Hacettepe Journal of Mathematics and Statistics
Volume 34 (2005), 67 – 74

GROUP SEQUENTIAL TEST OF NON-PARAMETRIC STATISTICS FOR SURVIVAL DATA

Yaprak Parlak Demirhan* and Sevil Bacanlı*

Received 21 : 02 : 2005 : Accepted 29 : 11 : 2005

Abstract

In this study a group sequential test of non-parametric statistics is examined in order to compare two groups of survival data. A new general form for a group sequential test of non-parametric statistics is given. The distribution of test statistics, obtained at the end of each stage, have been derived for this general form. In addition, an example based on a simulated data set is used to illustrate the test process that covers the group sequential test of non-parametric statistics in the given general form.

Keywords: Group sequential test, Survival data, Log-rank test, Wilcoxon test, Tarone-Ware family.

1. Introduction

In any experiment or survey, data is accumulated over a period of time especially in industrial acceptance sampling and clinical trials. Because of ethical, administrative and economic reasons, interim analyses of accumulated data are conducted. Sequential tests were proposed because a fixed sample test is not useful for such data. However, sometimes, continuous data monitoring can be a serious practical burden. So it is most convenient to analyze the data in groups, thus accumulating data is analyzed at intervals rather than at every new observation. This is the major difference between Sequential Tests and Group Sequential Tests (GSTs). GSTs are convenient to conduct, they support the early stopping a trial, and can achieve most of the benefits of sequential tests (namely, lower expected sample sizes and shorter average study lengths).

For instance, in a phase III clinical trial, if an early stopping occurs, the hypothesis that there is no difference between the effects of treatments on survival times of individuals is rejected and subsequent patients are assigned to the superior treatment. In a phase III clinical trial, the primary interest is to investigate the effects of alternative treatments

*Department of Statistics, Hacettepe University, Beytepe, 06532, Ankara, Turkey.

on survival rates. If the distribution of the survival data in a study is not known exactly, non-parametric statistics can be used for those comparisons.

The principle of GSTs is based on the tests given by Pocock [9] and O'Brien&Fleming [6]. These tests are used for equal group sizes, whereas the tests based on the alpha spending approach, given by Lan&DeMets [5] and Kim&DeMets [4], and on the beta spending approach given by Pampallona&Tsiatis [7], can be used for either equal or unequal group sizes.

Tsiatis [12] and DeMets&Gail [2] described the GST of the log-rank statistic and Slud &Wei [10] presented the Wilcoxon test. These papers formed the basis of the usage of GST methods for survival analysis.

In the literature, the expressions for the calculation of the GST of non-parametric statistics for survival data lack clarity, and the case of tied observations (which occur when two individuals have the same survival time) is not considered. The aim of this study is to give the GSTs of non-parametric statistics for survival data in a general form, including ties and censored observations. Furthermore the distribution of the general form of GSTs for non-parametric statistics is derived, and a simulation study carried out to clarify the application of the GSTs of these non-parametric statistics. Test statistics were calculated after a fixed number of failure times.

In the second section, the non-parametric tests, frequently used for the analysis of survival data have been presented, and the general form of non-parametric statistics, which can be used for the comparison of survival data under the proportional hazards assumption, is given.

In the third section, it is shown that the test statistics in the given general form obtained at the end of each analysis, follow a normal distribution. Thus, it is appropriate to make GSTs of non-parametric statistics in this general form. Hence, the hypothesis tests can be conducted by using group sequential boundaries.

Finally, the fourth section covers a simulated example and its results for a fixed event approach.

2. Non-parametric Test Statistics

Non-parametric statistics are frequently used for the comparison of survival distributions which come from different courses of treatment. Also these methods need less assumptions compared to other methods.

In clinical trials, two groups of survival data, including censored observations, can be compared with respect to their hazard function or survival function after monitoring a certain number of individuals in sequence. Non-parametric tests, developed for these comparisons, are the log-rank test, the Wilcoxon test, and tests of the Tarone-Ware and $G^{\rho,\gamma}$ families.

