東海大學統計學系研究所

碩士論文

指導教授：張玉媚 博士

兩存活中位數的非劣性檢定

Tests for Non-inferiority of Two Median Survival Times

研究生:呂理揚 撰

民國一零八年六月

東海大學碩士班研究生

論文口試委員審定書

統計學系碩士班呂理揚君所提之論文

兩存活中位數的非劣性檢定

經本委員會審議，認為符合碩士資格標準。

論文口試委員召集人 陳春樹 （簽章）

委員 [illegible]

[illegible]

中華民國 108 年 07 月 04 日

致謝

首先感謝辛苦指導我的恩師張玉媚老師，在完成碩士論文的這兩年多中，受到老師不辭辛勞的指點提攜，使我的研究能夠逐步完成，無論我的觀念理解多麼遲鈍，但老師仍很有耐心的帶著我一步步的去理解，在我遇到疑惑與不解時給予協助，謝謝老師付出的時間、精神以及這段時間對我的訓練與指導。同時，要感謝我的論文口試委員陳春樹老師以及沈葆聖老師在百忙之中抽空前來參與口試，並提供寶貴的想法及建議，在此獻上最誠摯的感激。

再來要感謝所有統計系的老師及助教們給予我課程上的協助，也感謝在我研究生涯中，同為東海統計研究所的夥伴們，在研究所的這段期間，我們共同分享了研究所生活的酸甜苦辣，有困難時大家也互相扶持，這份美好的回憶我會永存在心中，祝福各位未來的事業順利，鵬程萬里。

最後要謝謝我的家人，在這二十幾年來對我投注的心力，在人生的路上一直陪伴著我，支持我作的任何決定，並在低潮時給予支持與鼓勵，面對家人們無盡的關愛與溫暖，在此致上我最大的感激，謝謝你們。

呂理揚

謹致於東海大學統計研究所

中華民國 108 年 7 月

ABSTRACT

With the development of inexpensive treatment regimens and less invasive surgical procedures, we are confronted with non-inferiority study objectives. In the early days, statistical methods for showing the non-inferiority of a new (test) treatment compared to an established standard (reference) one had been mainly developed for binary or normal outcomes. Recently, more and more people have studied the issue of assessing drug efficacy in the case where the experimental indicator is the median failure time.

In medical practice, the median failure time is a popular measure for characterizing the survival experience and treatment effect of a group of patients. Therefore, the purpose of this paper is to compare the efficacy of the drug based on the ratio of the median failure time. We use the confidence interval method proposed by Su & Wei (1993) and the non-inferiority log-rank method to perform non-inferiority verification. Then we compare two methods for non-inferiority verification ability by simulation studies and calculate how many samples it needs to achieve the verification capabilities we want. Finally, the application of the method is explained based on a practical data.

Key words: non-inferiority test; median failure time; confidence interval

摘要

隨著廉價的治療方案和微創手術的發展，我們面臨越來越多非劣效性的研究問題，早期用於檢定新治療與標準治療所使用的非劣效性統計方法，較多的研究內容均針對二元或常態的假設下進行研究。隨著研究不斷發展，近期則越來越多人研究在實驗指標為存活中位數的情況下，評估藥效的問題。

在醫學實務中，經常根據存活中位數衡量患者的生存經歷及治療效果，因此本論文的目的是根據兩存活中位數的比值比較藥物的療效。我們利用 Su & Wei (1993) 中所提出信賴區間的方法與傳統的 log-rank 方法分別進行非劣性檢定，並透過模擬研究比較兩種方法對於非劣性的檢定能力，並且計算其需要多少樣本才能達到我們想要的檢定能力。最後根據一筆實際資料說明所提方法的應用。

關鍵詞:非劣性檢定；存活中位數；信賴區間

目錄

1. 研究背景與動機

在臨床試驗中，通常將實驗療法(第 2 組)的存活分佈與標準療法(第 1 組)的存活分佈進行比較，比較的方法依實驗目的而不同。在實驗中，我們假設比例風險模型。設$\Lambda_k(t)$為第 k 組（k=1,2）的累積危險函數和風險比率$\Delta = \Lambda_1(t)/\Lambda_2(t)$。若想知道實驗療法是否優於標準療法，我們通常假設$H_0$: $\Delta = 1$，即兩組具有相同的存活分佈，對上 H_a: $\Delta > 1$ 即實驗組的存活時間更長，我們稱之為優勢試驗(superiority trial)。

但由於醫學的不斷進步，許多廉價的新藥品和微創手術被開發出來，對於檢測實驗療法是否有不輸於標準療法的功效就變的越來越重要，若是實驗療法之療效與標準療法相差無幾，並具有其他優勢時(例如:成本較低，使用時較無風險)，就能為我們帶來更好的醫療品質。這種情況就是非劣效性試驗(non-inferiority trial)，只要有證據顯示實驗療法不比標準療法差，我們就接受實驗療法。因此，我們想要證明實驗療法等於或不低於標準療法，也就是說，給定非劣效性邊際δ，假設檢定可寫為H_0: $\Delta \leq \delta$ $v.s.$ H_a: $\Delta > \delta$ ($\delta < 1$)。

