EXTENSION OF THE LOG RANK TEST[1]

Ji-Hyun Kim
Dept. of Statistics, Soong Sil University
Seoul 156-743, Korea.

Key words and phrases: conditional independence; censored data; hazard ratio; hypergeometric; exact test.

ABSTRACT

For the comparison of two groups of survival times subject to censoring the log rank test is widely used. The log rank test is known to be asymptotically efficient for the proportional hazards model. But if the ratio of hazards changes, the log rank test may not detect well the difference between two groups. In this article an alternative test procedure is proposed, and the performance of two test procedures is compared by simulation study.

1. INTRODUCTION

The log rank test is widely used for comparing two survival distributions. It is a natural extension of the Mantel-Haenszel test for the conditional independence of categorical variables. Before the alternative to the log rank test is discussed, let us begin with the Mantel-Haenszel test and its weakness.

In the categorical data analysis it is important to test for conditional independence. Many epidemiological studies investigate whether an association exists between a binary risk factor X and a binary response variable Y. They analyze whether an observed association between X and Y persists when the level of another factor Z that might influence the association is controlled. This involves testing conditional independence of X and Y controlling for Z. The Mantel-Haenszel (MH) test is widely used to check conditional independence for sparse tables. But if the association between X and Y varies along the levels of Z, MH test does not detect the association well. Let us point out the weakness that MH test has through a real data set, which brings the necessity of a new test procedure.

Table I, taken from Mantel (1963), summarizes the effectiveness of immediate injection vs. 1.5-hour-delayed injection of penicillin in protecting rabbits

[1] This work was supported by Grant 1997-001-D00082 from the Korea Research Foundation.

against lethal injection with β-hemolytic streptococci. To test whether the injection method is conditionally independent of the response at each penicillin level, MH test statistic is calculated as,

$$M^2 = \frac{[(3-1.5)+(6-2)+(5-(-0.5))]^2}{0.614+0.728+0.25} = 5.66.$$

with p-value 0.017. However, if the effectiveness of the injection method changes as the level of penicillin increases, in other words, if the immediate injection method is more effective at low penicillin level, but the delayed method is more effective at high penicillin level, then the Mantel-Haenszel test using M^2 is no longer efficient. The data in Table I shows such trend.

Table I: Effectiveness of Penicillin by Injection Methods (Mantel, 1963)

Penicillin	Injection	Response	
Level	Method	Cured	Died
1/8	Immediate	0	6
	Delayed	0	5
1/4	Immediate	3	3
	Delayed	0	6
1/2	Immediate	6	0
	Delayed	2	4
1	Immediate	5	1
	Delayed	6	0
4	Immediate	2	0
	Delayed	5	0

Kim and Lim (1998) proposed an alternative test statistic for conditional independence, which is computed as (notations are explained in the next section),

$$M_A^2 = \frac{[\sum_1^5(|n_{11k} - m_{11k}| - E|n_{11k} - m_{11k}|)]^2}{\sum_1^5 V(|n_{11k} - m_{11k}|)} = 9.62$$

with p-value 0.002. The evidence of association becomes larger with this test statistic. Two test procedures have different p-values but the conclusions are the same under the significance level 0.05. However, if we have observed (1, 1) instead of (2, 0) in the ninth row of Table I with all the other rows the same, we would get different conclusions since $M^2 = 2.91$ and $M_A^2 = 11.61$. This implies that MH test should be complemented by the alternative test procedure when we have no prior knowledge about the three-factor interaction of X, Y and Z.

The log rank test is nothing but the MH test applied to the two-sample comparison problem for censored data. In this study the weakness of the log rank test is pointed out, and an alternative test statistic is proposed as in MH test. It will be shown that the alternative test procedure becomes much more powerful when the hazard ratio between two groups changes dynamically.

