Advances and Applications in Statistics
Volume 12, Number 1, 2009, Pages 1-9
Published Online: May 19, 2009
This paper is available online at http://www.pphmj.com

A SIMPLE NONPARAMETRIC APPROACH OF SURVIVAL DATA WITH TWO-STAGE OUTCOMES OF DIFFERENT SCALES

CHANGYONG FENG[1], CARLO ERCOLI[2] and HONGYUE WANG[1]

[1]Department of Biostatistics and Computational Biology
University of Rochester
Rochester, NY 14642, U. S. A.
e-mail: feng@bst.rochester.edu

[2]Estman Dental Center
University of Rochester
Rochester, NY 14642, U. S. A.

Abstract

In this paper, we construct a simple nonparametric test to analyze the survival data consisting of observations from two sequential stages with different scales. Simulation studies show that this method has very good control of the type I error and turns out to be more powerful than the log-rank test. We apply our method to a study of dental research.

1. Introduction

In survival analysis with censored data, the log-rank test (Peto and Peto [4]; Kalbfleisch and Prentice [2]) is widely used to compare the survival functions of treatment groups. This kind of test turns out to be asymptotically efficient under the hypothesis of proportional hazards (Peto and Peto [4]; Cox [1]). On the other side, if there is no censoring, the

2000 Mathematics Subject Classification: 62G20, 62G35.

Keywords and phrases: survival analysis, Wilcoxon rank-sum test, log-rank test.

Received January 14, 2009

two-sample t-test (under the normality assumption) or the Wilcoxon rank-sum test can be used. In this paper, we develop a rank-sum test for a new kind of survival data with two different kinds of time scales from two sequential stages.

In the first stage, the experimental units were put on an instrument at the same time. The trial was designed to be terminated at a fixed time. The variable of interest in this stage is the time to failure. However, by the end of the first stage, some units may have not experienced the event of interest. This means that their survival times are censored. Then all units surviving the first stage experiment go to the second stage.

The second stage experiment is destructive. The interest is in the force needed to break the experiment units. After these two stages, we are interested in comparing the strength of two groups of units made from two kinds of materials.

This is not a typical survival problem. If we only have the data from the first stage, we can use the survival time as the measurement of strength in units. Then the log-rank test or other nonparametric methods can be easily used to test the difference between the two groups. On the other hand, if we only have the data in the second stage, the two sample t-test or some nonparametric methods can be easily used. The data from these two stages have different scales. In the first stage, the survival time is measured by days, hours, etc. In the second stage, the force is measured by Newton, etc.

This problem offers challenges to the current statistical methods. There are two scales to measure the strength of the material. One possible solution is to define the strength as a function of the survival time in the first stage and the force in the second stage. This method transforms the two-dimensional two-different scale data into one-dimensional. The two-sample t-test or the Wilcoxon rank-sum test can be used for the collapsed data. However, there are some problems with this method. First, we do not know the form of this function. Second, the censoring in the first stage makes it difficult to evaluate this function.

In the next section, we describe the data and the model used in our analysis.

2. Data and Model

Here we describe the structure of the data. Suppose there are two groups in the study. Let T_{ij}, $i = 1, 2$, $j = 1, ..., n_i$ denote the underlying survival times in the first stage. Here n_i is the number of units in group i. Define

$$\delta_{ij} = \mathbf{1}_{\{T_{ij} \le C\}}, \quad X_{ij} = \min(T_{ij}, C) = T_{ij}\delta_{ij} + C(1 - \delta_{ij}),$$

where $\mathbf{1}_A$ is the indicator function of set A and C is the pre-specified termination time in the first stage. Let R_{ij1} be the ranking of X_{ij} in the pooled data from two groups, i.e.,

$$R_{ij1} = \sum_{k=1}^{2}\sum_{l=1}^{n_k} \mathbf{1}_{\{X_{kl} \le X_{ij}\}}.$$

For the second stage data, only the units surviving the first stage $(\delta_{ij} = 0)$ will have observations in this stage. For these individuals, let the tolerable force to destroy the unit be U_{ij}. Then we define the tolerable force for all individuals in the second stage as

$$Y_{ij} = \begin{cases} 0, & \text{if } \delta_{ij} = 1, \\ U_{ij}, & \text{if } \delta_{ij} = 0. \end{cases}$$

