
第 1 章: 存活分析概論

1: Introduction to Survival Analysis

在醫學研究中, 醫師通常對於某特定疾病的人口死亡率或存活率感到興趣, 先讓我們考慮以下簡單的問題:

例 1.1. 一個關於死亡率與存活率的簡單問題

某一家醫院 在 2006 年 9 月 7 日, 有 100 位病人入院, 隔一天, 2006 年 9 月 9 日, 有 99 位病人出院, 再一天, 2006 年 9 月 10 日, 有 1 位病人仍然住院, 且當天有 1 位病人死亡, 請問 2006 年 9 月 10 日的住院死亡率為何? $\frac{1}{100}$? $\frac{1}{1}$? 其實這個問題答案, 要看你如何收集你的資料與如何定義住院死亡率!

存活分析 方法是用來研究或分析樣本所觀察到的某一段時間長度之分配. 一段時間長度通常是從一特定事件起始之 **時間原點 (起始時間點) (starting event time point, zero time point)** 直到某一特定事件發生的時間點. 例如: 從進入研究時間點, 確認診斷出癌症的時間點, 或手術日期等, 通常觀察直到死亡事件為止. 換言之, 在醫學研究中, 並非所有進入此研究的觀察對象 (通常是病人), 其被追蹤 (**follow-up**) 的時間都足夠長, 所以研究者並不能夠觀察到所有研究對象最後的結果, 及所有的特定事件發生時間點. 這樣的資料通常稱為 **設限資料 (censored data)**. 此外, 這樣的時間資料通常有高度的偏分配 (**skewed distribution**), 必須要有特殊的統計方法來分析這樣的資料.

· 2 · 1.1 存活研究與存活資料

更精確地來說, 存活分析 是對 生命時間 (失敗時間) (**life time, failure time**) 的分佈做研究. 存活分析是一個籠統定義的統計名詞, 此名詞包含分析各種正的隨機變數 (**positive random variable**) 的統計技巧. 通常, 此隨機變數的數值是一個初始事件到某些終止事件的時間, 如從出生之時間點 (治療開始的時間點) 到死亡 (疾病復發的時間點).

這類 **事件發生時間 (time-to-event)** 的資料出現在不同領域中, 譬如醫學中的存活率 (**survival rate**), 公共衛生中的死亡率 (**mortality**), 流行病學中的生命量表 (**life table**), 保險統計學及人口統計學中的生命統計資料 (**vital statistics**), 工程學中的可靠度分析 (**reliability**), 社會學中的事件歷史分析 (**event history analysis**), 市場中的消費者對特定商品購買時間, 公司企業的存活時間, 以及經濟學中的失業率等. 觀察存活的研究及存活的資料是了解存活分析最好的方法, 接下來的章節會從醫學上簡單的資料到更符合實際情況的存活資料, 一一陳述幾個假設性的例子.

1.1 存活研究與存活資料

1.1.1 粗死亡率及死亡發生比率

Crude Mortality Rate and Incidence Death Rate

在 **橫斷面研究 (cross-sectional study)**, 疾病的 **盛行率 (prevalence)** 是透過某個給定的時間點下, 受到該疾病影響的人口 **分率 (proportion)**, 以 (數學方程) 式子 (1.1.1) 表達.

$$\text{點盛行率} = \frac{\text{某給定時間點下感染疾病人數}}{\text{某給定時間點下暴露於危險因子的人口數}} \quad (1.1.1)$$

例如, 我們想在一個社區中某給定時間點 (在 1980 年 1 月 1 日) 下, 確認肺癌人數, 計算發生肺癌的分率。

風險 (風險機率) (risk) 通常是指在一段特定的時間下, 發生疾病的機率。例如在時間起始點 (1990 年 1 月 1 日), 對沒有發生肺癌疾病的一群人, 追蹤這些未發病的人, 到最後的追蹤時間為 1990 年 12 月 31 日, 在這段特定的時間中, 觀察哪些人會發生肺癌, 計算發生肺癌的機率, 我們通常稱這些在時間起始點沒有肺癌的人, 暴露在可能發生肺癌的 **風險** 之下 (**at risk**)。同樣的時間起始點, 對有肺癌的一群人, 追蹤一段特定的時間下, 計算因為肺癌而死亡的死亡率。常用來計算死亡率為 **粗死亡率 (crude mortality rate)**, 與 **死因別死亡率 (cause-specific mortality rate)**, 這可藉由盛行率的概念來定義如下式子 (1.1.2) 與 (1.1.3)

$$\text{粗死亡率} = \frac{\text{特定一段時間內, 具有某特定疾病中且死亡的人數}}{\text{特定一段時間內, 具有某特定疾病的總人數}}; \quad (1.1.2)$$

$$\text{死因別死亡率} = \frac{\text{特定一段時間內, 具有某特定疾病且因該疾病而亡的人數}}{\text{特定一段時間內, 具有某特定疾病的總人數}}. \quad (1.1.3)$$

世代研究 (cohort study) 是一種與橫斷面研究不同的研究, 我們在研究開始時間點, 確認暴露及未暴露風險因子 (**risk factor**) 的兩個群體, 經過一段時間, 比較兩組發生疾病的比率, 例如, 我們在研究開始時間點, 例如 1996 年 1 月 1 日, 確認了抽煙及不抽煙的兩個群體, 並且比較兩組在研究開始後的十年之間, 到 2005 年 12 月 31 日), 罹患肺癌的比率, 我們將這些比率視為 **累計發生率 (cumulative incidence rate, CI)**, 換言之, 這是一個在研究開始時間點之前, 沒有疾病的人, 在某段時間內罹患疾病的機率, 定義為 **風險 (risk)**, 也可稱做 **累計發生率 (cumulative incidence)**。在流行病學上,

· 4 · 1.1 存活研究與存活資料

我們通常以 **發生率 (incidence)** 表示成 **累計發生率** 的縮寫。累計發生率的概念與盛行率相似, 兩者皆以樣本的 **分率 (proportion)** 估計母體的死亡率。在醫學上, **風險 (risk)**, 通常為發生某特定疾病之 **機率, (probability)** 以樣本的 **分率** 做估計。

