

長期存活資料比率之統計分析

蘇秀媛^{1*}、蘇志雄²、謝鑫能¹

¹國立台灣大學農藝學系

²致理技術學院會計學系

摘要

在存活資料中，若試驗期間能無限延長，有部份資料將能永久存活，故稱長期存活資料。本文中長期存活資料為區間設限資料型式，假設存活母體為指數分布與 Weibull 分布，定義區間設限資料下的概度函數，並導出長期存活資料比率之最大概度估計式及概度比檢定，同時利用 Kaplan—Meier 的估計方法，提出無母數方法來估計「區間設限長期存活資料」比率及其檢定方法。本文利用蒙地卡羅統計模擬方法比較在最大概度法及無母數方法不同情形下的表現狀況，結果發現無母數方法在檢定「區間設限長期存活資料」比率表現優於最大概度法，此與長期存活資料為右方設限資料型式之結果一致。

關鍵詞：長期存活資料、區間設限、最大概度估計、概度比檢定。

Statistical Analysis of Proportion of Long-term Survivor Data

Hsiu-Yuan Su^{1*}, Chih-Hsiung Su², and Hsin-Neng Hsieh¹

¹ Department of Agronomy, National Taiwan University, Taipei 10617, Taiwan ROC

* 通信作者, suwong@ntu.edu.tw

投稿日期：2006 年 5 月 10 日

接受日期：2006 年 7 月 30 日

作物、環境與生物資訊 3:337-344 (2006)

Crop, Environment & Bioinformatics 3:337-344 (2006)

189 Chung-Cheng Rd., Wufeng, Taichung Hsien 41301, Taiwan ROC

² Department of Accounting, Chihlee Institute of Technology, Banciao, Taipei Hsien 22050, Taiwan ROC

ABSTRACT

The long-term survival data means some of the data can survive in definite time. The likelihood function with interval-censored data was defined when the survival population was assumed to be exponential or Weibull distribution. The MLE (maximum likelihood estimator) and LRT (likelihood ratio test) can be derived from long-term survival data. Furthermore, a nonparametric method for estimating proportion of long-term survival data was derived based on Kaplan-Meier's estimation. In this paper, we compared the MLE and nonparametric method by Monte Carlo simulation method for long-term survival data. We found that the performance of nonparametric test was better than maximum likelihood ratio test, which was similar to the results obtained with right-censored data.

Key words: Long-term survivor data, Interval-censored, Maximum likelihood estimator, Likelihood ratio test.

前言

統計上對於不完整資料 (incomplete data) 的分析，大致上可以分為缺失值資料 (missing data) 及設限資料 (censoring data) 兩大類。對觀察的試驗資料記錄事件開始 (time begin) 及結束 (time end) 之試驗觀察時間稱為存活時間 (survival time)。若實際的存活時間受到試驗時間的結束造成資料的不完整，此類資料雖造成缺失，但仍會留下訊息，稱具有此種資料之試驗觀察資料為設限資料。



(censored data)。Kaplan and Meier (1958) 提出設限資料估計存活函數 (survival function) 之無母數方法。Gehan (1965)、Mantel and Haenszel (1959)、Peto and Pike (1973)、Charkraborti and Desu (1991) 及 Chen (1994) 相繼提出比較分析之統計方法。

長期存活資料(long-term survival data) 指在整個存活時間分布具有長久存活之個體存在 (Maller and Zhou 1996)，若 $F_T(t)$ 為存活時間之累積分布函數 (cumulated distribution function)，則通常 $F_T(\infty)=1$ ，但若擁有長期存活資料時，則 $Fr(\infty)=p, 0 < p < 1$ ，而 $1-p$ 即為母群中之長期存活資料比率，又稱免疫比例 (immune proportion)。Boag (1949) 是最早提出「長期存活資料」統計分析的學者 (Boag 稱「長期存活資料」為「治癒資料」)。Boag 利用英國某醫院中自 1929 至 1938 年之 121 名乳癌病患，記錄這 9 年來接受治療後的存活時間，但 9 年後仍有存活之病患，Boag 認為此資料不能完全視為設限資料，亦有可能屬於被完全治癒之病患。由於 Boag 利用卡方適合度檢定，認為存活時間 lognormal 分布較指數分布適合，提出一混合模型(mixed model)，並導出參數的最大概度估計量。

Haybittle (1965) 利用指數型式來表示『被治癒』之比率 $1-p=c=e^{-\frac{t}{\beta}}$ ，推導出 Gompertz 模式，Haybittle 利用最大概度估計法推估 c, β ，但並沒有直接檢定 c 之值，只是估算出其標準誤差(standard error)。Goldman (1984) 引用指數分布作為存活時間之分布，建立混合模式，並分析一組自 1973 至 1979 年之 33 個攝護腺癌病患(1973-1979) 之存活時間，及利用蒙地卡羅模擬，發現其最大概度估計量具有不偏的性質。