In Section 2, MH test of conditional independence for categorical data and its alternative are formally stated and summarized. In Section 3, the alternative test statistic is extended to the two-sample comparison problem for censored data. And by simulation results it is claimed that the log rank test should be complemented by the alternative test.

2. MH TEST AND ITS ALTERNATIVE

Assume we have categorical variables X, Y and Z. The control variable Z has K levels and both X and Y have binary responses. Then the observations can be represented as K strata of 2×2 tables. Here K strata can be either K levels of one variable or K all possible combinations of levels of several potentially confounding variables. Let n_{ijk} denote the count at the i-th level of X, the j-th level of Y and the k-th level of Z with π_{ijk} as its probability. Also we define $n_{i+k} = \sum_{j=1}^{2} n_{ijk}$, $n_{+jk} = \sum_{i=1}^{2} n_{ijk}$, $n_{++k} = \sum_i \sum_j n_{ijk}$, $\pi_{i+k} = \sum_j \pi_{ijk}, \pi_{+jk} = \sum_i \pi_{ijk}, \pi_{++k} = \sum_i \sum_j \pi_{ijk}$. Given the marginal totals n_{+1k}, n_{+2k}, n_{1+k} and n_{2+k} in each 2×2 table, n_{11k} is known to have hypergeometric distribution. Hence the conditional mean and variance of n_{11k} are

$$m_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

$$V(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}$$

Cell counts from different strata are independent. Thus $\sum_{k=1}^{K} n_{11k}$ has mean $\sum_k m_{11k}$ and variance $\sum_k V(n_{11k})$. Mantel and Haenszel (1959) proposed the test statistic

$$\frac{(|\sum n_{11k} - \sum m_{11k}| - \frac{1}{2})^2}{\sum V(n_{11k})}$$

Under the null hypothesis of conditional independence (i.e. $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ for all i, j, k), this statistic has approximately the chi-squared distribution with the degree of freedom 1. We do not consider the continuity correction term 1/2 and define the test statistic as

$$M^2 = \frac{(\sum n_{11k} - \sum m_{11k})^2}{\sum V(n_{11k})} = \frac{[\sum(n_{11k} - m_{11k})]^2}{\sum V(n_{11k})} \tag{2.1}$$

An advantage of MH test is its applicability to sparse tables. For example, when each table is obtained from paired sample, i.e. when $n_{1+k} = n_{2+k} = 1$, M^2 still has approximately the chi-squared distribution with df=1 for moderately large K.

M^2 gets larger when $n_{11k} - m_{11k}$ is consistently positive or consistently negative for all strata. Agresti (1990) points out that MH test is inappropriate when the association changes dynamically across strata with alternating signs of $n_{11k} - m_{11k}$ across k. Kim and Lim (1998) proposed an alternative test statistic M_A^2 which depends only on the magnitude of $n_{11k} - m_{11k}$, not on its sign:

$$\begin{aligned} M_A^2 &= \frac{[\sum |n_{11k} - m_{11k}| - E(\sum |n_{11k} - m_{11k}|)]^2}{V(\sum |n_{11k} - m_{11k}|)} \\ &= \frac{[\sum(|n_{11k} - m_{11k}| - E|n_{11k} - m_{11k}|)]^2}{\sum V|n_{11k} - m_{11k}|} \end{aligned} \tag{2.2}$$

In the denominator of equation (2.2)

$$\begin{aligned} V|n_{11k} - m_{11k}| &= E(n_{11k} - m_{11k})^2 - (E|n_{11k} - m_{11k}|)^2 \\ &= V(n_{11k}) - (E|n_{11k} - m_{11k}|)^2 \end{aligned}$$

Hence the conditional expectation $E|n_{11k} - m_{11k}|$ is the only term which need to be evaluated additionally in the equation (2.2). The evaluation of $E|n_{11k} - m_{11k}|$ can be easily done numerically if we have the algorithm to calculate the hypergeometric probability.