例 1.2. 盛行率和累計發生率

表 1.1 中, 研究者計算 1980 年 1 月 1 日, 到 12 月 31 日之間, 一個社區中肺癌的盛行率及累計發生率。在研究開始時間點 1980 年 1 月 1 日, 共有人口數為 3,500,000, 到 1980 年 12 月 31 日, 共有 96,250 肺癌病例, 其中有 1250 新發生肺癌病例。盛行率為 $\frac{96,250}{3,500,000} = 2.75\%$, 即每千人中有 27.5 人罹患肺癌; 累計發生率 $\frac{1250}{(3,500,000 - 96,250 + 1250)} = 0.362/1000$, 每年, 每千人中有 0.362 新發生肺癌病例。

表 1.1: 一個社區中肺癌的盛行率及累計發生率

人口數	3,500,000	(在研究開始時間點 1980 年 1 月 1 日)
肺癌病例數	96,250	(共有 1250 新發生肺癌病例數)
盛行率	2.75%	$96,250/3,500,000$ (在 1980 年 12 月 31 日)
累計發生率	0.362/1000	$1250/(3,500,000 - 96,250 + 1250)$ (在 1980 年中)

例 1.3. 粗死亡率 (死亡盛行率) 和累計發生率

圖 1.1 是 7 個個體自 1990 年 1 月 1 日到 12 月 31 日追蹤肺癌的時間, ● 定義為研究開始時間點, × 定義為死亡時間點, ○ 定義為最後一次追蹤時間點仍然活者, 那麼肺癌的粗死亡率 (肺癌死亡的盛行率) 可以合理表示整個資料嗎?

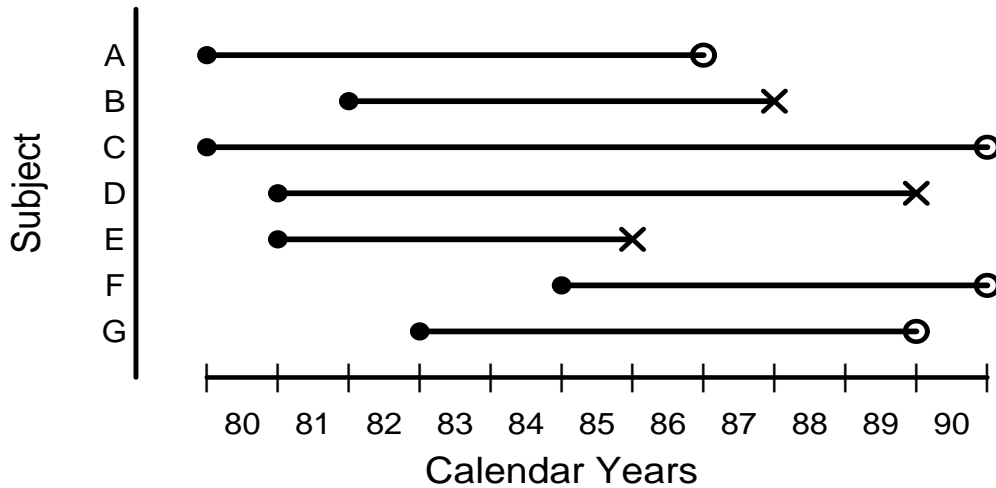


圖 1.1: 發生率與盛行率 (Incidence and Prevalence): ● initial follow-up, × death, ○ alive.

在臨床醫學研究中, 大多數接受研究的個體 (病患), 是在研究的起始一段時間內, 陸續進入研究, 在最後追蹤時間內, 有些人會失去追蹤聯繫, 因此這些人的最後資訊, 在研究結束之後, 無法使用, 每一位個體的最後追蹤時間亦不盡相同, 如上述例子 1.3. 每一位個體接受研究的時間長度, 或是最後追蹤的時間長度, 各不相同, 因而對研究的貢獻各不相同. 醫學研究上定義了 **人-時間 (person-time)**, 是爲了計算在研究中所觀察的樣本個體, 其未罹患疾病, 或未發生事件之時間的總和, 因此 **人-年 (person-years)** 是計算每一位參與研究的個體, 對研究的時間作出了不同的貢獻, 研究者以 “年” (years) 來做爲測量, 例如 人-年 可以下列式子計算.

$$\begin{aligned}
 10 \text{ person-years at risk} &= 1 \text{ people for 10 years} && (1.1.4) \\
 &= 2 \text{ people for 5 years} \\
 &= 10 \text{ people for 1 years} \\
 &= 100 \text{ people for } 1/10 \text{ years}
 \end{aligned}$$

· 6 · 1.1 存活研究與存活資料

另外在菸害的相關醫學研究中, 對吸菸量的定義, 常用於計算的是 **包-年 (pack-years)**,

$$\begin{aligned}
 1 \text{ pack-year} &= \text{每天 1 包, 共 1 年} & (1.1.5) \\
 &= \text{每天 0.5 包, 共 2 年} \\
 &= \text{每天 2 包, 共 0.5 年}
 \end{aligned}$$

發生比率, 發生密度 (incidence rate, incidence density) 定義為在確定的一時段時間內, 新發生事件 (或疾病) 的個數, 除以總共觀察到的人-時間數。

$$\text{發生比率 (Incidence Rate, I.R.)} = \frac{\text{在確定的一段時間內, 新發生疾病的個數}}{\text{觀察到的總共人-時間數}} \quad (1.1.6)$$

例 1.4. 粗死亡率, (粗死亡累計發生率), 和死亡發生比率

從圖 1.2, 可計算範例 1.3 中的“粗死亡率”, (粗死亡累計發生率) 與“死亡發生比率”。

$$\text{粗死亡率} = \frac{3}{7} = 0.428 \quad (1.1.7)$$

$$\text{死亡發生比率} = \frac{3}{7 + 6 + 11 + \cdots + 7} = \frac{3}{51} = 0.058 \quad (1.1.8)$$

發生比率數值範圍, 可以從 0 到無限大, $[0, \infty)$, 並不像 (累計) 發生率, 兩者之間的差別, 可由圖 1.2 與上述例子區分. 無論是風險 (發生危險機率), 累計發生率或發生密度之計算, 都需要注意追蹤族群人數, 發生事件的時間點, 追蹤時間之長短, 才可以正確地計算分子 (新發生事件數) 與分母 (人口數或人-時間數).

