응용통계연구 제7권 2호
1994년 9월 pp. 79-88

Weighted Log Rank Test for Late Differences[1)]

Gyu-Jin Jeong, Sang-Gue Park[2)]

Abstract

Weighted log rank test is a widely applicable test when one is interested in detecting the differences between two groups. In many clinical trials it is common to see no differences in early experiments and does show significant differences later. We propose new weighted log rank test and illustrate it through an example. We also examine the empirical powers and show that the proposed test is more sensitive to detect late differences.

1. Introduction

In clinical studies investigators frequently want to determine if subjects from one population live longer than those from the other population. Let F_1 and F_2 denote the cumulative distribution functions of two populations to be compared, then this problem can be formalized with the following hypotheses;

$$H_0: F_1(t) = F_2(t) \quad \text{v.s.} \quad H_1: F_1(t) \geq F_2(t) \tag{1}$$

for any t, $0 < t < \infty$, with strict inequalities for some t.

There are large literatures about testing (1). Medical experiments frequently involve some censored data, and many researchers studied nonparametric methods analyzing such data(Details in Miller(1981)). Among others Gehan(1965), Peto and Peto(1972) and Prentice(1978) generalized Wilcoxon test for testing (1). Mantel(1966) and Cox(1972) proposed the so-called log rank test. Tarone and Ware(1977) showed that these tests are unified as weighted log rank test, and proposed a weighted log rank test with another weight.

As pointed out in Tarone and Ware(1977), the tests presently in use may be insensitive to detect late differences because of their own weighing scheme. We consider, in this paper, the situation that there show no differences in early in experiment, but do show significant differences in late of experiment.

In Section 2, we will discuss several weighted log rank tests and propose two modifications for testing (1). We present an illustrated example in Section 3. In Section 4,

1) This research is supported by 1994 HANNAM University Research Fund.
2) Deptartmemt of Applied Statistics, Hannam University, Taejon, KOREA.

we compare the empirical powers of discussed tests under various situations via simulation studies, and give some concluding remarks in Section 5.

2. Weighted log rank test

Let N be the number of subjects in the combined sample, and let $t_1 < \ldots < t_k$ denote distinct failure times in the combined sample.

The well known test procedures for (1) is the weighted log rank test and the test statistic with weight wi is defined as follows;

$$WLR = \frac{\sum_{i=1}^{k} w_i(a_i - E(a_i))}{\sqrt{\sum_{i=1}^{k} w_i^2 Var(a_i)}}, \qquad (2)$$

where $a_i (i = 1,2,\ldots,k)$ is the number of failures in the first sample at t_i. And $E(a_i)$ and $Var(a_i)$ are given by

$$E(a_i) = \frac{m_{i1}n_{i1}}{N_i}, \qquad Var(a_i) = \frac{m_{i1}(N_i - m_{i1})}{N_i - 1}\frac{n_{i1}}{N_i}(1 - \frac{n_{i1}}{N_i}),$$

where m_{i1} is the number of failures at t_i in the combined sample, n_{i1} and N_i are numbers at risk at time t_i in the first sample and combined sample, respectively.

Since the weighted log rank statistic (2) is well known to be asymptotically normally distributed with mean 0 and variance 1(see, for example, Fleming and Harrington(1991), chapter 7), one can test the hypotheses (1) by the following rule; reject H_0 if WLR < $Z(\alpha)$, where $Z(\alpha)$ is the lower α quantile of standard normal distribution.

Many test statistics mentioned in the Section 1 can be expressed in the form of weighted log rank statistic. Tarone and Ware(1977) discussed the weighted log rank statistic (2). They showed that WLR becomes Gehan-Wilcoxon statistic when $w_i = N_i$ and WLR does the log rank statistic (also called Mantel-Haenszel(1959) statistic) when $w_i = 1$, and proposed one with $w_i = \sqrt{N_i}$. Later Prentice(1978) generalized Wilcoxon test, which was similar to the works of Peto and Peto(1972). Peto-Prentice test can also be expressed as (5) when $w_i = \prod_{j<i}(N_j + a_j - 1)/(N_j + a_j)$. This weight is, in fact, the product limit estimate (say, $S(t_i)$) of the survival function at time t_i.