Let R_{ij2} be the ranking of Y_{ij} in the pooled data in the second stage, i.e.,

$$R_{ij2} = \sum_{k=1}^{2}\sum_{l=1}^{n_k} \mathbf{1}_{\{Y_{kl} \le Y_{ij}\}}.$$

The strength of a unit is $S(X_{ij}, Y_{ij})$, which is a function of both the survival time at the first stage and the tolerable force at the second stage. Here we assume that S is an increasing function of two variables, i.e., if $X_{ij} \ge X_{i'j'}$ and $Y_{ij} \ge Y_{i'j'}$, then $S(X_{ij}, Y_{ij}) \ge S(X_{i'j'}, Y_{i'j'})$ (*consistency assumption*).

For each unit, we define the overall rank as the convex combination of the rank from two stages, i.e.,

$$Z_{ij}(\lambda) = \lambda R_{ij1} + (1 - \lambda) R_{ij2}, \quad \lambda \in [0, 1].$$

In this way we transfer a two-dimensional problem to a one-dimensional problem.

The overall rank is very intuitive and meaningful. Higher rank means higher strength. The weight λ can be chosen at our disposal. If we think the survival time is more important than the tolerable force, then we can put more weight to the first stage. We can also see that overall rank satisfies the consistency assumptions of strength. Therefore we can use the overall rank Z_{ij} (which is scale-free) as a measurement of the strength.

3. Wilcoxon Two-sample Test Based on Overall Ranking

Here we use the Wilcoxon two-sample test to compare the strength of two groups based on their overall rankings. Define the rank of Z_{ij} among those overall rankings by R_{ij}, which

$$R_{ij}(\lambda) = \sum_{k=1}^{2} \sum_{l=1}^{n_k} \mathbf{1}_{\{Z_{kl}(\lambda) \le Z_{ij}(\lambda)\}}.$$

The Wilcoxon two-sample test is defined as

$$W(\lambda) = \sum_{j=1}^{n_1} R_{1j}(\lambda) = \sum_{j=1}^{n_1} \sum_{k=1}^{2} \sum_{l=1}^{n_k} \mathbf{1}_{\{Z_{kl}(\lambda) \le Z_{1j}(\lambda)\}}.$$

First we prove the following theorem:

Theorem 1. *If* $0 < \lambda < 1$, *then* $W(\lambda)$ *is independent of* λ.

Proof. We only need to prove that $\mathbf{1}_{\{Z_{kl}(\lambda) \le Z_{ij}(\lambda)\}}$ is independent of λ. It is easy to see that $Z_{kl}(\lambda) \le Z_{ij}(\lambda)$ can only occur in the following three cases:

(a) if $\delta_{ij} = \delta_{kl} = 1$, then $Z_{kl}(\lambda) \le Z_{ij}(\lambda)$ iff $R_{ij1} \ge R_{kl1}$.

(b) if $\delta_{ij} = \delta_{kl} = 0$, then $Z_{kl}(\lambda) \le Z_{ij}(\lambda)$ iff $R_{ij2} \ge R_{kl2}$.

(c) if $\delta_{ij} = 0$, $\delta_{kl} = 1$, then $Z_{kl}(\lambda) \le Z_{ij}(\lambda)$ iff $R_{ij1} \ge R_{kl1}$.

In all cases, $\mathbf{1}_{\{Z_{kl}(\lambda)\le Z_{ij}(\lambda)\}}$ is independent of λ.