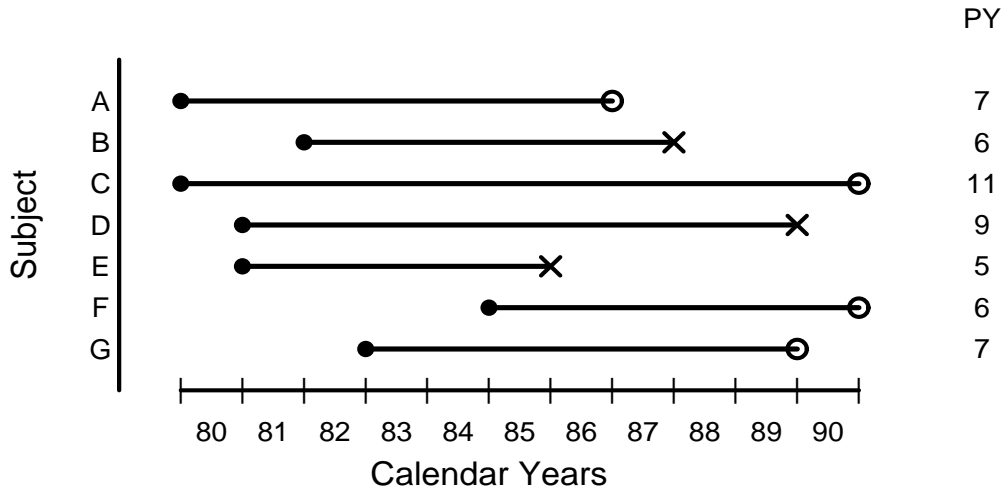


圖 1.2: 發生率與人-年 (Incidence and Person-Year): ● initial follow-up, × death, ○ alive.

例 1.5. 布瓦松 (Poisson) 分配與事件發生比率

假設 100 位最初無癌症而接受研究的個體, 追蹤這些個體直到肺癌發生, 或是失去追蹤. 假定平均的追蹤時間是 4.675 年, 並且觀察到 3 個原發性肺癌. 總共追蹤的人-年是 467.5 人-年, 因此發生比率估計為 $3/467.5$, 或者每 1000 人-年有 6.417 肺癌. 發生比率的信區間可以使用 **布瓦松分配** 計算. 在這個例子中, 我們假設在整個研究時期, 發生比率是固定的常數, 並且估計發生比率是每 1000 人年有 6.417.

流行病學研究常認為事件發生數目近似 **布瓦松機率分配 (Poisson distribution)**, 有恆定的發生比率之假設. 因分子為新發生事件或疾病的個數, 分母為總共觀察到的“人-時間”, 分子並不是分母的一部分. 這不是二項式機率分配 (binomial distribution), 不過, 因為發生比率接近於 0, 當二項式分配的發生事件機率參數接近於 0, 則這兩種機率分配近似.

但一個恆定的事件比率假設, 可能不是合理的, 或者可能與很多研究過程中, 觀察到的資料不符合. 在醫學研究過程中, 死亡或特定事件發生比率 (incidence rate), 在少數

情況下可以維持相同發生比率, 但大多述情況是, 死亡事件發生比率隨著醫學研究過程的時間增加而變化, 例如, 接受研究的個體, 隨著醫學研究過程的時間增加, 個體年紀增加, 死亡事件發生比率隨著增加. 另外, 接受研究的個體的追蹤時間, 經常是各不相同, 因此死亡事件發生比率不同, 如同在例題 1.5 中所見.

1.1.2 完整的觀察值 Complete Observations

在臨床醫學研究中, 大多數接受研究的個體或病患, 是在研究的起始一段時間內, 陸續進入研究, 在最後追蹤時間內, 多數人會失去追蹤聯繫. 讓我們先考慮完整的觀察情況.

例 1.6. 樣本數為 50 的皮膚癌動物實驗, 追蹤為期 1 年的研究

讓我們現在來考慮一個觀察時間長度為 1 年, 且樣本數為 50 的皮膚癌動物實驗, 在第 1 天讓所有動物接觸高劑量會導致皮膚癌的化學物質, 追蹤所有的動物到最後一天, 發現所有的動物皆存活, 而且有 20 隻動物有皮膚癌的病症, 現在想要估計 1 年中發生皮膚癌的發生分率 (proportion). 在第 1 年發生皮膚癌的分率可以估計為 $\hat{p} \times 100\% = 40\%$, 而 $\text{s.e.}(\hat{p}) = \pm \sqrt{\hat{p}\hat{q}/n} \times 100\% = \pm 6.9\%$, 信賴區間的近似值可由常態分配來計算. 而確切的信賴區間可利用二項分配所計算. 但是要注意的是, 如果有一半的動物在這 1 年中死於非皮膚癌的情形, 這樣的估計方法會有差異.

例 1.7. 樣本數為 50 的皮膚癌動物實驗, 追蹤動物一生命的研究

在另一個觀察樣本數為 50 的皮膚癌動物實驗, 追蹤 20 隻動物的一生 (life time), 結果有 20 隻動物在其一生命時間中, 發現有皮膚癌, 現估計皮膚癌症在生命期間的發生分率, 這結果幾乎與之前例子相同. 如果所有的觀察對象都在同一時間區間內 (interval of time) 被觀察, 則用簡單的分率近似值即可. 而在這個例子中時間的區間定義為一一生的生命時間 (life time).

例 1.8. 樣本數為 20 的兒童血癌存活時間完整追蹤研究 讓我們現在來考慮一個樣本數為 20 的兒童血癌追蹤研究, 觀察從兒童診斷出血癌到其死亡時間. 若我們有完整觀察值, 即 20 位兒童都有觀察到其死亡時間點, 其研究結果資料如表 1.2. 從診斷出血癌到其死亡時間的平均值 (mean) 為 19.7 月 (標準差 $S.D. = 15.3$), 而觀察時間範圍是從 3.1 到 51.2 個月, 其中有一半的觀察對象未滿 15 個月就死亡, 所以這時間的分配並非常態分配, 而是呈現高度偏態的分配.