Each weighing system of the weighted log rank statistic has own properties. First, it has

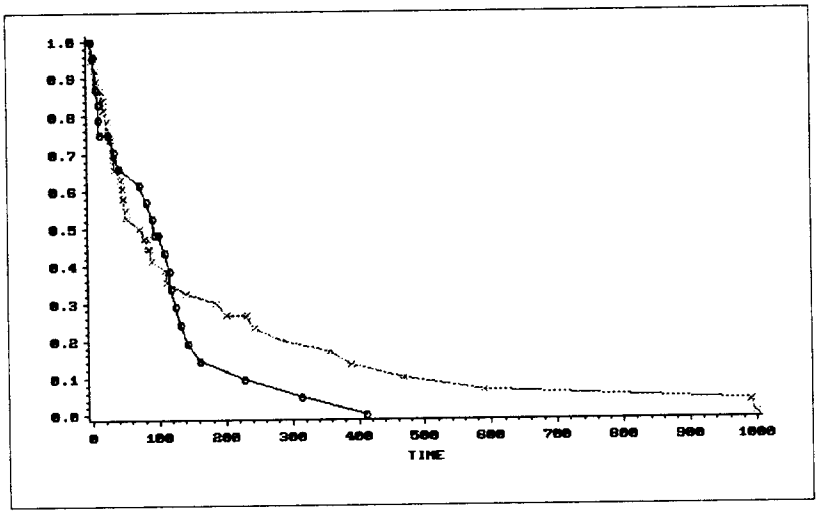
been known that the log rank test is a fully efficient rank test under Lehmann alternatives, and Peto-Prentice test is for logistic alternatives. On the other hand, Gehan-Wilcoxon is powerful while the log rank test is very insensitive to detect early differences. It becomes clear by noting that Gehan-Wilcoxon test puts more weight on early observations while the log rank test puts equal weight on each observation. In this point of view, Tarone and Ware(1977) proposed a test with a middle weight of these two tests, expecting reasonable power over the general alternatives even though it is not the best at certain alternatives. However, these tests including Peto-Prentice, which puts more weight on early observations, are still insensitive to detect late differences. The weights and its properties for detecting differences are summarized as follows;

Test statistics	weight	properties for detecting differences
Log rank test	1	more sensitive for late differences
Gehan-Wilcoxon	N_i	more sensitive for early differences
Tarone-Ware	$\sqrt{N_i}$	middle of log rank and Gehan-Wilcoxon
Peto-Prentice	$S(t_i)$	similar to Gehan-Wilcoxon

In the view of weights, it is natural that we give more weight on late observations to detect late differences. We now propose two weights, which are larger for late observations, $N-N_i$ and $\sqrt{N-N_i}$, expecting that these weights are more sensitive to detect late differences. The next example shows that the proposed tests are helpful in some situations.

3. Illustrated example

The 62 lung cancer patients data was given by Prentice(1973); 24 patients received standard treatment and 38 patients did test treatment. With this data we obtained the survival function estimates of two treatments, using the SAS lifetest procedure.



x : product limit estimates for standard treatment
o : product limit estimates for test treatment'

Figure 1. Estimated survival functions

From Figure 1 one can easily expect there are some differences in late time. The following table shows the results of testing (1);

Test statistics	values	P-values
Log rank test	-0.87	0.1922
Gehan-Wilcoxon	-0.08	0.4681
Tarone-Ware	-0.38	0.3520
Peto-Prentice	-0.14	0.4443
Proposed test with $N-N_i$	-1.63	0.0516
Proposed test with $\sqrt{N-N_i}$	-1.34	0.0901

The four tests presently in use fail to show significant differences between two treatments, even though two survival curves are showing some differences in late time.

However, the weighted log rank tests with weights $N-N_i$ or $\sqrt{N-N_i}$ is showing that differences under α=0.1.

This kind of situations occurs frequently in the medical experiments and the proposed tests are quite effective in these cases. We will compare the sizes and powers of the above mentioned six tests under various sample sizes and probability models through the simulation studies.

