The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 137-145

Random Permutation Test for Comparison of Two Survival Curves

Mi Kyung Kim[1], Jae Won Lee[2] and Myung Hoe Huh[3]

Abstract

There are many situations in which the well-known tests such as log-rank test and Gehan-Wilcoxon test fail to detect the survival differences. Assuming large samples, these tests are developed asymptotically normal properties. Thus, they shall be called asymptotic tests in this paper. Several asymptotic tests sensitive to some specific types of survival differences have been recently proposed. This paper compares by simulations the test levels and the powers of the conventional asymptotic tests and their random permutation versions. Simulation studies show that the random permutation tests possess competitive powers compared to the corresponding asymptotic tests, keeping exact test levels even in the small sample case. It also provides the guidelines for choosing the valid and most powerful test under the given situation.

Keywords : Survival distributions, Small sample, Versatile test, Permutation test, Monte Carlo simulation.

1. Introduction

For the comparison of two survival distributions, many test statistics are available for use such as the log-rank statistics (Mantel, 1966), the generalized Wilcoxon statistic (Gehan, 1965; Peto and Peto, 1972) and Lee's (1996) versatile statistics, among others. These tests are developed assuming large samples, so that, for samples of small or moderate size, their statistical performances such as the test level and power are not well validated. In this study, through Monte Carlo evaluation for small or moderate size samples, we get to know that type

[1] d2K Solutions Co., Ltd. 35-4 Youido-dong, Youngdungpo-gu, Seoul, 150-010, Korea.
E-mail : mkkimfds@d2ksolutions.com

[2] Dept. of Statistics, Korea University, Anam-dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : jael@mail.korea.ac.kr

[3] Dept. of Statistics, Korea University, Anam-dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : stat420@mail.korea.ac.kr

I errors are not controlled for many survival tests for the equality of two survival curves originated from the asymptotics. In contrast, we see that their random permutation testing versions show competitively good performance in power evaluation, keeping exact test levels.

Consider a survival experiment in which $n_1 + n_2$ subjects are randomly divided into two groups of size n_1 and n_2, subsequently receiving one of two treatments, say T (=1) or C (=2). We denote the survival data by (X_{ij}, δ_{ij}) for $i = 1, 2$ and $j = 1, \cdots, n_i$, which are assumed to be generated by a random censoring mechanism of independent failure and censoring time pairs (T_{ij}, C_{ij}):

$$X_{ij} = \min(T_{ij}, C_{ij}), \quad \delta_{ij} = I\{T_{ij} \leq C_{ij}\}.$$

The null and alternative hypothesis of this study are

$$H_0 : S_1(t) = S_2(t), \text{ for } \textit{all } t \geq 0 \quad \text{vs.} \quad H_1 : S_1(t) \neq S_2(t), \text{ for } \textit{some } t \geq 0,$$

where $S_i(t) = P\{T_{ij} > t\}$ for $i = 1, 2$. Censoring times C_{ij} are assumed to be generated independently from a common probability distribution, not depending on the treatment (T or C).

Log-rank test (Mantel, 1966) and the Gehan-Wilcoxon test (Gehan, 1965) have been widely used in comparing two survival distributions. Log-rank test gives the highest power under the proportional hazards model, and the Gehan-Wilcoxon test performs better than log-rank test in detecting early differences in survival distributions. It is crucial to apply the most powerful test under the given situation, and thus several tests have been proposed for comparing two survival differences. Fleming, O'Fallon and O'Brien (1980) proposed the modified Kolmogorov-Smirnov test. Their method modifies the classical Kolmogorov-Smirnov test based on the maximum distance between two empirical distribution functions to allow for the censored data, and uses the survival function instead of the empirical function. Pepe and Fleming (1989) proposed the weighted Kaplan-Meier test statistics which is based on the differences between two survival functions. Kaplan-Meier estimate gets more unstable near the end of the study due to censoring, and thus the weighted Kaplan-Meier method improves the power by giving the smaller weight to the late survival differences. Lee (1996) proposed some versatile tests based on the four individual weighted log-rank statistics of $G^{\rho,\gamma}$ class in which the weight functions cover a broad range of types of survival differences. He proposed to use the maximum or the linear combination of these weighted log-rank statistics, and this test is especially useful when the investigators are not able to specify in advance the specific type of the survival differences that may exist.