表 1.2: 樣本數為 20 的兒童血癌存活時間完整追蹤研究 (單位: 月)

3.1	5.6	7.1	9.6	6.4	34.3	18.5	51.2	14.1	17.3
5.2	7.8	46.3	25.0	8.8	29.1	23.7	33.9	4.7	43.9

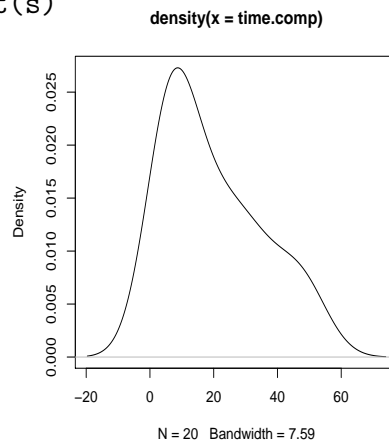
The decimal point is 1 digit(s)

to the right of the

```

0 | 35566789
1 | 0479
2 | 459
3 | 44
4 | 46
5 | 1
    
```

(a)



(b)

圖 1.3: 樣本數為 20 的兒童血癌存活時間完整追蹤研究: (a) 莖葉圖; (b) 機率密度估計圖.

相同的追蹤時間區間 (**equal follow-up interval**), 一般由固定的時間區間來定義, 例如追蹤 1 年, 但是也可以用任何 **時間單位 (time unit)** 測量, 如整個生命時間或

· 10 · 1.1 存活研究與存活資料

世代。若當所有的觀察對象在一個相同的追蹤時間區間內被觀察，如同上述例題 1.7，對在追蹤期間，對觀察到個體據有事件發生的追蹤時間，可用平均值和中位數來對事件時間作簡要的描述。對那些在生命時間區間中，追蹤“有發生癌症的個體之事件時間”而言，事件時間的平均值或許是一個可以解釋的統計量。反之，對那些僅在 1 年的追蹤時間中，追蹤“有發生癌症的個體之事件時間”而言，事件時間平均值是比較不易解釋的，另一更複雜的情形，若樣本的追蹤時間是在 1 至 5 年之間變動，則對那些 1 至 5 年之追蹤時間，追蹤“有發生癌症的個體之事件時間”而言，事件時間平均值是非常難解釋的。特別值得注意的是，不論上述何種情形，癌症發生時間的平均值和癌症發生時間的中位數對整個母體並不具有意義，這是因為並非所有的動物在這追蹤時間區間中發生癌症。

1.1.3 設限資料 Censored Data

設限資料 是由於部分觀察對象，在研究追蹤時間中途就失去追蹤而產生，或是觀察對象其追蹤時間長短不同，部分觀察對象追蹤時間短，未能觀察到特定事件結果而產生。對於一些容易解釋的統計描述及推論，設限資料會造成計算困難。

例 1.9. 一個樣本其觀察個體共 50 位，觀察的追蹤時間為 5 年，結果在表 1.3 中，你如何計算出這 50 位個體的 5 年死亡率？

表 1.3: 設限資料: 觀察 50 位個體之存活時間 (單位: 月)

人數	觀察結果
10	在 5 年之前，活著退出研究 (drop-out, alive before 5 years)
5	在 5 年研究中死亡 (die during the study)
35	在 5 年研究結束時仍然活著 (alive at 5 years)

對這些在追蹤時間內中途退出的 10 個觀察對象, 我們怎麼知道, 若他們沒有中途退出的話, 是否會在這 5 年中死亡? 若所有的觀察對象都在固定的相同觀察追蹤時間, 則我們可以在這個追蹤區間, 明確地估計出事件發生的分率 (危險分率, *proportion*); 然而, 當觀察對象在並不不同的時間區間長度, 則較適合用來估計事件發生率 (*event rate*, *risk*).

1.1.4 醫學研究中典型的存活資料

Typical Survival Data in Medical Researches

醫學研究中, 研究者通常針對某特定癌症 (例如肺癌), 自個體診斷的日期或手術的日期為時間起始點, 追蹤個體至死亡日期.

例 1.10. 日曆時間 (**Calender Time**): 從 1980 到 1990 追蹤 7 位有肺癌的個體情況

圖 1.4 顯示 1980 年 1 月 1 日至 1990 年 12 月 31 日, 長期追蹤 7 位有肺癌的個體的結果, ● 定義為研究開始時間點, × 定義為死亡時間點, ○ 定義為最後一次追蹤時間點仍然活著, 但研究者通常假設所有個體在相同日期的時間點進入研究, 並且計算自第 0 天 (時間零點, **day zero**) 至最後追蹤的時間長度.

圖 1.4 說明部份個體進入研究後一段時期的情況, 但某些人在追蹤時間內失去追蹤與聯繫, 以致於這些資訊在研究終了時無法使用, 每位個體的研究或是追蹤時間的長度不完全相同, 圖中觀測到個體 *B*, *D*, *E*, 在研究中死亡, 這些資料以日曆時間 (**calendar time**) 即日曆實際時間繪製, 因此, 研究者可以藉此圖調查某段時間內死亡人數或是存活率, 利用接下來的例子闡明此概念.

· 12 · 1.1 存活研究與存活資料

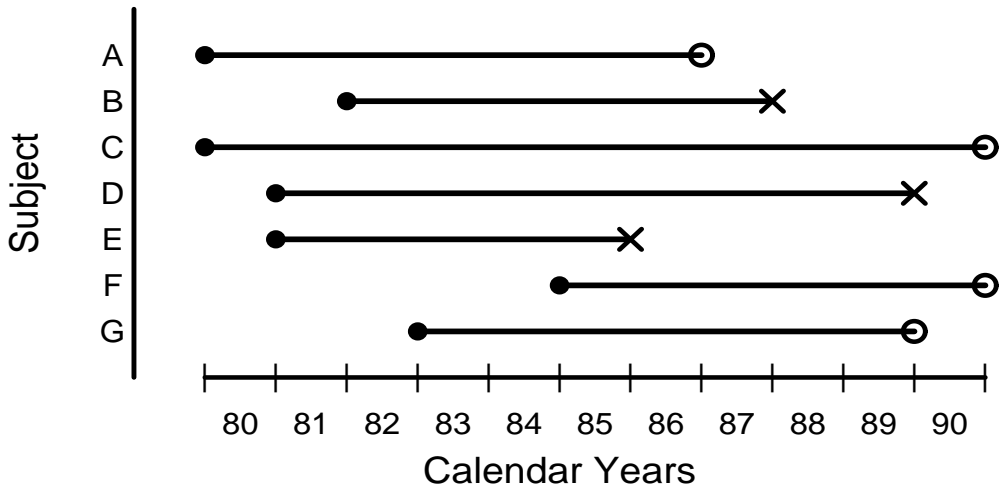


圖 1.4: 日曆時間 (Calendar Time): 從 1980 到 1990 追蹤 7 位有肺癌的個體情況. (● initial follow-up, × death, ○ alive.)