In these tests, two groups of survival times (death times) are considered together and ordered as $t_1 < t_2 < \cdots < t_\delta$. When tied observations occur, only δ' of the δ failure times are different. Let τ_{Aj} be the number of deaths in treatment A, η_{Aj} the number of individuals at risk in treatment A, τ_j the total number of deaths, η_j the total number of individuals at risk and ω_j the weight function at the time (t_j) of the jth death, $j = 1, \ldots, \delta$. Here 2×2 cross tables, structured as in Table 1. are constructed over all death times, which follow a hypergeometric distribution when τ_{Aj} or τ_{Bj} is given. Then the information obtained from these tables are combined to test whether the two survival data have the same distribution.

Table 1. Number of deaths at the time t_j of the j th death in each of the two groups of individuals.

Treatment	Number of deaths at t_j	Number of survivors beyond t_j	Number at risk just before t_j
A	τ_{Aj}	$\eta_{Aj} - \tau_{Aj}$	η_{Aj}
B	τ_{Bj}	$\eta_{Bj} - \tau_{Bj}$	η_{Bj}
Total	τ_j	$\eta_j - \tau_j$	η_{Bj}

The score statistic U can be computed over any treatment arm. In the presence of tied observations we have:

$$U = \sum_{j=1}^{\delta'} \omega_j (\tau_{Aj} - \varepsilon_{Aj}), \tag{1}$$

where $\varepsilon_{Aj} = \mathrm{E}(\tau_{Aj})$, and the variance of the score statistic (1) is:

$$V(U) = \sum_{j=1}^{\delta'} \omega_j^2 \vartheta_{Aj},$$

where $\vartheta_{Aj} = V(\tau_{Aj})$. Then the test statistic is as follows:

$$S = \frac{U^2}{V} \sim \chi_1^2.$$

Here $U/\sqrt{V} \sim N(0,1)$.

Proportional hazards can be evaluated by using curves obtained with the Kaplan-Meier estimators. Let $\ln \lambda = \theta$ under the assumption of proportional hazards, that is, $h_A(t) = \lambda h_B(t)$. The statistics are computed by using the relevant ω_j values that are given in Table 2.

Table 2. Non-parametric tests and their weights.

Tests	ω_j
Log-rank	1
P-P Wilcoxon	$\widehat{S}(t_j)$
Gehan Wilcoxon	n_j
Tarone-Ware family	$(n_j)^\rho$
$G^{\rho,\gamma}$ family	$[\widehat{S}(t_{j-1})]^\rho [1 - \widehat{S}(t_{j-1})]^\gamma$

As long as the curves satisfy the proportional hazards assumption, the log-rank test must be used. Otherwise, the Tarone-Ware test is less powerful than the Wilcoxon test, Tarone&Ware [11]. When the ratio of the hazard functions decrease, the tests based on the $G^{\rho,\gamma}$ family can be used.

3. A General Form for GSTs of non-parametric statistics

One may want to make a GST of $\mathrm{H}_0 : \theta = 0$ or $(\lambda = 1)$ against $\mathrm{H}_0 : \theta \neq 0$ or $(\lambda \neq 1)$. When the i th analysis is conducted, let the number of failure times be δ_i, where in the presence of tied observations, only δ_i' of these times are different, let $\tau_{Bj,i}$ denote the number of deaths in treatment B, $\eta_{Bj,i}$ the number of individuals at risk in treatment B;

