行政院國家科學委員會專題研究計畫 成果報告

區間刪減數據經由多次抽樣在兩個無母數檢定法上的比較研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 97-2118-M-366-001-
執 行 期 間 ：97 年 08 月 01 日至 98 年 08 月 31 日
執 行 單 位 ：樹德科技大學休閒事業管理系

計 畫 主 持 人 ：黎進三

計畫參與人員：博士班研究生-兼任助理人員：沈清福

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 98 年 08 月 29 日

目錄

The comparison of two nonparametric tests for interval-censored failure time data via multiple imputation

1 Introduction

Interval-censored (IC) data often arise from large scale clinical trials or cohort studies, such as AIDS cohort studies and cancer follow-up studies. A famous data from a study on early breast cancer patients can be found in Finkelstein and Wolfe (1985). Two treatments were used in the study: radiotherapy alone, and primary radiation therapy along with adjuvant chemotherapy. The patients were divided into two groups according to treatment and all were examined periodically every 4-6 months. The event of interest is the time until the appearance of breast retraction and the purpose of this study is to compare the effects between the two treatments. For many patients, some successive scheduled examinations just before the examination with a changed clinical status were missed, and so the type of the observations contributed by these patients is IC.

In this report, we will consider the problem of comparing two or more IC samples and assume that there are finitely many scheduled examination times in a medical study and the underlying survival function can be discrete or continuous. The IC sample in a study contain exact (non-censored), IC and right censored observations. An IC observation occurs when the event time is only known to lie in an interval and the interval contains at least one missed examination time. A right censored observation occurs when the event time has not been observed before the end of the study, and we denote the right endpoint of this observed interval to be the sign ∞. For exact or right-censored data the problem of multi-sample comparison has been extensively studied and some non-parametric tests have been proposed (see Kalbfleisch

and Prentice, 1980). The log-rank test is a popularly used non-parametric test and there have been some related studies which extended the log-rank test to IC data in the literature as well. For example Finkelstein (1986) developed a score test to fit the Cox proportional hazards model for IC data, and the score test can be viewed as a weighted log-rank test. Fay (1999) derived score tests for IC data with three different weight functions which respectively derived from proportional hazards, proportional odds, and logistic models. Pan (2000) proposed a multiple imputation algorithm based on the approximate Bayesian bootstrap for the problem of two-sample comparison. Sun (2001), Zhao and Sun (2004), and Huang *et al.* (2008) applied the concept of Turnbull's algorithm (1976) to estimate the numbers of failures and risks of IC data, and all proposed a multiple imputation procedure to develop a log-rank type test for comparing two or more IC sample. Huang *et al.* (2008) suggested that their proposed testing procedure has better performance than the other two log-rank type tests on power and size.

In this report we applied the the testing procedure proposed by Huang *et al.* (2008) to develop a weighted log-rank type test and a versatile test. The versatile test statistic is the maximum of two weighted log-rank type test statistics which are generated by the testing procedure proposed by Huang *et al.* (2008) with different weight functions.

This report is organized as follows. In Section 2 we review the concept of the log-rank test for exact data and then propose a weighted log-rank type test for IC data. Section 3 contains the results of simulation studies for investigating the performance of the proposed tests. Section 4 is a discussion.

2 Weighted nonparametric tests for IC data

2.1 Assumptions and notation

Consider a survival study in which there are n independent subjects who come from k different treatments. Let $T_i > 0$ be a discrete random variable to denote the failure time for the ith subject, $i = 1, \ldots, n$. In this report we apply the multiple imputation method proposed by Huang *et al.* (2008) to treat IC data on testing problem. Let $t_1 < t_2 < \cdots < t_{m-1} < t_m = \infty$ be the time points at which the probability function may have mass and $p_j = \mathrm{P}(T = t_j)$, $j = 1, \ldots, m$, denote the common probability function of the k treatments under the null hypothesis. Here $T_i = \infty$ is to denote that the failure of the ith subject occurs after the last scheduled examination time. In the case of exact data, $\{t_i\}$ is the union set of observations. Note that the observed failure time data in a clinical trial can be discretized if the underlying variable is continuous. We consider the case in which, for the ith subject, his/her observation is $[L_i, R_i]$, where $T_i \in [L_i, R_i]$, $L_i, R_i \in \{t_1, t_2, ..., t_m\}$ and $L_i \leq R_i$. If $L_i < R_i$, we call it an IC observation; in particular, if $L_i < R_i = \infty$, we especially call it a right-censored observation. If $L_i = R_i$ we call it an exact observation. $L_i = R_i = \infty$ is used for convenience of notation. Our goal is to determine whether the k treatments could have arisen from an identical failure time distribution.