例 1.11. 病患時間: 從 1980 到 1990 追蹤 7 位有肺癌的個體情況

圖 1.5 顯示 1980 年 1 月 1 日至 1990 年 12 月 31 日, 長期追蹤 7 位有肺癌的個體的結果, ● 定義為研究開始時間點, × 定義為死亡時間點, ○ 定義為最後一次追蹤時間點仍然活著, 研究者通常假設所有個體在相同日期的時間點進入研究, 並且計算自第 0 天 (時間零點, **day zero**) 至最後追蹤的時間長度.

當個體在最後一次追蹤仍存活時, 研究者將符號 + 置於時間長度數字的旁邊, 例如: 7+, 對每位個體而言, 自個體的 **時間零點 day zero**) 測量個體的研究時間長度, 例 1.11 及 圖 1.5 中常稱此個體研究時間為 **病患時間**, 或 **病患研究時間 (patient time)**. 必須注意到繪製的方式是, 每位個體在相同的日期 (時間原點) 開始紀錄, 追蹤到個別的結束時間點, 例如個體 B, D, E, 這段時間 (病患時間) 是從時間原點到死亡時間點, ×, 也就是個體的存活時間. 然而有些個體在最後一次追蹤仍然存活, 個體的存活時間未知, 只能

紀錄部份, 則存活時間為 右設限, 通常以 ○ 符號表示. 實際上, 真實資料紀錄著每位個體進入研究的日期至個體死亡, 或是最後追蹤仍活著的日期, 存活時間可以用最適合的單位, 日, 週, 月或是年作為計算單位.

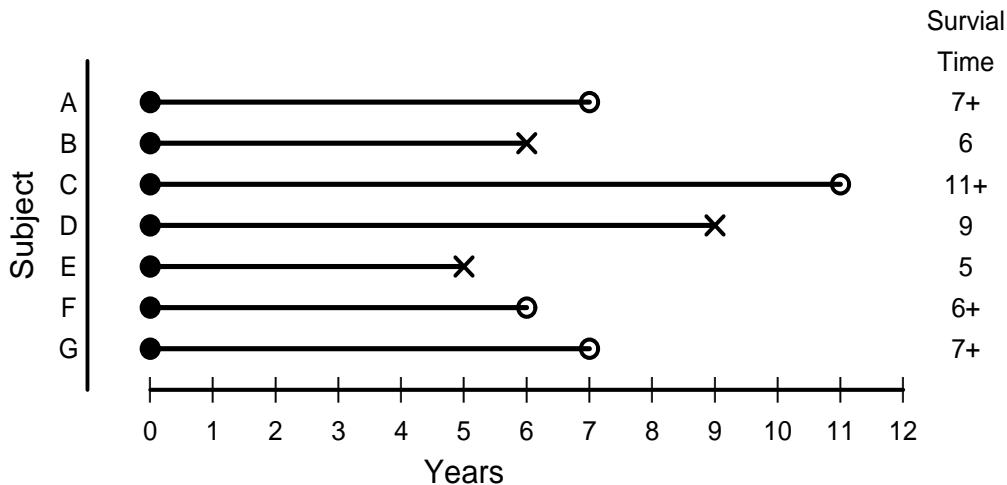


圖 1.5: 病患時間: 從 1980 到 1990 追蹤 7 位有肺癌的個體情況. (● initial follow-up, × death, ○ alive.)

1.2 存活資料的真實例子

接下來的章節中, 利用幾個真實研究資料, 且將其簡化改編而成的例題資料, 說明存活資料的本質, 並在第二章以後的章節中, 利用這些例子的資料數據, 闡述一些存活分析的統計技巧.

1.2.1 Gehan 急性白血病臨床試驗

例 1.12. Gehan 急性白血病臨床試驗

Remission Duration from a Clinical Trial for Acute Leukemia

Freireich 等 (1963) 和 Gehan (1965) 報告一個關於 6-MP (6-mercaptopurine) 的臨床試驗, 總共有 21 組配對, 42 位孩童, 治療分組的方式是每一配對內隨機分配的, 一位為 6-MP 新藥組, 一位為對照組 (使用安慰劑). 研究追蹤病人接受治療到白血病復發之間的緩解時間 (remission duration), 資料呈現於表 1.4, 而資料檔案為 **survGehan6MP.xls**. 測量時間單位為月, 若最後追蹤為死亡時, 設限指標 (censor indicator) 定義為 1, 若最後追蹤為存活, 設限指標則定義為 0, 且存活時間加註 +, 本研究主要的目的是研究這兩種治療, 對於急性白血病緩解存活率是否有差異.

表 1.4: Gehan 急性白血病臨床試驗: 比較 6-MP 組與對照組的緩解時間

配對	緩解狀態	存活時間	存活時間	配對	緩解狀態	存活時間	存活時間
1	1	1.0	10.0	11	0	2.0	11.0+
2	0	22.0	7.0	12	1	5.0	20.0+
3	0	3.0	32.0+	13	0	4.0	19.0+
4	0	12.0	23.0	14	0	15.0	6.0
5	0	8.0	22.0	15	0	8.0	17.0+
6	1	17.0	6.0	16	1	23.0	35.0+
7	0	2.0	16.0	17	1	5.0	6.0
8	0	11.0	34.0+	18	0	11.0	13.0
9	0	8.0	32.0+	19	0	4.0	9.0+
10	0	12.0	25.0+	20	0	1.0	6.0+
				21	0	8.0	10.0

1.2.2 白血病患者的白血球數與死亡時間

例 1.13. 白血病患者的白血球數與死亡時間

Feigl 與 Zelen (1965) 觀察白血病患者之存活, 並記錄如表 1.5. 研究所記錄的死亡時間, 以“週”為測量時間單位, 個體分成 2 組, AG+ 與 AG-, 研究同時記錄著個體在觀察時間零點時的白血球數量 (WBC). 這研究中, 所有個體皆死亡, 每一位個體紀錄著 2 個解釋變數, 其中 AG 離散型解釋變數, AG+ 表示病患在診斷時骨髓中出現 Auer rod 標記的白血球, 兩個不同組別的分組, AG+ 與 AG-, 並非隨機方式配置. 而解釋變數 WBC 是連續型解釋變數. 且每位個體的連續型解釋變數 WBC, 是非控制下的測量. 研究的目的是觀察與分析 2 個不同組別與 WBC, 對於存活時間的影響. 資料檔案為 `survLeukWBC.xls` (Cox, 1984).