$\tau_{j,i}$ the total number of deaths, $\eta_{j,i}$ the total number of individuals at risk and $\omega_{j,i}$ the weight functions for $i = 1, \ldots, N$. Then

$$\varepsilon_{Bj,i} = \frac{\eta_{Bj,i}\tau_{j,i}}{\eta_{Aj,i} + \eta_{Bj,i}}$$

and

$$\vartheta_{Bj,i} = \frac{\eta_{Aj,i}\eta_{Bj,i}\tau_{j,i}\big(\eta_{Aj,i} + \eta_{Bj,i} - \tau_{j,i}\big)}{\big(\eta_{Aj,i} + \eta_{Bj,i} - 1\big)\big(\eta_{Aj,i} + \eta_{Bj,i}\big)^2}.$$

The score statistic is,

$$U_{\delta'_i} = \sum_{j=1}^{\delta'_i} \omega_{j,i}\big(\tau_{Bj,i} - \varepsilon_{Bj,i}\big).$$

Variance of $U_{\delta'_i}$ is,

$$V_{\delta'_i} = \sum_{j=1}^{\delta'_i} \omega_{j,i}^2 \vartheta_{Bj,i}.$$

Let $\eta_{Aj,i} \cong \eta_{Bj,i}$ and $\theta \cong 0$ ($\lambda \cong 1$), then from the variance of the hypergeometric distribution, $\vartheta_{Bj,i} \cong \frac{\tau_{j,i}}{4}$, and the variance is obtained as,

$$V_{\delta'_i} = \frac{1}{4}\sum_{j=1}^{\delta'_i} \omega_{j,i}^2 \tau_{j,i}.$$

The differences $U_{\delta'_1}, U_{\delta'_2} - U_{\delta'_1}, \ldots, U_{\delta'_N} - U_{\delta'_{N-1}}$ are independent for the score statistics series $\{U_{\delta'_1}, \ldots, U_{\delta'_N}\}$, and the series forms a Markov chain with continuous state space and time. So it can be expressed with a Brownian-Motion process (Jennison [3]). Also $U_{\delta'_i} \sim N\big(\theta V_{\delta'_i}, V_{\delta'_i}\big)$, (Whitehead [13]).

By the property of the Brownian-Motion process with drift θ,

$$\begin{aligned}
X_1 &= U_{\delta'_1} - 0 \sim N\big(\theta t_1, t_1\big) \\
X_2 &= U_{\delta'_2} - U_{\delta'_1} \sim N\big(\theta(t_2 - t_1), (t_2 - t_1)\big) \\
&\vdots \\
X_N &= U_{\delta'_{N-1}} - U_{\delta'_N} \sim N\big(\theta(t_N - t_{N-1}), (t_N - t_{N-1})\big).
\end{aligned}$$

Now $U_{\delta'_i} \sim N\big(\theta t_i, t_i\big)$, where $U_{\delta'_i} = X_i + X_{i-1} + \cdots + X_1$. When this result is compared with the result given by Whitehead [13], it is evident that the series $\{U_{\delta'_1}, \ldots, U_{\delta'_N}\}$ can be viewed as a Brownian-Motion process with drift parameter θ observed at times $\{V_{\delta'_1}, \ldots, V_{\delta'_N}\}$. Therefore,

$$U_{\delta'_i} \sim N\left(\theta\frac{1}{4}\sum_{j=1}^{\delta'_i} \omega_{j,i}^2 \tau_{j,i}, \frac{1}{4}\sum_{j=1}^{\delta'_i} \omega_{j,i}^2 \tau_{j,i}\right).$$

Also it is evident that,

$$\mathrm{Cov}\big(U_{\delta'_i}, U_{\delta'_{i-1}}\big) = \frac{1}{4}\sum_{j=1}^{\delta'_{i-1}} \omega_{j,i}^2 \tau_{j,i}.$$