2.2 Weighted log-rank test for IC data

Firstly, we review the WLR test in the case of exact data. Suppose that d_j failures occur at t_j and that r_j subjects are at risk just before t_j. Let d_{jl} and r_{jl} be the corresponding numbers in the lth treatment, $l = 1, \ldots, k$, $j = 1, \ldots, m$. The WLR statistic U^0 is defined by $U^0 = (U_1^0, \ldots, U_k^0)^T$, where $(U^0)^T$ is the transpose of U^0,

$$U_l^0 = \sum_{j=1}^{m-1} w(t_j)[d_{jl} - d_j(r_{jl}/r_j)],\ l = 1, \ldots, k. \tag{1}$$

The covariance matrix of U^0 is defined by $V^0 = V_1^0 + \cdots + V_{(m-1)}^0$, where V_j^0 is a $k \times k$ matrix of rank $k-1$ with entries

$$(V_j^0)_{l_1 l_2} = \begin{cases} w(t_j)^2 r_{jl_1}(r_j - r_{jl_1}) d_j (r_j - d_j) r_j^{-2} (r_j - 1)^{-1} & \text{if } l_1 = l_2 = 1, \ldots, k, \\ -w(t_j)^2 r_{jl_1} r_{jl_2} d_j (r_j - d_j) r_j^{-2} (r_j - 1)^{-1} & \text{if } l_1 \neq l_2 = 1, \ldots, k, \end{cases} \quad (2)$$

$j = 1, \ldots, m-1$. The statistic

$$(U^0)^T (V^0)^{-1} (U^0) \quad (3)$$

is expected to have an approximate χ^2_{k-1} distribution under the null hypothesis that the k treatments have arisen from an identical failure time distribution. $(V^0)^{-1}$ is a generalized inverse of V^0. The statistic χ^2_{k-1} can be formed using any $k-1$ elements of U^0 and the corresponding $(k-1) \times (k-1)$ submatrix of V^0 (see Kalbfleisch and Prentice, 1980).Hereafter, we focus on the problem of comparing two samples and set $k = 2$.

Now consider the case of IC data, the following test procedure is based on the multiple imputation method presented by Huang *et al.* Initially, the collection $\{t_j\}$ can be chosen to be the union of the sets $\{L_i\}_{i=1}^n$ and $\{R_i\}_{i=1}^n$. Define $\alpha_{ij} = I(t_j \in [L_i, R_i])$, the indicator of the event $t_j \in [L_i, R_i]$. The maximum likelihood estimates of the p_j's can be easily obtained by using the following self-consistency equation (see Turnbull, 1976):

$$p_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{v=1}^{m} \alpha_{iv} p_v}, \ j = 1, \ldots, m \quad (4)$$

subject to the constraint $\sum_{j=1}^m p_j = 1$. Let the $\hat{p}_j$'s denote the convergent solution of the self-consistent equation. Our imputation procedure is to impute an exact failure time from an IC observation. The imputation procedure is as follows:

Let H be a prespecified positive integer and h be an integer satisfying $1 \leq h \leq H$.

Step 1: Let T_i^h be a realization drawn from the conditional probability function

$$P(T_i^h = t_j | T_i^h \in [L_i, R_i]) = \frac{\alpha_{ij} \hat{p}_j}{\sum_{v=1}^{m} \alpha_{iv} \hat{p}_v}, \ t_j \in [L_i, R_i], \ i = 1, \ldots, n, \ j = 1, \ldots, m.$$

Thus we have a set of exact data $\{T_i^h, i = 1, 2, \ldots, n\}$.

Step 2: Let the d_j^h's, r_j^h's, d_{jl}^h's and r_{jl}^h's denote the numbers of failures and risks defined by $\{T_i^h, i = 1, \ldots, n\}$. Let U^h and V^h be the statistics U^0 and V^0 defined by Eq.(1) and Eq.(2) but with the d_j's, r_j's, d_{jl}'s and r_{jl}'s replaced by d_j^h's, r_j^h's, d_{jl}^h's and r_{jl}^h's, respectively.

Step 3: Repeat Step 1 and Step 2 for each $h = 1, \ldots, H$.

Step 4: Let $\bar{U} = \sum_{h=1}^{H} U^h/H$. The variance of $\bar{U}$ is estimated by the following formula

$$\hat{V} = \frac{\sum_{h=1}^{H} V^h}{H} - \frac{\sum_{h=1}^{H}[U^h - \bar{U}]^2}{H-1} \tag{5}$$

Then the test statistic $\bar{U}^2/\hat{V}$ has approximately a χ_1^2 distribution under the null hypothesis.

2.3 Weight function

The first considered weight function is $w(t_j) = 1$ for all t_j. In many previous researches suggested that this weight function is very sensitive to detect a proportional hazards alternative for exact or right-censored data. Prentice (1978) proposed an extension of the Wilcoxon test to the case of right-censored data, and the test can be viewed as a weighted log-rank test with the weight function

$$w(t_j) = \prod_{l=1}^{j} \frac{r_l - d_l + 1}{r_l + 1}, \text{ for } j = 1, 2, \ldots, m.$$

This Wilcoxon test gives relatively more weight to earlier events than later ones, and then it is more sensitive to detect a early hazard difference alternative.

2.4 A versatile test

A weighted test performs bad if we choose an unsuitable weight function for the type of survival difference between two treatments. When we do not realize clearly the difference type, a versatile test can avoid happening the worst situation and it performs uniformly powerful for most cases of the survival difference type. Consider that a versatile test is based on the maximum value of two weighted log-rank type test statistics with different weight functions.