From Theorem 1 we can let $\lambda = 1/2$ and rewrite the Wilcoxon two-sample test as

$$W(\lambda) = \sum_{j=1}^{n_1}\sum_{k=1}^{2}\sum_{l=1}^{n_k}\mathbf{1}_{\{R_{kl1}+R_{kl2}\le R_{ij1}+R_{ij2}\}}.$$

Let X_{ij}^* and Y_{ij}^* denote, respectively, the numerical values of X_{ij} and Y_{ij}. For example, if $X_{ij}=5$ hours, and $Y_{ij}=200$ Newtons, then $X_{ij}^*=5$, $Y_{ij}^*=200$. Let $U_{ij} = X_{ij}^* + Y_{ij}^*$. It is easy to prove that

$$R_{kl1} + R_{kl2} \le R_{ij1} + R_{ij2} \quad \text{iff} \quad U_{kl} \le U_{ij}.$$

Using this fact we transform the Wilcoxon two-sample test based on Z_{ij} to a regular Wilcoxon two-sample test based on U_{ij}. This means that for our two-stage data based on two different scales, the Wilcoxon two-sample test can be constructed based on the sum of numerical values from two different scales.

4. Simulation Study

In the simulation study, we compare the empirical size and power of our Wilcoxon two-sample test based on the overall rank to that of the log-rank test based on the first stage data.

1. Empirical size

In the first stage, suppose the survival time T has an exponential distribution with rate 0.02. Therefore the mean survival time is 50. The censoring time C ranges from 40 to 100 with censoring rate from 44.9% to 13.5%.

For those units surviving the first stage, the destructive force of the second stage is assumed to have a normal distribution with mean 400 and standard deviation 30. Table 1 reports the empirical size of the log-rank test and the Wilcoxon two-sample test (theoretical significance level

is 0.05) after 2,000 Monte Carlo simulations. The sample size ranges from 10 to 40.

Table 1. Empirical size (2,000 simulations)

		Sample size					
		n = 10		n = 20		n = 40	
C	Censoring rate (%)	Log-rank	Wilcoxon	Log-rank	Wilcoxon	Log-rank	Wilcoxon
40	44.9	0.040	0.053	0.050	0.048	0.040	0.044
60	30.1	0.042	0.055	0.053	0.054	0.048	0.052
80	20.2	0.050	0.046	0.046	0.046	0.057	0.055
100	13.5	0.050	0.052	0.050	0.052	0.058	0.058

From Table 1, we can see that both tests have empirical size very close to the theoretical significance level even with high censoring rate (44.9%) and relatively small sample size n = 10.

2. Empirical power

We assume that the second group is statistically better than the first group. In the first stage, the distribution of T in two groups is exponential with rates 0.02 and 0.01, respectively. In the second stage, the tolerable forces have normal distributions with $N(400, 900)$ and $N(420, 900)$, respectively.

Table 2. Empirical power (2,000 simulations)

	Sample size					
	n = 10		n = 20		n = 40	
C	Log-rang	Wilcoxon	Log-rang	Wilcoxon	Log-rang	Wilcoxon
40	0.104	0.324	0.271	0.568	0.536	0.876
60	0.170	0.298	0.359	0.524	0.657	0.830
80	0.206	0.267	0.414	0.486	0.716	0.782
100	0.236	0.248	0.456	0.468	0.745	0.762

Table 2 reports the empirical power of two tests under different censoring times and different sample sizes. We can see that

(1) The powers of two tests increase with sample size. This is what we have expected.

(2) The Wilcoxon two-sample test based on overall rank is more powerful than the log-rank test based on the first stage data only.

(3) Given the sample size, as censoring time increases, the power of the log-rank test increases. However, the power of the Wicoxon decreases. As we know, the log-rank test is the most powerful nonparametric test in the case of proportional hazards. With higher censoring rates, more units go to the second stage. This makes the Wilcoxon two-sample test more powerful than the log-rank test. As the censoring rate decreases, the data in the first stage offer more and more information. This increases the power of the log-rank test.