In this case, under the null hypothesis,

$$\frac{U_{\delta_i'}}{\sqrt{V_{\delta_i'}}} \sim N(0,1).$$

Let $X_i = U_{\delta_i'} - U_{\delta_{i-1}'}$ be a new random variable. Then

$$\begin{aligned} X_i^* &= \frac{X_i}{\sqrt{V_{\delta_i'} - V_{\delta_{i-1}'}}} \\ &= 2\Bigg(\sum_{j=1}^{\delta_i'} \omega_{j,i}^2 \tau_{j,i} - \sum_{j=1}^{\delta_{i-1}'} \omega_{j,i}^2 \tau_{j,i}\Bigg)^{-1/2} \big(U_{\delta_i'} - U_{\delta_{i-1}'}\big). \end{aligned} \tag{2}$$

Consequently, the distribution of X_i^*, given by (2), is as follows:

$$X_i^* \sim N\Bigg[\frac{\theta}{2}\Bigg(\sum_{j=1}^{\delta_i'} \omega_{j,i}^2 \tau_{j,i} - \sum_{j=1}^{\delta_{i-1}'} \omega_{j,i}^2 \tau_{j,i}\Bigg)^{1/2}, 1\Bigg].$$

The test statistic is defined as follows:

$$S_i = \sum_{\ell=1}^{i} X_\ell^*. \tag{3}$$

As given in Parlak [8], the statistic in (3) follows a normal distribution. That is:

$$S_i \sim N\Bigg[\frac{\theta}{2} \sum_{\ell=1}^{i} \Bigg(\sum_{j=1}^{\delta_\ell'} \omega_{j,\ell}^2 \tau_{j,\ell} - \sum_{j=1}^{\delta_{\ell-1}'} \omega_{j,\ell}^2 \tau_{j,\ell}\Bigg)^{1/2},\ i\Bigg].$$

Because of this distributional property, the given general form for non-parametric statistics is appropriate to GSTs. Here the test statistic we are interested in is calculated using the weights given in Table 2. When the number of failure times, δ_i', are equal for each stage the following condition holds:

$$\sum_{j=1}^{\delta_i'} \tau_{j,i} - \sum_{j=1}^{\delta_{i-1}'} \tau_{j,i} = \delta_0'$$

The expected value of X_i^* in (2) is denoted by Δ. It is evident that $S_i \sim N(i\Delta, i)$. In this case the series $\{S_1, \ldots, S_N\}$ of test statistics can be used to test the null hypothesis using Pocock's or O'Brien&Fleming's critical values. By comparing the statistics S_i defined in (3) with Pocock's or O'Brien&Fleming's critical value (C_i), we can make a decision about the hypothesis as below:

1. We stop to reject H_0 if $\mid S_i \mid \geq C_i$, $i = 1, \ldots, N-1$. Otherwise, we continue to the next stage.
2. We stop to reject H_0 if $\mid S_N \mid \geq C_N$. Otherwise, we stop to accept H_0 (Jennison [3]).