In the case of exact data and two-sample problem, a consistent estimator of covariance between two weighted log-rank test statistics, w_1 and w_2, can be obtained by the following equation

$$CV^0 = \sum_{j=1}^{m-1} w_1(t_j)w_2(t_j)r_{j1}r_{j2}d_j(r_j - d_j)r_j^{-2}(r_j - 1)^{-1}.$$

By applying multiple imputation to IC data, the procedure of estimating the corresponding covariance is presented as follows:

Continuing the procedure presented in Section 2.2, for different weight functions w_1 and w_2, let U_1^h's, U_2^h's, $\bar{U}_1$, $\bar{U}_2$, $\hat{V}_1$ and $\hat{V}_2$ be the corresponding values of U^h's, $\bar{U}$ and $\hat{V}$, respectively.

Step 5: For each $h = 1, \ldots, H$, let CV^h be the statistics CV^0 but with the d_j's, r_j's, d_{jl}'s and r_{jl}'s replaced by d_j^h's, r_j^h's, d_{jl}^h's and r_{jl}^h's, respectively. The covariance of $\bar{U}_1$ and $\bar{U}_2$ is estimated by

$$\widehat{CV} = \frac{\sum_{h=1}^{H} CV^h}{H} - \frac{\sum_{h=1}^{H}[U_1^h - \bar{U}_1][U_2^h - \bar{U}_2]}{H-1}. \tag{6}$$

Step 6: The correlation coefficient ρ of $\bar{U}_1$ and $\bar{U}_2$ can be estimated by $\hat{\rho} = \widehat{CV}/\sqrt{\hat{V}_1\hat{V}_2}$. Let the versatile test statistic U_{max} be $Max(\bar{U_1}^2/\hat{V}_1, \bar{U_2}^2/\hat{V}_2)$.

rom Lee (1996), the random vector $(\bar{U}_1/\sqrt{\hat{V}_1}, \bar{U}_2/\sqrt{\hat{V}_2})$ has a bivariate Normal distribution with means $= (0, 0)$, variances $= (1, 1)$ and the correlation coefficient$=\rho$. We do not know the true value of ρ, so the true critical value is also unknown. For completing this testing process, we obtain an approximate value of the critical value by simulation method. We generate 10000 values of U_{max}, where $(\bar{U}_1/\sqrt{\hat{V}_1}, \bar{U}_2/\sqrt{\hat{V}_2})$ is sampling from a bivariate Normal distribution with $\hat{\rho}$, and then sort the 10000 values from small to big and take the $9500th$ value to be the desired value at a significance level of 0.05.

3 Simulation Studies

We conducted a simulation study to assess the size and power properties of the three tests under a proportional hazards model and four non-proportional hazards models.

For the proportional hazards model, failure times are generated from the exponential distributions with hazards 0.2 (mean=5) for population 1 and $0.2\exp(\beta)$ for population 2, $\beta = 0.6, -0.6$, and there are n_1 observations $\{[L_i, R_i], i = 1, \ldots, n_1\}$ from population 1 and n_2 observations $\{[L_i, R_i], i = n_1 + 1, \ldots, n_1 + n_2\}$ from population 2. For a non-proportional hazards model, each group of failure times is generated from a piecewise exponential distribution. The relation between two survival curves $(S_1(t), S_2(t))$ are characterized by early hazard difference alternative, late hazard difference alternative, and crossing hazard difference alternative I and II. Let $\lambda_i(t)$ is the corresponding hazard function for $S_i(t), i = 1, 2$. The form of survival function is

$$S(t) = \exp\{-\sum_{j=1}^{m} \lambda(t_j)(t_j - t_{j-1})I(t_j \leq t)\}, \text{ where } t_0 = 0.$$

The non-proportional hazards model configurations are:

(i) Early hazard difference alternative:

$\lambda_1(t) = 0.1I(t \leq 3) + 0.2I(t > 3)$

$\lambda_2(t) = 0.3I(t \leq 3) + 0.2I(t > 3)$

(ii) Late hazard difference alternative:

$\lambda_1(t) = 0.2I(t \leq 4) + 0.1I(t > 4)$

$\lambda_2(t) = 0.2I(t \leq 4) + 0.5I(t > 4)$

(iii) Crossing hazard difference alternative I:

$\lambda_1(t) = 0.1I(t \leq 3) + 0.3I(3 < t \leq 6) + 0.2I(t > 6)$

$\lambda_2(t) = 0.3I(t \leq 3) + 0.1I(3 < t \leq 6) + 0.2I(t > 6)$

(iv) Crossing hazard difference alternative II:

$\lambda_1(t) = 0.2I(t \leq 2) + 0.1I(2 < t \leq 5) + 0.5I(5 < t \leq 8) + 0.2I(t > 8)$

$\lambda_2(t) = 0.2I(t \leq 2) + 0.5I(2 < t \leq 5) + 0.1I(5 < t \leq 8) + 0.2I(t > 8)$

Figure $1 \sim 5$ display the survival functions for these five alternatives.

Let $n = n_1 + n_2$ and $n_1 = n_2 = n/2$. Take $n = 100$. The estimated size and power are based on 50,000 and 10,000 replications, respectively, at a significance level of 0.05. $H = 200$ is used.