4. Empirical powers

We compare the empirical powers of the six tests; 1. Logrank test, 2. Gehan-Wilcoxon test, 3. Tarone-Ware test, 4. Peto-Prentice test, 5. proposed test with weights $N-N_i$, 6. proposed test with weights $\sqrt{N-N_i}$.

We use three types of population distributions under exponential and uniform censoring(10% or 20%) under various parameter configurations;

1) exponential distributions with the scale parameters θ_1 = 0.01 v.s. θ_2 = 0.01, 0.008, 0.007, 0.006,
2) Weibull distributions with the scale parameters θ_1 = 0.4 v.s. θ_2 = 0.4, 0.35, 0.3 and common shape parameter of value 2,
3) mixed exponential distributions, one of which has a scale parameter θ_1 = 0.01 and the scale parameter of the other one is changed from θ_2 to θ_2' at $F_2(t)$ = 0.4. In this case we use θ_2 = 0.01 and θ_2 = 0.009, 0.008, 0.007, 0.006. The mixed exponential distribution was chosen to show larger late differences(see, for details, Flemming et al.(1980)).

For sample sizes, we fix the sample size (30, 30) to see the effects of power behavior under various population models, and then we change the size to (35, 25), (40, 20) and (20, 40) to see the effect of different allocation of sample size. Simulation results are based on 1000 repetitions, by Fortan program.

Table I. Monte Carlo Power Estimates, n1 = n2 = 30, Exponential Populations, Exponential Censored (10%).

Parameters		Test Statistics					
θ_1	θ_2	1	2	3	4	5	6
.01	.01	.052	.05	.052	.047	.061	.056
.01	.008	.207	.174	.191	.175	.184	.199
.01	.007	.352	.302	.332	.313	.316	.332
.01	.006	.553	.476	.552	.483	.506	.546

Table II. Monte Carlo Power Estimates, n1 = n2 = 30, Exponential Populations, Uniform Censored (10%).

Parameters		Test Statistics					
θ_1	θ_2	1	2	3	4	5	6
.01	.01	.045	.049	.051	.049	.052	.051
.01	.008	.199	.17	.191	.174	.188	.201
.01	.007	.347	.303	.329	.307	.309	.338
.01	.006	.555	.473	.517	.483	.496	.527

Table III. Monte Carlo Power Estimates, n1 = n2 = 30, Weibull(θ,2) Populations, Uniform Censored (20%).

Parameters		Test Statistics					
θ_1	θ_2	1	2	3	4	5	6
.4	.4	.051	.046	.05	.048	.057	.053
.4	.35	.242	.207	.229	.209	.218	.241
.4	.3	.627	.535	.591	.551	.575	.623

Table IV. Monte Carlo Power Estimates, n1 = n2 = 30, Mixed Exponential Populations(change θ_2 to θ_2' at F(t)=0.4), Uniform Censored (10%).

Parameters			Test Statistics					
θ_1	θ_2	θ_2'	1	2	3	4	5	6
.01	.01	.01	.049	.050	.048	.048	.057	.050
.01	.01	.009	.105	.082	.089	.083	.119	.110
.01	.01	.008	.190	.127	.152	.129	.215	.206
.01	.01	.007	.292	.183	.237	.192	.336	.335
.01	.01	.006	.439	.270	.352	.289	.528	.547

Table V. Monte Carlo Power Estimates, n1 = 35 and n2 = 25, Mixed Exponential Populations(change θ_2 to θ_2' at F(t)=0.4), Uniform Censored (10%).

Parameters			Test Statistics					
θ_1	θ_2	θ_2'	1	2	3	4	5	6
.01	.01	.01	.033	.042	.037	.040	.037	.033
.01	.01	.009	.076	.07	.072	.071	.087	.082
.01	.01	.008	.133	.103	.121	.105	.163	.161
.01	.01	.007	.272	.162	.210	.17	.321	.328
.01	.01	.006	.436	.243	.323	.256	.533	.527

Table VI. Monte Carlo Power Estimates, n1 = 40 and n2 = 20, Mixed Exponential Populations(change θ_2 to θ_2' at F(t)=0.4), Uniform Censored (10%).

