Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6557–6564
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the versatility of the combination of the weighted log-rank statistics

Seung-Hwan Lee*

Department of Mathematics and Computer Science, Illinois Wesleyan University, Bloomington, IL 61702, USA

Received 30 June 2006; received in revised form 2 March 2007; accepted 6 March 2007
Available online 12 March 2007

Abstract

In many applied situations, it is difficult to specify in advance the types of survival differences that may exist between two groups. Therefore, it is tempting to use some tests that emphasize these differences, but are sensitive to a wide range of the survival differences. In this paper such versatile tests are considered, whose procedures are based on the simultaneous use of the weighted log-rank statistics that are asymptotically normal under the null hypothesis of no difference between two groups. Simulations are performed to examine power of the tests in small and moderate sample sizes when the data are uncensored to heavily censored. Implementation of the procedures are discussed in a real data example for illustration.

Keywords: Survival distribution; Versatile test; Weighted log-rank statistics

1. Introduction

Weighted log-rank tests are commonly used for testing the equality of two survival distributions in the presence of arbitrary right censoring because of their sensitivity to survival difference. With a suitable choice of weight function, these tests may assign more weight on either early, middle, or late survival differences. A particularly important class is the $G^{\rho,\gamma}$ family (Fleming and Harrington, 1991, p. 257) with weights of the form $\{\hat{S}(t-)\}^{\rho}\{1-\hat{S}(t-)\}^{\gamma}$ for $\rho \geqslant 0, \gamma \geqslant 0$, where $\hat{S}(t-)$ is the left-continuous Kaplan–Meier (Kaplan and Meier, 1958) estimate, also referred to as the product-limit estimate of the survival function based on the pooled survival data. Various weighted log-rank statistics, as individual members of the $G^{\rho,\gamma}$ family, can be obtained by the selection of ρ and γ. For example, when $\gamma = 0$, $\rho = 0$ and 1 correspond to the log-rank statistic (Cox, 1972; Mantel, 1966) and the Prentice–Wilcoxon statistic (Gehan, 1965), respectively. The log-rank statistic assigns equal weights throughout the study, and is known to have optimum power in detecting proportional hazards in two groups. The Prentice–Wilcoxon statistic assigns more weights on early survival differences. On the contrary, the $G^{0,1}$ statistic with $\rho = 0$ and $\gamma = 1$ emphasizes late differences.

In clinical studies with censored data, test statistics are frequently selected to assess treatment effects because of their sensitivity to ordered hazards alternatives (Kosorok and Lin, 1999). For example, statistics that give greater (lesser) weight to events occurring early in the study will be more (less) sensitive to early differences in survival differences, as mentioned in Wu and Gilbert (2002). Due to this fact, properly chosen weight functions should be used, since a

* Tel.: +1 309 556 3421; fax: +1 309 556 3864.
E-mail address: slee2@iwu.edu.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.03.006

poorly chosen weight function may lead to decreased sensitivity to the actual observed differences. However, in many situations, it is very difficult to specify in advance the types of survival differences that will actually occur. Consequently, misspecification of the weights in weighted log-rank statistics may result in undesirable outcomes in power of tests.

To remedy the above-mentioned problems, in this paper some versatile test procedures are considered, which emphasize the survival differences that may exist between two groups, but are sensitive to a wide range of types of survival differences. Special attention will be given to early and/or late survival differences that frequently occur in clinical studies. Various methods regarding versatile test procedures have been proposed by many authors, including Gill (1980), Tarone (1981), Fleming and Harrington (1981), and Fleming et al. (1987). More recent works include Self (1991), Lee (1996), Kosorok and Lin (1999), Shen and Cai (2001), and Wu and Gilbert (2002). Fleming et al. (1987) studied the versatility of the supremum versions of the log-rank and Wilcoxon statistics across a wide range of survival differences, and Wu and Gilbert (2002) proposed some linear rank tests based on quadratic weighted log-rank statistic, among others.

Some versatile tests with a broad range of sensitivity properties are discussed based on the combination of the statistics $G^{1,0}$ and $G^{0,1}$ that weight early and late survival differences, respectively. The weighted log-rank statistics, however, can be highly correlated, especially when there is heavy censoring (Fleming and Harrington, 1991). Because of this, the versatility of combined tests, based on the weighted log-rank statistics, can severely decline as the censoring increases. Thus, the combined tests may not be more versatile than either test used separately. As a way to circumvent this problem, Yang et al.'s (2005) method is used in the procedures, which provides good power of combined tests, especially when the correlation between tests is high. It is reasonable from Yang et al.'s (2005) method that combined tests, retaining good versatility, would be obtained. We consider three combined tests: the absolute value of their average, the average of their absolute values, and the maximum of their absolute values. Important insights into the properties of the proposed tests are obtained from simulations conducted for different types of alternatives in small and moderate sample sizes with various censoring proportions. Simulation results show that the maximum of the statistics $G^{1,0}$ and $G^{0,1}$ appears to have an overall best performance across a wide range of the alternatives compared with other proposed statistics while being nearly as sensitive as the corresponding $G^{1,0}$ and $G^{0,1}$ in detecting early and late survival differences, respectively.

This paper is organized as follows: Section 2 describes the proposed procedures. Numerical studies are discussed in Section 3, including simulations and illustration in a real data example. Section 4 concludes this paper.

2. Some versatile tests

We assume the two-sample general random censorship model. Consider two censored samples with sample sizes n_1 and n_2. Let T_{ij}, $i=1,2$, $j=1,\ldots,n_i$, be independent, positive random variables and C_{ij} be independent censoring variables that are also independent of the survival variables T_{ij}. For the censored two-sample model, we observe (X_{ij}, Δ_{ij}), $i=1,2$, $j=1,\ldots,n_i$, where $X_{ij}=\min(T_{ij}, C_{ij})$ and $\Delta_{ij}=I(T_{ij}\leqslant C_{ij})$, where I is the indicator function. These data can be expressed in terms of counting process notation. Let $N_{ij}(t)=I(X_{ij}\leqslant t, \Delta_{ij}=1)$ and $Y_{ij}(t)=I(X_{ij}\geqslant t)$. Also let

$$\bar{N}_i(t)=\sum_{j=1}^{n_i} N_{ij}(t), \quad \bar{Y}_i(t)=\sum_{j=1}^{n_i} Y_{ij}(t) \quad \text{for } i=1,2.$$

Note that $\bar{N}_i(t)$ is the number of failures in group i before or at time t and $\bar{Y}_i(t)$ the number at risk in group i at time $t-$, $i=1,2$. Define survival functions $S_i(t)=P(T_{ij}\geqslant t)$, $i=1,2$, which are assumed to be absolutely continuous throughout the paper. The hypothesis of interest is that two survival functions are same. That is $\mathrm{H}_0: S_1(t)=S_2(t)$ for all t. The commonly used two-sample censored data rank statistics, called the weighted log-rank statistics, can be expressed as

$$W_k=\sqrt{\frac{n_1+n_2}{n_1 n_2}}\int_0^\infty W(t)\frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}\left\{\frac{\mathrm{d}\bar{N}_1(t)}{\bar{Y}_1(t)}-\frac{\mathrm{d}\bar{N}_2(t)}{\bar{Y}_2(t)}\right\},$$

where $W(t)$, $t\geqslant 0$, is a bounded nonnegative weight function. Since the weight function is sensitive to the alternative hypothesis, in applications, carefully chosen weight functions should be used. We consider a family of censored data