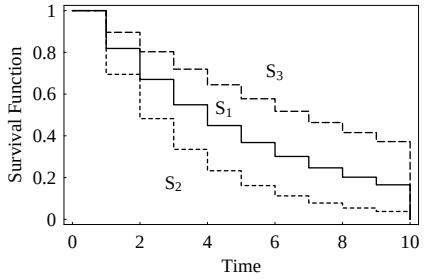


Figure 1: Proportional hazard difference alternative
$\lambda_1(t) = 0.2$ for all t
$\lambda_2(t) = 0.2 \exp(0.6)$ for all t
$\lambda_3(t) = 0.2 \exp(-0.6)$ for all t

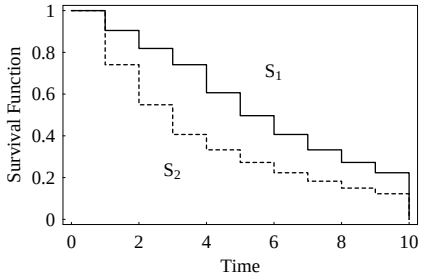


Figure 2: Early hazard difference alternative
$\lambda_1(t) = 0.1I(t \leq 3) + 0.2I(t > 3)$
$\lambda_2(t) = 0.3I(t \leq 3) + 0.2I(t > 3)$

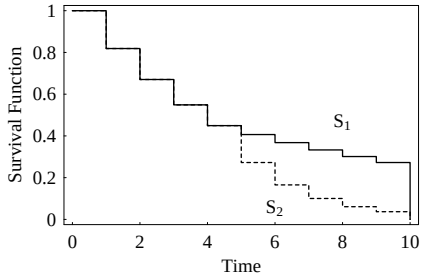


Figure 3: Late hazard difference alternative
$\lambda_1(t) = 0.2I(t \leq 4) + 0.1I(t > 4)$
$\lambda_2(t) = 0.2I(t \leq 4) + 0.5I(t > 4)$

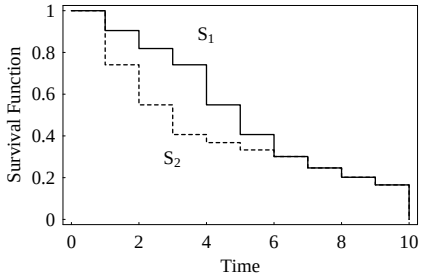


Figure 4: Crossing hazard difference alternative I
$\lambda_1(t) = 0.1I(t \leq 3) + 0.3I(6 \geq t > 3)$
$+ 0.2I(t > 6)$
$\lambda_2(t) = 0.3I(t \leq 3) + 0.1I(6 \geq t > 3)$
$+ 0.2I(t > 6)$

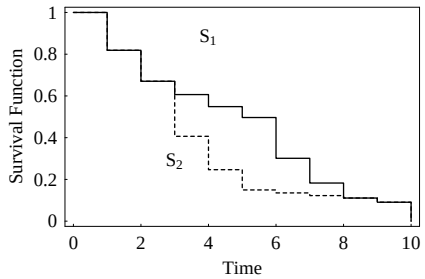


Figure 5: Crossing hazard difference alternative II
$\lambda_1(t) = 0.2I(t \leq 2) + 0.1I(5 \geq t > 2)$
$+ 0.5I(8 \geq t > 5) + 0.2I(t > 8)$
$\lambda_2(t) = 0.2I(t \leq 2) + 0.5I(5 \geq t > 2)$
$+ 0.1I(8 \geq t > 5) + 0.2I(t > 8)$

To generate IC data we first assume that a subject is examined periodically, say at 1, ..., 9, but that he/she may miss several scheduled examinations and the probability of missing a scheduled examination is q ($0 \leq q < 1$), which is independent of all other missing patterns and failure times. Under this assumption we discretize the exponential distributions at 1, 2, ..., 9 and ∞, and take $q = 0.5$. We say that an IC observation has length j if $(R - L + 1) = j$. It is easy to understand that the proportion of observations that have a long length is larger as q is larger.

In the following two Tables, LR-MI denotes the weighted log-rank type test with $w(t_j) = 1$; Wilcoxon-MI denotes the weighted log-rank type test with $w(t_j) = \prod_{l=1}^{j} \frac{r_l - d_l + 1}{r_l + 1}$; WLR-Max denotes the proposed versatile test. Table I presents the estimated size of the three tests. The nominal levels of the three tests for all alternatives are close to the specified significant level 0.05, but the nominal level is slightly higher than the specified significant level for crossing hazard difference alternative I.

Table II presents the estimated power of the three tests. The LR-MI test has better performance than the other two tests on power for proportional hazards model and late hazard difference alternative, and it is very powerful for late hazard difference alternative. The Wilcoxon-MI test has better performance than the other two tests for early hazard difference alternative and crossing hazard difference alternative I, and it is very powerful for crossing hazard difference alternative I. For crossing hazard difference alternative II, the three tests have similar power, but the Wilcoxon-MI test and WLR-Max test have slightly more power than the LR-MI test. From some simulation results in the literature, the Wilcoxon-MI test is known to be more sensitive to early hazard difference alternative and crossing hazard difference alternative I than the LR-MI test, but the LR-MI test is more sensitive to proportional hazards model and late hazard difference alternative than the Wilcoxon-MI test. From our simulation results, the same situation also occurs between the LR-MI test and the Wilcoxon-MI test. The WLR-Max test performs

consistently well and has performance close to the most powerful test of the three tests on each alternative.