Kim and Lim (1998) have shown by simulation that the test using M_A^2 is more efficient for dynamically changing association of X and Y across strata. They proved that the test statistic M_A^2 has the approximate chi-squared distribution with df=1 for large K under the null hypothesis of conditional independence. They have also shown by simulation that for small number of strata K, say less than 5, the chi-squared distribution may not give a good approximation for the null distribution of M_A^2. In the next section, a resampling method estimating the exact p-value is provided.

3. LOG RANK TEST AND ITS ALTERNATIVE

The log rank test for the comparison of two survival distributions is an extended MH test. The way to construct a sequence of 2×2 tables from two samples of survival times subject to right censoring will be briefly summarized in our context (See p.96 in Miller, 1981). First, make the combined ordered sample from two samples of survival times. At each failure time 2×2 table is constructed, with n_{1+k} and n_{2+k} as the risk sets of sample 1 and 2 at the k-th failure time, respectively. And n_{11k} is the number of failures in sample 1 at the k-th failure time. n_{11k} takes 0 or 1 value if there are no ties. The number of strata K becomes the number of distinct uncensored observations in the combined sample. These K 2×2 tables are not independent. But the asymptotic normality of M, or the asymptotic chi-squared distribution of M^2 still holds (Gill, 1980).

The alternative test to MH test can be directly extended to two-sample problem for censored data as MH test can. The log rank test is appropriate for two-sample problem when two groups have the proportional hazards. If the hazard ratio changes dynamically, the log rank test may have the same weakness as MH test. The test statistic M_A^2 in (2.2) can be an alternative in such case. We conjecture that the asymptotic null distribution of M_A^2 would be chi-squared distribution with df 1 like M^2 when the two survival distributions are identical. The conjecture is based on the normality of M_A for independent cases of Section 2 as proved in Kim and Lim (1998). Since the dependent structure does not affect the asymptotic chi-squared distribution of the log rank test statistic M^2, we conjecture that the same thing would happen to M_A^2. The simulation results to be explained later support this conjecture. In this section, we compare the performance of the alternative test procedure with the log-rank test by simulation.

A Resampling Method to Estimate the Exact p-value

First, we present a resampling method to estimate the exact p-value for the alternative test statistic M_A^2. The resampling method is motivated and validated by Fisher's exact test for 2×2 tables. Instead of considering all possible tables to get the exact p-value as in Fisher's exact test, we resample with the appropriate null probability and estimate the exact p-value.

If the two survival distributions are homogeneous, row and column in each given 2×2 table are conditionally independent. And the converse is true. (This claim easily stands to reason once you recall that at each uncensored observation in the combined sample we have a 2×2 table.) Hence the null hypothesis of homogeneity of two groups is equivalent to the null hypothesis of conditional independence of row and column for each stratum given.

Given K $2{\times}2$ tables $\{(n_{++1}, n_{1+1}, n_{+11}, n_{111}), \ldots, (n_{++K}, n_{1+K}, n_{+1K}, n_{11K})\}$, the resampling method can be described as follows.

Step 1 Under the null hypothesis of conditional independence generate one set of K $2{\times}2$ tables $\{(n_{++1}, n_{1+1}, n_{+11}, n^*_{111}), \ldots, (n_{++K}, n_{1+K}, n_{+1K}, n^*_{11K})\}$ and calculate $M_A^{2(1)}$ as in (2.2). Here n^*_{11k} $(k = 1, \ldots, K)$ are hypergeometric random numbers with probability

$$P(n^*_{11k} = x) = \frac{\binom{n_{+1k}}{x}\binom{n_{++k}-n_{+1k}}{n_{1+k}-x}}{\binom{n_{++k}}{n_{1+k}}}$$

where $\max(0, n_{1+k} + n_{+1k} - n_{++k}) \leq x \leq \min(n_{1+k}, n_{+1k})$. SAS function PROBHYPR was used to get the probability. Discrete random number can be easily generated if its probability distribution is given. The computational load of generating hypergeometric random numbers is not heavy since the upper bound of x is 1 if there are no ties.