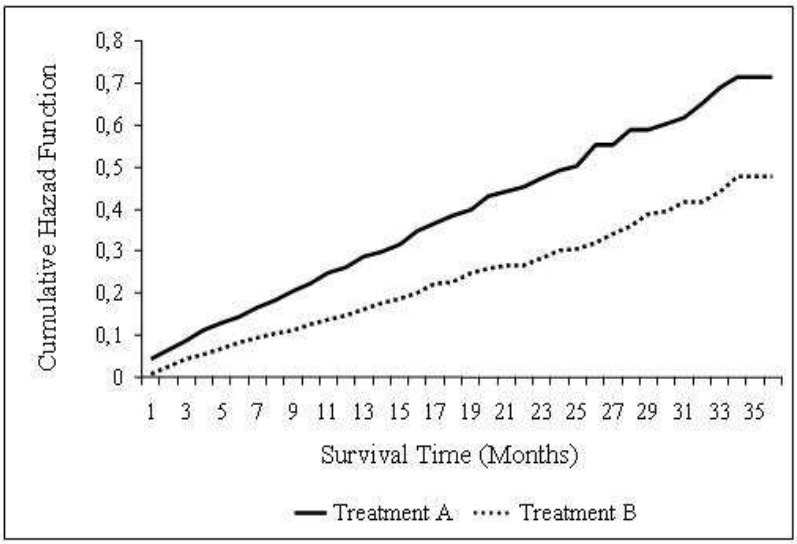
4. A Simulated Example

In this section, we present an application of GSTs for survival data. Survival analysis requires data from observations that is collected over a period of time. So, using a simulated data set was the most convenient way of obtaining data for this study. A Minitab macro was prepared to generate the data. The monitoring time was taken to be a three year period, namely 1095 study days. All parameters were determined as exceptional simulation conditions. Entrance of patients into the study was assumed

to follow a Poisson distribution with a mean of 0.95 in order to allow the entrance of approximately one patient for each study day. Because patients are assigned to the treatment arms randomly, a Bernoulli distribution with a probability of success of 0.5 was chosen; survival times from the beginning of the treatment to the censoring or to the failure of individuals on treatments A and B, were taken to follow an exponential distribution with means 2114 and 1332, respectively, in order to find evidence against the null hypothesis stating the equivalence of the two survival distributions, and also to permit right censored observations. Occurrences at the end (right-censoring/failure) of the period follow a Bernoulli distribution with a failure probability of 0.002. So the failure number will be small and at the end of the 1095 th study day there will be individuals alive who are at risk, permitting the test to conclude properly. Under these conditions, 1000 iterations were made and 1551 individuals generated. The entrance time of the first observation was taken as the beginning of the study.

The aim was to make a GST of the hypothesis that the effects of the treatments on survival times are equivalent, and to determine the superior treatment with a minimum number of failures. From Figure 1, which was obtained from retrospective simulated data, it is seen that the hazard functions for the two treatment arms do not cross. Because the proportional hazards assumption is satisfied, non-parametric statistics can be used, particularly the log-rank test.

Figure 1. Cumulative hazard versus time



It is assumed that,

$$H_0 : \theta = 0 \Longleftrightarrow (ln\lambda = 0) \Longleftrightarrow (\lambda = 1)$$
$$H_1 : \theta = \mp 0.470 \Longleftrightarrow (ln\lambda = \mp 0.470) \Longleftrightarrow (\lambda = 1.6 or \lambda = 0.625).$$

While testing those hypothesis, one may want to determine the design for a particular type one error probability, power and stage number from the beginning. Then, the number of failures for each stage can be determined with Pocock's or O'Brien &Fleming's designs. On the other hand, sometimes the analysis may be conducted after a predetermined random number of failures. In this situation each Pocock's or O'Brien &Fleming's critical value is used, but the power of the test will be different for each test.

Let $\alpha = 0.05$, $(1-\beta) = 0.90$ and $N = 5$. For the log-rank test, in O'Brien &Fleming's and Pocock's test designs, for $i = 1, \ldots, 5$, $\delta_i = 40$ and $\delta_i = 46$ failures, respectively, must have occurred at each stage to conduct the test. Otherwise, if it is decided to use a fixed sample size corresponding to $N = 1$, 190 failures would be expected to conduct the test. However, as O'Brien and Fleming's test design was chosen, the analysis has been conducted after every 40 failures. This is the major advantage of group sequential designs.

In Table 3. L-R, P-PW, W and T-W are used instead of log-rank, Peto-Peto Wilcoxon, Gehan Wilcoxon and Tarone-Ware statistics, and also O-F and P represent O'Brien& Fleming's and Pocock's GST boundaries, respectively.

Table 3. Results of GSTs of non-parametric statistics.