表 1.5: 白血病患者的白血球數與死亡時間

AG+ (N = 17)			AG- (N = 16)		
病患	WBC	死亡時間	病患	WBC	死亡時間
1	2300	65	18	4400	56
2	750	156	19	3000	65
3	4300	100	20	4000	17
4	2600	134	21	1500	7
5	6000	16	22	9000	16
6	10500	108	23	5300	22
7	10000	121	24	10000	3
8	17000	4	25	19000	4
9	5400	39	26	27000	2
10	7000	143	27	28000	3
11	9400	56	28	31000	8
12	32000	26	29	26000	4
13	35000	22	30	21000	3
14	100000	1	31	79000	30
15	100000	1	32	100000	4
16	52000	5	33	100000	43
17	100000	65			

1.2.3 美國退伍軍人醫院肺癌臨床試驗

例 1.14. 美國退伍軍人醫院肺癌臨床試驗 VA Lung Cancer Clinical Trial

Prentice (1973) 報告一個研究, 關於年長退伍軍人罹患肺癌, 且無法接受手術之臨床試驗, 病患在榮民醫院隨機接受標準治療方法或新的化學治療方法, 比較治療的主要結果變數為死亡時間, 變數名稱列在表 1.6, 資料檔案為 `survVATrial.xls`. 研究者主要是想要比較兩種治療的存活機率. 這是一個隨機分配的臨床試驗, 病患治療分組是隨機分配, 但其它變數, 如診斷時的年齡, 細胞型態, 診斷時身體狀態表現的分數都可能影響存活結果.

表 1.6: 美國退伍軍人醫院肺癌臨床試驗

變數	描述
<code>treat (therapy)</code>	治療組別: 0 = 標準; 1 = 新治療.
<code>cellcode (cell)</code>	細胞型態; 1 = 鱗狀細胞; 2 = 小細胞; 3 = 腺體細胞; 4 = 大細胞.
<code>time (t)</code>	存活時間, 診斷日期至死亡日期, 單位以日計.
<code>censor</code>	設限狀態: 0 = 設限; 1 = 死亡.
<code>KPS (status) (x₁)</code>	Karnofsky performance score, 診斷時身體狀態表現的分數.
<code>diagtime (mfd) (x₂)</code>	診斷到隨機分配的時間, 以月計.
<code>age (x₃)</code>	診斷時的年齡 (以年計).
<code>prior (x₄)</code>	先前是否接受治療; 0 = 無; 1 = 有.

1.2.4 非小細胞肺癌手術後存活研究

例 1.15. VGH 第一期非小細胞肺癌手術後存活研究

Wu 等 (2003) 發表的第一期非小細胞肺癌回溯性研究中, 評估第一期肺癌手術後追蹤結果. 研究者收集從 1970 到 1995 年, 在台北榮民總醫院之病患罹患第一期非小細胞肺癌, 共有 300 位接受完整的外科切除手術, 而表 1.7 列出其中 7 位病患部分研究結果. 變數 **OPDATE** 是手術日期, 變數 **FUDATE** 是死亡日期或最後一次追蹤日期, 變數 **FUMO (FUDAY)** 是手術日期至死亡日期或最後一次追蹤日期的存活時間, 以月份 (日) 作為測量單位, 設限變數 **DEATH** 為 1 表示該個體死亡, 2, 3, 4, 則表示個體死於其他原因, 5 表示個體失去追蹤, 0 則表示該個體在最後一次追蹤時仍然存活, 實際上的真實的資料, 從個體進入研究日期開始紀錄, 直到個體死亡或是最後一次追蹤仍存活的日期結束, 存活時間可以用任何適合的單位, 日, 週, 月或是年作為計算單位, 死亡原因也可以有許多不同類型. 此研究主要目的是對於整體存活 (overall survival) 和癌症存活 (cancer-specific) 狀況做追蹤分析; 研究之次要目的, 在於探查手術中組織內移除的總淋巴結數目 (LND), 以及其他的解釋變數與共變異數, 是否會影響病人的存活時間. 部分的資料存於檔案 **survVGHlungca95300.xls**, 表 1.8 則列出變數的敘述.

表 1.7: VGH 第一期非小細胞肺癌手術後存活研究: 7 位個體部分研究結果

Subject	DEATH	OPDATE	FUDATE	FUMO	FUDAY
1	2	1980/08/02	1985/05/11	58.10	1743
2	3	1980/09/01	1993/04/15	153.63	4609
3	3	1980/09/03	1980/11/14	2.40	72
4	1	1980/09/15	1981/04/05	6.73	202
5	0	1980/10/14	1995/03/04	175.13	5254
6	1	1980/10/14	1989/02/08	101.30	3039
7	2	1980/12/11	1983/09/22	33.83	1015

表 1.8: VGH 第一期非小細胞肺癌手術後存活研究: 變數說明

變數	描述
age	年紀 (歲數以年為單位) (age)
sex	性別: 1 = 男 (Male); 2 = 女 (female). (Gender)
cough	肺癌診斷時是否有咳嗽: 0 = no; 1 = yes. (Cough at diagnosis)
smoke	抽菸量: 以“包-年”為單位 (pack-years). (Smoking amount)
hemopty	肺癌診斷時是否有咳血: 0 = no; 1 = yes. (Hemoptysis at diagnosis)
chestpain	肺癌診斷時是否有胸痛: 0 = no; 1 = yes. (Chest pain at diagnosis)
wtloss	肺癌診斷時是否有體重減輕: 0 = no; 1 = yes. (Body weight loss)
surgeonid	執行手術醫師: 1, 2, 3, 4, 6, 9. (Surgeon indicator)
opdate	手術日期 (Date of operation)
cell	細胞種類: 1 = 扁平上皮 squamous cell; 2 = 其他 others. (Cell type)
tumorsize	腫瘤大小吋吋: 公分. (Tumor size measured in cm)
lnd	移除的總淋巴結數. (Total number of the removed lymphoid nodes)
metadate	診斷腫瘤轉移日期. (Date of metastasis)
fudate	最後觀察日期. (Date of death or date of last follow-up visit)
death	最後觀察的狀況: 0 = 存活; 1 = 死於肺癌; 2, 3, 4 = 死於其他原因; 5 = 失去追蹤. (Censor status at last visit)

1.2.5 心臟移植存活時間

例 1.16. 心臟移植存活時間 Survival Times of Heart Transplantation

一個早期心臟移植的研究, 是關於病人接受心臟移植後的存活分析 (Cox 與 Oakes, 1984), 表 1.9 列出前 5 位病人的研究結果, 而表 1.10 列出資料內變數的描述, 資料存於檔案 `survCoxHeart.xls`. 研究自 1967 年計畫開始, 至 1980 年 2 月結束, 此段時間內原總共有 249 位病人願意接受移植, 但是僅 184 位病人實際接受移植, 其餘 65 位病人中, 有 8 位死亡, 相對之下, 實際接受移植的 184 病人中, 有 65 位到研究時間結束時仍存活.

