生死過程 (Birth-Death Processes) 的提出, 隨機過程 (Stochastic Processes) 理論的發展, 對第二次世界大戰後生物統計學在人之生、病、老、死過程的描述研究裨益甚鉅。

戰後美、英等主要國家, 除致力於經濟復甦外, 似爲掃除爭戰塗生的罪惡, 在科學研究上愈重視生命科學, 冀望從擺脫病魔的威脅, 來重拾對人類尊嚴的重視。英國政府內的 MRC (Medical Research Council) 與美國政府內的 NIH (National Institutes of Health) 等醫學研究機構, 在提昇基礎研究外, 更強調科際整合研究。逐漸吸引大量的生物統計人才, 與醫學研究者合作。對於重要疾病, 如肺結核、小兒麻痺症、癌症、心臟血管疾病及糖尿病等, 展開流行病學研究與臨床試驗。當時這些機構的許多生物統計學家後來成爲近三、四十年來領導這領域研究取向的重要學者。尤其, 聯合多所大學醫學中心的合作研究中, 生物統計學家在研究方案設計、資料蒐集與管理、分析及報告居不可或缺的

角色。近十多年來, 這些重要的研究機構對於心理疾病、愛滋病、遺傳病及老年病等慢性疾病的研究, 也不遺餘力。

從疾病自然史的演進而看, 人從出生之後可能由基因與環境的單獨或相互作用, 而產生病灶。若早期檢查, 可能先行診療, 如以子宮頸篩檢爲例, 此階段可偵測出臨床前期的原位癌, 嚴格來說, 這尙未達到癌的定義。但若不察覺, 拖到病徵出現才就醫, 這時候可能達到臨床期的侵襲癌, 這就是一般所稱的癌病了。然而, 科學研究的突破多在從現象 (即數據) 去歸納出規則。所以, 醫學研究的方向也循著與疾病自然史逆向在發展。當時面臨癌症, 醫學臨床研究極需在治療方法有所突破, 以挽救垂危的病人。

英國名統計學家 Austin Bradford Hill 自 1940 年代開始領導 MRC 的生物統計部門, 即成功的鼓吹隨機化、有對照組的臨床試驗 (Randomized Controlled Clinical Trials, 簡記爲 RCTs)。美國 NIH 在 1950 年代也全面展開 RCTs 來研究各重要疾病。以心臟病的臨床試驗而言, 統計評估的病人結果變項是各種心臟功能的測量。當時, 對癌症而言, 沒有令人滿意的療效評估指標, 延長癌症病人之生命期遂成爲很自然的評估變項。病情相若的患者從進入 RCT 後到其死亡的存活期, 即「時程」(Duration), 便成爲生物醫學統計近年來熱門的研究素材,「存活分析」(Survival Analysis) 成了重要課題。

另一方面, 從較大規模的社區研究, 跨出醫院, 去調查重要疾病的盛行率, 可估計各地區已罹病人口的分布; 另有對致病因子的病例對照研究, 以篩檢出高危險群患者, 是早先流行病學研究方式。逐漸的發現橫斷式的研究無法窺知疾病自然史。疾病各期隨致病因子在不同年齡層有極不同的新個案發生率, 所以近廿年來的流行病學研究漸改採世代追蹤或更新的病例世代研究法。對於追蹤期間皆未發病的研究對象, 在研究結束時被稱爲設限資料 (Censored Data)。因此, 原活躍於臨床試驗研究的存活分析, 也被採用於這類流行病學研究之中。

一個癌症 RCT 方案結束時, 常還有病人活著。這種病人的死亡事件尙未觀察到, 其進入試驗後的存活期也就無法完全知曉。譬如說, 甲病人在試驗方案一開始即參加此研究, 一年底即不幸死亡; 一年底時, 乙病人參加此研究, 到兩年底方案結束時仍活著; 若以追蹤時程計, 兩人皆爲一年, 但甲歿乙存, 意義大不相同。在這情況下, 甲爲完全觀察到的數據, 乙爲不完全觀察的設限數據, 因爲其眞正存活期一定大於一年, 我們稱乙被右方設限 (Rightly Censored) 在一年。若在小於一年的任何時間點計算存活機率, 甲、乙皆須歸算於分母及分子。在估計一年死亡機率時, 甲、乙皆應被歸算於分母內; 甲則在分子也出現, 但乙則不然。超過一年後, 因無法確知乙活至下一時間點, 後來的條件存活機率的計算中, 不再將乙歸算於分母, 也因不知其何時死亡, 乙也不出現於分子。將各個完全觀察到的死亡時間點切隔在時間軸上, 依上述方法可求得每個切隔區間的條件存活機率。這樣便可如同生命統計表一樣的求得在各死亡點的存活機率。