Farewell (1977) 利用 Logistic 函數作為共變數 $\bar{X}$ 之連結函數(link function)，成為 Bernoulli 分布「發生」之機率，並假設存活時間服從指數分布，討論實驗設限時間及 Bernoulli 分布「發生」之機率大小對該混合

模式之效率。Farewell (1982) 引用相同觀念，考慮存活時間服從 Weibull 分布，分析魚在不同的銹含量中其存活時間之差異。

Gordon (1990) 提出一組資料中有兩種可能的致命因素，觀察樣本資料中可以明確判定是屬於何種因素致命，Gordon 做一假設對某種特定因素來作分析時，將所有病人區分為兩個部分，一部份為未治癒組與另一部份皆視為治癒組。並提出混合式 Gompertz 分布，利用最大概度估計法估計參數並予以分析。

Su *et al.* (2000, 2003) 利用存活分析的概念，將種子發芽時間視為存活資料來處理分析，將溫度視為可解釋之變數時，引用加速失敗時間模式 (accelerated failure time model) 進行分析。Su *et al.* (2005) 延用 Boag 的方法，透過存活分析(survival analysis) 的統計方法，建立種子發芽比率的統計推論，將種子發芽時間視為統計上一組第 I 類型的右方設限存活資料(type I right censored survival data)，定義在右方設限資料下概度函數，並假設存活母體為指數分布與 Weibull 分布，導出「右方設限長期存活資料」比率之最大概度估計式及概度比檢定，同時利用 Kaplan-Meier 的估計方法，提出無母數方法來估計「長期存活資料」比率及檢定方法。

本文是 Su *et al.* (2005) 的延伸，同樣嘗試以存活分析的概念，將種子發芽資料視為存活資料來處理。種子發芽的時間只知道介於兩個時間點之內，實際上確切時間並無法觀察到，故將每一顆種子之發芽時間視為區間設限長期存活資料 (interval censored long-term survivor data)，建立種子發芽比率的統計推論。透過存活分析的統計方法，來建立種子長期存活資料的發芽比率之統計推論。

存活分析方法之數學函數及其估計

假設存活時間 T 為連續型之隨機變數，則可定義以下數個重要存活分析之數學函



數：

(1) 存活機率密度函數(probability density function)：

$$f_T(t) = \lim_{\varepsilon \rightarrow 0} \frac{P(t < T < t + \varepsilon)}{\varepsilon}, \quad t \geq 0.$$

(2) 存活累積分布函數(accumulated distribution function)：

$$F_T(t) = P_r(T \leq t) = \int_0^t f_T(u) du, \quad t \geq 0.$$

(3) 存活函數(survival function)：

$$S_T(t) = P_r(T \geq t) = \int_t^\infty f_T(u) du = 1 - F_T(t), \quad t \geq 0.$$

(4) 故障函數(hazard function)：

$$h_T(t) = \lim_{\varepsilon \rightarrow 0} \frac{P_r(t < T < t + \varepsilon | T > t)}{\varepsilon} = \frac{f_T(t)}{S_T(t)}, \quad t \geq 0.$$

Kaplan and Meier (1958) 利用 product limit 的無母數方法，提出存活函數 $S_T(t)$ 之估計式為：

$$\hat{S}_T(t) = \prod_{t_j < t} \left(\frac{n_j - d_j}{n_j} \right)$$

其中 n_j 表示存活時間 t_j 前確定存活之個數， d_j 表示存活時間 t_j 前之「發生」之個數，將存活時間分成 k 個區間，並利用 Greenwood 法則推導出 $\hat{S}_T(t)$ 之標準誤差：

$$S.E.\{\hat{S}_T(t)\} = \hat{S}_T(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}}$$

長期存活資料比率之統計推論

假設存活時間為 T ，其機率密度函數為 $f_T(t)$ ， $t \geq 0$ ，存活函數為 $S_T(t)$ ， $t \geq 0$ ，長期存活資料比率為 $S_T(\infty) = 1 - p$ ， $0 < p < 1$ 。

考慮由 $t_0 = 0 < t_1 < t_2 < \dots < t_k < \infty$ ， $(k+1)$ 個時間點，將時間切割成 k 段，第 i 段時間區間為 (t_{i-1}, t_i) ， $i = 1, 2, \dots, k$ ；設 n_i 表示在第 i 段時間區間內發生之個體總數， $i = 1, 2, \dots, k$ ， $\sum_{i=1}^k n_i = n_u$ ； n 表所有個數總體。則資料之概度函數為

$$\ell = \left\{ \prod_{i=1}^k [p(S_T(t_{i-1}) - S_T(t_i))]^{n_i} \right\} \cdot [p \cdot S_T(t_k) + 1 - p]^{n - n_u}$$