表 1.9: 心臟移植存活時間: 5 位病患部分研究結果

subject	wait	transplant	survtime	status
1	49	0		1
2	5	0		1
3	0	1	15	1
4	35	1	3	1
5	17	0		1

表 1.10: 心臟移植存活時間

變數	描述
wait	等待時間: 接受研究建議日期到移植日期 (或研究結束), 以日計算. (waiting time, in days, before transplantation)
transplant	移植指標 (transplantation indicator): 1 = 接受移植; 0 = 無.
survtime	移植後存活時間, 以日計算. (post-transplantation survival time, in days.)
status	最後的生命狀態指標 (final status): 1 = 死亡, 0 = 存活.

適當地了解移植研究的設計,對於了解資料具有決定性,當內科醫師做出適當的診斷後建議病人接受心臟移植研究,若病人接受心臟移植研究計畫建議,並且成為移植的候選者,當出現一個捐贈者的心臟時,透過醫學判斷,挑選接受心臟的適合候選者。對於接受移植的病人或是仍然在等待移植的病人,研究者會觀察他們接受移植研究之建議日期,實際之移植日期,以及移植後的死亡日期或存活時間,並以指標函數表示最後的生命狀態;至於未移植的病人,其接受研究建議日期起算的存活時間,及研究結束時最後的生命狀態的指標函數仍是有效的。

1.3 存活分析的特徵

Characteristic of Survival Analysis

研究某一樣本的存活經驗通常是有用的,若此研究樣本是一個龐大母體的代表,則這樣的存活經驗特別有用,研究樣本的存活經驗是龐大母體存活經驗的估計值,當研究樣本是從目標母體中抽取的隨機樣本,則估計值精確地用機率測量是可以計算的。存活分析方法是為了充分運用存活資料中獨有的特徵,以及存活研究中特別的個體而量身訂做。

1. 通常存活資料與其他型態資料的差異之處在於設限的現象, **設限資料** 是指我們無法完全得到事件發生時間的觀測值,而妨礙我們使用標準的統計方法及推論,尤其是右設限資料描述了實際未觀測事件時間的下界,若依變數 (independent variable), 或結果變數是事件的時間,你要如何處理這樣的實例?
 2. 母體的存活時間通常是非常態的分配,在大多的統計推論中異於高斯(常態)分配,許多標準或近似統計方法無法精確描述這樣的資料。
 3. 我們通常對於整體存活時間的分配有興趣,許多標準的統計方法以平均存活時間 μ 和標準差 σ 作為推論的方向。但是,事件時間在分配之極端處的百分位值表
-

現, 通常是存活分析中令人較感興趣的, 舉例而言, 許多人希望自己能夠活到第 95 個百分位以上, 而不是只活到第 50 個百分位之上的存活時間. 存活分析中, 關注於每個個體在治療或手術後單位時間事件的發生率.

4. 研究過程中某些解釋變數, 例如血壓值, 年紀, 會隨著時間改變, 你如何利用迴歸分析中的概念, 處理這些解釋變數與其他的 **時間相依的共變數 (time-dependent covariates)**?

1.3.1 收集存活資料 Survival Data Collection

發展存活分析方法, 是爲了評估時間單一方向的流動與存活資料收集所引的幾個議題, 存活分析方法以分析事件時間之隨機長度分配爲目的, 相當於我們檢查每單位時間事件的發生率, 事件時間不僅與我們關心事件的定義有關, 也與每位個體的時間原點定義有關, 研究中, 由於研究時間限制, 無法觀測到過長的事件時間觀測值, 這會產生右設限資料, 標準的研究設計與資料收集對於不同的個體會出現不同潛在的追蹤長度, 若個體事件時間比潛在追蹤時間長, 則個體的事件時間是右設限, 其他的個體登記, 中途退出研究, 研究計畫終止或是追蹤過程中觀測值的頻率會導致不同型態的設限. 接下來的表 1.11 摘要出關於存活分析的資料收集.

在一些研究中, 個體並非在時間零點 (time zero) 登記, 這樣的樣本受限於事件時間必須至少與登記時間一樣長, 此樣本型態稱爲 **截斷樣本 (truncated sample)**, 統計推論必須考慮觀測資料的實際情況, 以及樣本的條件. 某些研究中, 觀測時間是不連貫的 (時間非連續), 這會造成 **區間設限資料 (interval censored data)**, 表 1.12 摘要出存活分析與基礎統計分析的差異.

· 22 · 1.3 存活分析的特徵

表 1.11: 資料收集中重要的時間及日期

名稱	時間軸屬性	描述
研究開始	資料收集者	開始收集資料的日曆日期.
事件時間原點	個體	個體事件時間的測量時間零點 (time 0).
進入追蹤點	個體	開始追蹤的時間. (entry point).
觀測時間	資料收集者	個體觀測到的日曆時間.
最後追蹤點	個體	退出研究時間點.
研究結束點	資料收集者	結束收集資料的日曆日期, 通常在這個時間點, 結束所有個體的追蹤.
事件時間	個體	僅部份個體有觀察到事件時間, 部分為設限時間.