除了設限資料的存活函數估計外, 對於療效評估、預後因子的探討, 存活分析的方法論從多樣本檢定到迴歸模式的估計已發展的相當成功。工業上的可靠度研究, 與存活分析相似。但前者多採參數型存活機率函數來處理, 而後者的進展皆在處理無母數型之存活機率函數。因爲生物醫學研究的存活資料難似工業實驗之受測時間能被給予物理機轉的意義, 因此無母數統計方法在存活分析也成了主流。

其實存活分析的視野, 已被擴充到其他研究。若研究的「事件」是腫瘤治癒後的再復發 (Relapse), 則治癒後到復發的期間, 稱爲緩解時程 (Remission Duration)。這「事件」與至事件發生所經歷的時程, 猶如古典存活分析術語的「死亡」與「存活時間」之間的關係。所以近廿年來, 其在人口學、公共衛生學及社會科學的研究也更廣爲應用, 而泛稱爲「事件史分析」(Event History Analysis)。所關心的「事件」可包含結婚、生育、離婚、發病、職業災害等, 甚可跳出以「人」爲研究對象, 而研究公司破產的「事件」與其經營長短的「存活期」。

從設限的機轉而言, 也有左方設限 (Left Censoring) 與區間設限 (Interval Censoring) 的現象。如以原住民的酒癮問題調查爲例, 發現某原住民在受調查時已患酒癮, 卻記不起其酒精濫用的起始年齡, 這時這人的酒精濫用起始年齡則被調查年齡左方設限。另若某原住民在兩次調查之前次是正常而後次才被知已患酒癮, 則其酒癮起始年齡被其兩次調查時之年齡區間設限了。這兩種設限機轉的無母數分析方法, 較右方設限更難處理, 也還是近年來方法論研究仍常見的題目。

三、愛滋病研究與截缺資料

在不完整資料中, 與設限資料類似的是截缺資料 (Truncated Data)。以第一節所述歐洲教區登錄的資料而言, 在受洗前去世的新生兒沒被登錄, 也就是其存活時間被截缺了, 這類新生兒就不存在於教區登錄的檔案內, 而登錄的人之出生年月日必小於其受洗年齡。由被登錄者的出生與受洗兩個年齡, 我們可用來估計被截缺的年齡機率分布函數。

本世紀之最大病敵首推愛滋病, 由於其罹患族群無法以公開調查研究。所以, 有些潛伏期較短即發病的患者, 常就無法在世代觀察研究中被納入。若不注意這種截缺的現象, 就可能低估了愛滋病潛伏期。工業上的資料常有些測量因高於儀器可測之範圍而被截缺; 經濟上的繳稅資料因低收入者不須繳稅也有截缺的資料。近年來生物醫學統計由於愛滋病研究而興起對截缺資料的重視, 也是在對無母數統計方法的探討方面。

有趣的是, 處理左方截缺資料的無母數方法在技術上與處理右方設限的有雷同之處。在處理設限又有截缺資料的研究上, 統計問題則更複雜了。

四、結語

存活分析從古典的生命統計表演變到關心設限、截缺資料的事件史分析, 是生物統計

在這研究課題上三個世紀的學術研究之累積經驗。而亙古不變的, 縱貫整個課題的是仿如 John Graunt 的「先知預言」出現於其名著的標題: 人群生命的研究是自然科學也是社會科學研究者應共同關懷的。

參考文獻

1. Chiang, C. L. (1968). Introduction to Stochastic Processes in Biostatistics. John Wiley, New York.
2. Cox, D. R. and Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall, New York.
3. Kalbfleisch, J. D. and Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. John Wiley, New York.
4. Kruskal, W. H. and Tanur, J. M. Eds. (1978). International Encyclopedia of Statistics, Volumes 1 and 2. The Free Press, New York.
5. Woodroofe, M. (1985). "Estimating a Distribution with Truncated Data," The Annals of Statistics, 13, 163-177.

—本文作者任職於中央研究院統計科學研究所—