Step 2 Repeat Step 1 B times, getting $M_A^{2(1)}, \ldots, M_A^{2(b)}, \ldots, M_A^{2(B)}$. From these values we can approximate the distribution of M_A^2 under the assumption of the conditional independence. Hence we can estimate the exact p-value.

This estimate of p-value is free from the assumption of asymptotic chi-squared distribution of the test statistic under the null hypothesis of homogeneity. Hence for the use of the test statistic M_A^2 we actually do not need the conjecture about the chi-squared null distribution of it. This resampling method also can be applied to approximate the exact null distribution of M^2.

Simulation Study

We consider four situations for simulation study:

1. Two groups are homogeneous: Two groups have an exponential distribution with the same hazard rate.
2. Hazard ratio between two groups is constant: Two groups are exponential with different hazard rates.
3. Hazard ratio changes across 1: One group is exponential, the other is Weibull with increasing failure rate (IFR)
4. The hazard ratio changes more dramatically: One group is Weibull with decreasing failure rate (DFR), the other is Weibull with IFR.

For each situation, we compare the performance of two test statistics M^2 and M_A^2 by the rejection probabilities, i.e. the probabilities of rejecting the null hypothesis of homogeneity.

Let us describe the simulation study for the situation 1. Two groups are homogeneous with the exponential distribution of hazard rate 1. For one generated set of data from two homogeneous groups, the observed statistics, M^2 and M_A^2, are calculated. From the chi-squared distribution with df 1, the asymptotic p-values are obtained, so that we can conclude whether the null hypothesis of homogeneity of two groups should be rejected or not. To see whether the chi-squared distribution provide a good approximation, an (approximately) exact test using resampling method is also implemented. We resample $B = 400$ sets as described in Steps 1-2. From these 400 resampled values of test statistic, estimate of the exact p-value of each test is obtained. Now we can conclude whether the null hypothesis should be rejected or not by comparing the estimate of the exact p-value with the nominal significance level 0.05. (By the discreteness of the test statistics M^2 and M_A^2, it does matter for a small combined sample size, say less than 40, whether we include or not the observed value of the test statistic in the rejection region. Thus we report the mid p-value, i.e. we count 1/2 instead of 1 in the evaluation of the p-value when the resampled value of the test statistic, $M_A^{2(b)}$ or $M^{2(b)}$, equals the observed one.) This whole step are iterated 400 times, giving the estimates of the rejection probability or the actual significance level for the situation 1. Since two groups are homogeneous in the situation 1, the actual significance level should coincide with the nominal one for both statistics M^2

and M_A^2. Table II (i) shows the result. (Only the estimates of the exact p-value are reported since the asymptotic and the exact p-value show little difference, which substantiates the conjecture about the null distribution of the test statistic M_A^2.) We consider no censoring case and the case with the average censoring weight 20%. Uniform distribution was used for generating the censoring time. As expected, whether asymptotic or exact, most of the actual significance levels are within the error bound of the nominal significance level. (Assuming $B = \infty$, the error bound with the confidence coefficient 0.95 is about $0.05 \pm 2\sqrt{(0.95)(0.05)/400}$ or 0.05 ± 0.02).

In the situation 2, the proportional hazards model is assumed, where two groups are exponential with different hazard rates 1 and 2, respectively. The log rank test is expected to perform better than its alternative for this Lehmann-type difference. Table II (ii) gives the result as expected.