Table I: Estimated size of tests for IC data

	Test		
Model	LR-MI	Wilcoxon-MI	WLR-Max
Proportional hazards	0.049	0.049	0.049
Early hazard difference	0.051	0.050	0.050
Late hazard difference	0.052	0.051	0.052
Crossing hazard difference I	0.053	0.053	0.053
Crossing hazard difference II	0.052	0.050	0.051

Table II: Estimated power of tests for IC data

	Test		
Model	LR-MI	Wilcoxon-MI	WLR-Max
Proportional hazards			
$\beta = \quad 0.6$	0.771	0.696	0.750
$\beta = -0.6$	0.686	0.641	0.670
Early hazard difference	0.690	0.829	0.803
Late hazard difference	0.571	0.212	0.524
Crossing hazard difference I	0.205	0.494	0.447
Crossing hazard difference II	0.433	0.465	0.470

4 Discussion

The number of multiple imputation must be taken enough to ensure stability of simulation results. We suggest that the number of multiple imputation should be greater than 100 for LR-MI test or Wilcoxon-MI test and be greater than 200 for WLR-Max test. Our proposed method is attractive for its simplicity and generality.

For the problem of comparing two or more IC samples, we can apply it to each existing tests of comparing two or more exact samples. The variance or covariance estimators via our method for IC data only needs to use the formula of variance or covariance estimator derived from exact data, and the formula is not complicated in most cases of exact data, so the variance or covariance estimator via our method is easy to compute. For the future work, we can try the other forms of weight function to find a more powerful test for each alternative.

References

1. Fay, M. P. (1999). Comparing several score tests for interval-censored data. *Statistics in Medicine*, **18**(3), 273-285.

2. Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, **41**(4), 933-945.

3. Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**(4), 845-854.

4. Lee, J. W. (1996). Some Versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, **52**(2), 721-725.

5. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data.* Wiley: New York, 1980, 16-18.

6. Pan, W. (2000). A two-sample test with interval-censored data via multiple imputation. *Statistics in Medicine*, **19**(1), 1-12.

7. Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika*, **65**, 167-179.

8. Schick, A. and Yu, Q. Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

9. Sun, J. (1996). A non-parametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine*, **15**(13), 1378-1395.

10. Sun, J. (2001). Nonparametric test for doubly interval-censored failure time data *Lifetime data analysis*, **7**, 363-375.

11. Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistics Society. Series B*, **38**(3), 290-295.

12. Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine*, **23**(10), 1621-1629.

成果自評：

從模擬結果來看，將 Huang (2008)等人所提出應用於區間刪減數據下的多次抽樣法推廣至加權對數秩型態檢定法(weighted log-rank type test)，其型 I 誤差(type I error)的估計值非常接近預設值，顯示此一套用方式的效果非常好。另外我們也提出一多功能式的檢定法，模擬結果顯示此一檢定法在所討論的四種存活函數差異模型下具有均勻優異的鑑別力表現。目前正嘗試不同於計畫所提類型的加權函數，並且套用於其他不同型式的檢定法，研究他們套用多次抽樣法的鑑別力表現與型 I 誤差的估計準確度如何。近期會將相關論文投稿至國際期刊。

參加 2009 年國際統計研討會 JSM 報告

國際統計研討會是整個國際統計界最大規模的學術研討會，今年主辦單位將研討會的地點選在美國首都華盛頓. 出席此次研討會的有來自世界各國統計界的先進，他們服務的單位包括政府機構、研究單位、藥廠、大學，還有許多即將從研究所畢業，急著找事的碩士生、博士生。

會議一共安排六天，從八月一日開始到八月六日，整個會議分為 611 個 session，每個 session 有五至七個左右的演講, 所以僅僅 speaker 就超過 3000 人，其它純粹與會而沒給演講的人大約亦有 2000 人左右，所以整個會議期間至少有五六千人參加。 此次人數比往年多，會議時程比往年長，可能是因為在首都的關係，政府機構林立，又幾乎都跟統計離不開關係，例如 F.D.A. (Food and Drug Administration), U.S.D.A (United States Department of Agriculture), N.I.H. (National Institute of Health) 等, 幾乎裡面的研究人員都來參加了。我的講題是 Generalized Wilcoxon test for interval-censored failure time data. 另外余岐青教授（紐約州立大學 Binghamton 教授）也在另一個 session 中介紹我們和另外二位博士班學生合作的論文。

在會議期間碰到許多從台灣來的同行，大約有五六個。也碰到一些過去在美國一起唸書而現在在美國政府機構服務的朋友, 他們研究的問題都和我有密切的關係，所以和他們在次見面收益良多。