而針對非劣效性的問題已有學者研究過，Freitag (2005)整理許多針對右設限存活資料之非劣性問題的現有方法，並在文章中提到如何選擇非劣效性邊界。Jung (2017)也提出了三種針對不同情況下

的非劣效性設計，並計算了每種情況下所需之樣本數量。

本篇的目的則是想比較兩種非劣性檢定方法，一種方法源自於 Su & Wei(1993)所提出之存活中位數比值的無母數區間估計方法。我們將其方法應用於非劣性檢定，並將之與傳統的非劣性 log-rank 方法進行比較。

接下來我們會在第 2 章回顧我們想要比較的兩種方法和計算樣本的方式，並在第 3 章中給出比較的步驟，再來第 4 章我們模擬了一些情況並比較，第 5 章中我們應用於一筆實際資料做分析並且在最後給出結論。

2. 文獻回顧

2.1 信賴區間法(Su & Wei, 1993)

一般來說，治療效果的參數或半參數建模是困難的。比例風險模型不適合描述治療差異;類似地，加速失效時間的模型不適合這些數據（Freitag＆Munk, 2004）。在這種情況下，治癒模型似乎更合適，然而，不能做出相等的長期生存概率的假設。最常用的非參數方法是使用兩條存活曲線的單個時間點比較，另一種選擇是使用存活中位數的差異或比率來定義非劣效性。在 Su＆Wei（1993）中，提出了估計這些量並計算相應信賴區間的過程。

Su & Wei 提出了一種簡單且純粹的無母數推理程序，用於比較兩個存活中位數，即使兩種基礎分布的函數形狀不同，過程也是漸進有效的，而且能在不使用任何非參數密度估計的情況下導出兩存活中位數的比率和信賴區間。

假設$S_i(\cdot)$為組 i（i=1, 2）的存活函數，且設θ_{i0}和$\hat{S}_i(\cdot)$是對應的中位數和 Kaplan-Meier 的估計值。透過求解$\hat{S}_i(\theta_i) = 0.5$，可以很容易地獲得$\theta_{i0}$的一致估計$\hat{\theta}_i$。假設我們有興趣推斷$\tau_0 = g(\theta_{10}, \theta_{20})$，對於某些給定的函數$g$。例如，$\tau_0$可以是比率$\theta_{20}/\theta_{10}$或差異$\theta_{20} - \theta_{10}$。此外，假設$\theta_{20}$可以表示為$h(\tau_0, \theta_{10})$。考慮統計量

$$W(\tau_0,\theta_1)=\frac{(\hat{S}_1(\theta_1)-\frac{1}{2}))^2}{\sigma_1^2(\hat{\theta}_1)}+\frac{(\hat{S}_2\big(h(\tau_0,\theta_1)\big)-\frac{1}{2})^2}{\sigma_2^2(\hat{\theta}_2)}$$

其中$\sigma_1^2(t)$通常是$\hat{S}_i(t)$的 Greenwoods variance (Greenwoods, 1926)估計值。

這裡，θ_1是一個干擾參數，消除這個參數的一種常見方法是相對於θ_1讓$W(\tau_0,\theta_1)$最小化，然後將得到的統計量用$G(\tau_0)$表示，這是最小離散檢驗統計量(minimum dispersion test statistic)。可以證明$G(\tau_0)$漸近卡方分佈且自由度為 1。

我們可以通過反轉$G(\cdot)$來建構具有 level$(1-\alpha)$的τ_0信賴區間 I，其中 I=$\{\tau: G(\tau) < x_1^2(\alpha)\}$和$x_1^2(\alpha)$是$X_1^2$的上$\alpha$百分位數。對於任一連續分佈函數$S_1$和$S_2$，上述過程漸近有效。此外，允許兩組的設限分配不同，因為對於任何固定的τ_0，$W(\tau_0,\theta_1)$是θ_1的函數，可以很容易的獲得 I。

2.2 非劣性 log-rank 檢定

令n_k為第 k 組的樣本數$(n = n_1 + n_2)$且T_{ki}為第 k 組第 i 個病人的事件時間$(1 \le i \le n_k; k = 1,2)$。對於第 k 組，事件時間$T_{k1}, \cdots, T_{k,n_k}$為 IID，其風險函數為$\lambda_k$。在風險成比例假設下，$\Delta = \lambda_1/\lambda_2$ 表示風險比值。

我們觀察到的資料為(X_{ki}, d_{ki})，其中X_{ki}為T_{ki}和設限時間C_{ki}之最小值且d_{ki}為事件之指標 1 為有事件發生 0 為其他。$Y_k(t) = \Sigma_{i=1}^{n_k} I(X_{ki} \ge t)$ 和 $N_k(t) = \Sigma_{i=1}^{n_k} d_{ki} I(X_{ki} \le t)$ 分別為第 k 組的涉險人數和發生事件總人數，$N(t) = N_1(t) + N_2(t)$。