故資料之對數概度函數為

$$\ell = \log L = \sum_{i=1}^k n_i (\log p + \log(S_T(t_{i-1}) - S_T(t_i))) + (n - n_u) \log(p \cdot S_T(t_k) + 1 - p)$$

且其參數向量為 $\bar{\theta} = (\theta_1, \theta_2, \dots, \theta_s, p)'$ ，其中 $\theta_1, \theta_2, \dots, \theta_s$ 為存活時間母體分布之參數， p 為總發生率參數。求出 $\bar{\theta}$ 之最大概度估計量 $\hat{\bar{\theta}}$ ，並導出其 p 之概度比檢定及 $(1 - \alpha) 100\%$ 信賴區間。

1. 有母數方法

(1) 當長期存活資料為區間設限資料且母體為指數分布(exponential(λ))

其概度函數為

$$\ell = \left\{ \prod_{i=1}^k [p(e^{-\lambda t_{i-1}} - e^{-\lambda t_i})]^{n_i} \right\} [pe^{-\lambda t_k} + 1 - p]^{n - n_u}$$

對數概度函數為

$$\ell = \log L = \sum_{i=1}^k n_i [\log p + \log \lambda + \log(e^{-\lambda t_{i-1}} - e^{-\lambda t_i})] + (n - n_u) \log(pe^{-\lambda t_k} + 1 - p)$$

參數向量為 $\bar{\theta} = (\lambda, p)'$

(i) $\bar{\theta}$ 之點估計：

$$\frac{\partial \ell}{\partial \theta} = 0 \Rightarrow \begin{cases} \frac{\partial \ell}{\partial \lambda} = \frac{n_u}{\lambda} - \sum_{i=1}^k \left[\frac{t_{i-1} e^{-\lambda t_i}}{e^{-\lambda t_{i-1}} - e^{-\lambda t_i}} \right] - \frac{(n - n_u)(pe^{-\lambda t_k} t_k)}{pe^{-\lambda t_k} + 1 - p} = 0 \\ \frac{\partial \ell}{\partial p} = \frac{n_u}{p} + \frac{(n - n_u)(e^{-\lambda t_k} - 1)}{pe^{-\lambda t_k} + 1 - p} = 0. \end{cases}$$

利用 Newton-Raphson 數值方法可求得近似解， $\hat{\bar{\theta}} = (\hat{\lambda}, \hat{p})'$ 。

(ii) $\bar{\theta}$ 之共變數矩陣(covariance matrix)之估計：

$$U = \begin{bmatrix} \frac{\partial^2 \ell}{(\partial \lambda)^2} & \frac{\partial}{(\partial p)} \left(\frac{\partial \ell}{\partial \lambda} \right) \\ \frac{\partial}{\partial \lambda} \left(\frac{\partial \ell}{\partial p} \right) & \frac{\partial^2 \ell}{(\partial p)^2} \end{bmatrix}$$

其中



$$\begin{aligned}\frac{\partial^2 \ell}{(\partial \lambda)^2} &= \frac{n_u}{\lambda^2} - \sum_{i=1}^k \frac{e^{-\lambda(t_{i-1}+t_i)}(t_{i-1}-t_i)^2}{(e^{-\lambda t_{i-1}} - e^{-\lambda t_i})^2} + \\ &\quad (n-n_u)p(1-p) \frac{e^{-\lambda t_k} t_k^2}{(pe^{-\lambda t_k} + 1 - p)^2} \\ &= c_{11}(\lambda, p)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell}{(\partial \lambda)^2} &= \frac{-n_u}{p^2} - \frac{(n-n_u)(e^{-\lambda t_k} - 1)^2}{(pe^{-\lambda t_k} + 1 - p)^2} \\ &= c_{22}(\lambda, p)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial p} \left(\frac{\partial \ell}{\partial \lambda} \right) &= \frac{\partial}{\partial \lambda} \left(\frac{\partial \ell}{\partial p} \right) = -(n-n_u) \frac{t_k e^{-\lambda t_k}}{(pe^{-\lambda t_k} + 1 - p)^2} \\ &= c_{12}(\lambda, p)\end{aligned}$$

$$\begin{aligned}\Delta_c &= \begin{vmatrix} c_{11}(\hat{\lambda}, \hat{p}) & c_{12}(\hat{\lambda}, \hat{p}) \\ c_{12}(\hat{\lambda}, \hat{p}) & c_{22}(\hat{\lambda}, \hat{p}) \end{vmatrix} \\ &= c_{11}(\hat{\lambda}, \hat{p})c_{22}(\hat{\lambda}, \hat{p}) - (c_{12}(\hat{\lambda}, \hat{p}))^2\end{aligned}$$

$$\text{則 } \hat{\Sigma}_{\hat{\theta}} = U^{-1}(\hat{\theta}) = \begin{bmatrix} \hat{Var}(\hat{\lambda}) & \hat{Cov}(\hat{\lambda}, \hat{p}) \\ \hat{Cov}(\hat{p}, \hat{\lambda}) & \hat{Var}(\hat{p}) \end{bmatrix}$$