Generalized Wilcoxon test for interval-censored failure time data

Chinsan Lee

National Sun Yat-sen University, Kaohsiung,Taiwan

Shu-Te University of Technology, Kaohsiung,Taiwan

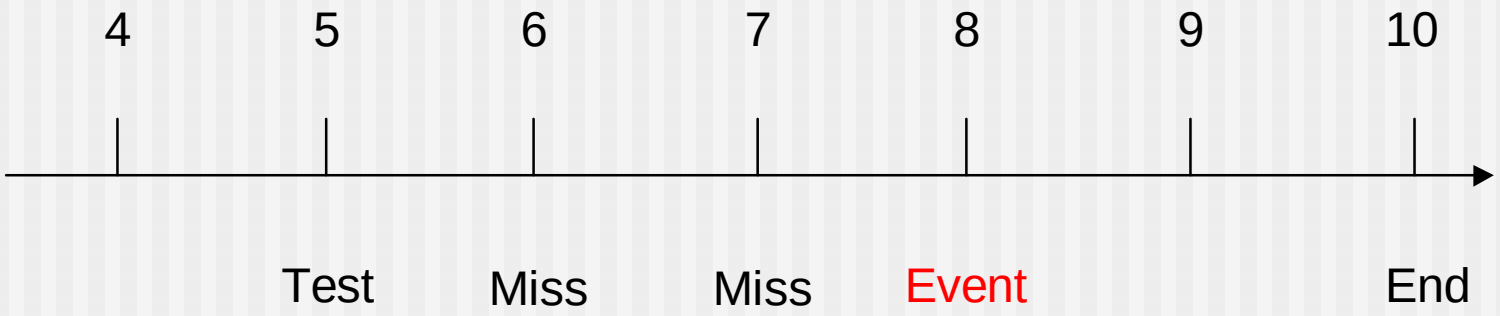
1 Introduction

- Interval censored data: $(X_L, X_R]$

  Interval censored failure time data often occur in follow-up studies where subjects can only be followed periodically and the failure time can only be known to lie in an interval.

Intreval-censored data:

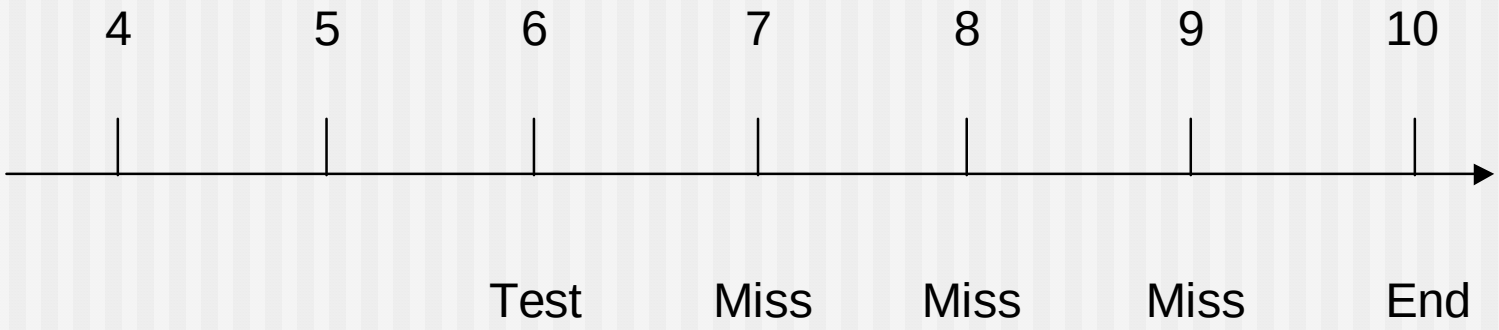
The event time is only known to lie in an interval and the interval contains at least one missed examination time



Observation: (5, 8]

Right censored data:

The event time has not been observed before the end of the study



Observation: $(6, \infty)$

- Turnbull (1976) has introduced a self-consistency equation to compute the maximum likelihood estimator of the survival function for arbitrarily censored and truncated data.

Generalized Wilcoxon test

Mantel (1967) extends Gehan's (1965) generalized Wilcoxon test to interval censored data.

Peto and Peto (1972) give a different version of generalized Wilcoxon test.

2 Rank tests for interval censored data

Assume that X is measured in discrete units and takes values $0 < x_1 < x_2 < \cdots < x_m$

Let $U = \{(0, x_1], (x_1, x_2], (x_2, x_3], ..., (0, x_m]\}$

be the collection of all $m(m+1)/2$ admissible intervals and define

$$p_j = P(X = x_j) \text{ where } \sum_{j=1}^{m} p_j = 1$$

If we have two samples for X and Y are respectively

$$(X_L^i, X_R^i], i = 1, 2, \cdots, n_1 \text{ and } (Y_L^i, Y_R^i],\ i = 1, 2, \cdots, n_2$$

Test the equality of survival function

$$H_0 : S_X(t) = S_Y(t), \forall t \geq 0$$

2.1 Mantel's generalized Wilcoxon test

$$W = \sum_{k=1}^{n_1} V_k, \text{ where } V_k = \sum_{h=1}^{n_1+n_2} V_{kh},$$

$$V_{kh} = \begin{cases} 1 & \text{if we know for sure obs-k>obs-h,} \\ -1 & \text{if we know for sure obs-k<obs-h,} \\ 0 & \text{if not sure.} \end{cases}$$