In this article, we propose the random permutation survival tests for comparing survival differences. We briefly introduce random permutation versions of the survival tests by taking a numerical example in section 2. In section 3, we investigate the performances of the above asymptotic tests and their random permutation versions with small or moderate size samples such as $n_1 = n_2 = 10$ or 20.

2. Random Permutation Survival Tests

Before turning to Monte Carlo evaluation of several tests under various situations in the next section, we briefly introduce random permutation versions of the survival tests. Suppose that we have the following two-group survival data of which the first-group fifteen patients have low grade tumor while the second-group twenty patients have high grade tumor (Fleming, O'Fallon and O'Brien, 1980).

Group 1: 28, 89, 175, 195, 309, 377+, 393+, 421+, 447+, 462, 709+, 744+, 770+, 1106+, 1206+.
Group 2: 34, 88, 137, 199, 280, 291, 299+, 300+, 309, 351, 358, 369, 369, 370, 375, 382, 392, 429+, 451, 1119+.

Mantel's log-rank statistic T is computed as 5.664 in this case. From a large sample theory, it can be referred to the chi-squared distribution with single degree of freedom under the null hypothesis that two distributions are equal, giving the asymptotic p-value 0.018. One may concern whether $(n_1, n_2) = (15, 20)$ is the case to apply the large sample theory or not.

Randomization testing approach is different. Under the null hypothesis, the underlying stochastic mechanisms are equal. Thus, Group 1 observations can be regarded as 15 observations chosen at random out of 35 (=15+20) observations. So, the observed $T = 5.664$ is just one case out of equally likely $M = \binom{35}{15}$ = 3,247,943,160 cases. In randomization test theory, T can be referred to the distribution of M such test statistics values T_0^*, yielding the p-value = $P_M\{T_0^* \geq T\}$. Since M is often too large, it is a formidable task to compute T_0^* for all cases. If the computation of p-value is the main thing, it is sufficient to compute T_0^* for a number m of randomly chosen cases, say 1,000, yielding an approximate p-value = $P_m\{T_0^* \geq T\}$ which has the estimated error limit $\pm 0.014 (= 1.96*\sqrt{0.05*0.95/1000})$ at 95% confidence level for true p-value near 0.05. Random choice of m cases is achieved using a sub-list of full $(n_1 + n_2)!$ permutations, assuming the beginning part of n_1 observations to form Group 1 and the remaining part Group 2. We call such randomization test by random permutation test. Random permutation test, even though it needs intensive computing, does not require a sampling theory that is often mathematically intractable for the finite samples.

For the above data, 24 cases out of $m = 1,000$ cases yielded T_0^* greater than or equal to $T = 5.664$, giving an approximate p-value = $P_{1000}\{T_0^* \geq 5.664\}$ = 0.020. With $m = 10,000$, more reliable approximate p-value can be obtained. In one trial, we get an approximate p-value = $P_{10,000}\{T_0^* \geq 5.664\}$ = 0.019. Therefore, we may reject the null hypothesis at 5% significance level.

Such randomization test has its origin at R. A. Fisher (1935), who noted the randomization can be utilized at the testing stage if it was adopted at the design stage. Even though it has certain conceptual advantages compared to random sampling approach in biomedical research (Ludbrook and Dudley, 1998), randomization tests have not been adopted routinely because of the intensive computation required. But, recent development of computer technology revived randomization testing methodology and they appeared several monographs on the topic (Good, 1994; Edgington, 1995; Manly, 1997). See Edgington and Gore (1986) for the first application of randomization tests to the survival data.