其中

$$\hat{Var}(\hat{\lambda}) = \frac{1}{\Delta_c} c_{22}(\hat{\lambda}, \hat{p})$$

$$\hat{Var}(\hat{p}) = \frac{1}{\Delta_c} c_{11}(\hat{\lambda}, \hat{p})$$

(iii) p 之 $(1-\alpha)100\%$ 信賴區間估計為：

由於最大概度估計量具有漸進常態之特性，故 $(\hat{\theta} - \bar{\theta}) \xrightarrow{n \rightarrow \infty} M.N.(\bar{0}, \Sigma_{\hat{\theta}})$ ，其中 $M.N.$ 表示多變量常態分布 (Multivariate Normal Distribution)，因此 p 之 $(1-\alpha)100\%$ 信賴區間為

$$\left[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{p})}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{p})} \right]$$

(iv) p 之概度比檢定 (likelihood ratio test)：

$$H_0: p=1 \quad \text{v.s.} \quad H_1: p<1$$

若 $H_0: p=1$ 成立之下：（即無『長期存活資料』），則概度函數為

$$\ell(H_0) = \left\{ \prod_{i=1}^k (e^{-\lambda t_{i-1}} - e^{-\lambda t_i})^{n_i} \right\} (e^{-\lambda t_k})^{n-n_u}$$

對數概度函數為

$$\ell(H_0) = \log \ell(H_0) = \sum_{i=1}^k n_i \log(e^{-\lambda t_{i-1}} - e^{-\lambda t_i}) - (n-n_u) \lambda t_k$$

則 λ 之最大概度估計量為 $\frac{\partial \ell(H_0)}{\partial \lambda} = 0$ ，可得

方程式

$$\sum_{i=1}^k n_i \left[\frac{t_{i-1} e^{-\lambda t_{i-1}} - t_i e^{-\lambda t_i}}{e^{-\lambda t_{i-1}} - e^{-\lambda t_i}} \right] - (n-n_u) t_k = 0$$

其根即為 λ 之最大概度估計值， $\lambda \hat{H}_0$ 。因此， $\ell(\hat{H}_0) = \sum_{i=1}^k n_i \log(e^{-\lambda \hat{H}_0 t_{i-1}} - e^{-\lambda \hat{H}_0 t_i}) - (n-n_u) \lambda \hat{H}_0 t_k$ 。若 $H_0 \cup H_1$ 成立之下，對數概度函數值為：

$$\begin{aligned}\ell(H_0 \cup H_1) &= \sum_{i=1}^k n_i [\log \hat{p} + \log \hat{\lambda} + (e^{-\lambda \hat{H}_0 t_{i-1}} - e^{-\lambda \hat{H}_0 t_i})] \\ &\quad + (n-n_u) \log(\hat{p} e^{-\lambda \hat{H}_0 t_k} + 1 - \hat{p})\end{aligned}$$

因此，檢定統計量為

$$IG_E = -2[\ell(\hat{H}_0) - \ell(H_0 \cup H_1)]$$

其決策法則如下：當 $IG_E > \chi_{\alpha}^2(1)$ ，則拒絕 H_0 ；若 $IG_E \leq \chi_{\alpha}^2(1)$ ，則接受 H_0 。

(2) 當長期存活資料為區間設限資料且母體為 Weibull (α, λ) 分布

概度函數：

$$\ell = \left\{ \prod_{i=1}^k [p(e^{-\lambda t_{i-1}^{\alpha}} - e^{-\lambda t_i^{\alpha}})]^{n_i} \right\} [pe^{-\lambda t_k^{\alpha}} + 1 - p]^{n-n_u}$$