根據 partial likelihood 定理，Jung and Chow (2012) 提出檢定 $H_0: \Delta = \Delta_0$ *v.s.* $H_a: \Delta = \Delta_1$ $(\Delta_1 > \Delta_0)$ ，若是檢定統計量 $W(\Delta_0)/\sigma_n(\Delta_0) > Z_{1-\alpha}$ ($Z_{1-\alpha}$為上α百分位數)則拒絕H_0，α則為型 I 錯誤率，其中

$$W(\Delta) = \int_0^\infty \frac{Y_1(t)Y_2(t)}{\Delta Y_1(t) + Y_2(t)} \{d\hat{\Lambda}_1(t) - \Delta d\hat{\Lambda}_2(t)\}$$

$$\sigma_n^2(\Delta) = \Delta \int_0^\infty \frac{Y_1(t)Y_2(t)}{\{\Delta Y_1(t) + Y_2(t)\}^2} dN(t)$$

$$\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(t) dN_k(t)$$

當$\Delta_0 = 1$是標準 log-rank 檢定，當$\Delta_0 < 1$ 且$\Delta_1 = 1$ 時就是 Jung 等人 (2005)所提出的非劣性檢定。

2.3 樣本計算

根據 2.2 節，在指定型 I 錯誤率、power 及對立假設H_a: $\Delta = \Delta_1$ $(\Delta_1 > \Delta_0)$下便可估計樣本大小 n。Jung 等人(2005)提出了樣本大小公式，其中$\Delta_0 < 1$ 且$\Delta_1 = 1$，用於設計非劣性試驗。

令$p_k = {n_k}/{n}$表示第 k 組的分配比例，$n = n_1 + n_2$。本節中的漸近結果是在H_a下導出的。令$S_k(t)$和$f_k(\mathrm{t}) = -\partial S_k(t)/\partial t$分別表示$H_a$下第 k 組的存活和機率密度函數。注意到在H_a下$S_1(\mathrm{t}) = S_2(t)^{\Delta_1}$且$f_1(\mathrm{t}) = \Delta_1 f_2(t) S_2(t)^{\Delta_1 - 1}$, 對於設限時間C，令$\mathrm{G(t)} = \mathrm{P(C \geq t)}$表示在兩組中共同的設限分佈的存活函數，由 Jung 等人(2005)這篇文章可知，$\sigma_n^2(\Delta)$漸近等價於$\mathrm{n}\sigma^2(\Delta)$，其中

$$\sigma^2(\Delta) = \Delta p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t) \{p_1 f_1(t) + p_2 f_2(t)\}}{\{p_1 S_1(t) + \Delta p_2 S_2(t)\}^2} dt \qquad (1)$$

根據$W(\Delta)$的定義可得

$$W(\Delta_0) - W(\Delta_1) = n^{-1/2}(\Delta_0 - \Delta_1) \int_0^\infty \frac{Y_1(t) Y_2(t)}{\{Y_1(t) + \Delta_0 Y_2(t)\}\{Y_1(t) + \Delta_1 Y_2(t)\}} dN(t)$$

上式會漸近等價於$n\omega$，其中

$$\omega = (\Delta_0 - \Delta_1) p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t) \{p_1 f_1(t) + p_2 f_2(t)\}}{\{p_1 S_1(t) + \Delta_0 p_2 S_2(t)\}\{p_1 S_1(t) + \Delta_1 p_2 S_2(t)\}} dt \qquad (2)$$

式子(1)(2)之積分值則由數值方法獲得。

在H_a下，廣義 log-rank 檢定統計量(Jung 等人(2005))可被展開為

$$\frac{W(\Delta_0)}{\sigma_n(\Delta_0)} = \frac{W(\Delta_1)}{\sigma_n(\Delta_1)} \times \frac{\sigma_n(\Delta_1)}{\sigma_n(\Delta_0)} + \frac{W(\Delta_0) - W(\Delta_1)}{\sigma_n(\Delta_0)}$$

藉由式子(1)(2)可以近似為

$$\frac{W(\Delta_1)}{\sigma_1} \times \frac{\sigma_1}{\sigma_0} + \frac{\omega\sqrt{n}}{\sigma_0}$$

其中$\sigma_l^2 = \sigma^2(\Delta_t)$,$l = 0,1$。

假設我們想要利用廣義 log-rank 檢定統計量(Jung 等人(2005))來計算在H_a: $\Delta = \Delta_1$的樣本大小，給定 power 為$1-\beta$ 且有單尾α，則

$$1-\beta = P(\frac{W(\Delta_0)}{\sigma_n(\Delta_0)} > Z_{1-\alpha}|H_a) \approx P(\frac{W(\Delta_1)}{\sigma_1} \times \frac{\sigma_1}{\sigma_0} + \frac{\omega\sqrt{n}}{\sigma_0} > Z_{1-\alpha}|H_a)$$