3. A Monte Carlo Study For Test Level And Power Evalution

We will consider one null configuration and four alternative configurations for Monte Carlo generation of survival data as in Lee (1996). Figure 1 presents graphs of the survival configurations in the following alternative cases:

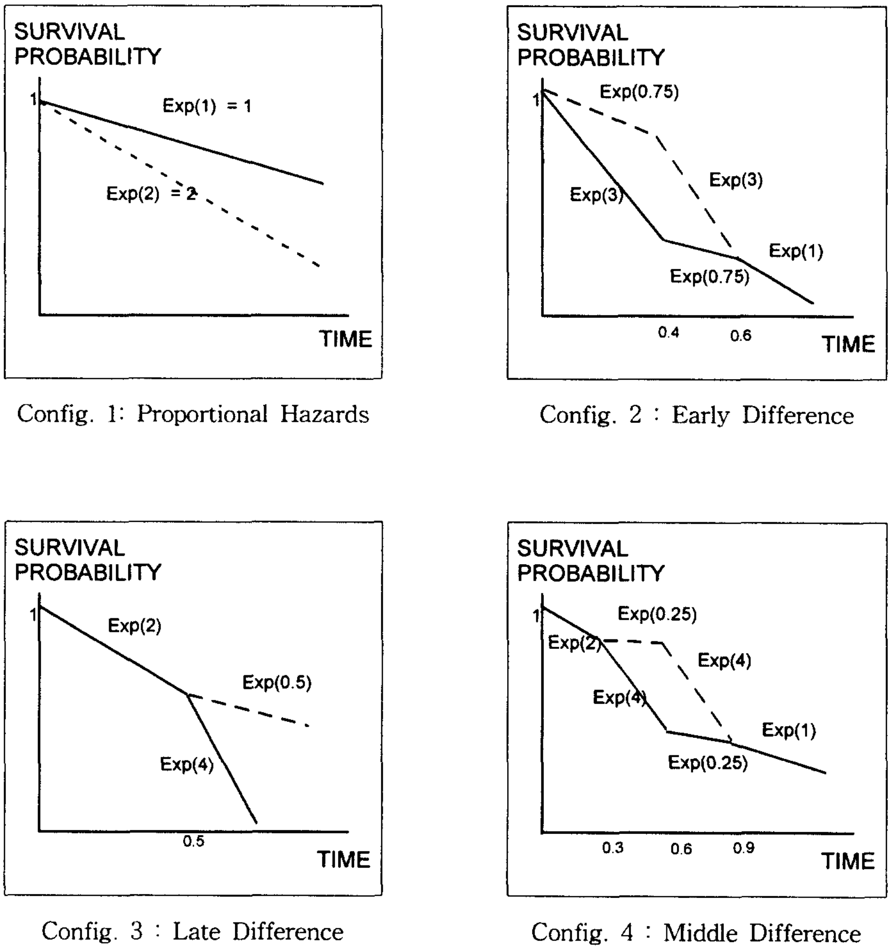
· Config. 0 (Null):	Exp(1) both for C and T.
· Config. 1 (Proportional Hazards):	Exp(1) for T, and Exp(2) for C.
· Config. 2 (Early Difference):	Exp(3/4) → Exp(3) → Exp(1) for T, and Exp(3) → Exp(3/4) → Exp(1) for C.
· Config. 3 (Late Difference):	Exp(2) → Exp(1/2) for T, and Exp(2) → Exp(4) for C.
· Config. 4 (Middle Difference):	Exp(2) → Exp(1/4) → Exp(4) → Exp(1) for T, and Exp(2) → Exp(4) → Exp(1/4) → Exp(1) for C.

Also, we consider both no censoring and censoring mechanism, where censoring times C are generated from Uniform(0,2) distribution. Note that, with censoring, no observations are greater than 2 and that both failure and censoring times have the same expected value under Config. 0. The test statistics used for Monte Carlo simulation are log-rank test, Gehan-Wilcoxon test, modified Kolmogorov-Smirnov test, weighted Kaplan-Meier test and Lee's (1996) test based on the maximum and the average (as a linear combination) of the weighted log-rank test statistics.