對數概度函數：

$$\ell = \log \ell = \sum_{i=1}^k n_i [\log p + \log(e^{-\lambda t_{i-1}^{\alpha}} - e^{-\lambda t_i^{\alpha}})] +$$

$$(n-n_u) \log[pe^{-\lambda t_k^{\alpha}} + 1 - p]$$

參數向量： $\vec{\theta} = (\alpha, \lambda, p)'$



(i) $\bar{\theta}$ 之點估計：

$$\frac{\partial \ell}{\partial \bar{\theta}} = 0 \Rightarrow \begin{cases} \frac{\partial \ell}{\partial \alpha} = -\lambda \sum_{i=1}^k \frac{e^{-\lambda t_{i-1}^\alpha} \log t_{i-1} - e^{-\lambda t_i^\alpha} \log t_i}{e^{-\lambda t_{i-1}^\alpha} - e^{-\lambda t_i^\alpha}} \\ \quad - (n - n_u) \lambda p \frac{e^{-\lambda t_k^\alpha} \log t_k}{pe^{-\lambda t_k^\alpha} + 1 - p} = 0 \\ \frac{\partial \ell}{\partial \lambda} = -\lambda \sum_{i=1}^k n_i \frac{e^{-\lambda t_{i-1}^\alpha} - e^{-\lambda t_i^\alpha}}{e^{-\lambda t_{i-1}^\alpha} - e^{-\lambda t_i^\alpha}} - (n - n_u) p \frac{e^{-\lambda t_k^\alpha}}{pe^{-\lambda t_k^\alpha} + 1 - p} = 0 \\ \frac{\partial \ell}{\partial p} = \frac{n_u}{p} + (n - n_u) \frac{e^{-\lambda t_k^\alpha} - 1}{pe^{-\lambda t_k^\alpha} + 1 - p} = 0 \end{cases}$$

利用 Newton-Raphson 數值方法可求得近似解， $\bar{\theta} = (\hat{\alpha}, \hat{\lambda}, \hat{p})$ 。

(ii) $\bar{\theta}$ 之共變數矩陣(covariance matrix)之估計：

$$U = \begin{bmatrix} \frac{\partial^2 \ell}{(\partial \alpha)^2} & \frac{\partial}{\partial \lambda} \left(\frac{\partial \ell}{\partial \alpha} \right) & \frac{\partial}{\partial p} \left(\frac{\partial \ell}{\partial \alpha} \right) \\ \frac{\partial}{\partial \alpha} \left(\frac{\partial \ell}{\partial \lambda} \right) & \frac{\partial^2 \ell}{(\partial \lambda)^2} & \frac{\partial}{\partial p} \left(\frac{\partial \ell}{\partial \lambda} \right) \\ \frac{\partial}{\partial \alpha} \left(\frac{\partial \ell}{\partial p} \right) & \frac{\partial}{\partial \lambda} \left(\frac{\partial \ell}{\partial p} \right) & \frac{\partial^2 \ell}{(\partial p)^2} \end{bmatrix}$$

$$\hat{\Sigma}_{\bar{\theta}} = U^{-1}(\hat{\theta}) = \begin{bmatrix} \text{Var}(\hat{\alpha}) & \text{Cov}(\hat{\alpha}, \hat{\lambda}) & \text{Cov}(\hat{\alpha}, \hat{p}) \\ \text{Cov}(\hat{\lambda}, \hat{\alpha}) & \text{Var}(\hat{\lambda}) & \text{Cov}(\hat{\lambda}, \hat{p}) \\ \text{Cov}(\hat{p}, \hat{\alpha}) & \text{Cov}(\hat{p}, \hat{\lambda}) & \text{Var}(\hat{p}) \end{bmatrix}$$

$$\Delta = \begin{bmatrix} B_{11}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{12}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{13}(\hat{\alpha}, \hat{\lambda}, \hat{p}) \\ B_{12}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{22}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{23}(\hat{\alpha}, \hat{\lambda}, \hat{p}) \\ B_{13}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{23}(\hat{\alpha}, \hat{\lambda}, \hat{p}) & B_{33}(\hat{\alpha}, \hat{\lambda}, \hat{p}) \end{bmatrix}$$

其中， $\text{Var}(\hat{p}) = \frac{1}{\Delta} (\text{Var}(\hat{\alpha}) \cdot \text{Var}(\hat{\lambda}) - \text{Cov}(\hat{\alpha}, \hat{\lambda})^2)$ 。

(iii) p 之 $(1-\alpha)$ 100% 信賴區間估計為：

$$\left[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{p})}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{p})} \right]$$

(iv) p 之概度比檢定(likelihood ratio test)：

$$H_0: p=1 \quad \text{v.s.} \quad H_1: p < 1$$

檢定統計量為

$$G_w = -2[\ell(\hat{H}_0) - \ell(H_0 \hat{\cup} H_1)]$$

決策法則：當 $G_w > \chi_{\alpha}^2(1)$ 則拒絕 H_0 ；若 $G_w \leq \chi_{\alpha}^2(1)$ 則接受 H_0 。