由$W(\Delta_1)/\sigma_1$在H_a下漸進$N(0,1)$我們可得

$$-Z_{1-\beta} = (Z_{1-\alpha} - \frac{\omega\sqrt{n}}{\sigma_0})\frac{\sigma_0}{\sigma_1}$$

因此，我們獲得所需的樣本大小公式為

$$n = \frac{(\sigma_0 Z_{1-\alpha} + \sigma_1 Z_{1-\beta})^2}{\omega^2}$$

3. 統計方法

3.1 以信賴區間法作檢定

假設$S_i(\cdot)$分別為標準療法(i=1)與實驗療法(i=2)的存活函數，且設θ_{i0}和$\hat{S}_i(\cdot)$是對應的中位數和 Kaplan-Meier 的估計值。透過求解$\hat{S}_i(\theta_i) = 0.5$，我們獲得$\theta_{i0}$的一致估計$\hat{\theta}_i$。為了接下來的比較，我們假設$\tau_0 = {\theta_{20}}/{\theta_{10}}$，並透過信賴區間的方法先計算$W(\tau_0, \theta_1)$值，其中

$$W(\tau_0, \theta_1) = \frac{(\hat{S}_1(\theta_1) - \frac{1}{2}))^2}{\sigma_1^2(\hat{\theta}_1)} + \frac{(\hat{S}_2\big(h(\tau_0, \theta_1)\big) - \frac{1}{2})^2}{\sigma_2^2(\hat{\theta}_2)}$$

再透過θ_1使$W(\tau_0, \theta_1)$最小來取得最小離散檢驗統計量$G(\tau_0)$，並通過反轉$G(\cdot)$來建構具有 level$(1-\alpha)$的τ_0信賴區間 I，其中 I=$\{\tau: G(\tau) < X_1^2(\alpha)\}$。我們的目的是進行非劣性檢定，假設檢定可寫為

$$H_0: {\theta_{20}}/{\theta_{10}} \le \delta \;\; v.s. \;\; H_a: {\theta_{20}}/{\theta_{10}} > \delta \quad (\delta < 1)，$$

其為單尾檢定，所以我們只取區間的下界做為檢定依據，I 則改為 I=$\{\tau: G(\tau) < x_1^2(2\alpha)\}$。若檢定結果為拒絕$H_0$，則我們可以稱實驗療法不比標準療法來的差。

3.2 以 log-rank 方法作檢定

承 2.2，在風險成比例假設下，我們令風險比值$\Delta=\lambda_1/\lambda_2$。觀察到的資料為$(X_{ki}, d_{ki})$，其中$X_{ki} = min(T_{ki}, C_{ki})$，$d_{ki} = I(T_{ki} < C_{ki})$。令$Y_k(t) = \Sigma_{i=1}^{n_k} I(X_{ki} \geq t)$ 和 $N_k(t) = \Sigma_{i=1}^{n_k} d_{ki} I(X_{ki} \leq t)$ 分別為第 k 組的涉險人數和發生事件總人數。根據 partial likelihood 定理，假設檢定為

$$H_0: \Delta = \Delta_0 \; v.s. \; H_a: \Delta = \Delta_1 \quad (\Delta_1 > \Delta_0)(\Delta_0 = \delta < 1)，$$

在型 I 錯誤率為α下，我們計算$W(\Delta_0)/\sigma_n(\Delta_0)$，其中

$$W(\Delta) = \int_0^{\infty} \frac{Y_1(t)Y_2(t)}{\Delta Y_1(t) + Y_2(t)} \{d\hat{\Lambda}_1(t) - \Delta d\hat{\Lambda}_2(t)\}$$

$$\sigma_n^2(\Delta) = \Delta \int_0^{\infty} \frac{Y_1(t)Y_2(t)}{\{\Delta Y_1(t) + Y_2(t)\}^2} dN(t)$$

$$\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(t) dN_k(t)$$

若是檢定統計量$W(\Delta_0)/\sigma_n(\Delta_0) > Z_{1-\alpha}$ 則拒絕H_0，表示實驗療法不比標準療法來的差。

想要比較兩種方法，必須在兩組之存活時間均來自指數分配的假設下進行，因為時間為指數分配時，其存活中位數之比率會剛好等於風險比值之倒數，就可以進行兩種方法之比較。

通過信賴區間方法，檢定的假設為

$$H_0: {}^{\theta_{20}}/_{\theta_{10}} \leq \delta \ \ v.s. \ \ H_a: {}^{\theta_{20}}/_{\theta_{10}} > \delta \quad (\delta < 1)$$，

其中δ為我們給定的門檻，當存活時間為指數分配時，可改寫為

$$H_0: {}^{\lambda_1}/_{\lambda_2} \leq \delta \ \ v.s. \ \ H_a: {}^{\lambda_1}/_{\lambda_2} > \delta \quad (\delta < 1)$$，

風險成比例假設下$\Delta=\lambda_1/\lambda_2$，檢定的假設也可寫為

$$H_0: \Delta = \Delta_0 \ v.s. \ \ H_a: \Delta = \Delta_1 \quad (\Delta_1 > \Delta_0)(\Delta_0 = \delta < 1)$$，