	Total Number of Failures		Test Statistics				Boundaries(C_i)	
Stage	A	B	L-R	P-PW	W	T-W	O-F	P
1	17	23						
279 th day	***40***		1.635	1.635	1.695	1.665	4.555	2.413
2	32	48						
408 th day	***80***		2.905	2.905	2.932	2.919	3.221	*2.413*
3	54	66						
526 th day	***120***		2.164	2.163	2.222	2.194	2.630	2.413
4	66	94						
608 th day	***160***		3.183	3.182	3.239	3.212	*2.277*	*2.413*
5	81	119						
709 th day	***200***		3.736	3.736	3.781	3.759	*2.037*	*2.413*

According to the results given in Table 3, when O'Brien&Fleming's GST boundaries are used, there is enough evidence to reject the null hypothesis at the 4th analysis immediately after 160 deaths for the two treatment arms on the 608th day of study. At that point there where 66 deaths on Treatment A and 94 deaths on Treatment B, with a hazard ratio of $\lambda = 0.70$ at a power of 90%. However, when Pocock's GST boundaries were used the null hypothesis is rejected at the 2nd analysis immediately after 80 deaths for the two treatment arms on the 408th day of study, at which point there were 32 deaths on Treatment A and 48 deaths on Treatment B with a hazard ratio of $\lambda = 0.67$ at a power of approximately 99%.

5. Conclusion

The log-rank test is commonly used for comparing survival distributions. In this study a general expression, consisting of a single formula describing GSTs of non-parametric statistics, is given. Hence, the distribution of the general representation is derived and the suitability of making a GST of non-parametric statistics in the given general form in the presence of ties and censored observations is discussed.

We found that if we use a fixed sample design for our simulated example we would have to wait for 190 failures before deciding about the hypothesis, where the 190th failure occurred on the 682th day of the study. But when we used a GST design we had to wait for only 40 failures, and so we conducted the test after every 40 failures. Here we could decide about the hypothesis at the 2nd stage - the 408th day of the study - (according to

the Pocock's critical value) or at the 4th stage - the 608th day of the study - (according to the O'Brien&Fleming's design). So the result of the GSTs is to enable the early stopping of the trial with a minimum number of failures supported with this example.

Especially in clinical trials, GSTs ensure a minimum number of deaths before a decision can be taken. Because of these advantages it is appropriate for researchers to use GSTs.

References

[1] Collet, D. *Modeling Survival Data In Medical Research*, (Chapman&Hall, 1994).
[2] DeMets, D. L. and Gail, M. H. *Use of logrank tests and group sequential methods at fixed calendar times*, Biometrics **41**, 1039–1044, 1985.
[3] Jennison, C. and Turnbull, B. W. *Group Sequential Methods with Applications to Clinical Trials*, (Chapman&Hall, 1999).
[4] Kim, K. and Demets, D. L. *Design and analysis of group sequential tests based on the type 1 error spending rate function*, Biometrika **79**, 149–154, 1987.
[5] Lan, K. K. G. and Demets, D. L. *Discrete sequential boundaries for clinical trials*, Biometrika **70** (3), 659–663, 1983.
[6] O'Brien, P. C., and Fleming, T. R. *A multiple testing procedure for clinical trials*, Biometrics **35**, 549–556, 1979.
[7] Pampallona, S. and Tsiatis, A. A. *Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis*, J. Statist. Planning and Inference **42**, 19–35, 1994.
[8] Parlak,Y. *Comparison of the Group Sequential Test Methods in Survival Analysis*, (Unpublished M.Sc. Thesis, Hacettepe University, Institute of Natural Sciences, Ankara, 2004).
[9] Pocock, S. J. *Group sequential methods in the design and analysis of clinical trials*, Biometrika **64** (2), 191–199, 1977.
[10] Slud, E. V. and Wei, L. J. *Two-sample repeated significance tests based on the modified Wilcoxon statistics*, JASA **77**, 862–868, 1982.
[11] Tarone, R. E. and Ware, J. *On distribution free tests of the equality of survival distributions*, Biometrika **64**, 156–160, 1977.
[12] Tsiatis, A. A. *Repeated significance testing for a general class at statistics used in censored survival analysis*, JASA **77**, 855–862, 1982.
[13] Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, (John Wiley, 1983).