2. 無母數方法

由 $0 = t_0 < t_1 < t_2 < \dots < t_k < \infty$ ， $(k+1)$ 個時間

點，將時間切割成 k 段時間區間，第 j 段時間區間為 $[t_{j-1}, t_j]$ ， $j=1, 2, \dots, k$ ；若所觀察值之資料並不確定其真正發生時間，但可以知道在時間 $[L_i, R_i]$ ， $i=1, 2, \dots, n$ 區間內發生，利用指標函數 α_{ij} 表示：

$$\alpha_{ij} = \begin{cases} 1 & \text{假如第 } i \text{ 筆區間資料會落在} \\ & [t_{j-1}, t_j] \text{ 時間區間內} \\ 0 & \text{假如第 } i \text{ 筆資料完全不會落在} \\ & [t_{j-1}, t_j] \text{ 時間區間內} \end{cases}$$

$p_j = S(t_j) - S(t_{j-1})$ 為第 j 段時間區間發生之機率。

Turnbull (1976) 對於此類區間設限資料，對於存活函數提出一遞迴估算公式演算法，其步驟摘要如下：

步驟 1：給定 $S_0(t_j)$ 之起始值 $j=1, 2, \dots, k$

步驟 2：計算 $p_j = S_0(t_{j-1}) - S_0(t_j)$ ， $j=1, 2, \dots, k$

步驟 3：估算在 t_j 時之發生之總個數：

$$d_j = \sum_{i=1}^n \frac{\alpha_{ij} p_j}{\sum_{s=1}^k \alpha_{is} p_s}$$

步驟 4：估算在 t_j 前尚未發生之總個數：

$$r_j = \sum_{s=j}^k d_s$$

步驟 5：利用 product limit 公式重新估計 $S(t)$

$$S_n(t) = \prod \left(\frac{r_j - d_j}{r_j} \right)$$

步驟 6：若 $S_n(t)$ 與 $S_0(t)$ 相近，則停止；否則利用 $S_n(t)$ 取代 $S_0(t)$ ，遞迴步驟 1~

步驟 5，直到 $S(t)$ 收斂為 $S_{tb}^{\wedge}(t)$ 。

(1) p 之點估計：

$$\hat{p}_{n2} = 1 - S_{tb}(t_k)$$

(2) p 之假設檢定

$$H_0: p=1 \quad \text{v.s.} \quad H_1: p < 1$$

檢定統計量： $\hat{p}_{n2} = 1 - S_{tb}(t_k)$



決策法則：當 $\hat{p}_{n2} < C_{\alpha 2}$ 則拒絕 H_0 ；若 $\hat{p}_{n2} > C_{\alpha 2}$ 則接受 H_0 。其中，當 H_0 成立之下 $P(\hat{p}_{n2} > C_{\alpha 2}) \leq \alpha$ ， $C_{\alpha 2}$ 選取由模擬方式產生。

$C_{\alpha 2}$ 之決定的模擬方式如下：

我們利用有界樣本估算區間半徑 (interval radius)，並且利用有界樣本之組中點及無界樣本之端點，再利用所求之區間半徑估算平均數。產生 n 組之 (L_i, R_i) ；其中每組 (L_i, R_i) 產生方法如下：

首先產生獨立之 $L_i \sim \text{exponential}(1/\lambda)$ ， $C_{1i}, C_{2i} \sim \text{uniform}(0, \text{區間半徑})$ ；令 $L_i = T_i - C_{1i}$ ； $R_i = T_i + C_{2i}$ ；若 $L_i < 0$ 則令 $L_i = 0$ 。計算 $\hat{p}_{n2}$ 之值。反覆執行 5000 次，可得 5000 筆 $\hat{p}_{n2}$ 值。計算第 $(100 \times \alpha)$ 個百分位數，可得 α 水準下之臨界值。

統計模擬

1. 模擬方法

我們欲對總發生率 (total occur rate) (亦即 $F(\infty) = p$ 或是 (1-長期存活比率)) 的最大概度檢定法探討其檢定力的表現，故利用蒙特卡羅方法 (Monte Carlo Method) 來模擬最大概度檢定法之檢定力的表現情形，透過模擬的結果來檢查在顯著水準上的維持 (即 $p=1$ 時之犯型 I 錯誤之比率) 及各種 $p < 1$ 之檢定力 (正確拒絕 H_0 之比率) 表現。其模擬代號如下：

MLE(IG)：區間設限資料之最大概度檢定法

NON(I)：區間設限資料之無母數檢定法
模擬及分析步驟摘要如下：

(1) MLE(IG) 方法之模擬：

(i) 此種方法是針對區間設限整群資料所提之方法，因此在模擬上，只考慮存活時間分布即可，故沿用指數分布及 Weibull 分布，[$\text{expornential}(\lambda=1)$ 及 $\text{Weibull}(\alpha=3, \beta=3)$]。