Under H_0,the test statistic is normal distributed with mean 0 and variance

$$Var(W) = n_1 n_2 \sum_{k=1}^{n_1+n_2} \frac{V_k^2}{(n_1+n_2)(n_1+n_2-1)}$$

2.2 Peto and Peto's generalized Wilcoxon test

Define the score of *i*th observation as

$$U_i = \frac{f(\hat{S}(X_L^i)) - f(\hat{S}(X_R^i))}{\hat{S}(X_L^i) - \hat{S}(X_R^i)},$$

where $\hat{S}$ is the estimated survival function

$f(y) = y^2 - y$ and $U_i = \hat{S}(X_L^i) + \hat{S}(X_R^i) - 1.$

Test statistic

$$Z^2 = (\frac{Y_1^2}{n_1} + \frac{Y_2^2}{n_2})/s^2,$$

where $Y_1 = \sum_{i=1}^{n_1} U_i,\ Y_2 = \sum_{i=n_1+1}^{n_1+n_2} U_i,$

and $s^2 = \sum_{i=1}^{n_1+n_2} U_i^2/(n_1 + n_2 - 1).$

Under H_0, the test statistic

Z^2 is distributed as χ_1^2

2.3 Weighted rank test

Let R_i be the rank given to $x_i, i = 1, 2, ..., m.$

For instance, $R_i = i$ corresponds to the Wilcoxon rank.

Rewrite any observation,

$$(Y_L^j, Y_R^j] \text{ say, as } (Y_L^j, Y_R^j] = (x_{u^{(j)}}, x_{v^{(j)}}],$$

$$\text{where } x_{u^{(j)}}, x_{v^{(j)}} \in \{0, x_1, x_2, ..., x_m\} \text{ and } x_{u^{(j)}} < x_{v^{(j)}}$$

We give the weighted rank

$$\text{rank}\left((Y_L^j, Y_R^j]\right) = \sum_{l=u^{(j)}+1}^{v^{(j)}} \frac{p_l}{p_{u^{(j)}+1} + ... + p_{v^{(j)}}} R_l$$

Let W_1, W_2 be respectively the average weighted rank of the X and Y samples, that

$$W_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \text{rank}\left((X_L^i, X_R^i]\right) = \frac{1}{n_1}\sum_{i=1}^{n_1}\Big(\sum_{l=u^{(i)}+1}^{v^{(i)}} \frac{p_l}{p_{u^{(i)}+1} + \ldots + p_{v^{(i)}}} R_l\Big)$$

$$W_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \text{rank}\left((Y_L^j, Y_R^j]\right) = \frac{1}{n_2}\sum_{j=1}^{n_2}\Big(\sum_{l=u^{(j)}+1}^{v^{(j)}} \frac{p_l}{p_{u^{(j)}+1} + \ldots + p_{v^{(j)}}} R_l\Big)$$

Then, we propose the test statistic

$$W_S = \frac{W_1 - W_2}{\sqrt{Var(W_1) + Var(W_2)}}.$$

By C.L.T, W_S is approximately a standard normal random variable.

Fay (1999) assume the probability of return to inspect at each time point $x_1, x_2, ..., x_m$ is q . The model assume:

$A_1, A_2, ..., A_{m-1}$ are iid random variables distributed with $bernulli(q)$, where $0 < q < 1$;

X, $(A_1, A_2, ..., A_{m-1})$ are independent;

The observable random vector is $(X_L, X_R] = (x_{s_j}, x_{t_j}]$,

where $s_j = \max_l\{0 \leq l < j : A_l = 1\}$,

$l_j = \min_l\{j \leq l \leq m : A_l = 1\}$ and $x_{s_j} < X \leq x_{t_j}$

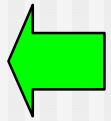
Intervals are selected by the following four rules:

$$(i)\ Q\{(0, x_r]\} = \sum_{j=1}^{r} p_j q(1-q)^{r-1},\ 1 \le r < m.$$

$$(ii)\ Q\{(0, x_m]\} = \sum_{j=1}^{m} p_j (1-q)^{m-1}.$$

$$(iii)\ Q\{(x_u, x_v]\} = \sum_{j=u+1}^{v} p_j q^2 (1-q)^{r-1},\ 1 \le u < v < m$$

$$(iv)\ Q\{(x_u, x_m]\} = \sum_{j=u+1}^{m} p_j q(1-q)^{m-u-1},,\ 1 \le u < m$$



Let $(U, 2^U, Q)$ be an arbitrary probability space as defined in section 2. We defined a random variable Z on this space.

$$Z\{(x_u, x_v]\} = \sum_{l=u+1}^{v} \frac{p_l}{p_{u+1} + ... + p_v} R_l, \qquad 0 \leq u < v \leq m. \quad (1)$$

where R_l is the rank given to x_l among $x_1, x_2, ..., x_m$.

Theorem 1: Suppose *Z* is the random variable defined on the probability space $(U, 2^U, Q)$ according to (1). Then the expectation of *Z*, E(*Z*), can be simplified as

$$E(Z) = \sum_{l=1}^{m} p_l R_l$$

which is independent of the choice of q.

The variance of *Z*, *Var(Z)*, is

$$Var(Z) = E(Z^2) - E^2(Z) = \sum_{i=1}^{m(m+1)/2} Q(I_i)R^2(I_i) - E^2(Z)$$

where $Q(I_i)$ and $R(I_i)$ are the selected probability

and the weighted rank of the *i*th admissible interval of I_i respectively, $I_i \in U$.