For sample sizes, we set $n_1 = n_2 = 10, 20, 50$ for typical values of small, moderate and large samples. 1,000 Monte Carlo samples are generated per situation, and 1,000 random permutations are selected for the randomization test per sample. Nominal significance level α is set to 5%. See Table 1 and Table 2 for Monte Carlo evaluation of test levels under Config. 0 and test powers under Config. 1, 2, 3 and 4.

We first focus on the comparison of the asymptotic tests and their random permutation versions. For Config. 0 (null case) with no censoring in Table 1, we may note that test levels are not kept well at the target value of 5% with small or moderate size samples for several asymptotic tests such as log-rank test, average test, maximum, and modified Kolmogorov-Smirnov test. Hence, direct power comparison between the asymptotic test and its random permutation version does not make sense in the case of small or moderate size samples for these four tests. For instance, under Config. 1, the asymptotic log-rank test with samples of size (20, 20) has larger power 0.613 than the random permutation log-rank test's 0.514, but under Config. 0 the former's level is 0.083 while the latter's level is 0.048. In contrast, all the random permutation tests automatically keep test levels at the target value.

Both the asymptotic Gehan-Wilcoxon test and the asymptotic weighted Kaplan-Meier test keep their test levels at the target value well even when $n = 10$ or 20, but give almost the same power under the alternative cases as their random permutation versions. For large samples of size (50, 50), power comparison between asymptotic test and random permutation



Config. 1: Proportional Hazards

Config. 2 : Early Difference

Config. 3 : Late Difference

Config. 4 : Middle Difference

Figure 1. Survival configuration for the alternative cases

Table 1. Monte Carlo evaluation of test levels and test powers when there is no censoring.(*)

Test statistic	Samples	Type(**)	Config. 0	Config. 1	Config. 2	Config. 3	Config. 4
Log-rank	10, 10	A	0.097	0.435	0.196	0.240	0.126
		R	0.053	0.285	0.120	0.124	0.075
	20, 20	A	0.083	0.613	0.220	0.405	0.101
		R	0.048	0.514	0.162	0.288	0.070
	50, 50	A	0.060	0.940	0.405	0.761	0.103
		R	0.047	0.921	0.362	0.723	0.097
Gehan-Wilcoxon	10, 10	A	0.055	0.302	0.178	0.095	0.090
		R	0.049	0.265	0.156	0.086	0.077
	20, 20	A	0.059	0.489	0.309	0.116	0.085
		R	0.057	0.467	0.298	0.111	0.080
	50, 50	A	0.044	0.856	0.689	0.206	0.122
		R	0.043	0.853	0.684	0.200	0.128
Average	10, 10	A	0.072	0.437	0.186	0.294	0.083
		R	0.054	0.422	0.163	0.261	0.105
	20, 20	A	0.046	0.479	0.210	0.488	0.100
		R	0.046	0.638	0.219	0.489	0.102
	50, 50	A	0.034	0.944	0.395	0.824	0.124
		R	0.046	0.962	0.469	0.867	0.159
Maximum	10, 10	A	0.093	0.484	0.194	0.418	0.152
		R	0.058	0.368	0.136	0.351	0.122
	20, 20	A	0.066	0.648	0.293	0.816	0.114
		R	0.047	0.579	0.210	0.758	0.083
	50, 50	A	0.046	0.936	0.628	0.997	0.198
		R	0.043	0.932	0.608	0.995	0.181
Modified K-S	10, 10	A	0.071	0.298	0.197	0.218	0.125
		R	0.053	0.246	0.148	0.188	0.098
	20, 20	A	0.057	0.419	0.309	0.372	0.152
		R	0.059	0.410	0.310	0.377	0.142
	50, 50	A	0.044	0.812	0.745	0.927	0.375
		R	0.045	0.821	0.758	0.929	0.398
Weighted K-M	10, 10	A	0.050	0.249	0.267	0.072	0.188
		R	0.046	0.234	0.263	0.070	0.182
	20, 20	A	0.052	0.427	0.429	0.103	0.315
		R	0.052	0.419	0.418	0.103	0.307
	50, 50	A	0.045	0.832	0.825	0.158	0.621
		R	0.044	0.827	0.825	0.157	0.614

(*) Based on 1,000 Monte Carlo replications. Error limits are ± 0.014 with 95% confidence under Config. 0.