In the situation 3, one group is exponential with hazard rate 1, the other is Weibull with the shape parameter 2 and the scale parameter 1. The hazard ratio between two groups changes dynamically, i.e. from below 1 to above 1. In the situation 4, the change is more dynamic since one is DFR (Weibull with the shape parameter 1/2) and the other is IFR (Weibull with 2). These are the situations where the alternative test using M_A^2 is claimed to be used. Table II (iii) and (iv) show the results. As expected, log rank test does not detect the difference well while the alternative does. Even for large sample ($n = 100$) with no censoring, the log rank test hardly detects the difference while the alternative practically always does. This case strongly advocates the use of M_A^2 in complement to the log rank test to detect the difference between two groups of survival times. The log rank test cannot be solely depended on when the proportional hazards model is dubious.

Application to Real Data

Between January, 1974 and May, 1984, the Mayo Clinic conducted a randomized trial in primary biliary cirrhosis of the liver (PBC), comparing the drug D-penicillamine (DPCA) with a placebo. A total of 312 cases (158 in DPCA group and 154 in placebo group) participated in the randomized trial. By the end of study, 125 of the 312 patients had died, with 122 distinct failure times. Appendix D of Fleming and Harrington (1991) contains the data,

Table II: Rejection Probabilities of the Log Rank Test and Its Alternative

		no censoring		20% censoring	
	n^*	log rank	alternative	log rank	alternative
(i)	20	.060	.063	.073	.048
	40	.055	.065	.045	.060
	60	.055	.070	.048	.048
	80	.068	.048	.065	.058
	100	.058	.038	.038	.050
(ii)	20	.308	.210	.168	.113
	40	.610	.368	.333	.153
	60	.735	.470	.523	.250
	80	.878	.568	.573	.228
	100	.938	.728	.663	.293
(iii)	20	.093	.073	.060	.060
	40	.085	.173	.045	.110
	60	.093	.275	.060	.190
	80	.108	.470	.065	.303
	100	.143	.538	.055	.350
(iv)	20	.085	.125	.078	.103
	40	.085	.380	.083	.270
	60	.145	.783	.073	.505
	80	.208	.908	.065	.648
	100	.183	.993	.050	.833

NOTE: The rejection probability is based on 400 number of simulations.
* n is the size of the combined sample. We have $n/2$ observations for each group.

and Example 0.2.2 displays the Kaplan-Meier estimates of the survival functions for the DPCA and placebo groups. The observed value of the log rank test statistic M^2 is 0.103 with the asymptotic p-value 0.748. The observed value of the alternative test statistic M_A^2 is 1.829 with the asymptotic p-value 0.176. (The estimates of the exact p-value using the resampling method with $B = 2,000$ are 0.754 and 0.216, respectively.) Although either test does not reject the null hypothesis of homogeneity, the degrees of evidence against the null hypothesis are quite different, which is anticipated by the display of the two Kaplan-Meier estimates.

4. CONCLUSION

The log rank test is probably the most widely used for the comparison of two groups of survival times subject to censoring. As the simulation study shows, the log rank test detects well the Lehmann-type difference. But it has serious fault if the hazard ratio between two groups changes dynamically, i.e. if the hazard ratio changes across 1. The newly proposed test statistic overcomes the weakness of the log rank test. However, the test using M_A^2 is less efficient for detecting the Lehmann-type difference. Hence it should be a complement to the log rank test. It is recommended that both test statistics M^2 and M_A^2 should be used to test for difference between two groups of censored observations if there is little prior knowledge about the hazard ratio between two groups.

BIBLIOGRAPHY

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts 124 (Mathematicsch Centrum, Amsterdam.)

Kim, J. and Lim, H. (1998). "Testing conditional independence in K 2×2 tables." (written in Korean). *The Korean Journal of Applied Statistics*, To appear.

Mantel, N. (1963). "Chi-square tests with one degree of freedom: extension of the Mantel-Haenszel procedure." *J. Amer. Statist. Assoc.*, 58, 690-700.

Mantel, N., and Haenszel, W. (1959). "Statistical aspects of the analysis of data from retrospective studies of disease." *J. Natl. Cancer Inst.*, 22, 719-748.

Miller, R. G. (1981). *Survival Analysis*. Wiley, New York.