(ii) 觀察時間採用固定時間模式，記錄落在此區間之個數。

$T \sim \text{exponential}(\lambda=1)$ ：

取 0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.3, 1.4, 1.5, 2.0 (10 段區間)

$T \sim \text{Weibull}(\alpha=3, \beta=3)$ ：

取 0, 0.4, 0.8, 1.1, 1.4, 1.7, 2.0, 2.4, 2.8, 3.2, 4.0 (10 段區間)

(iii) 樣本數 $n=20, 50, 100$ 。

(iv) 在檢定力之模擬方面，必須考慮 p 值，故先產生 $U \sim \text{uniform}(0,1)$ ，若 $U < 1-p$ ，則屬於永久存活資料，令觀察時間 $X_i = 10$ (一定不屬於觀察區間內)；若 $U \geq 1-p$ ，則在取觀察時間為 $X_i = T_i$ 。

(v) 重複 5000 次，計算其平均發生拒絕率，當 $p=1$ 時，其值為顯著水準；當 $p < 1$ 時，則為檢定力。

(2) NON(I) 方法之模擬：

(i) 由於本法是針對區間設限資料非整群資料，因此考慮其存活時間分布為指數分布及 Weibull 分布。對於觀察時間，在此是利用 *uniform* 分布產生，而且區間長度將會影響其樣本資料結構，故考慮下列情形：

$T_i \sim \text{exponential}(\lambda=1/30)$

$C_{1i} \sim \text{uniform}(0,2)$ ， $C_{2i} \sim \text{uniform}(0,2)$

$C_{1i} \sim \text{uniform}(0,4)$ ， $C_{2i} \sim \text{uniform}(0,4)$

$C_{1i} \sim \text{uniform}(0,8)$ ， $C_{2i} \sim \text{uniform}(0,8)$

則 $(L_i, R_i]$ 為區間資料，其中 $L_i = T_i - C_{1i}$ ， $R_i = T_i + C_{2i}$

$T_i \sim \text{Weibull}(\alpha=0.000873, \beta=2)$

$C_{1i} \sim \text{uniform}(0,2)$ ， $C_{2i} \sim \text{uniform}(0,2)$

$C_{1i} \sim \text{uniform}(0,4)$ ， $C_{2i} \sim \text{uniform}(0,4)$

$C_{1i} \sim \text{uniform}(0,8)$ ， $C_{2i} \sim \text{uniform}(0,8)$

則 $(L_i, R_i]$ 為區間資料，其中 $L_i = T_i - C_{1i}$ ，



$$R_i = T_i + C_{2i}$$

- (ii) 樣本數 $n=20, 50, 100$ 。
- (iii) 在檢定力之模擬方面，必須考慮 p 值，故先產生 $U \sim \text{uniform}(0,1)$ ，若 $U < 1-p$ ，則屬於永久存活資料，則令 $R_i = \infty$ （設為 99）， $L_i = T_i - C_{1i}$ ；若 $U \geq 1-p$ ，則 $L_i = T_i - C_{1i}$ ， $R_i = T_i + C_{2i}$ 。但由於起始時間不可能為負，故當 $L_i < 0$ 則令成 $L_i = 0$ 。
- (iv) 重複 5000 次，計算其平均發生拒絕率，當 $p=1$ 時，其值為顯著水準；當 $p < 1$ 時，則為檢定力。

2. 模擬結果

設 T 為指數分布時，Table 1 及 Table 2 分別為有母數方法 MLE(IG) 及無母數方法 NON(I) 之模擬結果。當 $p=1$ 時，表中所示為其顯著水準，當 $p < 1$ 時，則表中之結果為其檢定力。不論 MLE(IG) 或 NON(I) 之檢定力均不是很好，只有在 n 大時，且 $p=0.5$ 時，檢

定力表現較佳，且 NON(I) 在區間半徑低、中時的表現比大區間半徑的表現為佳。整體而言，利用無母數方法來檢定「區間設限長期存活資料」比率的表現較最大概度法為穩健。

設 T 為 Weibull 分布時，Table 3 及 Table 4 則分別為有母數方法及無母數方法之模擬結果。其結果與 T 為指數分布時之結果類似，只有在 n 大時，且 $p=0.5$ 時，檢定力較高。

討論

本文利用種子發芽行為資料，將觀察等待發芽成功的時間，解釋成存活時間，對於未能在觀察時間內發芽之種子資料，則視為一種設限資料。但種子發芽的確切時間無法確實觀察到，只知道種子發芽的時間介於兩個時間點之內，實際上確切時間並無法觀察到，於是本文將每一顆種子發芽存活時間都視作區間設限資料，比 Su *et al.* (2005) 視作右設限資料較為合理。

Table 1. Results of MLE(IG) when distribution of T is an exponential function.

