Statistics and Probability Letters 80 (2010) 1459–1466

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A class of semiparametric rank-based tests for right-truncated data

Pao-sheng Shen

Department of Statistics, Tunghai University, Taichung, 40704, Taiwan

ARTICLE INFO

Article history:
Received 11 November 2009
Received in revised form 1 April 2010
Accepted 23 May 2010
Available online 2 June 2010

Keywords:
Retrospective sampling
Semiparametric model
Two-sample tests

ABSTRACT

A class of semiparametric rank-based tests is proposed for the two-sample problem with right-truncated data, where the truncation distribution is parameterized, while the lifetime distribution is left unspecified. The class contains as special cases the extension of the semiparametric Mann–Whitney test proposed by Bilker and Wang (1996) for right-truncated data. The asymptotic distribution theory of the test is presented. The small-sample performance of the test is investigated under a variety of situations by means of Monte Carlo simulations.

1. Introduction

Parallel to random censorship models, random truncation models have various applications in astronomy, demography, epidemiology and other fields. In epidemiologic studies, statistical data are sometimes observed subject to a retrospective sampling criterion resulting in pure truncated data. Consider the following application.

Example 1: AIDS blood transfusion data

The AIDS Blood Transfusion Data provides an example of pure right-truncated data. Data are collected by the Centers for Disease Control (CDC), which is from a registry data base, a common source of medical data. The data were retrospectively ascertained for all transfusion-associated AIDS cases in which the diagnosis of AIDS occurred prior to the end of the study (June 30, 1991). The data consist of the time in month and only cases having either one transfusion or multiple transfusions in the same calendar month were used. Thus the date of infection is known for those transfusion-associated AIDS patients (Kalbfleish and Lawless, 1989). Nevertheless, cases either diagnosed or reported after June 30, 1989, were not included to avoid bias resulting from reporting delay. Also, cases having the infecting transfusions prior to July 1, 1977 were not included because this is when adults started being infected by the virus from a contaminated blood transfusion. Hence, the sampling scheme consists of observing the chronologic times of infection and disease for a random sample of persons who are infected and develop the disease in some chronologic time interval $[0, \tau]$, where τ denotes the length of the observed period (i.e. from July 1, 1977 to June 30, 1989; $\tau = 144$ months). Let U^* denote the chronologic time of infection. The lifetime of interest (denoted by T^*) is the time from HIV infection to diagnosis of AIDS, i.e. induction time. Hence, the chronologic time of diagnosis with AIDS is $U^* + T^*$ and only the subjects with $U^* + T^* \leq \tau$ are included in the data. Note that $V^* = \tau - U^*$ is often referred to as right truncation times. A selection bias results from the exclusion of those patients who develop AIDS after the end of study.

In this article, we consider hypothesis testing for the comparison of two subgroups ($i = 1, 2$) for right-truncated data. For example, it is of interest to compare the time (denoted by T_i^*, $i = 1, 2$) from HIV infection to diagnosis of AIDS between two subgroups of age at HIV infection. In Example 1, the right truncation times V_i^* corresponds to the time from HIV infection to the end of study. Assume that V_i^* is independent of T_i^*. For right-truncated data, one can observe nothing if $T_i^* > V_i^*$ and observe (T_i^*, V_i^*) if $T_i^* \leq V_i^*$. Much work has been developed in the literature focusing on truncation data problems; see

E-mail address: psshen@thu.edu.tw.

0167-7152/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2010.05.013

papers by Woodroofe (1985), Lagakos et al. (1988), Lai and Ying (1991), and Wang (1989, 1991, 1992). The Mann–Whitney test (Mann and Whitney, 1947) and the Wilcoxon test (Wilcoxon, 1945) are two closely associated nonparametric two-sample tests for the case of complete data. Extensions of nonparametric two-sample tests have been developed for the right-censored data. The Gehan test (Gehan, 1965) is an extension of the Mann–Whitney test that allows right-censored data. The log-rank test (Peto and Peto, 1972) is an extension of the Mantel–Haenzel test (Mantel and Haenzel, 1959; Mantel, 1966). Based on the integrated weighted difference in Kaplan–Meier estimators (Kaplan and Meier, 1958), Pepe and Fleming (1989) proposed a class of distance tests for right-censored data. For truncated data, Lagakos et al. (1988) studied a weighted log-rank test, which is a nonparametric test, especially for the case of pure right- (or left-) truncated data. Therefore, both the lifetime and the truncation distributions are considered nonparametrically. For the case of right-truncated data, Chi et al. (2007) proposed a nonparametric test statistic based on the integrated weighted difference between two estimated survival functions. Shen (2009) proposed a class of rank-based tests for left-truncated and right-censored data. The class contains as special cases the extension of log-rank test and Gehan test.