From the simulation results, we can see that the Wilcoxon two-sample test based on the overall rank has good control of the significance level and better power than the log-rank test just based on the data from the first stage.

5. Real Example

In this section, we analyzed the data from a dental study at the Eastman Dental Center at the University of Rochester Medical Center.

The purpose of this study was to investigate the influence of veneering porcelain thickness for all-ceramic crowns on failure resistance after cyclic loading. All-ceramic crowns ($n = 20$) were fabricated on an implant abutment (RN solid abutment) for the study. Two different framework designs with 2 different incisal thicknesses of veneering porcelain (2 mm and 4 mm) were used for each all-ceramic in 2 experimental groups ($n = 10$ in each group) with identically shaped crowns. The all-ceramic crowns consisted of alumina frameworks and veneering porcelain. All crowns were cemented on the corresponding abutments using a resin cement. They were subjected to 1,000 cycles of thermal cycling.

In the first stage, each specimen was mechanically tested with a custom-designed cyclic loading apparatus. This apparatus delivered simultaneous unidirectional cyclic loading at 135 degrees to the long axis

of the tooth to simulate the force application to a maxillary incisor, at an average rpm of 250 with a load of 49 N. The load was applied to the lingual aspect of the specimens at 2.5 mm below the incisal edge, using a round stainless steel indenter of 6 mm diameter. The frequency was monitored at least once a day during each testing with a contact tachometer. Each specimen was kept continuously wet by applying saline solution with a custom made delivery system and was loaded for 1.2×10^6 cycles simulating 5 years of clinical service or until it failed. The specimens were thoroughly evaluated for the presence of cracks with an optical stereomicroscope at $\times 10$ magnification. The specimen was considered as 'failure' if there was bulk fracture or the crack occurred on the facial aspect of the crown. If these complications occur in a clinical situation, the crown would likely be replaced. The specimen that was not categorized as 'failure' was categorized as 'survival'.

The specimens that did not show bulk fracture were further tested (second stage). They were loaded on the incisal edge along the long axis of the tooth with an 8 mm diameter flat stainless steel piston until they fractured, using a universal testing machine (500 Kg load cell) at a crosshead speed of 1.5 mm/min. To decrease the possibility that a localized stress application could determine fracture of the porcelain, a 1 mm layer of tin was interposed between the crown and the loading apparatus.

After the first stage, there were 5 failures in each group. The mean and standard deviation of forces in the second stage for two groups are 1513.24 ± 410.77 and 1263.10 ± 124.40, respectively. The p-value of our Wilcoxon two-sample test is 0.39. This shows no significant difference of resistance between two groups.

6. Discussion

In this paper, we propose a simple method to deal with a new type of survival data. This kind of data consists of observations from two sequential stages with different scales. The Wilcoxon two-sample test

was generalized to the data consisting of the overall rank from two stages. Our method turns to have good control of the significance level and is more powerful than the log-rank test only based on the first stage data (survival time). The method in this paper can be easily generalized to the case of more than two groups which is the Kruskal-Wallis test (Kruskal and Wallis [3]) based on the overall ranks. In our methods we assume type I censoring in the first stage. One possible future work is to generalize our method to the case of random censoring and interval censoring in the first stage.

References

[1] D. R. Cox, Regression models and life tables (with discussion), J. Roy. Statist. Soc. B 34 (1972), 187-220.

[2] J. D. Kalbfleisch and R. L. Prentice, The Statistical Analysis of Failure Time Data, 2nd Edition, Hoboken, Wiley, New Jersey, 2002.

[3] W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, J. Amer. Statist. Assoc. 47 (1952), 583-621.

[4] R. Peto and J. Peto, Asymptotically efficient rank invariant procedures (with discussion), J. Roy. Statist. Soc. A 135 (1972), 185-206.

[5] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945), 80-83.