Total occur rate		$p=1.0$	$p=0.9$	$p=0.8$	$p=0.5$
n	20	0.011	0.216	0.296	0.592
	50	0.028	0.312	0.348	0.880
	100	0.029	0.366	0.448	0.922

Table 2. Results of NON(I) when distribution of T is an exponential function.

Total occur rate		$p=1.0$			$p=0.9$			$p=0.8$			$p=0.5$		
Interval radius		2	4	8	2	4	8	2	4	8	2	4	8
n	20	0.017	0.008	0.004	0.157	0.273	0.210	0.255	0.314	0.221	0.267	0.473	0.340
	50	0.023	0.013	0.036	0.233	0.423	0.340	0.353	0.540	0.320	0.573	0.733	0.480
	100	0.027	0.027	0.055	0.433	0.557	0.493	0.533	0.610	0.440	0.813	0.807	0.593

Table 3. Results of MLE(IG) when distribution of T is a Weibull function.

Total occur rate		$p=1.0$	$p=0.9$	$p=0.8$	$p=0.5$
n	20	0.011	0.196	0.201	0.300
	50	0.021	0.281	0.298	0.899
	100	0.048	0.401	0.444	0.912



Table 4. Results of NON(I) when distribution of T is a Weibull function.

Total occur rate Interval radius	$p=1.0$			$p=0.9$			$p=0.8$			$p=0.5$		
	2	4	8	2	4	8	2	4	8	2	4	8
n 20	0.024	0.015	0.012	0.163	0.273	0.123	0.245	0.314	0.202	0.207	0.468	0.340
50	0.027	0.032	0.032	0.251	0.423	0.210	0.352	0.514	0.320	0.563	0.713	0.377
100	0.030	0.022	0.055	0.476	0.557	0.493	0.533	0.610	0.440	0.913	0.824	0.382

Su *et al.* (2005)指出右方設限資料的最大概度檢定法其顯著水準相當小，尤其在低設限率時，很容易接受虛無假設，其主要原因是由於模式中的參數將訊息吸收了。本文以區間設限取代右方設限，模擬結果顯示，當 n 大且 $p = 0.5$ 時，區間設限資料的最大概度檢定法 MLE(IG)及無母數方法 NON(I)之檢定力表現較佳，而 NON(I)在區間半徑低、中時的表現比大區間半徑的表現為佳，整體來看，無母數方法仍較為穩健。

引用文獻

- Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **J. Royal Statist. Soc., Series B** 11:15-44.
- Chakraborti S, Desu MM (1991) Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship. **Statist. Neerlandica**. 45:227-254.
- Chen YI (1994) A generalized steel procedure for comparing several treatments with a control under random right-censorship. **Commun. Statist. Simul. Comp.** 23:1-16.
- Farewell VT (1977) A model for a binary variable with time censored observations. **Biometrika** 64:43-46.
- Farewell VT (1982) The use of mixture models for the analysis of survival data with long-term survivors. **Biometrics** 38:1041-1046.
- Gehan, EA (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. **Biometrika** 52:203-223.
- Goldman AI (1984) Survivorship analysis when curve is a possibility: a Monte Carlo study. **Statist. Medicine** 3:153-163.
- Gordon NH (1990) Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. **Statist. Medicine** 9:397-407.
- Haybittle L (1965) A two-parameter model for the survival curve of treated cancer patients. **J. Amer. Statist. Assoc.** 60:16-26.
- Kaplan EL, P Meier (1958) Nonparametric estimation from incomplete observations. **J. Amer. Statist. Assoc.** 53:457-481.
- Maller RA, S Zhou (1996) Survival Analysis with Long-time Survivors. p.67-79. p.97-104. Wiley, New York.
- Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. **J. Natl. Center Inst.** 22:719-748.
- Peto R, NC Pike (1973) Conservatism of the approximation $\sum(O-E)^2/E$ in the logrank test for survival data or tumor incidence data. **Biometrics** 29:579-584.
- Su HY, CH Su, HL Wang (2000) Survival Analysis of the Effect of Temperature on Germination of Seeds. (in Chinese) **Chinese Agron. J.** 10:71-82.
- Su HY, IC Huang, HN Hsieh (2003) The Interval Censored Survival Analysis for the Effect of Temperature on Germination of Seeds. (in Chinese) **Chinese Agron. J.** 13:141-149.
- Su HY, CH Su, HN Hsieh (2005) The proportion of long-term survivor data in right censored case. (in Chinese) **Crop Environ. Bioinform.** 2:95-104.
- Turnbull BW (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. **J. Royal Statist. Soc., Series B** 38:290-295.