(**) Test type 'A' is the asymptotic test. Test type 'R' is the random permutation test.

Table 2. Monte Carlo evaluation of test levels and test powers when observations are censored by Uniform(0,2) distribution.(*)

Test statistic	Samples	Type(**)	Config. 0	Config. 1	Config. 2	Config. 3	Config. 4
Log-rank	10, 10	A	0.0069	0.262	0.182	0.161	0.139
		R	0.043	0.194	0.122	0.108	0.084
	20, 20	A	0.052	0.462	0.237	0.304	0.134
		R	0.043	0.411	0.190	0.204	0.105
	50, 50	A	0.060	0.793	0.415	0.635	0.197
		R	0.055	0.781	0.375	0.625	0.114
Gehan-Wilcoxon	10, 10	A	0.046	0.184	0.181	0.071	0.093
		R	0.047	0.171	0.156	0.063	0.076
	20, 20	A	0.048	0.355	0.296	0.109	0.098
		R	0.049	0.348	0.293	0.064	0.094
	50, 50	A	0.057	0.696	0.632	0.159	0.128
		R	0.060	0.686	0.625	0.156	0.136
Average	10, 10	A	0.053	0.292	0.174	0.231	0.145
		R	0.058	0.299	0.165	0.228	0.140
	20, 20	A	0.035	0.480	0.214	0.416	0.151
		R	0.046	0.509	0.238	0.459	0.181
	50, 50	A	0.038	0.815	0.360	0.775	0.222
		R	0.047	0.862	0.307	0.666	0.311
Maximum	10, 10	A	0.030	0.240	0.160	0.333	0.161
		R	0.047	0.269	0.153	0.301	0.148
	20, 20	A	0.029	0.425	0.269	0.675	0.173
		R	0.040	0.446	0.282	0.667	0.198
	50, 50	A	0.037	0.778	0.568	0.974	0.259
		R	0.049	0.816	0.628	0.892	0.297
Modified K-S	10, 10	A	0.034	0.154	0.185	0.151	0.115
		R	0.035	0.161	0.153	0.138	0.096
	20, 20	A	0.048	0.359	0.333	0.369	0.173
		R	0.048	0.357	0.364	0.350	0.173
	50, 50	A	0.046	0.704	0.714	0.846	0.353
		R	0.050	0.702	0.711	0.849	0.358
Weighted K-M	10, 10	A	0.043	0.203	0.308	0.062	0.135
		R	0.040	0.194	0.290	0.057	0.124
	20, 20	A	0.051	0.359	0.552	0.064	0.206
		R	0.051	0.358	0.541	0.064	0.197
	50, 50	A	0.060	0.718	0.650	0.086	0.305
		R	0.059	0.717	0.659	0.083	0.300

(*) Based on 1,000 Monte Carlo replications. Error limits are ± 0.014 with 95% confidence under Config. 0.

(**) Test type 'A' is the asymptotic test. Test type 'R' is the random permutation test.

test is fair since their test levels are similar. By looking up Table 1 for Config. 1, 2, 3 and 4, we can find that the power loss of random permutation test compared to asymptotic test is negligible. For instance, under Config. 1, the asymptotic log-rank test with samples of size (50, 50) has the power 0.940, while the random permutation log-rank test's power is 0.921. In Table 2 for the censoring case, we observe that the test levels of the above asymptotic tests are kept better than no censoring cases in small or moderate size samples. We also observe there is almost no difference in power between the asymptotic test and its random permutation version. Power comparison between asymptotic test and random permutation test in the case of large samples of size (50, 50) leads to the similar conclusion to that for no censoring case.