Let $Z_1, Z_2, ..., Z_{n_j}$, $j = 1, 2$, be the random sample from Z, then, we can know the expectations and the variances of

W_1 and W_2 are $E(W_j) = E(\bar{Z}_j) = \sum_{l=1}^{m} p_l R_l$

$$Var(W_j) = \frac{1}{n_j} Var(Z) = \frac{1}{n_j} \left(\sum_{i=1}^{m(m+1)/2} Q_j(I_i) R^2(I_i) - E^2(Z) \right), \ j = 1, 2,$$

where $\bar{Z}_j$ and $Q_j(I_i)$ are the sample mean of the rank and the selected probability of the ith admissible interval in each population respectively.

$$W_S = \frac{W_1 - W_2}{\sqrt{Var(W_1) + Var(W_2)}} = \frac{\bar{Z}_1 - \bar{Z}_2}{\sqrt{\frac{Var(Z_1)}{n_1} + \frac{Var(Z_2)}{n_2}}}$$

is approximately a standard normal random variable.

In practical computation, we replace $Var(Z_1)$ and $Var(Z_2)$ by $\hat{Var}(Z_1)$ and $\hat{Var}(Z_2)$ respectively, where $\hat{Var}(Z_1)$ and $\hat{Var}(Z_2)$ are the estimated variances with $\hat{q}_1$, $\hat{q}_2$, and $\hat{p}$ replacing q_1, q_2 and p

2.3.1 m.l.e of p and return probability q

- Turnbull's algorithm, $$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_j^i p_j}{\sum_{l=1}^{m}\alpha_l^i p_l},\ j = 1, 2, ..., m.$$

- Consider the formulas, ➡

the likelihood function can be written as

$$L(p_1, p_2, ..., p_m, q) = P(p_1, p_2, ..., p_m)G(q), \text{ where } G(q) = q^{k_1}(1-q)^{k_2}$$

q can also be estimated by $\frac{k_1}{k_1+k_2}$ trivially,

where k_1 and k_2 are determined by the sample.

3 Simulation study

We set $x_i = i,\ i = 1, 2, ..., 5$ and $x_6 = \infty$.

When the failure times are generated from the exponential distribution, we set the population 1 with hazards 0.3 and $0.3e^{\beta}$ for population 2.

When the failure times are generated from the lognormal distribution, we set the parameter μ and σ with 1 for population 1, and $\mu = 1 + \beta, \sigma = 1$ for population 2.

Table 1:

dis exp	q	test	β	-0.4	-0.2	0	0.2	0.4
						n=100		
	0.8	WRT		0.364	0.145	0.047	0.140	0.410
		HLY		0.375	0.156	0.054	0.145	0.430
		Mantel		0.353	0.144	0.049	0.131	0.381
		Peto		0.354	0.144	0.048	0.126	0.369
	0.5	WRT		0.341	0.126	0.053	0.136	0.379
		HLY		0.352	0.126	0.055	0.144	0.387
		Mantel		0.325	0.124	0.051	0.134	0.359
		Peto		0.332	0.123	0.048	0.131	0.347
						n=200		
	0.8	WRT		0.615	0.207	0.049	0.234	0.683
		HLY		0.647	0.219	0.052	0.242	0.700
		Mantel		0.608	0.201	0.049	0.225	0.648
		Peto		0.606	0.198	0.049	0.222	0.637
	0.5	WRT		0.553	0.204	0.048	0.219	0.641
		HLY		0.566	0.200	0.049	0.227	0.654
		Mantel		0.542	0.203	0.047	0.212	0.618
		Peto		0.545	0.200	0.046	0.209	0.599

					n=100		
lognormal							
	0.8	WRT	0.439	0.164	0.054	0.160	0.429
		HLY	0.425	0.154	0.054	0.160	0.408
		Mantel	0.441	0.165	0.052	0.158	0.427
		Peto	0.438	0.165	0.053	0.157	0.430
	0.5	WRT	0.398	0.140	0.047	0.124	0.395
		HLY	0.390	0.135	0.050	0.116	0.372
		Mantel	0.397	0.140	0.047	0.123	0.388
		Peto	0.399	0.142	0.048	0.119	0.385
					n=200		
	0.8	WRT	0.724	0.265	0.049	0.259	0.724
		HLY	0.706	0.251	0.047	0.248	0.700
		Mantel	0.730	0.258	0.050	0.255	0.722
		Peto	0.732	0.259	0.051	0.251	0.726
	0.5	WRT	0.654	0.221	0.056	0.221	0.666
		HLY	0.624	0.204	0.054	0.214	0.643
		Mantel	0.658	0.219	0.049	0.221	0.658
		Peto	0.665	0.227	0.055	0.223	0.665

4 An application to AIDS cohort study

Consider the data of 262 hemophilia patients, 105 patients who received at least 1,000 μg/kg of blood factor for at least one year between 1982 and 1985, and 157 patients who received less than 1,000 μg/kg in each year.

The failure time of interest is the time of HIV seroconversion. The object is to test the difference of the failure times between the two treatments.

Applying our test, Mantels test, Peto and Peto's test and H.L.Y test to this data set, the values of these test statistics are -7.815, -7.352, 56.476 and 59.480 respectively, and all the four p-values are closed to 0. The four tests all conclude that the HIVseroconversion appeared significantly different.