表 1.12: 存活分析中常見的統計差異

事件時間的主題	統計方法
對稱性分配	不會發生, 且不適當.
常態分配	通常與存活研究的母體無關, 且不是存活樣本分配的基礎.
典型的估計值	百分位 (並非平均值)
存活常見的分配	右設限資料的右端分配
存活常見的分配	截斷資料的左端分配
存活的累積分配	存活函數 = 1 - 累積分配函數
密度型態的函數	危險函數 = 密度函數 / 存活函數
迴歸參數	常以危險比值表示

1.3.2 存活迴歸模型 / 邏輯斯迴歸 / 布瓦松模型 的區別

Distinguishing Survival / Logistic / Poisson Models

首要之處必須避免型 0 錯誤 (Type 0 Error), 無法辨別出研究問題本身的型態, 也要避免型 III 錯誤, 正確的答案回答錯誤的研究問題, 更要避免型 VI 錯誤, 錯誤的答案回答錯誤的研究問題. 研究者通常想要研究事件的發生率, 資料包含個體追蹤以及紀錄事件的發生, 存活迴歸模型分析產生存活曲線的估計值及每單位時間事件發生率. 邏輯斯迴歸模型分析發生或是未發生的事件分率 (event proportion) 之估計值. 布瓦松迴歸模型產生每單位時間事件發生比率 (event rate) 的估計值. 我們應使用哪一個? 通常我們會適當地結合這些分析, 以下摘要使用三種分析方法時的參考, 但仍應對每一個實例個別地研究問題本身使用與解釋.

1. 使用邏輯斯迴歸模型分析的條件為: 每位個體追蹤一段相同的時間, 並且對於時間原點有明確定義, 僅對事件第一次發生感到興趣, 當事件在時間原點之後, 很快就發生時 (例如, 心肌梗塞發生後 30 天內死亡), 通常使用此方法.
 2. 使用布瓦松迴歸模型分析的條件為: 每位個體的追蹤週期不同, 追蹤過程中事件率是一常數 (因此時間原點不是問題), 例如, 公開發表的死亡比率通常是建立在布瓦松分析之上.
 3. 使用存活迴歸模型分析的條件為: 每位個體追蹤不同的一段時間, 且每位個體的時間原點可能並沒有明確定義, 且事件發生比率在追蹤期間會隨著時間而改變, 通常會有失去追蹤或是設限的線索. 長期追蹤也會建議使用存活分析, 因為事件發生比率可能會在一個長期的時間區段變化.
-

1.4 資料分析練習

1. 胃賁門癌存活分析 (Carcinoma of Gastric Cardia Survival Analysis)

胃賁門癌發生於胃的頂部與食道交接處, 由於胃在腹腔內, 而食道在胸腔內, 胃賁門癌的根治性切除手術治療 (radical resection), 常常須同時進行開胸術與剖腹術, 並且進行各種食道與胃腸接合手術; 胃賁門癌早期診斷不易, 探查手術時, 若發現腫瘤無法完全切除, 常須進行各種姑息性切除 (palliative resection), 或者是繞道手術. 研究者長期追蹤在台北榮民總醫院之病人罹患胃賁門癌, 接受外科手術, 共有 403 位. 此研究主要目的是對於整體存活 (overall survival) 和癌症存活 (cancer-specific) 狀況做追蹤分析; 研究之次要目的, 在於探查病患其他的共變異數, 是否會影響病人的存活時間. 部分的資料存於檔案 `survVGHcardiac400.xls`, 表 1.13 則列出變數的敘述.

- (a) 請對本資料進行描述性統計.
 - (b) 將最後追蹤的存活時間, 視為連續型反應變數, 分析治療性質變數 (CURATIVE) 的存活狀態是否有差異?
 - (c) 將最後追蹤的存活時間, 視為連續型反應變數, 考慮所有的共變數, 配適共變異分析 (ANCOVA) 模型, 說明必要之統計假設, 並分析其結果.
 - (d) 將最後追蹤的存活狀態 (STATUS), 視為二元類反應變數, 治療性質變數 (CURATIVE) 的存活狀態是否有差異?
 - (e) 將最後追蹤的存活狀態 (STATUS), 視為二元類反應變數, 考慮所有的共變數, 配適邏輯斯迴歸, 說明必要之統計假設, 並分析其結果.
 - (f) 考慮追蹤時間與存活狀態, 以布瓦松分配假設, 分別比較兩組治療性質的事件發生率否有差異?
 - (g) 考慮追蹤時間與存活狀態, 以布瓦松分配假設, 考慮所有的共變數, 配適布瓦
-

松迴歸, 說明必要之統計假設, 並分析其結果.

表 1.13: 胃賁門癌存活分析: 變數說明

變數	描述
SEX	性別: 1 = 男 (Male); 0 = 女 (female).
AGE	年紀 (歲數以年為單位) (age).
ALB	血漿蛋白值, 單位: mg/dL.
ALBGR	血漿蛋白值分組: 1 = [Alb < 3.7] 不正常, 2 = [Alb ≥ 3.7] 正常.
DIET	病人手術前可吃的食物: 0 = 無法進食, 1 = 液態, 2 = 軟性, 3 = 正常.
OPSTOMA	食道與胃腸接合處: 1 = 腹腔內, 2 = 右胸腔, 3 = 左胸腔, 4 = 頸部.
OP1	胃的主要手術 0 = 無, 1 = 賁門切除, 3 = 近端胃切除, 4 = 全胃切除. 2, 5 = 無法切除.
OP2	食道的手術. 1 = 遠端食道切除, 2 = 亞全食道切除, 0 = 無.
OP3	重建手術. 1 = 大腸, 2 = 胃管, 3 = 小腸, 0 = 未做重建手術.
CTPRE	手術前化學藥物治療, 0 = no, 1 = yes .
RTPRE	手術前放射線治療, 0 = no, 1 = yes
CTPOST	手術後化學藥物治療, 0 = no, 1 = yes
RTPOST	手術後放射線治療, 0 = no, 1 = yes.
T	腫瘤本體侵犯範圍分期: 1, 2, 3, 4, 9 = missing.
N	腫瘤淋巴侵犯轉移分期: 0, 1, 2, 9 = missing.
M	腫瘤遠處轉移分期: 0, 1, 9 = missing.
STAGEW	腫瘤 TMN 分期: 1.1 = Stage Ia, 1.2 = Stage Ib, 2 = Stage II, 3.1 = Stage IIIa, 3.2 = Stage IIIb, 4 = Stage IV.
CURATIVE	治療性質: 0 = 姑息性治療, 1 = 根治性治療.
COMPLICA	產生併發症: 0 = no, 1 = yes.
STATUS	最後追蹤的存活狀態: 0 = 存活, 1 = 因癌症死亡, 3 = 死於其他原因.
FUMO	最後追蹤時間, 單位: 月.