Parameters			Test Statistics					
θ_1	θ_2	θ_2'	1	2	3	4	5	6
.01	.01	.01	.043	.033	.036	.033	.046	.043
.01	.01	.009	.065	.058	.063	.060	.091	.084
.01	.01	.008	.141	.102	.123	.109	.169	.165
.01	.01	.007	.248	.157	.206	.166	.287	.292
.01	.01	.006	.411	.231	.318	.243	.491	.495

Table VII. Monte Carlo Power Estimates, n1 = 20 and n2 = 40, Mixed Exponential Populations(change θ_2 to θ_2' at F(t)=0.4), Uniform Censored (10%).

Parameters			Test Statistics					
θ_1	θ_2	θ_2'	1	2	3	4	5	6
.01	.01	.01	.058	.058	.061	.059	.072	.061
.01	.01	.009	.094	.081	.086	.083	.133	.115
.01	.01	.008	.161	.117	.139	.124	.223	.198
.01	.01	.007	.278	.162	.213	.172	.360	.342
.01	.01	.006	.431	.241	.321	.259	.545	.546

The following results are observed;

(1) As expected, log rank test is efficient under exponential models. But it should be noted that Tarone-Ware test and the test with weights $\sqrt{N-N_i}$ are also very powerful regardless of censoring patterns and parameter configurations.

(2) Under Weibull model, log rank test and the test with weights $\sqrt{N-N_i}$ are powerful.

(3) Under mixed exponential model, the tests with weights $N-N_i$ and $\sqrt{N-N_i}$ are more powerful than the rest.

(4) The size of each test discussed here is below the nominal level when the first sample size is bigger than the second sample size, while it is above the nominal

level in the opposite sample size case. The balanced sample sizes are recommended.

5. Concluding remarks

Using weighted log rank test, the choice of weights seems very important to detect the real differences. The proposed tests are useful to detect late differences, while Gehan-Wilcoxon, Tarone-Ware and Peto-Prentice test are useful for early differences. Log rank test can be viewed as a test with middle weights between two kinds of differences. From this point of view we recommend to calculate several test statisitcs when we one would two survival curves.

References

[1] Cox, D.R. (1972). Regression Models and life tables, *Journal of Royal Statistical Society* Ser. B, 34, 187-220.

[2] Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly cen -sored samples, *Biometrika,* Vol. 52, 203-223.

[3] Fleming, T.R. and Harrington, D.P. (1991). *Counting processes and survival analysis,* John Wiley & Sons, New York.

[4] Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. and Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right censored data, *Biometrics,* Vol. 36, 607-626.

[5] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotheraphy Reports* 50, 163-170.

[6] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of National Cancer Institute,* Vol. 22, 719-748.

[7] Miller, R.G. (1981). *Survival analysis,* John Wiley & Sons, New York.

[8] Peto, R. and Peto, J. (1972). Asymptotic efficient rank invariant test procedures(with discussion), *Journal of Royal Staitstical Society, Ser.* A, Vol. 135, 185-207.

[9] Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables, *Biometrika,* Vol. 60, 279-288.

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

[11] Tarone, R.E. and Ware, J. (1977). On the distribution free tests for equality of survival distributions, *Biometrika,* Vol. 64, 156-160.

후기 차이 검출을 위한 가중 로그 순위 검정[3)]

정규진, 박상규[4)]

요약

가중 로그 순위 검정은 생존분석에서 두 집단의 차이를 검정하는데 가장 많이 사용되는 검정이다. 생존실험에서 실험 초기에는 별다른 차이를 보이지 않다가 실험 중반을 넘어가며 집단간의 차이를 보이는 경우가 많은데 이러한 경우에 기존에 제시되어 있는 여러 가중 로그 순위 검정은 비효율적이다. 이 문제에 적당한 새로운 가중 로그 순위 검정을 제시하며 예를 통하여 제시된 검정의 유용성을 살펴보고 모의실험을 통해 이들의 검정력도 조사한다. 제시된 검정은 후기 차이 검출에 매우 효율적이다.

3) 이 연구는 1994년도 한남대학교 교비연구비로 수행되었음.

4) (300-791) 대전직할시 대덕구 오정동 133, 한남대학교 응용통계학과.