所以時間為指數分配時，就能利用 log-rank 方法進行檢定並與信賴區間法所推得的非劣性檢定做比較，最簡單的方式為分別計算出兩種方法各個組合下的型 I 錯誤及 power，並在相同條件下，比較兩者之間的優劣。

4. 模擬研究

4.1 模擬設定

對於上述方法，我們進行了一些模擬研究，我們假設兩組存活函數的時間均來自隨機的指數分配，並且兩組的設限分配均為 U(0, 5)，在固定第二組的風險$\lambda_2 = 1$及型 I 錯誤率$\alpha = 0.05$下，我們模擬在不同的λ_1及不同門檻(δ)下兩種方法的檢定結果，如表一。

接下來，我們在不同檢定力下考慮我們所需要的樣本數，假設我們有興趣的 power 為 0.8 及 0.9，則在不同Δ下，我們推估了達到期待 power 下所需要的樣本數，其結果也隨著n_1及n_2所取的比例不同而有所改變，結果如表二。

4.2 模擬結果

表一中的每個項目都是由 3000 次迴圈所產生之結果。例如，考慮$n_1 = n_2 = 100$，$\delta = 0.8$，$\lambda_1 = 1.3$ 的情況，兩組的設限比例分別為 (0.153 , 0.2)。在 3,000 次迴圈中的每次迴圈中，分別生成每組 100 個觀察值，其設限比例由事先給定的設限分配 U(0, 5)來確定，利用上述條件代入兩種方法並模擬出其結果，就本案而言，兩種方法獲得的檢定力(power)分別是 0.705 及 0.932，可以看出利用信賴區間方法所做的非劣性檢定其結果不如傳統的 log-rank 檢定來的好。

觀察表一中的其他情況發現，不管λ_1和δ如何改變，所有結果都顯示 log-rank 方法能有較佳的檢定能力，不過這種情況會在樣本數增加的時候有所改善，當樣本數夠大時，兩種方法的檢定力會很接近。

再來，我們由前面所提到的方法來模擬樣本數，在設限分配為 U(0, 5)且$S_2(t) = e^{-\lambda_2 t}$，我們可以推得在$H_a$下

$$\sigma_0^2(\Delta) = \Delta_0\lambda_2 p_1 p_2 \int_0^5 \frac{G(t)e^{-\lambda_2(1+\Delta_1)t}(p_2 e^{-\lambda_2 t} + \Delta_1 p_1 e^{-\lambda_2\Delta_1 t})}{(p_1 e^{-\lambda_2\Delta_1 t} + \Delta_0 p_2 e^{-\lambda_2 t})^2} dt$$

$$\sigma_1^2(\Delta) = \Delta_1\lambda_2 p_1 p_2 \int_0^5 \frac{G(t)e^{-\lambda_2(1+\Delta_1)t}(p_2 e^{-\lambda_2 t} + \Delta_1 p_1 e^{-\lambda_2\Delta_1 t})}{(p_1 e^{-\lambda_2\Delta_1 t} + \Delta_1 p_2 e^{-\lambda_2 t})^2} dt$$

$$\omega(\Delta) = (\Delta_0 - \Delta_1)\lambda_2 p_1 p_2 \int_0^5 \frac{G(t)e^{-\lambda_2(1+\Delta_1)t}(p_2 e^{-\lambda_2 t} + \Delta_1 p_1 e^{-\lambda_2\Delta_1 t})}{(p_1 e^{-\lambda_2\Delta_1 t} + \Delta_0 p_2 e^{-\lambda_2 t})(p_1 e^{-\lambda_2\Delta_1 t} + \Delta_1 p_2 e^{-\lambda_2 t})} dt$$

最後我們由

$$n = \frac{(\sigma_0 Z_{1-\alpha} + \sigma_1 Z_{1-\beta})^2}{\omega^2}$$

來推得我們所估計之樣本數，結果如表二所示。

例如取δ為 0.7，在$(\Delta_0, \Delta_1) = (0.7, 1.2)$且$n_1 : n_2 = 1:1$的前提下，我們可獲得估計之樣本數在 power=0.8 下為$n_1 = n_2 = 52$。

表一. 兩種非劣性方法之比較結果

a.Type I error or power for n1=n2=50						
	Type I error or power					
(λ_1, λ_2)	Interval Method			Log-Rank Method		
	δ=0.7	δ=0.8	δ=0.9	δ=0.7	δ=0.8	δ=0.9
(1.3 , 1)	0.609	0.425	0.273	0.872	0.692	0.500
(1.2 , 1)	0.500	0.327	0.210	0.784	0.569	0.354
(1.1 , 1)	0.361	0.212	0.139	0.647	0.393	0.237
(1 , 1)	0.266	0.139	0.070	0.469	0.251	0.118
(0.9 , 1)	0.159	0.071	**0.033**	0.278	0.124	**0.048**
(0.8 , 1)	0.082	**0.031**	**0.011**	0.139	**0.045**	**0.015**
(0.7 , 1)	**0.041**	**0.012**	**0.003**	**0.057**	**0.011**	**0.004**