In some scenarios, there is enough information on the form of the truncation distribution to determine a well-fitting parametric form. In Example 1, the variable U^* reflects the growth of HIV, the rate of infections can be increasing exponentially beginning in the time of onset (July 1, 1977) until the recruitment time (June 30, 1991). In this case, the truncation distribution U_i^* may follow an exponential distribution. Hence, the truncation distribution of $V_i^* = \tau - U_i^*$ can be fitted well by a parametric form, a more powerful test for pure right-truncated data, as compared to the weighted log-rank test, can result from incorporating this information into the testing procedure. For pure right-truncated data, Bilker and Wang (1996) derived a semiparametric testing procedure that extends the Mann–Whitney test (1947) in which truncation distribution is assumed. Their simulation results indicate that for pure-right truncation data, the semiparametric test is more powerful than the weighted log-rank test. In Section 2, a class of semiparametric rank-based tests is proposed for the two-sample problem with pure right-truncated data, where the truncation distribution is parameterized, while the lifetime distribution is left unspecified. The asymptotic distribution theory of the proposed test is derived. In Section 3, the small-sample performance of the test is investigated under a variety of situations by means of Monte Carlo simulations.

2. A class of rank-based tests

For subgroup i ($i = 1, 2$), let F_i^* and G_i denote the distribution function of T_i^* and V_i^*, respectively. Let $a_{F_i^*} = inf\{t : F_i^*(t) > 0\}$ and $b_{F_i^*} = inf\{t : F_i^*(t) = 1\}$ denote the left and right support of F_i^*, respectively. Note that the left and right support of G_i is equal to 0 and τ, respectively. If $a_{F_i^*} = 0$ ($i = 1, 2$) and $b_{F_i^*} \le \tau$ ($i = 1, 2$), then F_i^* and G_i are all identifiable (see Woodroofe, 1985). However, when $b_{F_i} > \tau$, $F_i^*(t)$ is non-identifiable and only $F_i(t) = P(T_i^* \le t | T_i^* \le \tau)$ is identifiable. Hence, it is only possible to test the null hypothesis $H_0 : F_1(t) = F_2(t)$ for all $t \le \tau$, and the alternative hypothesis can be two-sided $H_a : F_1(t) \ne F_2(t)$ or one-sided, e.g. $H_a : F_1(t) < F_2(t)$. For $i = 1, 2, j = 1, \ldots, n_i$, let (T_{ij}, V_{ij}) denote the right-truncated sample from subgroup i.

2.1. Semiparametric rank-based tests

Suppose that the distributions of V_i^* ($i = 1, 2$) are parameterized as $G_i(x; \theta_i)$ for $0 < x \le \tau$, where $\theta_i \in \Theta_i \subset R^{q_i}$, and θ_i is a q_i-dimensional vector. Let $\beta_i = P(T_i^* \le U_i^*) = \int_{a_{F_i^*}}^{b_{F_i^*}} \bar{G}_i(x; \theta_i) dF_i(x)$ denote the probability of un-truncation. Based on the observed data (T_{ij}, V_{ij}), the likelihood can be written as $L = L_1 L_2$, where

$$L_1 = \prod_{j=1}^{n_1} \frac{g_1(V_{1j}; \theta_1)}{\bar{G}_1(T_{1j}; \theta_1)} \prod_{j=1}^{n_2} \frac{g_2(V_{2j}; \theta_2)}{\bar{G}_2(T_{2j}; \theta_2)}$$

and

$$L_2 = \prod_{j=1}^{n_1} \frac{\bar{G}_1(T_{1j}; \theta_1) dF_1(T_{1j})}{\int \bar{G}_1(T_{1j}; \theta_1) dF_1(T_{1j})} \prod_{j=1}^{n_2} \frac{\bar{G}_2(T_{2j}; \theta_2) dF_2(T_{2j})}{\int \bar{G}_2(T_{2j}; \theta_2) dF_2(T_{2j})},$$

where $\bar{G}_i(x; \theta_i) = 1 - G_i(x; \theta_i)$ ($i = 1, 2$) denotes the survival function of V_i^* and $g_i(x; \theta_i)$ ($i = 1, 2$) denotes the probability density function of V_i^*.

Subject to each fixed $\theta = (\theta_1, \theta_2)$, the unique nonparametric maximum likelihood estimates of $F_1(t)$ and $F_2(t)$ derived by maximizing the marginal likelihood L_2 can be explicitly expressed, respectively, as (see Vardi, 1985; Qin and Wang, 2001)

$$\hat{F}_1(t; \theta_1) = \frac{\sum_{j=1}^{n_1} I_{[T_{1j} \le t]} / \bar{G}_1(T_{1j}; \theta_1)}{\sum_{j=1}^{n_1} 1 / \bar{G}_1(T_{1j}; \theta_1)}, \quad t \le \tau$$