Now we examine the behavior of the above test statistics for each configuration of survival distributions. Since each of the above asymptotic tests and its random permutation version have almost the same power under the given alternative configuration, we only compare the powers among the random permutation tests. For Config. 1 (proportional hazards case), the random permutation version of the log-rank test, the average test and the maximum test give higher power than other tests. Especially in small or moderate size samples, random permutation version of the average test and the maximum test gives higher power than that of log-rank test. For Config. 2 (early difference case) the random permutation version of the modified Kolmogorov-Smirnov test and the weighted Kaplan-Meier test give the highest power, and that of the Gehan-Wilcoxon test and the maximum test also give good power. In small or moderate samples, the random permutation version of the weighted Kaplan-Meier method gives higher power than that of the modified Kolmogorov-Smirnov test. For Config. 3 (late difference case) in large sample case ($n=50$), the random permutation version of the maximum test and the modified Kolmogorov-Smirnov test gives the highest power and that of the average test also give good power. In the case of small or moderate size samples ($n=10$ or 20), however, the power of the modified Kolmogorov-Smirnov test decreases rapidly compared to other tests and is lower than that of the average test. Finally, for Config. 4 (middle difference case) with no censoring in Table 1, the random permutation version of the weighted Kaplan-Meier test and the modified Kolmogorov-Smirnov test gives the highest and the second highest power, respectively. However, the censoring naturally lowers the power of these two tests significantly compared to the other tests, and there is almost no difference among the random permutation versions of the Gehan-Wilcoxon test, average test, maximum test and the weighted Kaplan-Meier test.

4. Concluding Remarks

As observed by Ludbrook and Dudley (1998), in most biomedical experiments, subjects are randomly assigned to treatments, rather than randomly sampled from populations. Hence, randomization tests or random permutation tests have a strong merit, compared to conventional asymptotic tests. As applied to two-group survival data in particular, random permutation tests possess competitive powers compared to asymptotic tests, keeping the significance level exact, with samples of all sizes.

Log-rank test and Gehan-Wilcoxon test have been widely used in comparing two survival distributions, but there are many situations in which these heavily used tests fail to detect survival differences successfully. Several asymptotic tests have been recently proposed for comparing two survival differences, and the properties of these recent asymptotic tests and their random permutation versions are investigated by simulations in this paper. Simulation studies provide some insights into these tests across several types of survival differences and degree of censorship.

References

[1] Edgington, E. S. (1980). *Randomization Tests*, Dekker, New York.

[2] Edgington, E. S. and Gore, A. P. (1986). Randomization Tests for Censored Survival distribution, *Biometrical Journal*, vol. 6, 673-681.

[3] Fisher, R. A. (1935). *Design of Experiments*, Edinburgh, Oliver and Boyd.

[4] Fleming, T. R., O'Fallon, J. R. and O'Brien, P. C. (1980). Modified Kolmogorov-Smirnov Test Procedures with Application to Arbitrarily Right-Censored Data, *Biometrics*, vol. 36, 607-625.

[5] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing Arbitrarily singly-censored samples, *Biometrika*, vol. 52, 203-223.

[6] Good, P. (1993). Permutation Tests, Springer-Verlag, New-York.

[7] Lee, J. W. (1996). Some Versatile Tests Based on the Simultaneous Use of Weighted Log-Rank Statistics, *Biometrics*, vol. 52, 468-472.

[8] Ludbrook, J. and Dudley, H. (1998). Why Permutation Tests are Superior to t and F Tests in Biomedical Research, *The American Statistician*, vol. 52, 127-132.

[9] Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd ed., Chapman & Hall, London.

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

[11] Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data, *Biometrics*, vol. 45, 497-507.

[12] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant procedures (with discussion), *Cancer Chemotherapy Report*, vol. 50, 163-170.