b.Type I error or power for n1=n2=100						
	Type I error or power					
(λ_1, λ_2)	Interval Method			Log-Rank Method		
	δ=0.7	δ=0.8	δ=0.9	δ=0.7	δ=0.8	δ=0.9
(1.3 , 1)	0.875	0.705	0.495	0.990	0.932	0.762
(1.2 , 1)	0.780	0.559	0.341	0.960	0.825	0.593
(1.1 , 1)	0.649	0.400	0.210	0.889	0.649	0.353
(1 , 1)	0.475	0.223	0.094	0.726	0.395	0.164
(0.9 , 1)	0.273	0.100	**0.040**	0.457	0.183	**0.050**
(0.8 , 1)	0.130	**0.038**	**0.010**	0.217	**0.046**	**0.010**
(0.7 , 1)	**0.037**	**0.012**	**0.001**	**0.048**	**0.011**	**0.003**

c.Type I error or power for n1=n2=200						
	Type I error or power					
(λ_1, λ_2)	Interval Method			Log-Rank Method		
	δ=0.7	δ=0.8	δ=0.9	δ=0.7	δ=0.8	δ=0.9
(1.3 , 1)	0.992	0.938	0.770	1.000	0.997	0.957
(1.2 , 1)	0.965	0.837	0.581	0.999	0.978	0.839
(1.1 , 1)	0.902	0.648	0.350	0.993	0.885	0.565
(1 , 1)	0.738	0.407	0.148	0.936	0.645	0.241
(0.9 , 1)	0.457	0.183	**0.033**	0.707	0.290	**0.043**
(0.8 , 1)	0.203	**0.041**	**0.006**	0.332	**0.046**	**0.003**
(0.7 , 1)	**0.038**	**0.003**	**0.001**	**0.046**	**0.001**	**0.000**

d.Type I error or power for n1=n2=300						
	Type I error or power					
(λ_1, λ_2)	Interval Method			Log-Rank Method		
	δ=0.7	δ=0.8	δ=0.9	δ=0.7	δ=0.8	δ=0.9
(1.3 , 1)	0.999	0.991	0.905	1.000	1.000	0.993
(1.2 , 1)	0.996	0.942	0.734	1.000	0.997	0.944
(1.1 , 1)	0.975	0.812	0.462	0.999	0.970	0.721
(1 , 1)	0.890	0.529	0.171	0.989	0.793	0.290
(0.9 , 1)	0.638	0.214	**0.037**	0.850	0.355	**0.051**
(0.8 , 1)	0.270	**0.042**	**0.003**	0.422	**0.051**	**0.001**
(0.7 , 1)	**0.046**	**0.004**	**0.000**	**0.051**	**0.001**	**0.000**

表二.不同 power 下所需之樣本估計

$n_1:n_2=1:1$		power			
		0.8		0.9	
Δ_0=δ	Δ_1	n_1	n_2	n_1	n_2
0.7	0.8	937	937	1302	1302
0.7	0.9	252	252	350	350
0.7	1	121	121	169	169
0.7	1.1	74	74	103	103
0.7	1.2	52	52	72	72
0.7	1.3	40	40	55	55
0.8	0.9	1142	1142	1584	1584
0.8	1	310	310	430	430
0.8	1.1	150	150	208	208
0.8	1.2	93	93	128	128
0.8	1.3	65	65	90	90
0.9	1	1390	1390	1926	1926
0.9	1.1	380	380	526	526
0.9	1.2	185	185	256	256
0.9	1.3	115	115	158	158

(續)

$n_1:n_2=1:2$		power			
		0.8		0.9	
Δ_0=δ	Δ_1	n_1	n_2	n_1	n_2
0.7	0.8	601	1202	832	1664
0.7	0.9	172	344	237	474
0.7	1	87	174	120	240
0.7	1.1	56	112	76	154
0.7	1.2	41	82	56	112
0.7	1.3	33	66	44	88
0.8	0.9	791	1582	1092	2184
0.8	1	227	454	312	624
0.8	1.1	115	230	158	316
0.8	1.2	74	148	101	202
0.8	1.3	54	108	74	148
0.9	1	1030	2060	1422	2844
0.9	1.1	296	592	406	812
0.9	1.2	150	300	206	412
0.9	1.3	97	194	132	264

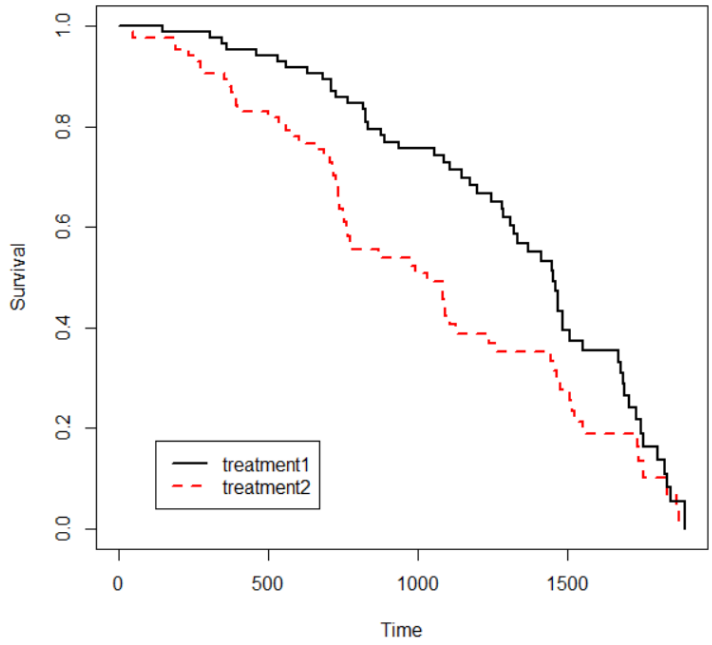
(續)

$n_1:n_2=2:1$		power			
		0.8		0.9	
Δ_0=δ	Δ_1	n_1	n_2	n_1	n_2
0.7	0.8	1624	812	2266	1133
0.7	0.9	414	207	580	290
0.7	1	188	94	266	133
0.7	1.1	110	55	156	78
0.7	1.2	74	37	104	52
0.7	1.3	54	27	76	38
0.8	0.9	1850	925	2576	1288
0.8	1	476	238	666	333
0.8	1.1	220	110	308	154
0.8	1.2	130	65	182	91
0.8	1.3	88	44	122	61
0.9	1	2110	1055	2934	1467
0.9	1.1	548	274	764	382
0.9	1.2	256	128	356	178
0.9	1.3	152	76	212	106

5. 實際資料分析

我們使用一筆有關服用健膽舒錠(Ursodeoxycholic acid)的資料(R內建資料 udca1)說明所提方法的應用。這筆資料記錄了 170 位實驗對象，他們進入實驗的時間、存活的時間，以及他們所接受的治療是實驗療法(投放安慰劑)還是標準療法(服用健膽舒錠)，是否發生事件(出現病情惡化或死亡)等等，實驗療法(trt2)及標準療法(trt1)的人數及設限人數分別是(86,30)及(84,23)人。

首先，我們畫出 Kaplan-Meier 圖形，觀察兩組的存活率是否有差異，由圖一可以看出兩者確實存在差異，可見服用健膽舒錠是有助於延長存活時間的。



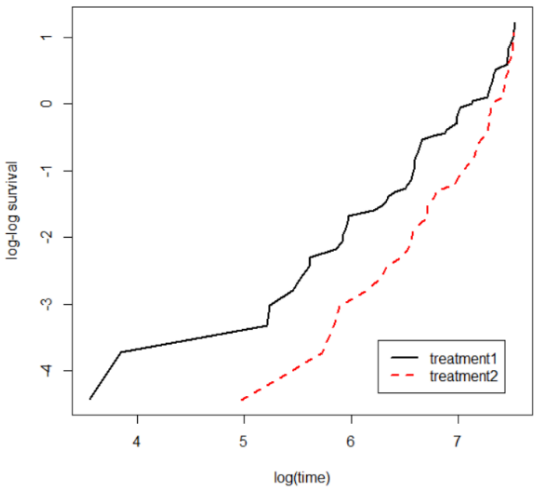
圖一、時間 v.s.存活函數曲線

接下來，我們需要確認資料是否符合指數分配的假設，因此我們使用 log-log 轉換對資料進行配適，在 Weibull 分配中，存活函數為 $S(t) = e^{-\lambda t^{\gamma}}$，函數經過 log-log 轉換後會變為

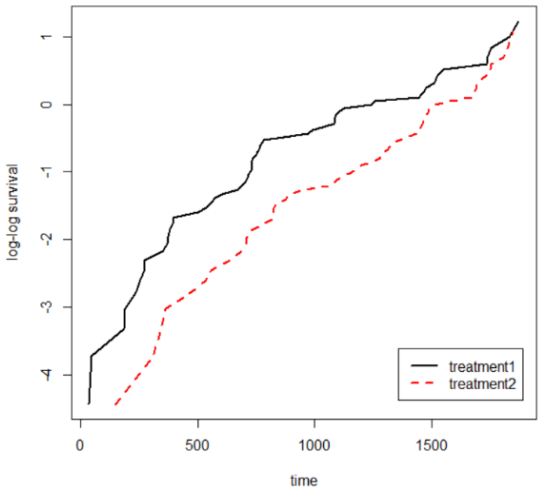
$$log\big(-log\big(S(t)\big)\big) = log(\lambda t^{\gamma}) = log(\lambda) + \gamma log(t)$$

在指數分配情況下，γ會趨近於 1，資料經過轉換後，我們利用 $\log(-log\big(S(t)\big))\, v.s.\, log(time)$圖形來判別資料是否符合指數分配，由圖二我們發現兩組資料皆很接近指數分配，所以可以用來進行檢定。

檢定前，我們還須確認資料是否符合風險成比例的假設，我們利用$\log(-log\big(S(t)\big))\, v.s.\, time$的圖形來判別，若是兩條線接近平行的狀態，則風險成比例。由圖三我們發現兩條曲線相近，因此可以推斷大致上兩組治療的風險是成比例的。接下來我們就利用信賴區間法及 log-rank 檢定非劣性，並比較結果。



圖二、log(-log(S(t))) v.s. log(t)



圖三、log(-log(S(t))) v.s. t

我們將資料用於信賴區間與 log-rank 兩種檢定方法，計算出兩組存活中位數分別為$\theta_{10} = 1453, \theta_{20} = 1033$，存活中位數比值為$\theta_{20}/\theta_{10} = 0.7109$，所得到的信賴區間下界為 0.5215。在假設型 I 錯誤率$\alpha = 0.05$情況下，當我們把門檻δ設為 0.8 時，信賴區間方法檢定結果為不拒絕虛無假設，而 log-rank 檢定之檢定統計量$W(\Delta_0)/\sigma_n(\Delta_0) = -0.662 < 1.645$也為不拒絕虛無假設，也許是門檻太過嚴苛的緣故，導致兩種方法均顯示實驗療法皆不如標準療法來的好。而當$\delta = 0.53$時，信賴區間檢定結果為不拒絕虛無假設，log-rank 檢定統計量為$W(\Delta_0)/\sigma_n(\Delta_0) = 1.661$拒絕虛無假設;另外在$\delta = 0.5$時，log-rank 檢定統計量為$W(\Delta_0)/\sigma_n(\Delta_0) = 2.016$，兩種方法同時拒絕虛無假設，也就是說兩種檢定均顯示實驗療法擁有不輸於標準療法的效果。總和來說，在這筆資料中 log-rank 方法的檢定結果的確比信賴區間法來的準確，造成兩種方法檢定的結果不同可能是因為樣本數太小時，信賴區間法較沒有足夠證據顯示實驗療法不輸於標準療法。

6. 結論

隨著醫療不斷的發展，新創醫療或藥物的非劣性議題也越來越被重視，在實際的應用上，大多是因爲現存藥物雖然療效較佳，但使用上不夠便利，或是安全性不佳；基於這些原因，有必要去開發新的用藥，證明其療效不比現存的差，但能解決上述所提到的問題。在進行非劣性試驗時，首先，就是要先定出新藥與現存的藥物可接受的療效差異值，也就是門檻δ，以進行樣本數的估算及後續的試驗。

本文僅在最簡單的情況下，比較信賴區間法與 log-rank 檢定之優劣。在指數分配且風險成比例的特定情況下，兩種方法能拿來做比較，結果顯示，在小樣本下，log-rank 方法具有較佳的檢定能力，隨著樣本的增加，兩種方法趨於一致，其結果也可以推廣到優勢試驗及等效性試驗來做比較。

在本篇文章中僅使用了前輩們所提出之檢定方法做比較及探討，因此只能在特定的資料型態下進行，若是在資料非指數分配下則無法進行比較。對於後續的議題延伸，可以考慮在更廣義的模型中、其他分配下又或是更複雜的區間設限資料下探討存活中位數的比值或差異在不同檢定方法下的檢定力，將其更普遍的應用於各種資料型態中。

參考文獻

1. John Q.Su, L.J.Wei. (1993). Nonparametric Estimation for the Difference or Ratio of Median Failure Times. Biometrics, 49:603-607.
2. Sin-Ho Jung, Shein-Chung Chow. (2012). On Sample Size Calculation for Comparing Survival Curves under General Hypothesis Testing. J Biopharm Stat. 2012;22(3):485-495.
3. Sin-Ho Jung. (2017). Design of phase II non-inferiority trials. Comtemporary Cliinical Trials Communications 7:23-27.
4. Gudrun Freitag. (2005). Methods for Assessing Noninferiority with Censored Data. Biometrical Journal 47(1),88-98.
5. Kallappa M.Koti. (2013). New Tests for Assessing Non-Inferiority and Equivalence from Survival Data. Open Journal of Statistics,2013,3,55-64.
6. Wellek S. (1993). A log-rank test for equivalence of two survivor functions. Biometrics. 49:877-881.
7. Elvis E. Martinez, Debajyoti Sinha, Wenting Wang, Stuart R. Lipsitz, and Richard J. Chappell (2017). Tests for Equivalence of Two Survival Functions: Alternative to The Tests Under Proportional Hazards. Stat Methods Med Res. 2017 February; 26(1): 75–87.
8. Efron, B. (1967). The two-sample problem with censored data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV, 831–853.
9. Freitag, G. and Munk, A. (2004). On Hadamard differentiability in k-sample semiparametric models – with applications to the assessment of structural relationships. Journal of Multivariate Analysis (to appear).

10. Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. Biometrika 68, 105–112.